Collaborative Sciences Center for ROAD SAFETY

FINAL REPORT

Completing the Picture of Traffic Injuries: Understanding Data Needs and Opportunities for Road Safety

November 2018

Christopher Cherry (Principal Investigator)
Amin Mohamadi Hezaveh
Melany Noltenius
Asad Khattak
University of Tennessee, Knoxville

Louis Merlin
Eric Dumbaugh
Florida Atlantic University

David Ragland
University of California, Berkeley

Laura Sandt
University of North Carolina, Chapel Hill

THE UNIVERSITY OF TENNESSEE KNOXVILLE · Berkeley UNIVERSITY OF CALIFORNIA · FAU FLORIDA ATLANTIC UNIVERSITY · Duke UNIVERSITY · THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

U.S. DOT DISCLAIMER

ACKNOWLEDGMENT OF SPONSORSHIP

This project was supported by the Collaborative Sciences Center for Road Safety, www.roadsafety.unc.edu, a U.S. Department of Transportation National University Transportation Center promoting safety.

TECHNICAL REPORT DOCUMENTATION PAGE

General instructions: To add text, click inside the form field below (will appear as a blue highlighted or outlined box) and begin typing. The instructions will be replaced by the new text. If no text needs to be added, remove the form field and its instructions by clicking inside the field, then pressing the Delete key twice. Please remove this field before completing form.

1. Report No. CSCRS-R4	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Completing the Picture of Traffic Injuries: Understanding Data Needs and Opportunities for Road Safety		**5. Report Date** April 30, 2018 **6. Performing Organization Code**
7. Author(s) Christopher Cherry, Ph.D., https://orcid.org/0000-0002-8835-4617 Amin Mohamadi Hezaveh, Ph.D. candidate, https://orcid.org/0000-0003-2480-1956 Melany Noltenius, Ph.D., https://orcid.org/0000-0001-7887-6182 Asad Khattak, Ph.D., https://orcid.org/0000-0002-0790-7794 Louis Merlin, Ph.D., https://orcid.org/0000-0002-9267-5712 Eric Dumbaugh, Ph.D., https://orcid.org/0000-0002-5254-9711 David Ragland, Ph.D., https://orcid.org/0000-0002-8996-1320 Laura Sandt, Ph.D., https://orcid.org/0000-0001-9468-7891		**8. Performing Organization Report No.**
9. Performing Organization Name and Address University of Tennessee, Knoxville, TN Florida Atlantic University, Boca Raton, FL University of California, Berkeley, CA University of North Carolina, Chapel Hill, NC		**10. Work Unit No.** **11. Contract or Grant No.** Grant #: 69A3551747113
12. Sponsoring Agency Name and Address Collaborative Sciences Center for Road Safety 730 Martin Luther King Jr. Blvd., Suite 300, Chapel Hill, NC 27599-3430		**13. Type of Report and Period Covered** Final Report (February 2017-April 2018) **14. Sponsoring Agency Code**

15. Supplementary Notes
Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.

16. Abstract
Although traditional police recorded crash data has improved over time, providing additional data and analytics will demonstrate a more "complete picture" of crashes and injuries. In this study, we examined this complete picture of traffic crashes and determined which data elements of conventional crash data can be used to perform linkage between multiple data sources and eventually create a data linkage map of a more "complete picture" of traffic crashes. This study stands as a quick-reference guide for practitioners and researchers to understand existing databases and how datasets relate to each other. The main body of the report is meant to frame the issue of data that can be relevant to understanding crashes and show how they interface with conventional crash datasets. Furthermore, we provided a comprehensive analysis of pre-crash, crash, environment, and post-crash datasets which includes linkages between multiple datasets and their implications for road safety analysis. This study concludes with five case studies. These case studies focus on 1) underreporting of pedestrian and bicyclist injuries in police crash data, 2) a review of pre-hospital response time and traumatic injury, 3) exploring the relationship between residential neighborhood accessibility and safety, 4) development of a Home-Based Approach as a complementary method for measuring road safety at residential address of the road users, and 5) examining factors influencing seat belt use rates at the neighborhood level. This series of studies assists policy makers and contribute to visualization of linked data that helps tell compelling safety stories that guide safety improvements.

17. Key Words	**18. Distribution Statement**
Safe Systems; Data Linkage; EMS response; Underreporting; Home-Based Approach; Seat Belt Use Rates; Accessibility	No restrictions. This document is available through the Collaborative Sciences Center for Road Safety (roadsafety.unc.edu), Chapel Hill, NC.

19. Security Classif. (of this report)	**20. Security Classif. (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified	Unclassified	137	

Executive Summary

Background

Police crash data has been the primary source of information for transportation safety analysis for decades. There has been continual effort to improve police crash data's ability to assist monitoring and evaluation of road safety and also improve understanding of strategies to improve safety. Police crash data is limited, primarily focusing on the physical circumstances surrounding the crash and excluding the many pre-crash and post-crash factors that affect crash rates and injury outcomes. Many complementary datasets exist and can provide a more complete picture of road safety. This report identifies how those datasets can be integrated into road safety analysis to answer questions that police crash data alone is ill-suited to answer.

Recent efforts expand datasets available to improve understanding of safety. Most of the data integration efforts and successes have revolved around using Post-Crash data, like hospitalization data to better understand injury outcomes. Specifically, hospitalization data and associated health datasets are vast and contain much more specific injury assessments than police crash reports. Improved care and access to appropriate hospital facilities is a substantial contributor to improved injury outcomes for traffic crashes. In the United States, the Crash Outcome Data Evaluation System (CODES) program was a state-by-state initiative organized by the National Highway Traffic Safety Administration (NHTSA) to formulate processes to consistently link hospital datasets with police crash datasets. These datasets have differences between states, so the CODES program encouraged states to develop processes to link their state-specific databases. This program, though no longer overseen by NHTSA, initiated a wave of linkage efforts between state police crash programs and counterpart hospitalization datasets. The biggest challenges with these efforts is acquiring access to sensitive health data and developing reliable linkage methods to integrate the two datasets.

Completing the picture of crashes with added data

A complete picture of crashes can be conceptualized in Exhibit 1, where crash outcome data (e.g., police crash or near miss data) is improved by adding complementary *Pre-Crash*, *Environment*, and *Post-Crash* datasets. In this report, we describe a broad set of *Pre-Crash, Environment, and Post-Crash* datasets (Exhibit 1). *Post-Crash* datasets include conventional hospitalization data, but also expand to the many other datasets that are important to understanding crash outcomes. Some examples include Medical Insurance Claims databases, Emergency Medical Systems (EMS) data, and Vital Statistics. Many of these datasets are managed by state departments of health and have been the focus of many previous linkage efforts.

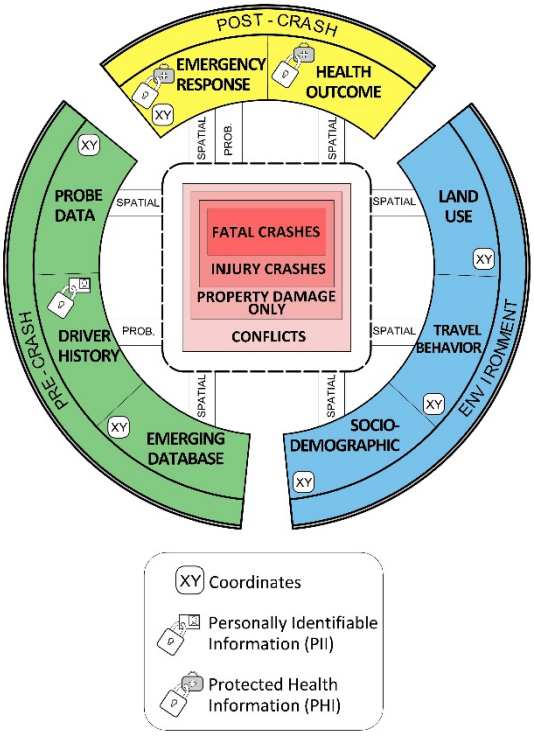

Exhibit 1. Complete Picture

Many other datasets are used with police crash data to provide greater safety insights. Two broad categories encompass most of the datasets that exist (non-*Post-Crash*). *Pre-Crash* datasets include data that could predict factors that influence crashes. Some pre-crash datasets are driver or vehicle oriented, like citation histories or other predictors of crashes derived from Department of Motor Vehicle (DMV) datasets. Pre-crash data could also include meteorological data, naturalistic data (e.g., SHRP2), or crowdsourced data (e.g., WAZE or others).

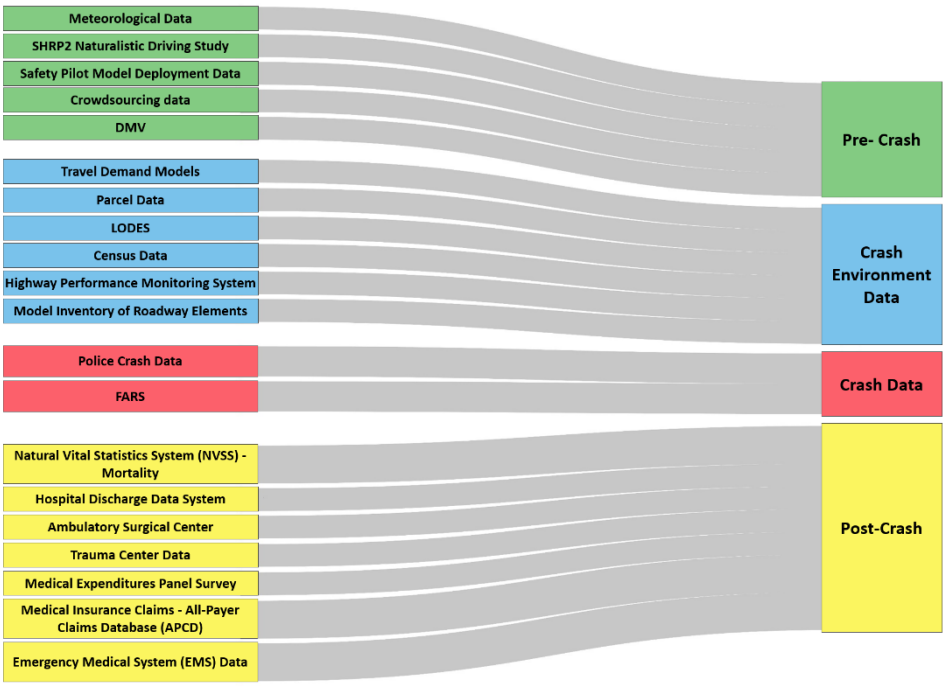

Exhibit 2. Datasets Described In The Report

Environment datasets broadly include datasets that describe the crash environment of specific crashes or, at the aggregate level, the environment in which many crashes occur (e.g., city level metrics). Environment data generally include the built environment (roadway and land use characteristics), socioeconomic information derived from census data, or information from travel demand models. These datasets contribute to understanding important exposure variables and factors of the built environment that contribute to crash rates and outcomes. The datasets and linkages described in this report are shown in Exhibit 2.

Methods of linking datasets

Reliably linking datasets is critical to exploiting the benefits of utilizing multiple datasets. In this project, we establish police crash data as the core dataset from which to establish linkage procedures[1]. All datasets described in this document have explicit safety implications but require linkage to crash databases through one of three methods: Deterministic Linkage, Probabilistic Linkage, or Spatial/Temporal Linkage (See Exhibit 1).

Deterministic linkages rely on a unique identifier to match records in a secondary database (e.g., hospital) with a primary database (e.g., police crash). Deterministic linkages can include a unique identification code for each crash victim, but more often include a series of identifiers that reliably link two records. These identifiers could include social security numbers, names, gender, and date and the analyst assigns linkages. Deterministic linkages can be used to link any person- or vehicle-oriented database with crash databases.

Often, the linking process includes some ambiguity, or data elements do not match (e.g., name variants). A probabilistic linkage approach aims to remove subjectivity in the data matching process and rely on probabilistic

[1] We also acknowledge that near-misses are a type of sub-crash incident that shares many common characteristics of crashes. The work here could also be paired with near-miss data or used to identify and analyze unreported crashes that are not in the police crash database.

approaches to link data. Each data element is matched probabilistically with elements in the other datasets. The data are included and matched if they meet a threshold probability. In almost all cases, deterministic or probabilistic matching is used when complementary datasets contain information about specific agents in a crash (e.g., crash victims or vehicles).

Many datasets do not include specific characteristics of crash victims, but rather rely on the environment of the crash or pre-crash conditions. Often these datasets are linked to police crash data by spatial and temporal identifiers. Geocoded crash data can be linked with surrounding data through geographic matching to pair data. Spatial matching allows characteristics that surround the crash to be paired with the crash data to identify factors that are not in the crash database that likely contributed to the crash. A few examples include roadway design information, surrounding land use characteristics, or weather conditions. This report details how each dataset relates to police crash data and how the linkage is applied to road safety.

Case Studies

To illustrate how data linkage can aid in completing the picture for transport safety analysis, we present five case studies. Each case study illustrates different approaches and advantages to utilizing a subset of linked data.

Case 1: Linked Crash & Health Data to Assess Bicycle and Pedestrian Safety, Under Reporting, and Injury Assessment

This case reviews the literature on bicycle and pedestrian underreporting and injury assessment in police crash data. This study focuses on how hospitalization data can provide better insight and safety for these road users. Specifically, one outcome of linking hospitalization data is understanding levels of underreporting among bicyclists and pedestrians. Crashes are highly underreported in police databases, and hospital data includes many records of traffic-related injuries that are not included in police crash databases. Additionally, hospitalization data enables better assessment of types and severity of injuries to cyclists and pedestrians that are not included in crash databases. These findings have implications for accurate injury surveillance programs and transportation countermeasures.

Case 2: EMS Response Time and Crash Outcomes

Seemingly similar crashes have different health outcomes, depending on a number of factors. Often rural crashes have higher fatality rates than similar urban crashes. One of the factors that likely contributes to crash outcomes is the quality and speed of Emergency Medical System (EMS) response. EMS response time varies from minutes to hours after a crash with significant implications on the survivability of a severe crash. This work reviews the literature on linking police crash data with EMS data. The implication is that one can find the extent to which EMS response time contributes to better health outcomes from crashes, which could lead to better EMS policies and help explain disparities in road safety between urban and rural areas.

Case 3: Neighborhood Accessibility and Traffic Safety

There is a growing literature on the relationship between road safety and urban form. In this case study we investigate how neighborhood accessibility levels, calculated through travel demand modeling approaches, affect safety at the neighborhood level. This is achieved by developing accessibility metrics at the Census Block Group or Traffic Analysis Zone (TAZ) level and predicting crashes, geocoded to the home location of the crash victim. We link police crash and planning level data to estimate whether there is a relationship between crashes and accessibility.

In this way, we can use accessibility as an exposure metric to understand how the regional distribution of land use changes crash and severity rates.

Case 4: Home-Based Approach: A Complementary Definition of Road Safety

In this study, we established a new definition to complement the traditional definition of the road safety; the traditional definition of the road safety attributes safety to the location of the crashes (i.e., Location-Based Approach). Attributing safety to the location of the crash is a reliable method for assigning engineering and enforcement resources. However, one may question whether using the location of traffic crashes is a useful tool for accessing individuals with higher risk of involvement in traffic crashes for educational purposes. In this study, we defined Home-Based Approach as the expected number of crashes that road users who live in a certain geographic area have during a specified period. This definition attributes traffic crashes to the home-address of the individuals. This Home-Based-specific assessment follows typical epidemiological approaches to the burden of disease monitoring and uses matched census data to understand factors that influence crash frequency of individuals at zonal level (e.g., census tract). In this approach, high-crash neighborhood hotspots can be identified, and focused interventions can be initiated to improve road safety.

Case 5: Neighborhood-Level Factors Affecting Seat Belt Use

Seatbelt rates have remained relatively constant in most states. In Tennessee, blanket seatbelt campaigns have had some effect on raising the average seatbelt rate. Following the Home-Based Approach (HBA) in Case 4, this case study identifies neighborhoods that have low seatbelt use rate based on crash data and correlating seatbelt rates with socioeconomic information derived from census data. In this approach, we geocode crashes to the victim's home address and estimate seatbelt use rates based on crash data. We find that seatbelt use rates, on average, match those from direct roadside observation studies, but the range of rates in the census tract level varies widely. Analysis indicated that some neighborhoods have very high seatbelt use rates while others have very low seatbelt use rates. Moreover, with this approach, we can identify rates by seating position and time of day, which are not well measured from direct observation studies. This approach can be used to target educational or enforcement activities in the neighborhoods with low seatbelt use rates to focus resources and achieve better overall outcomes.

Project Highlights

- Police Crash Data is an important, but flawed, source of information to comprehensively understand road safety
- Other datasets can help analysts and researchers create a "complete picture" of crashes
- There are several linkage methods that depend on data access, privacy, and types of analysis required.
- Most efforts to date at the state level have been to link post-crash (hospital-oriented) data with police crash datasets.
- This work provides an analysis of pre-crash, crash, environment, and post-crash datasets that inform road safety; including identifying linkages between datasets and implications for safety.
- We demonstrate the many opportunities for linking data with five case studies that are illustrative of the types of analysis possible.

Contents

Table of Figures

Acronym	Definition
ACPM	Aggregate Crash Predictions Models
ACS	American College Of Surgeons
ACS	American Community Survey
AHA	American Hospital Association
APCD	All-Payer Claims Database
ASC	Ambulatory Surgery Centers
ASOS	Automated Surface Observing System
ATIS	Automated Terminal Information Service
AWOS	Automated Airport Weather Observation System
BAC	Blood Alcohol Content
CAMH	Centers For Medicare & Medicaid Services (CMS) Alliance To Modernize Healthcare
CDC	Centers For Disease Control And Prevention
CMOD	Crash Medical Outcomes Data Project
CMS	Centers For Medicare & Medicaid Services
CODES	Crash Outcome Data Evaluation System
DAS	Data Acquisition System
DMV	Departments Of Motor Vehicles
DPPA	Driver's Privacy Protection Act
DUL	Data Use License
ED	Emergency Department
EDIS	Emergency Department Information System
EMS	Emergency Medical System
FAA	Federal Aviation Administration
FHWA	Federal Highway Administration
GIS	Geographic Information System
HC	Household Component
HDDS	Hospital Discharge Data System
HIPAA	The Health Insurance Portability And Accountability Act
HPMS	Highway Performance Monitoring System

Acronym	Definition
IC	Insurance Component
ICAO	International Civil Aviation Organization
IRB	Institutional Review Board
IRTAD	International Traffic Safety Data And Analysis Group
ISS	Injury Severity Score
LINCS	Nonfatal Crash Surveillance
LODES	Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics
MAIS	Maximum AIS
MEPS	Medical Expenditure Panel Survey
METAR	Meteorological Terminal Aviation Routine
MIRE	Model Inventory Of Roadway Elements
MMUCC	Guideline Model Minimum Uniform Crash Criteria)
MPC	Medical Provider Component
MPO	Metropolitan Planning Organizations
MVC	Motor Vehicle Crash
NASS	National Automotive Sampling System
NCHRP	National Cooperative Highway Research Program
NCIPC	National Center For Injury Prevention And Control
NDS	Naturalistic Driving Study
NEMSIS	National Emergency Medical Services Information System
NHC	Nursing Home Component
NHTSA	National Highway Traffic Safety Administration
NTDB®	National Trauma Data Bank®
NTDS	National Trauma Data Standard
NTSB	National Transportation Safety Board
NVSS	National Vital Statistics System
OECD	Organization For Economic Co-Operation And Development
PHI	Protected Health Information
PII	Personally Identifiable Information

Acronym	Definition
QHAPDC	Queensland Hospital Admitted Patients Data Collection;
QISU	Queensland Injury Surveillance Unit
QRCD	Queensland Road Crash Database
RTA	Road Traffic Accident
SFGH	San Francisco General Hospital
SHRP 2	Second Strategic Highway Research Program
SSN	Social Security Number
SWITRS	Statewide Integrated Traffic Records System
TAC	Technical Assistance Center
TAZ	Traffic Analysis Zone
TBI	Traumatic Brain Injury
TDM	Travel Demand Models
TRB	Transportation Research Board
VMT	Vehicle Miles Traveled

Data Integration Introduction

Police recorded crash data has improved over time, but still fails to report all aspects of crashes that are important to developing a full understanding of crash mechanism, injury burden, and ultimately total health outcomes. Traditionally, safety and injury analysis have occurred in isolated fields, with road safety researchers relying predominately on police-recorded crash reports, and public health researchers relying on health records (e.g., hospital, emergency department, ambulatory care data). Often, these records do not reflect the same findings, even for the same crash victims. By themselves, injury severity and crash reporting rates are often inconsistent between datasets. A "complete picture" of traffic crashes must be established to address these limitations. This complete picture needs to consider a multi-perspective approach to road safety instead of one point of view by considering multiple sources of data. This complete crash picture can be used to answer inconsistencies in findings and provide a better understanding of the nature of traffic crashes, injury outcome, and eventually direct cost of traffic crashes.

The study objectives are:

1) To briefly review data linkage methodology,
2) Review examples of linking databases,
3) Establish a framework for developing a complete picture of traffic crashes,
4) Identify databases that have potential to complete the picture of traffic crashes, and
5) Illustrate linkage potential through case studies.

History of Crash and Health Data Integration Efforts

Various researchers and organizations around the world have attempted to link different sources of traffic crashes. Most of them, however, are not systematic efforts. This section briefly discusses prominent examples of linking police crash reports with health data in the United States.

Crash Outcome Data Evaluation System (CODES)

The National Highway Transportation Safety Administration (NHTSA) initially created the Crash Outcome Data Evaluation System (CODES) to quantify and report on the benefits of safety equipment and legislation in terms of mortality, morbidity, injury severity, and health care costs. The effort was undertaken in response to Section 1031(b) of the Intermodal Surface Transportation Efficiency Act of 1991 (Martinez 1993), which required NHTSA to conduct a study and report to Congress on the benefits of safety equipment (i.e., seat belt use and motorcycle helmet use) in traffic crashes (Kindelberger and Milani 2015). In 1992, NHTSA sought grant applications from entities with existing statewide crash and injury data systems that were capable of generating crash, medical, and financial-outcome information if linked together. Any state agency, non-profit organization, or educational institution was eligible to develop and coordinate a coalition of data owners and users to perform the desired linkages (Milani et al. 2015).

By October 1992, seven states were awarded grants to establish CODES programs. CODES became institutionalized in the awarded states based on a series of partnerships among state traffic safety and public health agencies and NHTSA. State agencies, universities or affiliates, or non-profit institutions were the leading organizations in the CODES project. In some cases, lead organizations entered into agreements with support entities, such as universities, to conduct the actual data linkage and/or analyses. CODES cooperative agreements were administered through NHTSA's National Center for Statistics and Analysis.

NHTSA encouraged grantees to seek and secure other supplemental funding for their CODES programs and to move toward program institutionalization to ensure sustainability while CODES was developing. In 2013, NHTSA had CODES cooperative agreements with grantees in 15 states: Connecticut, Delaware, Georgia, Illinois, Kentucky, Maine, Maryland, Minnesota, Missouri, Nebraska, New York, Ohio, South Carolina, Utah, and Virginia.

Some states left the CODES network or retired their programs; nevertheless, some of the states continued to conduct the linkage independently. Other states started linkage projects independently which were not part of the CODES data network. Some of these states changed the nature of linking process, and instead of using probabilistic methods provided by CODES, they instead used matching identifiers and deterministic linkage to link health outcome data and crash database. Other states initiated projects to link Emergency Medical Services and trauma registry data, with the goal of adding crash and other data sets. Still, others have used commercial software that replicated the CODES methodology and established or piloted state projects similar to those seen in CODES (Milani *et al.* 2015). During transition, NHTSA did not become owners of state CODES data and does not host the data. CODES data reside in the states where the linkage originated, and NHTSA does not disseminate CODES data. Requests for data or information are handled by individual state CODES project sites (Kindelberger and Milani 2015).

Crash Medical Outcomes Data Project (CMOD)

The California Department of Public Health, Safe and Active Communities (SAC) Branch implemented a project to integrate police crash reports and medical data to gain a better understanding of traffic injury severity. The Crash Medical Outcomes Data (CMOD) Project is funded by NHTSA, under the auspices of the Traffic Records Coordinating Committee, and administered by California's Office of Traffic Safety. CMOD is included in California's Strategic Highway Safety Plan under Challenge Area 16: Improve Safety Data Collection, Access, and Analysis (CMOD 2017). The goal of the CMOD project is to focus on person-level risk factors and crash outcomes to better understand how to prevent Californians from being injured or killed in traffic crashes. This project electronically links police crash reports with health outcome data sources, and with death data. This project also responds to the need for instant online analysis of traffic injury data for policy evaluation and planning by making person-level crash data accessible via the web (CMOD 2017).

Importance of linking databases

Linking crash databases to medical data is of interest to various agencies and organizations for use in analysis. The Centers for Disease Control and Prevention (CDC) determined that linked data, such as that produced by CODES, was valuable to determining risk factors for motor vehicle injury and in designing and evaluating interventions to address these risk factors (Milani *et al.* 2015). As a result, in 2010, CDC and NHTSA administrators signed a memorandum of understanding for collaborative strategies. Between 2010 and 2012, the CDC and NHTSA agreed to explore the feasibility and benefits of an ongoing partnership in the CODES program. The CDC's continuing interest in crash-medical data linkage is a primary motivator of this study.

Currently, the CDC's National Center for Injury Prevention and Control (NCIPC) has enlisted the Centers for Medicare & Medicaid Services (CMS) Alliance to Modernize Healthcare (CAMH), a federally funded research and development center operated by the MITRE Corporation, to create linked information for the Nonfatal Crash Surveillance (LINCS) guide. LINCS expands on previous efforts and best practices for establishing and improving linkage programs at the state level. This guide provides direction in developing data linkage plans including the selection of variable(s), data linkage method(s), data linkage tool(s), and an approach to organize and select the match results.

In a study regarding single-unit truck crashes and their injury outcomes, the National Transportation Safety Board (NTSB) recommended that data linkage systems such as CODES be continued, and issued the following conclusion (NTSB 2013): "Data from the Crash Outcome Data Evaluation System provide detailed information on injury diagnoses and severities in relation to crash characteristics, cover a large proportion of the population of the participating states, are not available elsewhere, and provide useful insight into traffic safety problems."

In a Notice of Proposed Rulemaking published on March 11, 2014, the Federal Highway Administration (FHWA) included a proposed recommendation that states begin preparations so that no later than January 1, 2020, all states

use a medical record injury outcome reporting system that links injury outcomes from medical records to crash reports (Milani *et al.* 2015).

The Transportation Research Board (TRB), FHWA, NHTSA, and CDC are liaisons to TRB's National Cooperative Highway Research Program (NCHRP) project 17-57, *Development of a Comprehensive Approach for Serious Traffic Crash Injury Measurement and Reporting Systems.* The project's goals are to 1.) identify an improved injury scoring system for further consideration, 2.) create a roadmap to assist states in developing and implementing an interim system, and 3.) develop a state-based framework to perform comprehensive linkage of records related to motor vehicle crashes that resulted in serious injuries, and provide incremental steps and priorities for achieving the linkage. The project recommended linking police crash reports to health data as the best way to obtain an accurate, serious injury measurement (Flannagan *et al.* 2013).

The International Traffic Safety Data and Analysis Group (IRTAD group), an ongoing working group of the Joint Transport Research of the Organization for Economic Co-operation and Development (OECD) and the International Transport Forum, issued recommendations for reporting on serious road traffic casualties. This includes the recommendation that assessment of injury severity should preferably be conducted by medical professionals, that police data should be complemented by hospital data, and that a 'seriously injured road casualty' be defined as a person with injuries assessed at level 3 or more on the Maximum Abbreviated Injury Scale, i.e., "MAIS3+" (Amoros *et al.* 2009)

Linking Methodologies

This section briefly introduces the most common methods for linking observations in two sets of data that describe a particular event or individual. For more details about the linkage see Cook et al. (2015). This chapter introduces five linkage processes: interface, direct linkage, deterministic linkage, probabilistic linkage, and spatial join.

Interface Method

In the interface method, two data sources are able to interact with each other in real time seamlessly. For example, when a police officer scans a driver license, he or she is immediately provided with information regarding the driver and vehicle. Different hospitals under the same ownership may have their databases interfaced through the use of a medical record number. Interfaces should be highly reliable and usually support critical business practices. Interfaces between crashes in police databases and healthcare data rarely exist due to the complex nature of data ownership, privacy, and how the data are compiled. Crash data and hospital data are collected by different entities: crash data by public agencies such as police departments, and health information by general hospitals or private healthcare companies. Similarly, the recorded data, particularly the health data, is not archived in real time. A compilation of data from multiple law enforcement agencies or hospital systems is often conducted well after the event and by different state agencies, Departments of Public Safety or Departments of Transportation for crash records, and the Department of Health or State Hospital Association for hospital data.

Direct Linkage

The direct linkage method is usually used when two databases share a unique single identifier or a set of identifiers that enable databases to be joined. However, to perform this linkage, both databases must share similar data elements or collect their data in a similar way and format. For example, in linking traffic crashes and health outcome databases, use of a social security number (SSN) or license number for drivers enables researchers to make direct links. To protect personal information, police and hospital staff do not usually collect SSN, and if they do so, due to the Health Insurance Portability and Accountability Act (HIPAA), this data is not readily available to researchers. As a result, performing direct linkage is not usually possible using SSN. The quality of the linkage and success rate is

highly dependent on the presence of the identifiers. Moreover, if the success rate in the linkage process is low (due to lack of identifiers), it is necessary to use other linking methods.

Deterministic Linkage

Another method that can be helpful in the linkage process is to use individuals' unique identifiers in multiple sources of data. This linkage method is only viable if data containing personal information identifiers are available and are accessible to researchers. In this process, instead of using a unique identifier that attributes records to the crash record or hospital record, researchers use multiple quasi-unique fields that describe an individual who was involved in a crash. Date elements (time of the crash, date of birth), personal information identifiers (name, address), geographical data element (e.g., location of crash or admission to health facility), in addition to gender and age are the most commonly used.

A perfect match occurs when all the quasi-unique fields agree on a pair of records. However, for several reasons (e.g., lack of access to data or missing data), it is likely that only some of the quasi-unique data will agree. In order to select the best linkage between two records in the datasets, researchers usually develop a scoring system and choose the best pair, based on the highest score achieved and meeting some defined minimum threshold score.

Although use of such a scoring system seems plausible and reliable, its limitations must be considered. First, each researcher develops a scoring system based on individual preferences or personal judgment. Moreover, it is unlikely that two researchers will develop an identical scoring system. As a result, based on the scoring system that researchers use, two researchers might end up with different match rate and matched observation in one study. Another shortcoming is that the relative rarity or commonness of a specific value in a field is not considered. Agreement on a value that occurs in half of the records in a database is weighted exactly the same as agreement on a value that occurs on only a single record. Finally, there is no way to assess a researcher's confidence in a given pair of records beyond the fact that the pair achieved the threshold.

The simplest form of a deterministic linkage requires all quasi-unique fields to agree on a pair of records for the results to be considered a perfect match. However, this is not always the case. To improve on some of the limitations presented above, researchers often construct scoring schemes for their deterministic linkages by giving higher point totals to variables that are considered to be more reliable or specific and fewer points to more general or less reliable fields. Manual validation can identify if the scoring schemes are producing accurate results.

Probabilistic Linkage

The probabilistic linkage is the methodology used by the CODES data network. Probabilistic linkage addresses the limitations of the deterministic methods and judgment of researchers. Defining weight and threshold for the linkage procedure do not affect the process. The aim of probabilistic linkage is to generate the probability that a pair of records describe the same person and event. In this method, a pairwise comparison between every data record in two separates databases is conducted by comparing the data fields of two different files. Having numerous comparisons for one set of records in a specific database with all or part of the records in a second database leads to judgment about whether the two records refer to the same individual (can be linked) or not (cannot be linked). For more information on the subject of the mathematical models of probabilistic linkage and current state of practice of CODES, please see Cook *et al.* (2015) on the examination of methodologies and multi-state traffic safety applications.

Spatial Join

This method is appropriate for data that are stored with geographic locations, often coordinates represented in a Geographic Information System (GIS) file format (shapefile). The shapefile format is a digital vector storage format

for storing geometric location and associated attribute information. Databases that describe the surroundings of the crash site are usually available in shapefile format—the Highway Performance Monitoring System and the US Census are two examples. To link a crash to its corresponding environment, a spatial join must be performed. A spatial join is a special type of join operation in which fields from one layer's attribute table (feature layer) are attached to another layer's attribute table (target layer) based on the relative locations of the features in the two layers. By using a spatial join, researchers can easily join data from one feature class to another feature class.

There are two types of spatial join relationships: one-to-one relationship or one-to-many relationship. In a one-to-one relationship, each individual target feature is joined with exactly one join feature. In a one-to-many join, each target feature is joined with more than one join feature, with a specified logic for aggregating across join feature attributes into a single attribute in the target feature (ESRI 2009).

Linking Examples

Investigating the injury outcome of traffic crashes (e.g., severity, treatment cost) play an important role in road safety analysis; therefore, the following section focuses on linking police crash reports and health outcome data. The literature on road safety includes abundant examples of studies linking police crash reports with other sources to achieve a better understanding of road crash mechanisms. In this section, the studies are classified into two types. First, the linkage between police crash reports and health outcome data is discussed, the focus of which is to provide examples of different goals of linkage methods. The following section focuses on spatial linkages.

Linking Highway Patrol Crashes and Hospital Oriented Data

There are five primary categories of goals for linking police crash reports to medical records of traffic victims:

1- Comparison of the injury severity and reliability of police evaluation of traffic crashes.
2- Investigation of the relationship between factors influencing true injury severity of road users.
3- Investigation of the underreporting issue and identification of unreported traffic crashes, particularly among non-motorized road users.
4- Substance abuse by drivers,
5- The evaluation of the effectiveness of vehicle equipment for reducing the injury risk. Each of these categories is discussed in the following section.

Comparison of The KABCO Scale and AIS Injury Severity Scale.

There are typically two sources of injury severity distribution in road safety literature. The sources of injury severity are police reports and medical records of those injured. However, due to the restrictions on accessing medical records, accurate injury severity information is not always accessible. One of the limitations of crash reports is misclassification of injury by police officers (Sherman *et al.* 1976, Farmer 2003, Compton 2005). To achieve a more accurate level of injury severity, and a better evaluation of the factors affecting injury severity (e.g., crash modification factor), one solution is to use injury severity reported by hospital data. Police officers at a crash scene do not generally make accurate estimations the injured body part(s) and the scale of injury severity (McDonald *et al.* 2009, Tsui *et al.* 2009, Tarko and Azam 2011). This can be attributed to several factors. Responding officers are not medically trained and make their conclusions based on the circumstances and appearances of the victims at the crash scene. As a result, some of the injuries that are life-threatening and are not obvious are misclassified by officers (CDPH 2015). In some cases, visible injuries may appear more serious than they are (Compton 2005, McDonald *et al.* 2009), leading to misclassification of the injury severity by a police officer at the crash scene. The objectives of injury classification by police officers (e.g., classification) is not consistent with the objectives of medical personnel (e.g., treatment).

Another source of misclassification can be attributed to differences in state classification systems and in police officer training (Farmer 2003). On the other hand, clinical scales such as Abbreviated Injury Scale (AIS) are more accurate and reliable compared with police officer evaluation of injury severity. The AIS metric was developed based on the threat to life and survivability, not merely incapacitation, which is the basis for the KABCO scale (FHWA 2011); in KABCO scale K, A, B, C, and O respectively stand for a crash with fatal, Incapacitating, Non-Incapacitating Evident, Possible Injury, and No Injury. Second, the method of severity assignment has a significant impact on accuracy. AIS codes are determined by clinical personnel, who have access to medical records for the individual. Third, ability to further calculate Maximum AIS (MAIS) and Injury Severity Score (ISS) figures from the AIS codes enhances the capabilities of injury severity assessment. Analyses may focus on individual injuries or overall severity for the victim (Burch *et al.* 2014). However, the AIS is often limited to trauma injuries and reported in trauma registries, limiting its application for injuries with lower severity levels. Moreover, these studies also found that reported use of a standard medical metric was found to be more consistent between states than law enforcement crash injury scoring using the KABCO scale (Sherman *et al.* 1976, Burch *et al.* 2014).

Depending on the study context, sensitivity percentages for severe injury reported by law enforcement that matched medically determined severe injury cases were reported to range from 49% to 78% (Sherman *et al.* 1976, Popkin *et al.* 1991, Farmer 2003, Compton 2005, Burch *et al.* 2014). Results of the comparison indicated that law enforcement scales were better at classifying non-incapacitating or non-severe injuries (Popkin *et al.* 1991, McDonald *et al.* 2009) and usually overestimate severe injuries (Farmer 2003).

Agran *et al.* (1990) in a study in Orange County, California, revealed that the police injury severity scale was found to correlate poorly with a scale based on medical diagnoses, and substantial underreporting by police of serious injuries was demonstrated. Farmer (2003) in a study in Virginia used data from police crash reports gathered by the National Automotive Sampling System (NASS). The data included records of all towed 1995–2001 model passenger vehicles between 1996 and 2000 and NASS/CDS were examined (n = 10,860). A total of 9,939 of the observations either had AIS measures of injury severity or coding of “fatal” on the NASS/CDS treatment-mortality variable. Analysis indicated that police officers usually correctly identified fatal or non-injured victims. On the other hand, police coding of injury was inexact. Results indicated that male and elderly drivers are misclassified as severely injured less often than female and nonelderly drivers. In addition, misclassification of traffic injuries varied by region. The authors also concluded that “police-reported information is insufficient for determining vehicle impact and driver injury severity” (Farmer 2003).

In another study in California, data from 2008–2012 (n= 1,156,150) for traffic crashes resulting in non-fatal injuries in motor vehicle crashes were examined (CDPH 2015). The data was obtained from linked crash medical data (i.e., Statewide Integrated Traffic Records System [SWITRS] and health outcome data). Comparison of the injury severity indicated that there was a high level of agreement on the non-severe injuries. Analysis indicated that most injuries with a non-severe ISS were reported in SWITRS as “other visible injury” or “complaint of pain” (high specificity). On the other hand, officers reported fewer than two-thirds of injuries with a severe ISS as severe (low sensitivity). In addition, for every four severe injuries reported in SWITRS, only one had a corresponding severe ISS. When injuries that were reported in SWITRS as other visible injuries were combined with those reported as severe, the authors found that the matching of severe injuries increased (higher sensitivity). The authors also reported that the injuries with the highest levels of discrepancy were those to the torso, head and neck, and traumatic brain injuries (CDPH 2015).

Burch *et al.* (2014) compared traffic crashes injury severity between Maryland and Utah by using CODES data. Data from 2006 to 2008 retrieved from the highway patrol, hospital, and trauma centers which yielded samples of 735K and 436K observations respectively in Utah and Maryland. The data were linked by using a probabilistic linking method. Results indicated that 50% of all injured persons in Maryland were coded as level B or more severe. For

Utah, this figure was 40%. Analysis of health data indicated that the AIS score also observed some level of fluctuation over time within states. Additionally, MAIS scores were more comparable between the states, and the distribution of MAIS was much more comparable between states. Maryland had approximately 85% of hospitalized injured cases coded as MAIS = 1 or less severe. In Utah, this percentage was close to 80% for all three years. Burch *et al.* (2014) also recommended the AIS scale as a more reliable measure of injury severity, and linking crash and hospital data for more accurate evaluation of injury rates.

Rosman and Knuiman (1994) linked police crash reports to hospital (n = 18,544) data (n = 3,980) by using a probabilistic method in Western Australia. They concluded that police reported the level of severity for non-fatal injuries was often inaccurate. Aptel *et al.* (1999) studied traffic crashes on the French Island of La Réunion in the Indian Ocean by comparing police records and hospital data and concluded that police report overestimated injury severity. However, the authors failed to capture 13% of the traffic fatalities and also reported that police severity grades were not in agreement with the actual hospitalization lengths.

Lopez *et al.* (2000) in Australia reported that police injury classification was correct in 78% of cases. Crashes resulting in high injury severity were more likely to be classified correctly than those resulting in less severe injury. Male injury victims were more likely to be correctly classified than females. The authors also concluded that injury classification rates were not influenced by trauma location, road user type, and number of vehicles involved. Tsui *et al.* (2009) studied misclassification in China by comparing hospital data and police crash reports. The authors concluded that police report injury grading diverged noticeably from the definition of hospital stay and police remarkably overestimated the injury severity. Moreover, the authors concluded that injury severity, victim age and position of the victim in traffic crashes had significant associations with injury misclassification.

To conclude, correlation between police evaluation of injury severity and the corresponding level of injury in a health facility varies. Several factors such as study area, sample size, and target population affect this correlation. Linking police crash report and health outcome data would provide a more accurate evaluation of injury severity. A first step in a linkage process would be to evaluate disparities between police crash data and hospital data, subject to local or regional norms, before proceeding with further analysis.

Factors Correlated with Injury Severity

One application of data linkage is determining a more accurate correlates of injury severity for traffic crash victims. Instead of using the KABCO scale, as reported by police, the medical outcome of traffic crashes could be used for more detailed injury severity investigations. In one study, Tarko and Azam (2011) used data from the Indiana crash database, of which 4,822 pedestrians' crashes were linked to medical records. They investigated the impact of selectivity bias in linked police-hospital data. Selectivity bias happens when a person with an injury perceived as more severe (regardless of the accuracy) is more likely to be directed to a hospital and as a result linked with medical data in comparison with a person with less severe injuries (Tarko and Azam 2011). The authors confirmed the presence of the sample selection issue, underreporting low-severity crashes. The selectivity bias is considerable in predictions of low injury levels. Instead of using KABCO, they used the MAIS scale for injury severity. The authors used a bivariate ordered Probit to investigate factors affecting injury severity and concluded that male and older pedestrians are particularly exposed to severe injuries. Rural roads and high-speed urban roads are dangerous for pedestrians, particularly when crossing such roadways. Crossing a road between intersections was found to be particularly dangerous behavior. The size and weight of the vehicle involved in a pedestrian crash were also found to affect the pedestrian injury level.

Cook *et al.* (2000) also investigated factors affecting the medical outcomes (e.g., injury severity) of drivers over the age of 69 involved in traffic crashes by probabilistically linked hospital discharge data and police crash reports data between 1992 and 1995 in Utah (n = 14,466). They reported that older drivers were less likely to experience crashes

involving drug or alcohol use and high speed. However, older drivers were more than twice as likely to experience crashes involving a left-hand turn than were younger drivers. Also, older drivers were more likely to be killed or hospitalized than younger drivers. Among belted drivers, older drivers were nearly seven times more likely to be killed or hospitalized than younger drivers.

Using health outcome data as an index for evaluation of crash injury severity would provide a better assessment of the association between road design, traffic characteristics, and injury severity. The data linkage should be applied in studies to evaluate factors affecting injury severity such as estimating crash modification factors to get a more accurate of the impact of implementing a countermeasure on the expected number of crashes or severity on a road or intersection.

Underreporting of Traffic Crashes

Another challenge in addressing traffic crashes, particularly for vulnerable road users, is the problem of underreporting. For several reasons, it is likely that crashes with injury outcome are not recorded in police crash reports. This may be attributed to police minimum criteria for reporting a traffic crash or unwillingness of the road users to report a crash.

Dhillon *et al.* (2001) in a study in Long Beach, CA used data from two trauma hospitals (n = 474) and one non-trauma hospital (n = 1015) to investigate child pedestrians and bicyclists crashes. By using a capture-recapture model, they evaluated the degree of overlap between police and trauma data. They reported that 80% of hospital-reported cases were captured in the police database, while only 37% of police-reported cases were captured in the hospital database. In addition, hospital sources identified younger children, more Asian and Hispanic children, fewer bicyclists, and fewer African-American children compared with police sources.

Agran *et al.* (1990) studied traffic crashes of children under the age of 15 (i.e., pedestrian or bicyclists) in Orange County, California. They used California Highway patrol and hospital monitoring system data from April 1 to September 30, 1987, to evaluate underreporting. In a conservative estimate, they reported that police underreported 20% of pedestrian crashes and 10% of bicyclists crashes. One explanation for this number could be police agency reporting requirements (e.g., crash cost) and the fact that police are not likely to report non-motorized incidents on the roads (e.g., single bicycle crashes).

Sciortino *et al.* (2005) conducted a study in the San Francisco area to evaluate underreporting of pedestrian crashes. They used Statewide Integrated Traffic Reporting System (SWITRS) data, (n = 1991) and records of pedestrians treated at San Francisco General Hospital (SFGH, n = 1323) between 2000 and 2001. By using bivariate statistics, logistic regression and mapping they found that police collision reports underestimated the number of injured pedestrians by 21% (531/2442). Moreover, they concluded that some groups are more likely to have crash reports. Based on their reports, in cases of pedestrians treated at SFGH, African-Americans were less likely than whites (odds ratio = 0.55, p-value$\leq$0.01), and females were more likely than males (odds ratio = 1.5, p-value$\leq$0.01) to have a police collision report.

In another study, in Queensland, Australia, Watson *et al.* (2015) probabilistically linked 2009 data of the Queensland Road Crash Database (QRCD), the Queensland Hospital Admitted Patients Data Collection; (QHAPDC), Emergency Department Information System (EDIS), and the Queensland Injury Surveillance Unit (QISU). The number of road crash injuries was approximately 28,000, with about two-thirds not linking to any record in the police data. Moreover, the results also showed that underreporting was more likely for motorcyclists, cyclists, males, young people, and injuries occurring in remote and inner regional areas.

In an early attempt in Australia, Rosman and Knuiman (1994) linked police crash reports (n = 18,544) to hospital data (n = 3980) by using a probabilistic method. The link rate from hospital to police database for their study was

64%. The linkage rate was lowest for motorcyclists in single-vehicle crashes (29%) and highest for motor vehicle crashes (79%). They concluded that the linkage rate increased as the severity of the crashes increased. Moreover, they reported that linkage rate was lower for some ethnicities, and that the underreporting issue was a considerable problem, particularly in traffic crashes resulting in less severe injuries.

Lopez *et al.* (2000) in another study in Australia linked 1997 traffic crash injuries at one trauma center (n = 497) to police reports. Results indicated that only 82% of hospital crash injuries had a corresponding record in the police database. Older individuals and females were more likely to have a police report. However, motorcyclists had the lowest matching rate among road users, followed by pedestrians.

In a study in New Zealand, Wilson *et al.* (2012) investigated the underreporting problem in motorcycle crashes by using data from 2000 to 2004. They used a probabilistic linkage to link national hospital discharge records (n = 2685) with police traffic crash reports. Only 46% of cases could be linked to a police record; 60% of the serious injuries and 41% of the moderate injuries. The authors reported that the level of underreporting of cases involving serious threat-to-life injuries was alarming. Serious injury cases were less likely to be linked if only one vehicle was involved, or the injured riders and passengers were younger than age 20 or spent less than one week in the hospital following the crash. For moderate injury cases, there were also differences in linkage by injured body region and month of crash. The under-representation of young riders, and of a single vehicle and moderate injury crashes, suggests that the low linkage was most likely due to underreporting of crashes to the police.

In another study in Sussex, England, Cryer *et al.* (2001) linked non-fatal injury hospital admission data (n = 2666) to police road traffic accident reports for traffic crashes between April 1995 and March 1998. Due to limitations of their data, they used a manual method for linking the data based on the patient's name, date of crash/admission, and place of occurrence/admission. The overall linkage rate was 61%. The linkage rate was much lower for cyclists (31%) than for other road users: 67% for vehicle occupants, 69% for motorcyclists, and 72% for pedestrians. Cryer *et al.* (2001) used non-slight injury and non-fatal serious injury for their analysis. There were higher proportions of males among the police Road Traffic Accident (RTA) non-slight injury casualties than among the hospital admissions. The authors concluded that police reporting rates varied by age, road user type and injury severity. For the linked data, there was a smaller proportion of children under the age of 16 than in the hospital admissions data. There were smaller proportions of children and larger proportions of adults ages 16-64 among the police RTA non-slight injury casualties than among the hospital admissions. The proportions of casualties in each road user group appeared similar for the linked database and the hospital admissions data, with the exception of cyclists and motorcycle riders. For police RTA non-slight injury casualties relative to hospital admissions, there were lower proportions of cyclists, particularly among children, a higher proportion of motorcyclists, a higher proportion of young and adult drivers between the ages of 17 and 59, and a lower proportion of car passengers. For non-fatal serious injury, defined either using length of hospital stay or nature of the injury, when hospital admissions and the linked data were compared for the variables of age and gender, the proportions of males and females and age group categories were approximately the same.

Aptel *et al.* (1999) studied traffic crashes in La Réunion by comparing police records and hospital data. They reported that the police database was not representative of the crashes that occurred. Indeed, findings indicated that reporting rates varied according to the type of vehicle involved, the time of the crash and whether a physician was in charge of first aid.

Reviewing road safety literature indicates that underreporting of traffic crashes strongly correlated with study area and road user types. Underreporting is most significant with non-motorized road users, motorcycles, minor crashes, and is different between socioeconomic groups. Accordingly, to better understand this issue, it is necessary to compare the databases of crash reports by police and health data.

Substance Abuse and Motor Vehicle Crashes

Police officers have several ways to identify drivers who are under the influences of alcohol or drugs. However, due to the nature of roadside tests for such substances, officers can send the test samples to the hospital for further analysis. As a result, the police reports for most cases usually report pending status for drivers suspected for driving under the influence.

Miller *et al.* (2012) used a capture-recapture method to evaluate the underreporting of alcohol involvement in traffic crashes in the United States. They analyzed 550,933 CODES driver records from 2006–2008 police crash reports probabilistically linked to hospital inpatient and emergency department (ED) discharge databases for Connecticut, Kentucky, Maryland, Nebraska, New York, South Carolina, and Utah. Findings indicated that police correctly identified 32% of alcohol-involved drivers in non-fatal crashes and 48% in injury crashes. Police in all the states (excluding Kentucky) reported 47% of alcohol involvement for cases treated in EDs and released and 39% for admitted cases. However, hospitals reported 28% of involvement for ED cases and 51% for admitted cases. Police also reported alcohol involvement for 44% of those for whom hospital records reported alcohol-involvement, while hospitals reported alcohol involvement for 33% of those who police reported were alcohol-involved. Police alcohol reporting completeness rose with police reported driver injury severity. Analysis also indicated that 62% of alcohol involved-drivers were reported in at least one system (Miller *et al.* 2012).

In a study conducted in British Columbia, Canada, injured drivers who were admitted to a local trauma center or treated in the Vancouver General Hospital emergency department (1999–2003; n = 2,410) were linked to their corresponding police collision reports. The match rate in this study was 73.5%. Findings indicated that 35.6% of the drivers had a positive value of Blood Alcohol Content (BAC) and 30.5% of the drivers had BAC greater than 0.05%. Comparison of the two databases indicated that police documented alcohol involvement in 72% of injured drivers with BAC $\geq$ 0.05 percent. Police documentation of alcohol involvement was more common at higher BAC levels, in nighttime or single-vehicle crashes, for drivers who committed traffic violations or drove unsafe, and for drivers with a prior record of impaired driving (Brubacher *et al.* 2013).

Analysis of crash statistics in California between 2005 and 2014 indicated that 30,796 fatal motor vehicle collisions resulted in 33,775 fatalities. Drugs were involved in 19% (5,734) of these fatal collisions and alcohol was involved (i.e., blood alcohol content (BAC) > 0.01g/dL) in 27% (8,348). Analysis demonstrated that, although alcohol-involved crashes decreased from 2009 to 2014, drug-involved collisions continued to increase steadily, with an increase from 18% in 2009 to 21% in 2014. Among 19,543 drivers, 43% of all drivers involved in a fatal collision were tested for drugs by blood, urine, or other tests. At least one drug type was found in 5,821 drivers (30% of those tested) involved in 5,734 fatal collisions. The percentage of drug-positive drivers among those who were tested has increased steadily since 2007 from 26% to 37% in 2014. Drivers who tested positive for at least one drug were more likely to be male (80%) with a mean age of 37 years, 45% tested positive for two or more drugs, and 40% also had alcohol in their blood. Drivers between the ages of 45 and 64 (25% of all drug-positive drivers) had the highest portion of positive drug results followed by groups between the ages of 25 and 34 years (24%) (CMOD 2016).

Evaluation of Safety Equipment

Several studies used health outcome databases and traditional police crash reports to evaluate the effectiveness of safety equipment, including seat belts and helmets.

Seat Belt Use

Olsen *et al.* (2010) linked 1999 through 2004 data from motor vehicle crashes and emergency department from the Utah Health Data Committee/Office of Healthcare Statistics, using the probabilistic linking method. The study only considered child occupants between the ages of 0 and 12 traveling with a driver over the age of 21 with a known level of restraint use. They reported that child passengers of restrained drivers are more likely to be restrained

themselves, and that restrained drivers are less likely to sustain severe injuries in crashes than unrestrained drivers. Moreover, their statistical models showed a decreased risk of being evaluated in the ED for children riding with restrained drivers compared with children riding with unrestrained drivers. The authors also identified contributing factors to the association between driver restraint use and a decreased risk of ED evaluation for child passengers. These findings suggest that children are at a decreased risk of requiring a medical evaluation when riding with a restrained driver in a crash.

Han *et al.* (2015) also examined the relationship between seat belt use and injuries in terms of the body region and nature of injury among drivers over the age of 15 involved in motor vehicle crashes (n = 10,479). The data used in this study were from the Nebraska Crash Outcome Data Evaluation System (CODES) database for the years 2006–2011. The authors used Strategic Matching LinkSolv software to link both data sets with probabilistic record linkage techniques. The linkage variables included first name, last name, gender, date of birth, date of the crash, and location of the crash. The unbelted drivers sustained higher risk of brain injury in comparison with belted (10.4% vs. 4.1%). Moreover, the proportion of the head, face, and neck injury in unbelted drivers (29.3%) was higher than among belted drivers (16.6%). However, the proportion of the unbelted drivers who sustained spinal injury was lower for unbelted drivers in comparison with belted drivers (17.9% vs. 35.5%). The authors also reported that seat belt use was associated with decreased medical costs for traffic crashes.

In another study, Han *et al.* (2017) used data from Nebraska between 2004 and 2013 (n = 90,716). The findings indicated that the hospital cost due to the traffic crashes for motor vehicle occupants not wearing seat belts was significantly higher than for occupants wearing seat belts ($2,819 vs. $1,597). Moreover, they reported that in addition to seat belt use, race, gender, age, type of crash, time of crash, speed limit at the crash, alcohol-impaired driving, year of crash and type of health insurance also had a significant association with hospital cost.

Helmet Use

Cook *et al.* (2009) investigated the relationship between motorcycle helmet use and motorcycle crash outcomes (i.e., injury types and hospital charges). The authors used CODES data for motorcyclists in 18 states between 2003 and 2005. Their findings indicated that 57% of motorcyclists were helmeted at the time of the crash and that 43% were non-helmeted. For both groups, about 40% were treated at hospitals or died following the crash. A total of 6.6% of unhelmeted motorcyclists suffered moderate to severe head or facial injuries compared with 5.1% of helmeted motorcyclists. Fifteen percent of hospital-treated helmeted motorcyclists suffered traumatic brain injury (TBI) compared with 21% of hospital-treated unhelmeted motorcyclists. Almost 9% of unhelmeted and 7% of helmeted hospital-treated motorcyclists received minor to moderate TBI. More than 7% of unhelmeted and 4.7% of hospital-treated helmeted motorcyclists sustained severe TBI. The authors also reported that median costs for hospitalized motorcyclists who survived to discharge were 13 times higher for those incurring a TBI compared with those who did not sustain a TBI ($31,979 versus $2,461). Over 85% of hospital-treated motorcyclists without a TBI were discharged, compared with 56% of motorcyclists with severe TBI. Motorcyclists admitted to the hospital with TBI were more likely to die, be discharged to rehab, or transferred to a long-term care facility. While 17% of all hospital-admitted motorcyclists had TBI, they accounted for 54% of all admitted riders who did not survive.

Analysis of bicyclist crashes for riders between the ages of 0 and 17 in California using CMOD system indicated that young bicyclists treated at the emergency department (ED) or hospitalized for their injuries were significantly less likely to sustain a traumatic brain injury if they were wearing a helmet than if they were not (11% vs.17%) (CMOD 2013).

Spatial Linking Examples

The literature of the road safety provides abundant examples of spatial linking different databases to police databases. Due to the focus of case studies number 3, 4, and 5 on aggregate models, in this section we provided a brief review of aggregate crash prediction models.

Aggregate Crash Prediction Models

Aggregate crash prediction models (ACPM) assist agencies in identifying locations with high risk of crashes and their relationship with socio-demographic variables and transportation infrastructure. ACPMs enable researchers to identify hot spots to prioritize for safety considerations in medium and long-term transportation planning.

The following studies describe the traffic safety of specified geographic areas. Aggregate crash prediction models have been applied to several different spatial units from fine geographical areas such as census tracts (Ukkusuri *et al.* 2011, Wang and Kockelman 2013), block groups (Huang *et al.* 2016), and traffic analysis zones (i.e. TAZs[2]; (Hadayeghi *et al.* 2010, Abdel-Aty *et al.* 2011, Pulugurtha *et al.* 2013, Dong *et al.* 2014, Dong *et al.* 2015, Xu and Huang 2015) to vast geographic areas such as regions (Washington *et al.* 1999), and counties (Miaou *et al.* 2003, Aguero-Valverde and Jovanis 2006).

Several factors predict the relationship between socio-demographic, road infrastructure characteristics and road safety (e.g., crash frequency). Researches find that population density (Huang and Abdel-Aty 2010), age cohorts (Aguero-Valverde and Jovanis 2006, Hadayeghi *et al.* 2010, Pirdavani *et al.* 2012, Dong *et al.* 2015), incomes (Pirdavani *et al.* 2012, Xu and Huang 2015), employment (Quddus 2008, Hadayeghi *et al.* 2010), trip generation and distribution (Naderan and Shahi 2010, Abdel-Aty *et al.* 2011, Dong *et al.* 2014, Dong *et al.* 2015), and the number of persons with driving licenses (Pirdavani *et al.* 2012) have significant association with crash frequency at the zonal level. In addition, street built environment characteristics, such as intersection density (Huang and Abdel-Aty 2010, Xu and Huang 2015), road lengths with different speed limits (Abdel-Aty *et al.* 2011, Siddiqui *et al.* 2012), road lengths with different functional classifications (Quddus 2008, Hadayeghi *et al.* 2010), junctions and roundabouts (Quddus 2008), traffic patterns such as traffic flow and vehicle speed (Quddus 2008, Hadayeghi *et al.* 2010, Pirdavani *et al.* 2012), environment conditions such as total precipitation/snowfall, and the number of rainy/snowy days per year (Aguero-Valverde and Jovanis 2006), land use (Pulugurtha *et al.* 2013), and vehicle miles travelled (VMT) (Pirdavani *et al.* 2012) have association with safety at the zonal level.

Complete Picture of Traffic Crashes

To develop a complete picture of traffic crashes, a framework based on the Haddon matrix (Haddon Jr 1968) and road safety pyramid (Hydén 1987, Koornstra *et al.* 2002) was employed. The Haddon matrix is an analytical tool that helps to identify risk factors related to human, environment, and vehicles that are associated with traffic crashes. The road safety pyramid also provides information about road safety event frequency and severity.

The Haddon matrix provides information about the phases of a traffic crash and potential contributing risk factors at each phase. Haddon classified crash phases into three independent phases; pre-crash, crash, and post-crash events. Figure 1 presents an example of the Haddon matrix applied to traffic safety (Haddon Jr 1968, Barnett *et al.* 2005). Each phase depicts different risk factors. For example, the pre-crash phase includes the risk factors (e.g., exposure, behavior, and infrastructure) leading to traffic crashes. The post-crash phase represents the database that describes first-responder response to traffic incident and subsequent effects of injury treatment on overall health outcomes.

	Factors	

[2] Traffic Analysis Zone

Phase		Human	Vehicle And Equipment	Physical Environment	Social Environment
Pre-Crash	Crash Prevention	Information; Attitude; Behavior	Roadworthiness; Lighting; Braking	Road Design; Speed Limit; Pedestrian Facilities	Law Enforcement, Cultural Norms, and Attitudes
Crash	Injury Prevention During Crash	Use Of Restraint; Impairment	Occupant Restraint; Other Safety Device	Crash-Protective Road Side Objects; Forgiving Infrastructure	Good Samaritan Laws; Social Responsibility
Post-Crash	Life Sustaining	First Aid Skills; Access To Medics	Ease Of Access; Fire Risk	Rescue Facilities; Congestion	Insurance Coverage; Social or Family Support

Figure 1. An example of a Haddon Matrix application in road safety

Figure 2 presents a road safety pyramid (Hydén 1987) that describes the relationship between severity and frequency of elementary events in traffic. Hydén (1987) suggested the existence of some severity dimension common for all events[3] in traffic and proposed a model describing the relation between event severity and frequency (Figure 2). The scale ranges from traffic crash to undisturbed passages. Near-miss surrogate safety events require a rapid, evasive maneuver by the vehicle, pedestrian, cyclist, or animal, to avoid a crash (Guo *et al.* 2010). According to this model (Figure 2), the higher the severity (represented by the vertical position in the pyramid), and the lower the frequency (the volume of the pyramid slice at this height) of the events (Laureshyn *et al.* 2010). Several studies used near-crashes as a surrogate for traffic crashes to analyze safety of road users (Guo *et al.* 2010, Poulos *et al.* 2011). In general, "there is a strong relationship between the frequencies of contributing factors to crashes and for near-crashes." (Guo *et al.* 2010)

Considering the main goal of this study, which is creating a complete picture of traffic crashes by linking different databases, the Haddon matrix and road safety pyramid were used as a basis to create a framework for linking the databases. The aim is to classify and use datasets that contribute to understanding the spectrum of contributing factors and outcomes associated with a system-view of crashes. Moreover, the present study seeks to understand how different databases can be linked to a traffic incident (i.e., conflicts). The top layers of the pyramid presented in Figure 2 are road safety performance indicators and are elements of interest for this study. The police crash database is the leading source for estimating the crash outcome (i.e., number of killed, injured, and property damage only crashes). Moreover, the pyramid also includes layers regarding near miss (serious conflicts), or other safety-critical event (slight or potential conflicts). Therefore, in addition to the police crash database, near-crash data (i.e., serious, slight and potential conflicts) information, to the extent that it exists, is included, as a complementary safety outcome for a complete picture of a traffic crash (Figure 3). Since police crash report databases are the core source for analyzing road safety, the present study seeks to describe the connection between other databases and police crash report databases. Alternatively, near-miss databases are emerging with improved naturalistic methods and crowdsourced datasets. These datasets have many of the features and linking potential found in crash

[3] An encounter (a simultaneous arrival in a certain limited area) between two road users can be seen as an elementary event in the traffic process that has a potential to end up in a collision.

databases, but have yet to be systematically integrated into many safety analyses. Different relationships between secondary databases can be defined. However, those connections are beyond the principal focus of this study.

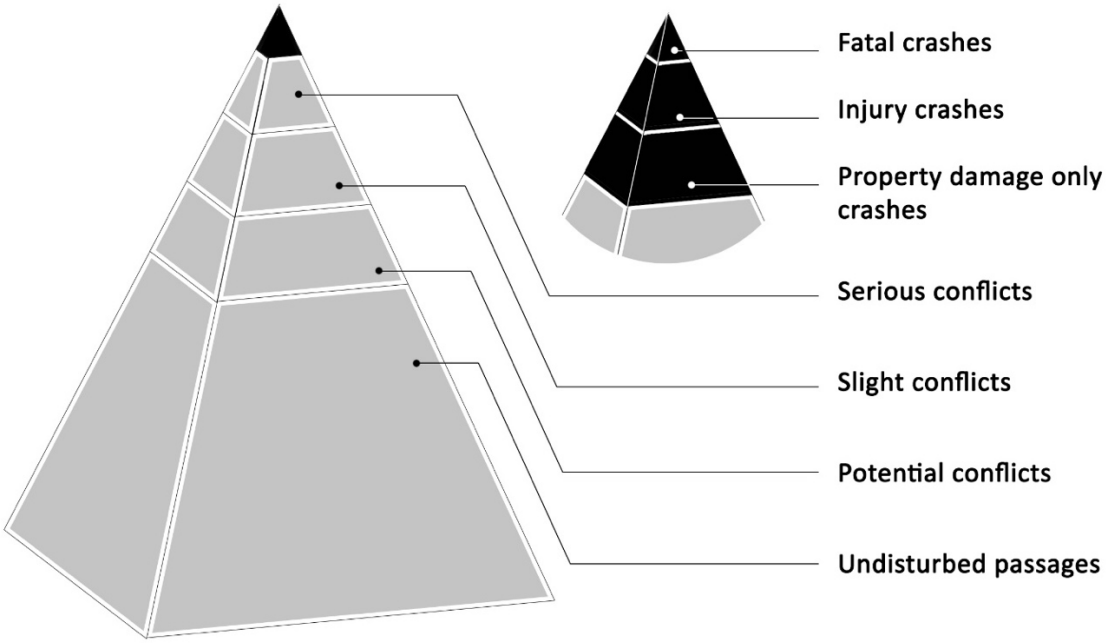

Figure 2. Road Safety Pyramid (Hydén 1987)

Regarding crash timeline, the databases can be classified as either pre-crash and post-crash for those events, and risk factors that occur before and after the crash incident. In addition to pre-crash and post-crash databases, there are databases that explain the general environment or geographic areas of traffic crashes (i.e., crash environment)—the data elements in these databases are not affected by traffic crashes and are exogenous to crash events and usually aggregated at temporal or spatial level. This class of data also can be classified into two sub-classes. The first explains sociodemographic information at aggregate level, while the second describes the built environment and traffic behaviors.

Figure 3 presents databases and classes are used to illustrate the complete picture of traffic crashes for the present study. A complete list of databases in each phase is described in the **Error! Reference source not found.** Furthermore, the databases were discussed from several points of view. First, we discussed the accessibility to the databases and whether the databases include Personally Identifiable Information (PII) or Protected Health Information (PHI) data elements. Second, we discussed different layers that were available in each database and common data elements in each layer. Last, the consistency of the measurement and forms were discussed across the states. The following sections include a concise list with abbreviated descriptions of potential databases that can be linked to develop a more complete picture of crashes.

Data quality is an important indicator that is important in any linkage method. Data quality benchmarks and definitions vary by dataset, and acceptable levels of quality vary by application. Some data quality indicators are based on subjective observations and are subject to human error or miscategorization (e.g., injury severity). Other quality indicators are technical and subject to limitations or specifications of technology (e.g., GPS latency or resolution). Notwithstanding the importance of considering data quality when merging datasets, this report does not assess the quality of mainstream or emerging datasets in isolation. Indeed, any dataset that is used for safety

analysis must do so with a clear understanding of the limitations and appropriate application of the data. Many applications are solely focused on assessing quality of data (e.g., underreporting).

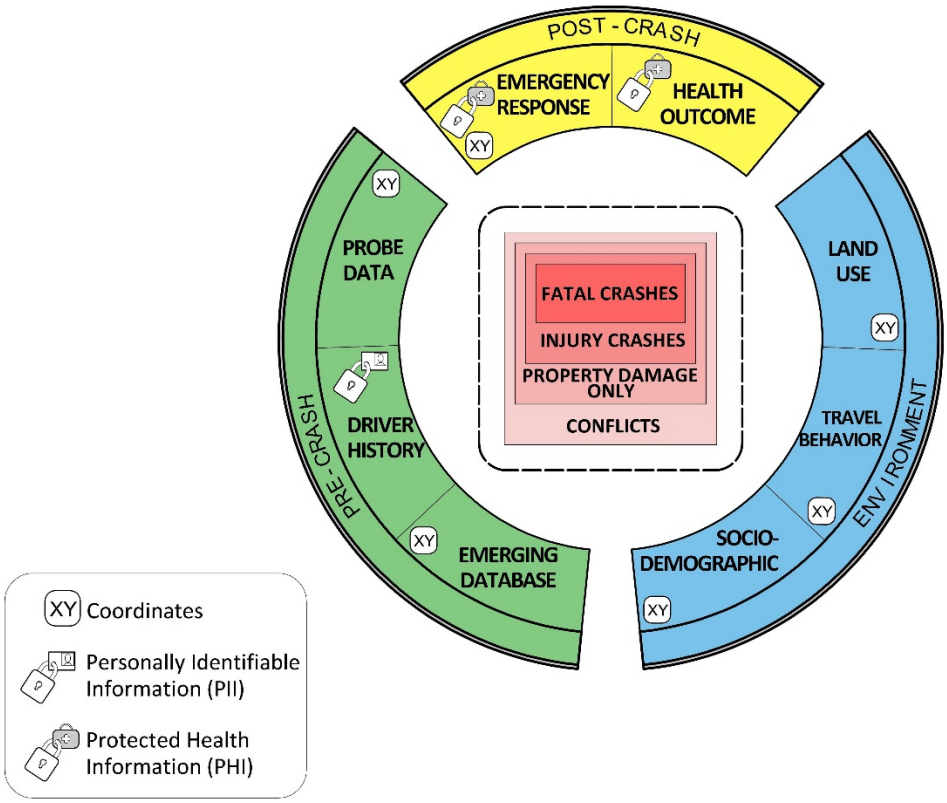

Figure 3. Complete Picture of Traffic Crashes

Crash Incident

Police Crash Reports

Police crash report data is the main core of road safety analysis and is collected by law enforcement officers at the crash scene. The database includes information about crash, individual, roadway, and vehicle involved at crash scene which can be used for variety of analysis. Police crash report databases are fairly consistent across the United States and follow a similar format which is based on MMUCC (Model Minimum Uniform Crash Criteria) (MMUCC 2012). Police crash databases generally include PII elements that require procedures to protect confidentiality of the police report. These elements are protected under the Federal Driver's Privacy Protection Act (DPPA) (18 U.S.C.A. 2721). Within agencies, data use agreements generally dictate terms of use of such data for carrying out its functions. Researchers, working with the owners of the data, are required to comply with Institutional Review Board protocols for the protection of human subjects if utilizing data with PII. The data elements that are available in police crash reports enable researchers and safety practitioners to spatially and probabilistically join to other groups of the databases. The linking methodology relies heavily on the level of access to PHI and the nature of the target datasets.

Fatality Analysis Reporting System (FARS) database

The Fatality Analysis Reporting System (FARS) is a nationwide database provided by NHTSA that provides annual data regarding fatal injuries sustained in motor vehicle traffic crashes for Congress and the American public. FARS database records information about the crash, vehicle, and individuals involved in fatal crashes. The database is

consistent across the nation. The database does not include any PII information, and it is available online for use by the public. The data elements in FARS database are similar to police crash database (e.g., crash coordinates) and enable researchers to perform spatial or probabilistic join to other databases.

Post-crash Data

Post-crash data consists of two data sub-classes: health outcome and emergency response. The health data sub-class consists of a database that describes the health outcomes of traffic crashes for injured individuals. Due to the nature and severity of the injury individuals might use different health facilities. The other subcategory includes databases that describe response (i.e., emergency response) to traffic incidents.

National Vital Statistics System (NVSS) – Mortality

Mortality data from the National Vital Statistics System (NVSS) are a fundamental source of demographic, geographic, and cause-of-death information. This is one of the few sources of health-related data that is comparable for small geographic areas and available for a long time period in the United States. The data are also used to present the characteristics fatalities, to determine life expectancy, and to compare mortality trends with other countries.

To access each state's data with identifiers, researchers and practitioners must contact state Department of Health or other agencies that are in charge of gathering mortality data. There are several layers in this database. The most important layers are patients' personal information, accident layer, insurance layer, admission layer and health outcome (TnDH 2014). In addition to National database, individual states also may provide databases regarding individual visiting emergency room visit. For more information, researchers need to contact local authorities and departments to access databases.

Hospital Discharge Data System (HDDS)

HDDS is a standard database created by the American Hospital Association (AHA) that records billing information of patients and can be used nationwide by institutional providers and payers for handling health care claims. Currently, this database uses the UB-04 data manual as a standard for billing paper institutional medical claims in the United States. The database includes information about patients' source of injury, health outcome (e.g., injury and cost of treatment), insurance information, and PHI. This database can be used to link to the police database to learn more about the injury outcome and social burden of traffic crashes. The HDDS database includes PHI. Any type of information regarding health status, provision of health care, or payment for health care which is created or collected by a covered entity by 45 CFR § 160.103 and can be linked to a specific individual is protected by HIPAA Privacy Rule provides federal protections. In order to access PHI data elements, one must follow a specific data agreement provided by the owner of database. The PHI elements that are available in HDDS enable researchers and safety practitioners to join to health outcome data and police crash reports probabilistically. The linking methodology highly relies on level of the access to PHI.

Ambulatory Surgical Center (ASC)

Ambulatory surgery centers (ASC) (also known as outpatient surgery centers or same day surgery centers) are healthcare facilities where surgical procedures not requiring an overnight hospital stay are performed. These types of surgeries are usually of a less complicated nature than those requiring hospitalization. The ASC database captures patient records using the CMS-1500 or UB-04 form. The consistency of the Data across States requires further investigation (TDH 2013). There are several layers in the ASC database. The most important layers are patients' personal information, insurance, and health outcome. However, depending on whether the CMS 1500 or UB-04 form is used, the variables in this layer vary. The data elements in this database include PHI, insurance and admission information.

Since this database includes PHI, access requires additional procedures. This database could be used to link to police crash reports to get a better estimate of less severe traffic crash injuries. The methodology for linking ASC database to police crash report database relies heavily on the level of access to the PHI.

Trauma Center Data

A trauma center is a hospital that is equipped and staffed to provide care for patients suffering from major traumatic injuries as a result of traffic crashes, gunshot wounds, and falls, among others. The National Trauma Data Bank® (NTDB®) is the largest aggregation of U.S. trauma registry data assembled. Participation is voluntary and is one of the leading performance improvement tools of trauma care (NTDB 2018). Since its inception in 1994, many researchers have published work based on data from the NTDB. In 2008, the NTDB implemented the National Trauma Data Standard (NTDS) to standardize data collection across all reporting hospitals. Currently, the NTDB contains detailed data on over 2.7 million cases from over 900 U.S. trauma centers (NTDB 2018). NTDB® research data may be used for informational and research purposes with approval from the American College of Surgeons (ACS) Committee on Trauma. Permission to use the NTDB dataset is required via an online data application form found on their website.

There are no general and all-inclusive guidelines for developing each of the trauma registry elements. Currently, each institution designs its trauma registry based on its needs or to meet mandatory state or city requirements (Zehtabchi *et al.* 2011). The NTDS provides definitions for variables and response codes. Institutional and state trauma registries often collect additional variables depending on their needs. Most trauma registries incorporate the following data: demographic information, mechanism of injury (external cause of injury codes), procedures, clinical diagnoses, based on International Statistical Classification of Diseases and Related Health Problems, length of stay, disposition, and in-hospital mortality. It may also include abbreviated injury scores (AIS), charges, payers, and information about complications and follow-up procedures if they occur at the same institution. Some important variables, such as longer-term mortality or functional outcomes, are rarely collected (Zehtabchi *et al.* 2011). The methodology for linking trauma center database to police crash report database relies heavily on the level of access to the PHI and the nature of local trauma center data.

Medical Insurance Claims – All-Payer Claims Database (APCD)

All-Payer Claims Databases (APCDs) are large-scale databases that systematically collect medical claims, pharmacy claims, and eligibility and provider files from private and public payers. ACPDs include information about healthcare prices, quality, and utilization. Moreover, the majority of states mandated ACPDs to report statutes requiring the submission of claims data to a state collecting agency. These databases could be used as sources to compare prices and utilization patterns throughout the complex healthcare system (Kelly and King 2017). The first statewide APCD system was established in Maine in 2003. By 2017, eighteen states had passed legislation and established APCDs. APCD systems collect data from existing claims transaction systems used by healthcare providers and payers. The information typically collected in an APCD includes patient demographics, provider codes, and clinical, financial, and utilization data.

Most states define the details about APCD data collection requirements (e.g., format and timing of submission, specifications of data elements, and thresholds for payers that are required to submit data) uniquely from other states (Porter *et al.* 2014). Each state assesses its own legislation to determine the most efficient way to design a flexible but comprehensive APCD as there is no universal model. States generally de-identify the data using encryption and statistical methods to mask the identity of the individuals in the database, though some states allow qualified users to access de-identified and research files. Each state must be contacted separately for access to the APCD database. The methodology for linking ACPD database to police crash database relies heavily on the level of access to the PHI in both databases.

Medical Expenditures Panel Survey

The Medical Expenditure Panel Survey (MEPS), which began in 1996, is a set of large-scale surveys of families and individuals across the United States, their employers and availability of health insurance, their medical providers (doctors, hospitals, pharmacies, etc.), and their use of nursing home services. MEPS collects data on the specific health services that Americans use, how frequently they use them, the cost of these services, and how they are paid for, as well as data on the cost, scope, and breadth of health insurance held by and available to U.S. workers. MEPS currently has four major components: The Household Component (HC), the Insurance Component (IC), the Medical Provider Component (MPC) and the 1996 Nursing Home Component (NHC). The Household Component provides data from individual households and their members; the Insurance Component is a separate survey of employers that provide data on employer-based health insurance; the Medical Provider Component is used to verify medical and financial characteristic information of medical events described in the HC survey from medical providers; and the 1996 Nursing Home Component comes from community sources with a minor amount of data collected from nursing home sources. The methodology for linking MEPS to other databases reliant on the available data and the nature of the secondary database. The MEPS could be used to calculate the social burden of traffic crashes.

Emergency Medical System (EMS) Data

The National Emergency Medical Services Information System (NEMSIS) is the national database that is used to store EMS data from the U.S. States. NEMSIS is a universal standard for how patient care information resulting from an emergency call for assistance is collected. NEMSIS is a collaborative system to improve patient care through the standardization, aggregation, and utilization of point of care EMS data at a local, state and national level. NEMSIS is a product of NHTSA's Office of EMS and in collaboration with the University of Utah is the host of the Technical Assistance Center (TAC). Traditionally EMS data has been housed in local trauma registry databases. More recently NHTSA has funded development of a national database for housing EMS data nationally, the National EMS Information System (NEMSIS). As of 2017, most states submit data to NEMSIS. (https://nemsis.org/). Access to the NEMSIS repository is through the University of Utah School of Medicine.

The primary elements of the NEMSIS database Version 2.2.1 are listed in Attachment A. The data dictionary for Version 2.2.1 can be found at https://nemsis.org/technical-resources/version-2/version-2-dataset-dictionaries/. Categories of variables are listed below. More detail can be added at each level to indicate the potential uses for each category.

Crash Environment Data

Travel Demand

Traffic analysis zone (TAZ) data and associated travel demand model data can be provided by any regional entity with an operational travel demand model. Travel demand models describe present and project future patterns of travel demand—the origins and destinations of trips, the modes used to reach those destinations, and the routes used for each individual trip.

Traffic analysis zone data is typically available from metropolitan planning organizations (MPOs), some states also manage travel demand models and may have such data at the state level. Travel demand models do not include personally identifiable information. Moreover, there are no consistent standards for Travel Demand Models (TDM), and these models' data elements varies by state. Often TAZ data includes elements from Census datasets, travel surveys, and employment surveys. Depending on the geographical unit analysis of the travel demand models, spatial joins can be used to link this database to other databases. Travel demand models provide valuable information about travel behavior at zonal and street segment levels which could be used to evaluate the association between travel habits, road infrastructures, and road safety.

Parcel Data

Parcel data describes the characteristics of individual land parcels. This data may include land use, land area, building area, building square footage, zoning, and land and building value information. Parcel data can be used to calculate land use distribution at the zonal or district level. It can also be used to calculate the level of mixed-use and the floor area ratio, or building density. Most jurisdictions in the United States have parcel data available in GIS form. The fields that are most typically available include taxable value for land, taxable value for buildings, land acreage, year structure built, and information about the most recent sale of the property including sale value, number of dwelling units and bedrooms. Less common fields include land use codes, zoning codes, and building area.

Parcel data are typically available only for the present year, but in some cases parcel databases are archived and available for previous years. One of the primary limitations of parcel data is that the fields available, as well as the quality of the data, can vary substantially by jurisdiction. Parcel data can be useful for fine-grain analysis of safety, identifying activity centers, understanding important local economic indicators such as land or building value, or identifying infrastructure improvements that have been made. Spatial joins can be used to link parcel data to police crash reports.

LODES

Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) data provide information on the residential location of workers, the workplace location of workers, and the flows between the two. The US Census synthesizes this data from multiple administrative sources including state unemployment records. The geographical unit of analysis for this data is the census block scale. LODES data are available going back to 2002, and most states are covered for most years. Out of 53 states and US territories, 50 were participating as of 2017. Personally identifiable information (PII) is not available from LODES as prevented by statutory requirements. Noise is added to the data in order to ensure that no PII is recoverable. LODES data include work area characteristics, residential area characteristics and origin-to-destination flows. Spatial joins can be used to link LODES database to police crash databases. This linking provides additional information regarding the work-related and residence-related activities at the location of the traffic crash.

Census Data

The US Census gathers data through the decennial census and the American Community Survey every year. The decennial census is an attempt at a complete count of the nation's residential population, but only covers age, race, household structure, and residence. The American Community Survey (ACS) covers a wide range of demographic and economic information for persons and households but is a sample and therefore cannot typically provide accurate counts at small levels of geography. The smaller the geography considered, the larger the margin of error is for ACS data. US Census data are available at many levels of geography, with the primary ones being state, county, census tract, census block group, and census block. lower levels of census geography nest within higher levels in the above scheme. US Census data is completely consistent across US states and territories. Therefore, the same data can be used nationwide. Regardless of the available census geographical unit, researchers and safety practitioners can use spatial join to link census data to police crash report or other databases. Linking census data in safety analysis provides information about the association between sociodemographic at zonal level variables and road safety patterns.

Highway Performance Monitoring System (HPMS) Data

The HPMS is a national program which includes inventory information for all of the nation's public roads as certified by the states' governors annually. HPMS contains data from 50 states and territories. This database is available online, and available free of charge. This database is available in shapefile format. Each state is required to annually

report and furnish all data per the reporting requirements that is specified in this HPMS Field Manual. This database has several layers—the layers are inventory, route, traffic, geometric, pavement, and special network (e.g., strategic highway network, national highway network).

HPMS is one of the most important databases in road safety analysis that provides information about vehicle miles traveled on the transportation network and road classifications. A spatial join is the main method to link HPMS to other databases. HPMS data are used to calculate following performance measures (DOT 2017):

- Rate of fatalities in 23 CFR 490.207(a)(2)
- Rate of serious injuries in 23 CFR 490.207(a)(2)
- Percentage of pavements of the Interstate System in Good condition in 23 CFR 490.307(a)(1)
- Percentage of pavements of the Interstate System in Poor condition in 23 CFR 490.307(a)(2)
- Percentage of pavements of the non-Interstate NHS in Good condition in 23 CFR 490.307(a)(3)

Percentage of pavements of the non-Interstate NHS in Poor condition in 23 CFR 490.307(a)(4)

- Rate of fatalities in 23 CFR 490.207(a)(2)
- Rate of serious injuries in 23 CFR 490.207(a)(2)
- Percentage of pavements of the Interstate System in Good condition in 23 CFR 490.307(a)(1)
- Percentage of pavements of the Interstate System in Poor condition in 23 CFR 490.307(a)(2)
- Percentage of pavements of the non-Interstate NHS in Good condition in 23 CFR 490.307(a)(3)
- Percentage of pavements of the non-Interstate NHS in Poor condition in 23 CFR 490.307(a)(4)

National, state and local transportation decision-making may also use the HPMS database in the transportation planning process for decision making to analyze trade-offs among the different modes of transportation.

Model Inventory of Roadway Elements

Model Inventory of Roadway Elements (MIRE) is a database provided by Federal Highway Administration (FHWA) that maintains listings of roadway inventory and traffic elements which are critical to safety management. The intention behind MIRE was to provide a guideline to help transportation agencies improve their roadway and traffic data inventories. MIRE provides a basis for a standard of what can be considered a good/robust data inventory and helps agencies move toward the use of performance measures to assess data quality (FHWA 2018). There are 202 elements that comprise MIRE Version 1.0. The MIRE elements are divided among three broad categories: roadway segments, roadway alignment, and roadway junctions. Data are collected in different states based on the importance of the data elements. The priority ratings are broken down into two major categories: critical and value-added. Critical elements are those that are necessary for states to collect in order to conduct basic safety management analysis. However, value-added elements are those whose presence is beneficial but which are not crucial to using current versions of safety analysis tools.

Pre-Crash Data

DMV data

All state departments of motor vehicles (DMV) maintain information about driver's license and non-driver ID applicants in electronic databases that share a core set of data elements. These departments have the demonstrated capability to share information with other government agencies and link to other databases (BCJ 2009). Every state department of motor vehicles database is different, but share similar information, such as name, address, date of birth, phone number, social security number and, in some cases, medical or disability information. These databases are designed to be shared with pre-authorized users on demand (e.g., police, etc.) and can link the identification dataset with other databases (e.g., voter registration, automobile titles, etc.).

While pre-authorized federal, state and local agencies (e.g., Police, Immigration, etc.) can freely access DMV data, federal and state laws specify the circumstances in which access is provided to entities (government and others) that have not been authorized, and these rules vary by state. There are numerous laws that govern how DMV

information is disclosed. For example, the Federal Drivers Privacy Protection Act requires all States to protect the privacy of personal information contained in a person's motor vehicle record with certain exceptions defined in Title 18, United States Code Section 2721 (BCJ 2009). However, some states sell DMV data access to private-sector users who qualify for the privilege under federal and state laws. The list includes law firms; insurance companies; auto sales, service and towing companies; private investigators and security firms; and companies and nonprofit organizations that employ drivers. Each state department of motor vehicle (DMV) should be contacted in order to obtain DMV data.

Meteorological Data

Meteorological data is governed by the National Oceanographic and Atmospheric Administration (NOAA). Weather station data can be archived and accessed through various data portals (e.g., Open Weather Map), which includes forecasts and historical data including precipitation, cloud cover, temperature, and wind. One global standard applies to weather stations associated with aviation. METAR (Meteorological Terminal Aviation Routine) weather data refer to a well-defined format for reporting weather information to pilots and meteorologists. Raw METAR data is standardized by the International Civil Aviation Organization (ICAO), which allows it to be understood throughout most of the world (NWS 2018). Three types of weather reporting systems are used to collect data for METAR formatting: Automated Terminal Information Service (ATIS), Automated Airport Weather Observation System (AWOS) and Automated Surface Observing System (ASOS) data.

AWOS data collection systems are maintained and operated by state or local governments under the authorization and certification of the Federal Aviation Administration (FAA). However, ASOS systems are maintained and operated by a joint effort between the National Weather Service and Department of Defense (NOAA 2017). The ASOS systems provide weather information beyond that of aviation. Weather stations from around the world report weather conditions every hour in a unified format. The Aviation Weather Center saves up to seven days of this data. METAR data can be downloaded from the Aviation Weather Center website at https://aviationweather.gov/, but it is archived on several other portals. The data can be accessed through an interactive display or from their data server. METAR raw data contains information that can be grouped into four categories: Data Station Layer, Wind Data Layer, Weather Conditions Layer and Visibility and Sky Conditions Layer.

SHRP2 Project NDS Data

Congress created the second Strategic Highway Research Program (SHRP 2) in 2005 to address the challenges of moving people and goods efficiently and safely on the nation's highways. SHRP 2 addresses four strategic focus areas—the role of human behavior in highway safety; rapid renewal of aging highway infrastructure; congestion reduction through improved travel time reliability; and transportation planning that better integrates community, economic, and environmental considerations into new highway capacity. Between 2010 and 2013 a total of 3,400 drivers participated in the Second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study (NDS), which produced over 4,300 vehicle-years of naturalistic driving data. Data were collected from six sites across the nation, the largest of which were in Seattle, Washington; Tampa, Florida; and Buffalo, New York.

Participant vehicles were instrumented with a data acquisition system (DAS) that collected four video views (driver's face, driver's hands, forward roadway, rear roadway), vehicle network information (e.g., speed, brake, accelerator position), and information from additional sensors included with the DAS (e.g., forward radar, accelerometers) (Hankey *et al.* 2016). In order to both uphold the privacy of study participants and promote the use of the data in important research, a process was created for researchers to establish a data use license (DUL) under which terms they may work with the data. In many cases, Institutional Review Board (IRB) approval of a qualified researcher's plans for data use will be required due to the involvement of human subjects.

There are several types of data in the SHRP 2 project. The variables can be classified into several layers namely, participant assessments, vehicle information, continuous data, trip summary data, event data, cell phone data and roadway data. Additionally, detailed investigations of selected crashes are also available for researchers.

Safety Pilot Model Deployment Data

The Safety Pilot Model Deployment Data (SPMD) program is a research initiative containing real-world implementation of connected vehicle safety technologies, applications, and systems in everyday vehicles and multimodal driving conditions. U.S. Department of Transportation (USDOT) National Highway Traffic Safety Administration, Intelligent Transportation Systems Joint Program Office, Federal Highway Administration, Federal Motor Carrier Safety Administration, and Federal Transit Administration sponsored the SPMD program. Research Data Exchange (RDE) also provides SPMD data for consumption. The RDE provides data and documentation of connected vehicle for research projects, collaboration with other users, and comments on hosted data sets.

The SPMD program is a comprehensive data collection effort under real-world conditions, with multimodal traffic and vehicles equipped with vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication devices that used DSRC to communicate Basic Safety Messages (BSMs). BSMs contains information regarding vehicle operation (e.g., speed, location) at a frequency of 10 messages per second.

RDS provides text-based (non-video, non-audio) data of the SPMD accompanied by a downloadable data dictionary and metadata document that provides information to support its use. The RDE SPMD data environment includes six data sets:

- Two driving datasets, consisting of data acquired using two types of DAS—DAS1 and DAS2
- One BSM data set, consisting of data generated by equipped vehicles
- One RSE data set, consisting of BSMs received by RSEs and signal timing and curve speed warning messages transmitted by RSEs
- One weather data set, consisting of weather information for the time periods corresponding to data collection.
- One network dataset containing traffic count data from Ann Arbor.

Several algorithms were applied to vehicle trajectories to protect the identities of the SPMD participants'. These algorithms truncated trip trajectories to remove the trip origins and destinations. Moreover, data elements including personally identifiable information were removed from the datasets.

Crowdsourcing data

Crowdsourcing involves the gathering of data for a particular purpose from a network of collaborators. Transportation crowdsourcing leverages the combined data gathered by a group of people utilizing smartphone apps and/or website interfaces. Transportation crowdsourced data can be provided by third-party vendors and dedicated platforms, as well as by social media and the internet. Crowdsourcing data can be gathered for different purposes, for examples these data sets gather information about mapping and navigation (e.g., Open Street Map); detailed traffic performance data (e.g., Waze, HERE Technologies); ride-sharing and public transit data (e.g., UBER Movement); pedestrian and bicycling data (e.g., STRAVA), and roadway infrastructure conditions (e.g., StreetBUMP). Each of the examples provides crowdsourced data related to some aspect of transportation. Depending on the nature of crowdsourcing database, one can use them to learn about factors influencing road safety that are not offered in traditional databases. Additional details about crowdsourcing databases and their application in transportation can be found in Appendix A.

Conclusion and Recommendations for Future Directions

Police crash report are the main source of data concerning traffic safety. Although police crash databases have hundreds of variables, they are unable to describe the causes and consequences of roadway crashes fully. This reflects one of many known limitations of such databases. Data at the crash scene is usually collected by a police officer at the time of the crash. Since officers have many competing objectives and are not experts in all contributing factors, much of the collected data only explains a part of the crash story. Moreover, police officer estimations of the severity of traffic injuries and crash costs are not medically precise and require careful, objective consideration later.

In order to provide qualitative and accurate estimates of the data elements, and learn more about the data elements that are not captured by police officers in standard police crash databases, it is necessary to create a more complete picture of traffic crashes. To do so, it is crucial to understand which data elements exist in various data sources (examples are provided in Appendix A) to be able to perform data linking. The remainder of this document describes the essential applications of linking different databases.

To link police crash data to other sources, first, it is necessary to identify the elements that can be related to other data elements in another database. Secondly, based on the nature of the target database(s) and accessible data elements, the proper linking method must be chosen (e.g., spatial vs. probabilistic). For example, if the goal is to perform the linking at a disaggregate level (i.e., individual level), PII or PHI is required that creates barriers to performing analysis. However, databases that store geographically aggregated data use the spatial join method for linking databases.

Due to the classification of the databases (explained in the previous section), databases with similar characteristics are presented in one class. Accordingly, a similar methodology can be used to link each class to police crash data. Figure 4, which is based on Figure 3-Complete Picture of Traffic –, illustrates both linking methods and the main data elements layers that are needed to link each class of data to police crash data.

Police Reporting and Crash Environment

As illustrated in Figure 1, the crash environment layer includes databases that describe the crash environment including and not limited to the built environment, traffic conditions, sociodemographic environment, weather conditions and aggregate travel behavior patterns. This class of data describes characteristics of the environment, therefore, to perform a linking process, there is a need to spatially join police crash report data and the target(s) databases.

Police officers at crash scenes do not usually collect detailed information about the crash environment, and if they do, the information has a qualitative nature rather than quantitative. One example is road design features. For example, officers report whether a crash occurred at a curve or not; the information does not typically include the radii of the curve. To analyze the effect of different curve radii on crash rate of a specific segment, police crash data must be linked to a database that includes geometry features of the road (e.g., MIRE, HPMS). Another example of the data that police do not record is traffic count—to get a better estimate of traffic count, police crash data can be linked to HPMS data or data from other local sources that store higher resolution traffic data. Other examples of the different linking databases and the methods to perform the linking are described in the previous section.

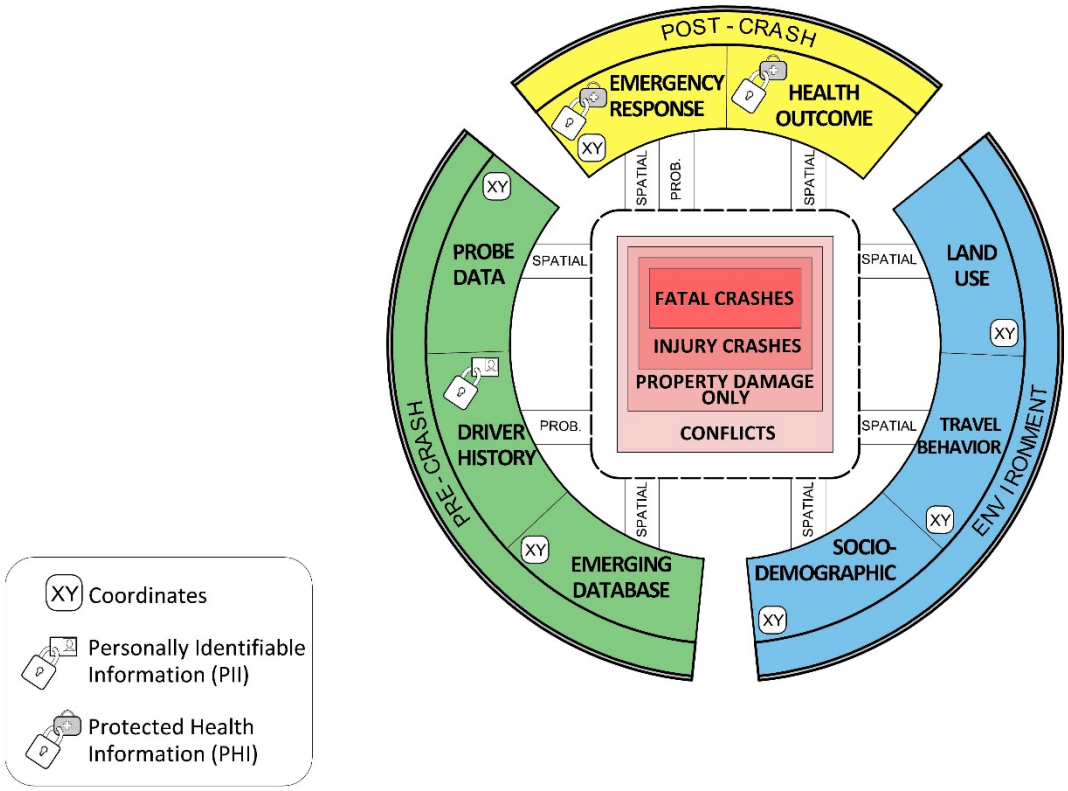

Figure 4. Complete Picture of Traffic Crashes With Linking Method

Police crash reports include data elements that describe the location (i.e., physical address or coordinates) of crashes and home address of individuals involved in crashes. The location of a crash is the data element that researchers/practitioner use to link environmental databases to a police database. Alternatively, home address of an individual also can be used as an alternative in analysis to capture wide-ranging information about sociodemographic factors of individuals who are involved in crashes. Address information can also be used to analyze the type of built or sociodemographic environments that comprise a disproportionate share of crashes. This approach can be used in analysis to identify neighborhoods whose residents have higher crash rates (see case study 3, 4, and 5 in the Appendix B for further details). In most cases, police crash report data must be aggregated at the zonal level in order to be joined with environmental data that is present at the zonal scale.

Police Reporting and Hospital-oriented Data

Hospital-oriented data are important in road safety analysis since they describe the health outcomes of individuals involved in traffic crashes and the social burden of traffic crashes. Depending on the goals of a particular study, researchers can analyze road safety at disaggregate (e.g., injury severity) or aggregate level (e.g., calculating social cost of traffic crashes).

Police officers' estimates of injury severity at the crash scene are not always accurate—for a better estimate of injury severity, analysis of health outcome data is helpful. An injured person can use a different health facility based on injury sustained in a traffic crash. Thus, in order to learn the true (or more accurate) outcome of crash-related injury, researchers must rely on a variety of databases. For example, individuals who suffer from less severe injuries that do not require overnight stays at the hospital are likely to be treated at ambulatory surgical centers, while those

who sustain major injuries requiring immediate attention are often transported to trauma centers or other medical facilities. This is also the case for tests related to driving under the influences of alcohol or other substances. Police officer reports in crash databases may not include accurate test results, while medical facilities have better diagnostic tools. Therefore, in order to learn more about driving under the influence of alcohol or drugs, it is necessary to link police crash databases to health outcome data. Another reason for seeking a better estimate of traffic crash outcomes is to calculate the direct cost of injuries resulting from traffic crashes, Medical Insurance Claims, Medical Expenditures Panel Survey, and the Hospital Discharge Data System are among databases that record the cost of injury treatment.

Underreporting is another challenge related to police crash databases, and can be attributed to several factors. Due to definition of police multi-vehicle crashes in police reports, officers generally do not report on single bicycle crashes (even in severe cases). Moreover, in some cases, pedestrians and bicyclists do not report their traffic crashes to the authorities for a number of reasons. Accordingly, police crash databases do not capture a significant portion of vulnerable road user crashes. Linking health outcome data to police crash database can provide additional details regarding the underreporting of pedestrians and bicyclist crashes. For more details on the subject of underreporting issue, please see case study 1 in Appendix B.

To perform the linking process at the individual level, researchers and practitioners must use PII in police crash databases and PHI in health outcome databases as inputs for the linking process. Depending on sample size and available data elements, the methodology for linking police crash database and health outcome data varies. The probabilistic method provided by CODES project is one of the common methods implemented in the United States.

Police Reporting and Pre-Crash Data (DMV)

Another source of data that can provide additional information about road users and their driving history (e.g., crash and citation) are the DMV state departments of motor vehicles (DMVs). They maintain information about driver's license and non-driver ID applicants in electronic databases that share a core set of data elements essential to voter registration, and that have the demonstrated capability to share information with other government databases (BCJ 2009). Every state DMV database is different, but they share similar information, such as name, address, date of birth, phone number, social security number, and, in some cases, medical or disability information. These databases are designed to be shared with pre-authorized users on demand (e.g., police departments) and can link the identification dataset with other databases (e.g., voter registration, automobile titles, etc.).

Conclusion

This study provided a brief review of a history of linking police crash reports and health data in the US as well as a review of the methodologies for linking databases and examples of the integrations' applications. Moreover, this report provided a complete picture of traffic crashes by linking different sources of the data that contribute to data science in road safety.

While this report is meant to provide a comprehensive list of main data sets that are relevant to safety, we cannot enumerate all possible use-cases or applications of safety data and potential datasets that can be integrated. Indeed, as road safety analysis becomes more system-oriented, the linkages between disparate data sources become more important. Similarly, as transportation systems have recently leveraged emerging and novel data sources, safety applications could evolve in ways that we cannot predict in this report. Importantly, researchers and practitioners should be able to apply any relevant dataset within the crash phase and linkage framework that we have presented here in order to provide a more complete picture of road safety.

References

Abdel-Aty, M., Siddiqui, C., Huang, H., Wang, X., 2011. Integrating trip and roadway characteristics to manage safety in traffic analysis zones. Transportation Research Record: Journal of the Transportation Research Board (2213), 20-28.

Agran, P.F., Castillo, D.N., Winn, D.G., 1990. Limitations of data compiled from police reports on pediatric pedestrian and bicycle motor vehicle events. Accident Analysis & Prevention 22 (4), 361-370.

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in pennsylvania. Accident Analysis & Prevention 38 (3), 618-625.

Amoros, E., Brosnan, M., Wegman, F., Bos, N., Perez, C., Segui, M., Heredero, R., Noble, B., Kilbey, P., Feypell, V., Year. Reporting on serious road traffic casualties, international traffic safety data and analysis group–irtad, organisation for economic co-operation and development (oecd). In: Proceedings of the International Transport Forum, Paris Google Scholar.

Aptel, I., Salmi, L.R., Masson, F., Bourdé, A., Henrion, G., Erny, P., 1999. Road accident statistics: Discrepancies between police and hospital data in a french island. Accident Analysis & Prevention 31 (1), 101-108.

Barnett, D.J., Balicer, R.D., Blodgett, D., Fews, A.L., Parker, C.L., Links, J.M., 2005. The application of the haddon matrix to public health readiness and response planning. Environmental health perspectives 113 (5), 561.

BCJ, 2009. Vrm department of motor vehicles databases. In: Justice, B.C.F. ed. Brennan Center for Justice New York, NY.

Brubacher, J.R., Chan, H., Fang, M., Brown, D., Purssell, R., 2013. Police documentation of alcohol involvement in hospitalized injured drivers. Traffic injury prevention 14 (5), 453-460.

Burch, C., Cook, L., Dischinger, P., 2014. A comparison of kabco and ais injury severity metrics using codes linked data. Traffic injury prevention 15 (6), 627-630.

CDPH, 2015. Exploratory analysis of injury classification of crash victims using crash-medical linked data in california. California Dempatment of Public Health, safe and Active ommunities Branch.

CMOD, 2013. Helmet use reduces traumatic brain injury among young bicyclists. In: Branch, C.D.O.P.H.S.a.a.C. ed.

CMOD, 2016. Drug presence in fatal motor vehicle collisions, california, 2005 - 2014. California Department of Public Health Safe and Active Communities Branch.

CMOD, 2017. Crash medical outcomes data project. California Department of Public Health.

Compton, C.P., 2005. Injury severity codes: A comparison of police injury codes and medical outcomes as determined by nass cds investigators. Journal of Safety Research 36 (5), 483-484.

Cook, L.J., Kerns, T.J., Burch, C.A., Thomas, A., Bell, E., 2009. Motorcycle helmet use and head and facial injuries: Crash outcomes in codes-linked data.

Cook, L.J., Knight, S., Olson, L.M., Nechodom, P.J., Dean, J.M., 2000. Motor vehicle crash characteristics and medical outcomes among older drivers in utah, 1992-1995. Annals of Emergency Medicine 35 (6), 585-591.

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.

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

Dhillon, P.K., Lightstone, A.S., Peek-Asa, C., Kraus, J.F., 2001. Assessment of hospital and police ascertainment of automobile versus childhood pedestrian and bicyclist collisions. Accident Analysis & Prevention 33 (4), 529-537.

Dong, N., Huang, H., Xu, P., Ding, Z., Wang, D., 2014. Evaluating spatial-proximity structures in crash prediction models at the level of traffic analysis zones. Transportation Research Record: Journal of the Transportation Research Board (2432), 46-52.

Dong, N., Huang, H., Zheng, L., 2015. Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects. Accident Analysis & Prevention 82, 192-198.

DOT, 2017. Highway performance monitoring system field manual. US Department of Transportation, Federal Highway Administration.

ESRI, 2009. Spatial join (analysis). ArcGIS Desktop Help 9.2. ESRI.

Farmer, C.M., 2003. Reliability of police-reported information for determining crash and injury severity.

FHWA, 2011. Kabco injury classification scale and definitions.

FHWA, 2018. What is mire? Federal Highway Administration, Wasgington DC, USA.

Flannagan, C., Mann, N.C., Rupp, J.D., 2013. Measuring serious injuries in traffic crashes.

Guo, F., Klauer, S.G., McGill, M.T., Dingus, T.A., 2010. Evaluating the relationship between near-crashes and crashes: Can near-crashes serve as a surrogate safety metric for crashes?

Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2010. Development of planning level transportation safety tools using geographically weighted poisson regression. Accident Analysis & Prevention 42 (2), 676-688.

Haddon Jr, W., 1968. The changing approach to the epidemiology, prevention, and amelioration of trauma: The transition to approaches etiologically rather than descriptively based. American journal of public health and the Nations health 58 (8), 1431-1438.

Han, G.-M., Newmyer, A., Qu, M., 2015. Seat belt use to save face: Impact on drivers' body region and nature of injury in motor vehicle crashes. Traffic injury prevention 16 (6), 605-610.

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.

Hankey, J.M., Perez, M.A., McClafferty, J.A., 2016. Description of the shrp 2 naturalistic database and the crash, near-crash, and baseline data sets. Virginia Tech Transportation Institute.

Huang, H., Abdel-Aty, M., 2010. Multilevel data and bayesian analysis in traffic safety. Accident Analysis & Prevention 42 (6), 1556-1565.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., Abdel-Aty, M., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of Transport Geography 54, 248-256.

Hydén, C., 1987. The development of a method for traffic safety evaluation: The swedish traffic conflicts technique. Bulletin Lund Institute of Technology, Department (70).

Kelly, A., King, J.S., 2017. All-payer claims databases: The balance between big healthcare data utility and individual health privacy.

Kindelberger, J., Milani, J.A., 2015. Crash outcome data evaluation system (codes): Program transition and promising practices.

Koornstra, M., Lynam, D., Nilsson, G., 2002. Sunflower: A comparative study of the development of road. Leidschendam: SWOV Institute for Road Safety Research.

Laureshyn, A., Svensson, Å., Hydén, C., 2010. Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation. Accident Analysis & Prevention 42 (6), 1637-1646.

Lopez, D.G., Rosman, D.L., Jelinek, G.A., Wilkes, G.J., Sprivulis, P.C., 2000. Complementing police road-crash records with trauma registry data—an initial evaluation. Accident Analysis & Prevention 32 (6), 771-777.

Martinez, R.E., 1993. The intermodal surface transportation efficiency act of 1991. Defense Transportation Journal.

McDonald, G., Davie, G., Langley, J., 2009. Validity of police-reported information on injury severity for those hospitalized from motor vehicle traffic crashes. Traffic injury prevention 10 (2), 184-190.

Miaou, S.-P., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping: A space-time modeling approach. Journal of transportation and statistics 6, 33-58.

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

Miller, T.R., Gibson, R., Zaloshnja, E., Blincoe, L.J., Kindelberger, J., Strashny, A., Thomas, A., Ho, S., Bauer, M., Sperry, S., Year. Underreporting of driver alcohol involvement in united states police and hospital records: Capture-recapture estimates. In: Proceedings of the Annals of Advances in Automotive Medicine/Annual Scientific Conference, pp. 87.

MMUCC, 2012. Model minimum uniform crash criteria. DOT HS 811, 631.

Naderan, A., Shahi, J., 2010. Aggregate crash prediction models: Introducing crash generation concept. Accident Analysis & Prevention 42 (1), 339-346.

NOAA, 2017. Federal meteorological handbook US Department of Commerce. National Oceanic and Atmospheric Administration

NTDB, 2018. National trauma data bank®ntdb research data set user manual and variable description list admission years 2002-2016. American College of Surgeons, Chicago, IL.

NTSB, 2013. Crashes involving single-unit trucks that resulted in injuries and deaths. National Transportation Safety Board, Washington DC, USA.

NWS, 2018. Aviation weather center metar data. National oceanic and atmospheric administration. In: Service, N.W. ed. National Weather Service

Olsen, C.S., Cook, L.J., Keenan, H.T., Olson, L.M., 2010. Driver seat belt use indicates decreased risk for child passengers in a motor vehicle crash. Accident Analysis & Prevention 42 (2), 771-777.

Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., Wets, G., Year. Developing zonal crash prediction models with a focus on application of different exposure measures.

Popkin, C.L., Campbell, B., Hansen, A.R., Stewart, R., 1991. Analysis of the accuracy of the existing kabco injury scale.

Porter, J., Love, D., Peters, A., Sachs, J., Costello, A., 2014. The basics of all-payer claims databases: A primer for states. Princeton, NJ: Robert Wood Johnson Foundation.

Poulos, R.G., Hatfield, J., Rissel, C., Grzebieta, R., McIntosh, A.S., 2011. Exposure-based cycling crash, near miss and injury rates: The safer cycling prospective cohort study protocol. Injury Prevention, injuryprev-2011-040160.

Pulugurtha, S.S., Duddu, V.R., Kotagiri, Y., 2013. Traffic analysis zone level crash estimation models based on land use characteristics. Accident Analysis & Prevention 50, 678-687.

Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of london crash data. Accident Analysis & Prevention 40 (4), 1486-1497.

Rosman, D.L., Knuiman, M.W., 1994. A comparison of hospital and police road injury data. Accident Analysis & Prevention 26 (2), 215-222.

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 (6), 1102-1113.

Sherman, H., Murphy, M., Huelke, D.F., Year. A reappraisal of the use of police injury codes in accident data analysis. In: Proceedings of the Proceedings: American Association for Automotive Medicine Annual Conference, pp. 128-138.

Siddiqui, C., Abdel-Aty, M., Choi, K., 2012. Macroscopic spatial analysis of pedestrian and bicycle crashes. Accident Analysis & Prevention 45, 382-391.

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

TDH, 2013. Ambulatory surgical treatment center data system user manual. Office Of Health Statistics Tennessee Department Of Health, Nashville, TN. USA.

TnDH, 2014. Death statistical file user manual. Division of Health Statistics Tennessee Department of Health.

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.

Ukkusuri, S., Hasan, S., Aziz, H., 2011. Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency. Transportation Research Record: Journal of the Transportation Research Board (2237), 98-106.

Wang, Y., Kockelman, K.M., 2013. A poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accident Analysis & Prevention 60, 71-84.

Washington, S., Metarko, J., Fomunung, I., Ross, R., Julian, F., Moran, E., 1999. An inter-regional comparison: Fatal crashes in the southeastern and non-southeastern united states: Preliminary findings. Accident Analysis & Prevention 31 (1), 135-146.

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

Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. Accident Analysis & Prevention 75, 16-25.

Zehtabchi, S., Nishijima, D.K., McKay, M.P., Clay Mann, N., 2011. Trauma registries: History, logistics, limitations, and contributions to emergency medicine research. Academic emergency medicine 18 (6), 637-643.

APPENDIX A

Detailed Description of Linked Safety Datasets

Crash Incident Databases

Police Crash Reports

Police crash reports are the core of road safety analysis. Law enforcement officials collect data either electronically or manually about every motor vehicle crash via police crash reports. The data in each crash report is then submitted to a state's centralized database where it is edited, reported and analyzed by a wide range of stakeholders. The crash reports follow a similar format which is based on the MMUCC (Model Minimum Uniform Crash Criteria) (MMUCC 2012).

How to Access the Database

Police crash reports are usually available for researchers and practitioners through highway patrol or Department of Homeland Security. Moreover, some of the programs such as the Highway Safety Information System provide crash reports to researchers at no charge. In order to access crash records with personal identifiers, it is necessary to check with the owner of the database to meet the minimum requirement for accessing the database.

Consistency of the Data across States

The database is of a similar format in every state, however, due to the voluntary nature of the MMUCC guideline, the data elements that describe a traffic crash can vary between states. In addition, each state has a financial threshold for reporting a traffic incident.

Typically, the data elements and their values (attributes) describe who was involved, where the crash took place, when and under what circumstances it took place, what the impacts of the crash were, and why the crash happened. When used by a reporting agency, MMUCC data elements record what happened during and after a crash. Since this data is so critical to state and local decision-making, state, and local agencies are encouraged to collect as many of the recommended MMUCC data elements and their attributes as possible (MMUCC 2012).

Layers of the Database

The MMUCC database has several layers of data that describe different aspects of the crash incident, namely crash, vehicle, person, and roadway data elements.

Person-Level

This layer includes information about the individual who was involved in a traffic crash and describes the characteristics, actions, and consequences to the individual. Regarding the accessibility of the data to the public, this layer can be further categorized into two sublayers. The first sublayer includes Personally Identifiable Information (PII) which needs certain permission or approval for retrieving the data. The second part includes personal information that are available to the public.

The PII data elements in this layer are the most vital information for performing the deterministic linkage. Names, family name, middle name, date of birth, complete resident's address, phone, and driving license information (e.g., statues and restrictions) are among the most vital PII data elements. Other data elements that would help the researchers in the linking (i.e., probabilistic and deterministic) process and describe individuals and are available to the public are age, gender.

The other sublayer contains data elements that describe individuals condition (regardless of the user type) at the crash scene. Among these data elements are violations, restrictions, and injury severity outcome of the traffic crash.

MMUCC recommended that this variable be recorded in the person level: name of person involved, date of birth, gender, person type, injury status, occupant's motor vehicle unit number, seating position, restraint systems/motorcycle helmet use, air bag deployed, ejection, driver license jurisdiction, driver license number, class, CDL and endorsements, speeding-related, driver actions at time of crash, violation codes, driver license restrictions, driver license status, distracted by, condition at time of the crash, law enforcement suspects alcohol use, alcohol test, law enforcement suspects drug use, drug test, transported to first medical facility, injury area, injury diagnosis, injury severity.

Crash Level

This level contains information that describes the general crash environment. Useful information that could be used in this study to perform linking includes coordinates of the crash location, address of the crash location (i.e., county, city, place, street address), time of the crash (i.e., year, month, day, hour, minute) which are useful for in performing linking. It is worth noting that not all the crashes have established coordinates; nevertheless, we can use the address provided by police to geocode the crash location. The quality of the geocoding depends on the detail of the address provided by a police officer in the electronic forms.

Moreover, this layer includes other time-related data elements that describe the time of the crash, notify and response of police officer and time of emergency response team which are useful in both probabilistic and deterministic methods.

MMUCC suggests that this variable be recorded at the crash level: crash identifier, crash classification, crash date and time, crash county, crash city/place (political jurisdiction), crash location, first harmful event, location of first harmful event relative to the traffic way, manner of crash/collision impact, source of information, weather conditions, light conditions, roadway surface conditions, contributing circumstances – roadway environment, relation to junction, type of intersection, school bus-related, work zone-related (construction/maintenance/utility), crash severity, number of motor vehicles involved, number of motorists, number of non-motorists, number of non-fatally injured persons, number of fatalities, alcohol involvement, drug involvement, and day of week.

Vehicle Level

The motor vehicle data elements describe the characteristics, events, and consequences of the motor vehicle(s) involved in the crash. Vehicle level also consists of two sub-layers. The first layer includes vehicle identifiers that are not accessible to the public such as Vehicle Information Number (VIN) and license plate number. These two data elements could be used to link database to DMV data.

The second sublayer includes information vehicle registration, state, make, model year, model, body type, number of occupants, and similar vehicle-related data elements. Moreover, this layer includes information about vehicle maneuvers prior to the crash. Data elements such as direction of travel before the crash, number of lanes, roadway alignments, traffic control device, motor vehicle maneuver, speed, and similar items are stored in this sub-layer.

MMUCC suggests that this variable be recorded in the vehicle level: VIN, motor vehicle unit type and number, motor vehicle registration state and year, motor vehicle license plate number, motor vehicle make, motor vehicle model year, motor vehicle model, motor vehicle body type category, total occupants in motor vehicle, special function of motor vehicle in transport, emergency motor vehicle use, motor vehicle posted/statutory speed limit, direction of travel before crash, traffic-way description, total lanes in roadway, roadway alignment and grade, traffic control device type, motor vehicle maneuver/action, vehicle damage, sequence of events, most harmful event for this motor vehicle, hit and run, towed due to disabling damage, contributing circumstances, motor vehicle

Roadway Level

Roadway data elements are generated by linking crash to roadway inventory and highway data. The data elements used for linkage include “crash location” and others as necessary, depending upon the type of roadway inventory system implemented by the state. When a state does not have a roadway inventory, the police officer should collect as many of the data elements as possible should at the crash scene (MMUCC 2012).

MMUCC recommends that this variable be recorded at the roadway level: bridge/structure identification number, roadway curvature, grade, part of national highway system, roadway functional class, annual average daily traffic, widths of lane(s) and shoulder(s), width of median, access control, railway crossing id, roadway lighting, pavement

markings, longitudinal, presence/type of bicycle facility, mainline number of lanes at intersection, cross-street number of lanes at intersection, total volume of entering vehicles.

Fatality Analysis Reporting System

FARS is a database that provides information regarding motor vehicle traffic crashes with fatal injuries. NHTSA, Congress and the American public are the primary users this data. FARS contains data on all fatal traffic crashes across the nation and Puerto Rico. In order for a crash to be included in FARS, it must meet specific criteria. First, the crash must occur on a traffic way customarily open to the public. Second, the crash must result in the death of a person within 30 days of the crash.

There are several state agencies in each state that work in agreement with NHTSA to provide information in a standard format on fatal crashes in the state. Each agency collects, codes and submits the data into a microcomputer data system transmitted to Washington, D.C. Quarterly files are produced for analytical purposes to study trends and evaluate the effectiveness highway safety programs.

How to Access Database with PII

FARS database does not include personally identifiable information (PII). The FARS database is available for the public. The database can be downloaded from FTP website (ftp://ftp.nhtsa.dot.gov/fars/). The FTP website included traffic crashes fatally between 1975 to 2016.

Consistency of the Data across States

The National Center for Statistics and Analysis (NCSA) which is a component of NHTSA, is the director of FARS database. NHTSA has a cooperative agreement with an agency in each State's government; this agency is responsible for providing information on all qualifying fatal crashes that occurred in the state. NCSA's FARS program staff are managing the agreement between NHTSA and state government agencies. Moreover, in each state, trained state employees, called FARS Analysts, are responsible for gathering, translating, and transmitting their state's data to NCSA in a standard format.

Layers of the Database

FARS also collects information on over 100 different coded data elements that characterize the crash, the vehicle and the people who were involved in a fatal traffic crash. As a result, the FARS database consisted of several layers namely: accident, vehicle, person, parkwork, pbtype, cevent, vevent, vsoe, damage, distract, factor, maneuver, violatn, vision, nmcrashm nmimpair, nmprior, safetyeq, vindecode. It is worth noting that the presence of this layer is consistent for all years. Moreover, the variable descriptions in many cases were subjected to change (NHTSA 2016). For more details, please visit the Fatality Analysis Reporting System (FARS) Analytical User's Manual (NHTSA 2016). The most important layers in this layer are accident, vehicle, person, and pbtype.

Accident

This data file contains information about crash characteristics and environmental conditions at the time of the crash. There is one record per crash. The most critical variables in this layer are: county, city, time of crash (month, hours, day, minute), traffic way identifier, route signing, land use, functional system, special jurisdiction, latitude, longitude, time of EMS arrival at hospital (hour, minute), notification time (hour, minute). This layer is present in all the years (NHTSA 2016).

Vehicle

This data file contains information that describes the in-transport motor vehicles and their corresponding driver who are involved in the crash. Each motor-vehicle has one record in the FARS database. The most important variables in this layers are: registration state, registered vehicle owner, vehicle information (make, body, model year), Vehicle Identification Number (VIN), Motor Carrier Identification Number (MCID), MCID Issuing Authority, MCID Identification Number, driver's license state, driver's ZIP Code. This layer is present in all the years (NHTSA 2016).

Person

This data file covers information describing all persons involved in the crash crashes. Road user types are motorists (i.e., drivers and passengers of in-transport motor vehicles) and non-motorists (e.g., pedestrians and pedal-cyclists). This layer includes data elements that describe individual characteristics (e.g., age, sex, vehicle occupant restraint use) and their injury severity. Each person has its unique corresponding record in the FARS database. The most important variables in this layer are age, sex, person type, injury severity, seating position, death time (year, month, day, time, hour), race, Hispanic origin. This layer is present in all the years (NHTSA 2016).

Pbtype

This layer contains information regarding crashes between people on personal conveyances and bicyclists, motor vehicles, and pedestrians. Data from the crash are entered into the Pedestrian and Bicycle Crash Analysis Tool (PBCAT). The output fields from PBCAT, including the pre-crash actions of the parties involved (crash type), are included in this dataset. Each of pedestrian, bicyclist or person on a personal conveyance have a unique record. The most important variables in this layer are age, sex, person type, crash location. This layer is also present in all the years (NHTSA 2016).

Post-Crash Database

Emergency Medical System (EMS) Data

Traditionally EMS data has been housed in local trauma registry databases. More recently NHTSA has funded development of a national database for housing EMS data nationally, the National EMS Information System (NEMSIS). As of 2017, most states are submitting data to NEMSIS. (https://nemsis.org/)

How to access EMS data

Access to data NEMSIS repository is through the University of Utah School of Medicine. The link to request access to NEMSIS data is https://nemsis.org/using-ems-data/request-research-data/). The following web link provides the status of the contribution of each state for Version 2 NEMSIS (https://nemsis.org/). Version 3 is now underway. A number of variables are only accessible at the state level.

In California data from CEMSIS (the California version of NEMSIS) can be accessed by submitting a request to the EMSA unit at the California Department of Public Health (California DPH). Procedures for accessing data vary from state to state.

Layers of the database

The primary elements of the NEMSIS database Version 2.2.1 are listed below. The data dictionary for Version 2.2.1 is at the following website: https://nemsis.org/technical-resources/version-2/version-2-dataset-dictionaries/. Categories of variables are listed below. More detail can be added at each level to indicate the potential uses for each category. Items within these categories are differentially available.

- Demographic (Agency) Dataset
- AGENCY GENERAL INFORMATION
- EMS agency variables have limited availability
- AGENCY CONTACT INFORMATION
- Agency zip code
- EMS (Event) Dataset
- RECORD INFORMATION
- Administrative record codes
- UNIT / AGENCY INFORMATION
- Agency codes and response delay codes
- UNIT / CALL INFORMATION

- Complaint report information
- TIMES
- Response time variables
- PATIENT
- Patient information including zip code (not name) (matching variables0
- BILLING
- Billing codes
- SCENE
- Number of patients and type of location
- Incident Zip Code
- SITUATION
- Injury and provider information
- SITUATION / TRAUMA
- Cause of Injury (IDC code)
- SITUATION / CPR
- Cardiac variables
- MEDICAL HISTORY
- Patient information
- INTERVENTION / MEDICATION
- Medication variables
- INTERVENTION / PROCEDURE
- Procedure variables
- DISPOSITION
- Disposition variables
- OUTCOME AND LINKAGE
- ED and hospital disposition

National Vital Statistics System (NVSS) – Mortality

Mortality data from the National Vital Statistics System (NVSS) are a fundamental source of demographic, geographic, and cause-of-death information. This is one of the few sources of health-related data that are comparable for small geographic areas and are available for a long time-period in the United States. The data are also used to present the characteristics of those dying in the United States, to determine life expectancy, and to compare mortality trends with other countries.

How to access database with PII

To access each state data with identifiers, researchers and practitioners need to contact state Department of Health or other agencies that are in charge of gathering mortality data.

Layers of the database

There are several layers in this database. The most important layers are patients' personal information, accident layer, insurance layer, admission layer and health outcome (TnDH 2014).

Patients' personal information

This layer includes personal information of the deceased. The variables in this layer are Certificate Number, Generational Identifier of Decedent, Surname of Decedent, Given Name of Decedent, Middle Initial of Decedent, Age at Death, Date of Birth, Race, Sex, Marital Status, Social Security Number, The Deceased was ever in US Armed Forces, State of Residence, County of Residence, City of Residence, Address of Residence, City Limits Indicator, Zip Code of Residence, State/Country of Birth, Hispanic Origin, Education of Decedent, Surname of Decedent's Father, Given Name of Decedent's Father, Maiden Name of Decedent's Mother, and Given Name(s) of Decedent's Mother.

Accident layer
This layer contains data elements about the cause and time of death. This layer also has information of traffic accident. Variables in this layer are If Injury is due to Transportation Accident, Underlying Cause of Death, Type of Vehicle, Injury at Work, Time of Injury, Date of Injury, Date of Death, Time of Death, County of Death, and City of Death.

Admission layer
This layer includes information about the facility. Variables in this layer are Facility Code, Type of Place of Death, Type of Certifier, Date Signed by Certifier, and License Number of Physician or Medical Examiner.

Hospital Discharge Data System

The first standard dataset for reporting a single billing form and standard data set that can be used nationwide by institutional providers and payers for handling health care claims was created by the American Hospital Association (AHA) brought the National Uniform Billing Committee (NUBC) in 1975 (Gilbreath and Dinkins 2003).

In Tennessee, all the hospitals are required by law to report patient-level discharge information to the Tennessee Department of Health. This report includes all incidents regarding discharges from rehabilitation hospitals, rehabilitation and psychiatric units within acute care hospitals, and free-standing ambulatory surgical treatment centers that are part of a hospital, only if they are from a TDH licensed hospital and meet the requirements for "Reportable Records." Moreover, discharges for charity or free care are included in the reporting requirement, and they are handled similarly (TDH 2017).

How to Access Database with PII

To access this database, one needs to contact their local state departments of health, the procedure for accessing the data is different in each state. The procedure for accessing the data with Protected Health Information (PHI) is also different from the dataset without PHI. The Protected Health Information under the US law is any information about health status, provision of health care, or payment for health care that is created or collected by a Covered Entity (or a Business Associate of a Covered Entity) and can be linked to a specific individual. Under the US Health Insurance Portability and Accountability Act (HIPAA), PHI that is linked based on the following list of 18 identifiers must be treated with special care (Anon 2018a). The PHI usually includes the following variables:

- Names
- Geographic data
- All elements of dates
- Telephone numbers
- FAX numbers
- Email addresses
- Social Security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers including license plates
- Device identifiers and serial numbers
- Web URLs
- Internet protocol addresses
- Biometric identifiers (i.e., retinal scan, fingerprints)
- Full face photos and comparable images
- Any unique identifying number, characteristic or code

Consistency of the Data across States

In 1982, the NUBC voted to accept the UB-82 and its associated data manual for implementation as a national uniform bill. Virtually all states adopted the use of the UB-82 data set. After an 8-year moratorium on change, the UB-82 was replaced by UB-92 and became the standard for billing paper institutional medical claims in the United States. This system was also replaced in 2007 with UB-04 form (current state of practice). This was a nationwide change which included state of Tennessee (TDH 2017).

Layers of the database

There are several layers in this database. The most important layers are patients' personal information, accident layer, insurance layer, admission layer and health outcome (TDH 2017).

Patients' personal information

This layer includes information about the individual who received service in ASTC. Regarding the accessibility of the data to the public, this layer can be further categorized into two sublayers. The first sublayer includes Personally Identifiable Information (PII) which needs certain permission or approval for retrieving the data. The second part includes personal information that is available to the public. The data elements in this layers are patient's name, family name, residential address, gender, date of birth, federal tax number (SSN) and patient's ethnicity/race.

Accident layer

This layer contains data elements about the source of injury (not necessarily the motor-vehicle crashes). The variables in this layer are accident date, accident location, accident hour, accident code and external cause of injury.

Insurance layer

This layer contains data elements regarding patients insurance. The variables are: Classification Classification of Payer(s), Health Plan Identification Number, Patient's Relationship to Insured(s), National Provider Identification (NPI), Insured's Unique ID Number, Insurance Group Number(s), Name of Insured's Employer Primary Insured Initials, Secondary Insured Initials, Tertiary Insured Initials Primary Insured Name – First and Last, Secondary Insured Name – First and Last, and Tertiary Insured Name – First and Last.

Health layer

This layer include information about diagnosis and hospital cost. The variables in this layer are: Principal Diagnosis Code with POA, Other Diagnosis Codes with POA, Admitting Diagnosis Code, Patient Reason for Visit Code, Diagnosis and Procedure Version Qualifier, External Cause of Injury Code (E-Code), Principal Procedure Code, Principal Procedure Date, Other Procedure Codes and Dates, Attending Provider ID Numbers, Operating Provider ID Numbers, Other Provider1 ID Numbers, Other Provider2 ID Numbers, Type of Emergency Department Visit, and Outcome of Emergency Department Visit.

Admission layer

This layer includes information regarding the admission event. The variables are: Admission Date, Admission Hour, Type of Admission/Visit, Point of Origin Service Date(s), Creation Date, Units of Service, Total Charges by Revenue Code Category, and Non-Covered Charges.

Ambulatory Surgical Center

Due to concern about medical charges at the hospital led to an increasing proportion of surgeries being performed on an outpatient basis (TDH 2013). This shift from Traditionally surgeries in hospitals to ASTC had arose a new trend of injury treatment. Ambulatory surgery centers (ASC)[4] are healthcare facilities where surgical procedures not requiring

[4] Also known as outpatient surgery centers or same day surgery centers

an overnight hospital stay are performed. These types of surgeries usually have less complicated nature than that requiring hospitalization.

Both ambulatory surgery center and a specialty hospital provide similar facilities and support similar types of procedures for their patients. However, the specialty hospital may provide the same procedures or slightly more complex ones, and the specialty hospital will often allow an overnight stay. ASCs do not routinely provide emergency services to patients who have not been admitted to the ASC for another procedure.

In most states, ASC facilities mandate to report their records to state's Department of Health. For example, in Tennessee, T. C. A. 68-1-119 requires each licensed ambulatory surgical treatment center (ASTC) in the state to report its claims data to the Tennessee Department of Health. ASTCs also need to keep personal identifies confidential. Patient Record Data Systems (PRDS), part of the Division of Health Statistics, has been charged by the Department to carry out this project. Each ASTC will transmit its records to the vendor it has selected. Each vendor will send the records it has received from its ambulatory surgical treatment centers to PRDS (TDH 2013).

How to Access Database

Depending on the state, the procedure for accessing the database varies.

Consistency of the Data across States

The national standard for the usage of the CMS-1500 form is the manual produced by the American Medical Association's National Uniform Claims Committee. The national standard for the usage of the UB-04 forms is established by the American Hospital Association's National Uniform Billing Committee (TDH 2013). Each record sent by the vendor to the Health Department is the electronic representation of one CMS-1500 or UB-04 form. Records are fixed format. All the records for a quarter for all facilities reporting through a vendor should be combined into two separate files, one of CMS-1500 records and the second of UB-04 records. The consistency of the data over the states needs further investigation (TDH 2013).

Tennessee established a standard layout for reporting. However, if another layout is acceptable to the center and its' vendor, it is acceptable to the Health Department if all reporting requirements are met. The vendor is also responsible for creating the vendor-generated fields. These may be created from raw data values or reported directly by the centers (TDH 2013).

Layers of the Database

There are several layers in this database. The most important layers are patients' personal information, insurance, and health outcome. However, depending CMS 1500 or UB-04 the variables in this layer varies. In the following, the data in the CMS 1500 will be discussed. The UB-04 form is similar to HDDS database. It is also worthy to mention that CMS 1500 does not report the health and diagnosis variables.

Patients' Personal Information

This layer includes information about patients who received service in ASTC. Regarding the accessibility of the data to the public, this layer can be further categorized into two sublayers. The first sublayer includes Personally Identifiable Information (PII) which needs certain permission or approval for retrieving the data. The second part includes personal information that is available to the public. The data elements in this layer are patient's name, family name, residential address (i.e., state, city, zip code), marital status, employment, federal tax id (SSN), race, Federal Tax ID Number, Federal Tax ID Number (SSN / EIN)

Insurance layer

This layer contains data elements regarding insurance, namely Type of Insurance, Insured ID Number Insured's Policy Group or FECA Number, Insured's Employer's or School Name, Insurance Plan or Program Name, Other Insured's Policy or Group Number, Insurance Plan or Program Name, Another Health Benefit Plan, Insurance Plan Program Classification Code (Primary), Insurance Plan Program Classification Code (Other)

Admission layer
This layer includes information regarding the admission event: the variables are ID of Referring Physician, NPI Number of Referring Physician, Hospitalization Dates Related to Current Services – From and Through, Diagnosis or Nature of Illness or Injury Date(s) of Service – From and Through Place of Service, Emergency Code, Patient's Account Number, Total Charges, Days or Units EPSDT.

Trauma Center Data

A trauma center is a hospital that is equipped and staffed to provide care for patients suffering from major traumatic injuries such as vehicle collisions, gunshot wounds, falls, etc. Depending on available resources of each trauma center, patients may require stabilization and transfer to another facility. The National Trauma Data Bank® (NTDB®) is the largest aggregation of U.S. trauma registry data ever assembled. Participation is voluntary and is one of the leading performance improvement tools of trauma care (NTDB 2018). Since its inception in 1994, many researchers have published work based on data from the NTDB. In 2008, the NTDB implemented the National Trauma Data Standard (NTDS) to standardize data collection across all reporting hospitals. Currently, the NTDB contains detailed data on over 2.7 million cases from over 900 U.S. trauma centers (NTDB 2018).

How to Access the Database

NTDB® data are maintained in a secure database with limited internal access. External users must gain permission to the database and data; users are then supplied data at the aggregate level only. The data set is de-identified, and no protected health information is provided. Use of NTDB data is in strict compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The NTDB does not distribute or report hospital information in any manner that allows the reporting hospital to be identified without the express written permission of the hospital. The dataset collected by NTDB is considered a limited dataset under HIPAA, and the research dataset that ACS releases is a de-identified dataset (ACS 2018).

National Trauma Data Bank® (NTDB®) research data may be used for informational and research purposes with approval from the American College of Surgeons (ACS) Committee on Trauma. Permission to use the NTDB dataset is required via an online data application form found on their website.

Consistency of the Data across States

There are no general and all-inclusive guidelines for developing each of the trauma registry elements. Currently, each institution designs its trauma registry based on its needs or to meet mandatory state or city requirements (Zehtabchi *et al.* 2011). The NTDS provides definitions for variables and response codes. Institutional and state trauma registries often collect additional variables depending on their needs. Most trauma registries incorporate the following data: demographic information, mechanism of injury (external cause of injury codes), procedures, clinical diagnoses (based on International Statistical Classification of Diseases and Related Health Problems), length of stay, disposition, and in-hospital mortality. Many also include abbreviated injury scores (AIS), charges, payers, and information about complications as well as follow-up procedures if they occur at the same institution. Some important variables, such as longer-term mortality or functional outcomes, are rarely collected (Zehtabchi *et al.* 2011).

Layers of the Database

The Research Data Set (RDS) provided by the NTDB is a set of relational tables and consists of 18-20 data files. These files are provided in ASCII-CSV (comma separated value) format, standard SAS (*.sas7bdat) data tables (for datasets Admission Year (AY) 2007 and later), and DBF format (DBASE version 2.0), which can be easily imported to most statistical software. The relational tables are too large to be analyzed in Microsoft Excel but have been used in Microsoft Access, SAS, STATA, SPSS, and Tableau.

Three different classes of tables exist in the data set:

1. Incident-based tables - Most of the data files include a unique incident identifier (inc_key) for merging the data files together.
2. Facility-based tables - One data file (RDS_FACILITY) includes the facility information for participating hospitals, and these data can be merged to RDS_ ED, RDS_DEMO, and RDS_DISCHARGE, by using the unique facility identifier (fac_key).
3. Lookup tables - The remaining data files (RDS_AISDES, RDS_ECODEDES, RDS_DCODEDES, and RDS_PCODEDES; RDS_DIAGNOSISDESC and RDS_PROCEDUREDESC in Admission Year 2002-2006 data) are look-up tables with the description of the AIS code, ICD-9-CM E-Code, ICD-9-CM diagnosis codes, and ICD-9-CM procedure codes, ICD-10-CM E-Code, ICD-10-CM diagnosis codes, ICD-10-CM procedure codes, and ICD-10-CM location codes. The look-up tables can be merged with the unique RDS_DCODE, RDS_ICD10_DCODE, RDS_ECODE, RDS_ICD10_ECODE, RDS_PCODE, RDS_ICD10_PCODE, and RDS_ICD10_LOC (RDS_DIAGNOS and RDS_PROCEDUR in AY 2002-2006 data) table

Of interest are the following files:

- RDS_DEMO includes information about patient demographics, including patient's birth year, age at time of injury, gender, race, and ethnicity;
- RDS_FACILITY provides information about the facility;
- RDS_SCENE pertains to information about the scene of the trauma, including year, county and site of injury (i.e., Home, Farm, Mine and Quarry, Industrial Places and Premises, Place for Recreation and Sport, Street and Highway, Public Building, Residential Institution, Other Specified Places, and Unspecified Places);
- RDS_PROTDEV refers to protective devices;
- RDS_SAFETY includes information pertaining to safety equipment used/worn at time of the injury;
- RDS_TRANSPORT includes the type and mode of transportation;
- RED_ECODE pertains to external cause of injury
- BODYREGION - Body region based on the AAAM (Association for the Advancement of Automotive Medicine) area on body, includes head, face, neck, thorax, abdomen, spine, upper extremity, lower extremity and unspecified.
- RDS_ED references the ED and Injury information including year when the patient was injured, first recorded time of patient's arrival at reporting hospital, and whether patient used alcohol or drugs.

Medical Insurance Claims – All-Payer Claims Database (APCD)

In recent years, a growing number of states have established databases that collect health insurance claims information from all healthcare payers into a statewide information repository. By December 2017 at least 18 states had enacted "all-payer claims databases" (or APCDs). APCDs are large-scale databases that systematically collect medical claims, pharmacy claims, dental claims (typically, but not always), and eligibility and provider files from private and public payers. The first statewide APCD system was established in Maine in 2003. By 2017, 18 states had passed legislation and established APCDs. Cooperative state-private sector databases are found in Colorado, Kansas, Minnesota, Tennessee, Maine, Maryland, Massachusetts, New Hampshire, Rhode Island, Utah and Vermont (NCSL 2017). States that have been implementing APCDs more recently include Connecticut, Nebraska, New York, Virginia and West Virginia. States that have had existing voluntary efforts to maintain an APCD include Virginia, Washington, and Wisconsin (NCSL 2017).

APCD systems collect data from existing claims transaction systems used by healthcare providers and payers. The information typically collected in an APCD includes patient demographics, provider codes, and clinical, financial, and utilization data. Because of the difficulties involved with the collection of certain information, most states implementing APCD systems typically have not included a number of data elements, such as denied claims, workers' compensation claims, and, because claims do not exist, services provided to the uninsured (Porter *et al.* 2014).

How to Access the Database

Most states define the details about APCD data collection requirements (e.g., format and timing of submission, specifications of data elements, and thresholds for payers that are required to submit data) uniquely from other states (Porter *et al.* 2014). Each state assesses its own legislation to determine the most efficient way to design a flexible but comprehensive APCD as there is no model. Determining what data and information will be released and to whom can be the most sensitive aspect of APCD implementation, and there is significant variation in policies and practices across states (Porter *et al.* 2014). States generally de-identify the data using encryption and statistical methods to mask the identity of the individuals in the database, though some states allow qualified users to access de-identified and research files. Regulations that specify data access and release policies vary according to state legal and political environments (e.g., Minnesota does not release data to external organizations because of privacy concerns; some states limit data access to state government only) (Porter *et al.* 2014). Each state will need to be contacted separately to gain access to the APCD database. Various state departments house the database (i.e., Department of Health (Utah, Minnesota); Independent state agency (West Virginia, Maine); Health and insurance departments with overlapping responsibilities (New Hampshire), and External, non-governmental agency (Colorado)).

Layers of the Database

Though not standardized, generally there are several layers in each state's APCD database. Generally, these would include information regarding the Patient, Subscriber, Claim, and Service. The Agency for Healthcare Research and Quality has performed basic database formatting of APCD data based on Accredited Standards Committee X12 (ASC X12) input (AHRQ 2018a). ASCX12 is a standards organization chartered by the American National Standards Institute to develop and maintain electronic data interchange architecture.

Patient Personal Information Layer

This layer includes personal information of the patient. These variables could include Encrypted Member First Name, Middle Initial, and Last Name; Patient Address, City, State and Zip Code; Date of Birth and Gender, and Patient Identification Code and Patient Relationship Code to Subscriber, and Subscriber Social Security Number.

Subscriber Personal Information Layer

This layer includes personal information of the patient. These variables could include Encrypted Subscriber First Name, Middle Initial, and Last Name; Subscriber Insured Group or Policy Number, Subscriber Plan Specific Contract Number, and Subscriber Social Security Number, which would link the Subscriber Data layer to the Patient Data layer.

Medical Insurance Claim Layer

This layer generally contains data elements about the medical insurance and payments claim. This layer can contain information on Insurance Product, such as Health Maintenance Organization (HMO), Preferred Provider Organization (PPO), Type of Insurance Contract (e.g., single, family, etc.), and Subscriber Insured Group or Policy Number, which can link to the Subscriber Personal Information Layer. The Medical Insurance Claim Layer can also include Health Plan Payments, Member Payment Responsibility, Type and Date of Bill Paid, Claim Status and Revenue Codes.

Medical Service Layer

This layer may contain data elements about the service and treatment. Variables in this layer could include Information on the Dates or Service, and the Date Service was Approved; Admission Date/Time, Type, Point or Origin and Billing Provider; Discharge Date/Time and Status; Admitting Diagnosis, E-Code and/or Other Diagnosis, and Billing Provider Information, such as Last Name, Organization Name and Number, and Billing Provider ID. Also, this layer contains Service Provider Information, such as Service Provider ID, City Type, Entity Type, Service Provider Name (i.e., First, Last and Middle Initial) and Number, State/Zip Code, Specialty, and Tax Identification Number. Other dataset fields include information regarding Charge Amount, Co-Pay Amount, Coinsurance Amount, Deductible Amount, Diagnosis Related Group Code, Drug Code, Procedure Code, and Revenue Code.

Medical Expenditures Panel Survey

The Medical Expenditure Panel Survey (MEPS), which began in 1996, is a set of large-scale surveys of families and individuals across the United States, their employers and availability of health insurance, their medical providers (doctors, hospitals, pharmacies, etc.), and their use of nursing home services. MEPS collects data on the specific health services that Americans use, how frequently they use them, the cost of these services, and how they are paid for, as well as data on the cost, scope, and breadth of health insurance held by and available to U.S. workers.

How to Access the Database

There are four parts to the MEPS data: The Household Component (HC), the Insurance Component (IC), the Medical Provider Component (MPC), and the 1996 Nursing Home Component (NHC). The HC data for households and IC estimates for national, regional, state and metropolitan areas are available on the MEPS Web site in data tables (i.e., tabular form), downloadable data files, via interactive data tools, and in publications using HC data. MPC data is used to supplement, verify, and/or replace information provided by household respondents about the charges, payments, and sources of payment associated with specific healthcare encounters. MPC data are collected once for each calendar year of data, during the year following the reference year. The most current data years of the MPC questionnaires are available online. MPC data from previous data years are available upon request to the MEPS project director. The purpose of the MPC data is to supplement household-reported data, and it is not intended to be an independent sample of providers for estimation purposes. The NHC was conducted in 1996 only. Because of confidentiality concerns, data collected from the 1996 MEPS-NHC questionnaire are only available at the Agency Healthcare Research and Quality (AHRQ 2018b) Data Center.

Layers of the database

MEPS currently has four major components: The Household Component (HC), the Insurance Component (IC), the Medical Provider Component (MPC) and the 1996 Nursing Home Component (NHC). The Household Component provides data from individual households and their members; the Insurance Component is a separate survey of employers that provide data on employer-based health insurance; the Medical Provider Component is used to verify medical and financial characteristic information of medical events described in the HC survey from medical providers, and the 1996 Nursing Home Component comes from community sources with a minor amount of data collected from nursing home sources.

Household Component Layer

The Household Component (HC) collects data from a sample of families and individuals in selected communities across the United States, drawn from a nationally representative subsample of households that participated in the prior year's National Health Interview Survey (conducted by the National Center for Health Statistics). Multiple rounds of household interviews are conducted allowing MEPS to collect detailed information for each person in the household on many topics. The round of household interviews covers approximately 45 sections, with each section comprising multiple questions. Listed below is a brief description of each of the 45 sections:

1. Access to Care (AC) identifies whether each household member has a medical provider, aspects of the providers, and problems a household may have in obtaining health care;
2. Alternative/Preventive Care (AP) gathers information on any preventive care received, such as frequency of dental and physical check-ups, flu shots, etc.;
3. Assets (AS) asks about household members' real estate, businesses, vehicles, investments, other assets, and debts;
4. Calendar Section (CA) monitors the use of a health events calendar provided to the respondent for use in recording visits to medical providers and medical places;
5. Condition Enumeration (CE) obtains a summary assessment of each person's physical and mental health;
6. Caregiver (CG) collects information on potential caregivers both inside and outside the household; the types and duration of care provided, etc.;

7. Closing (CL) is when participants are asked to provide written authorization for the MEPS to collect additional information from medical/insurance providers and employers;
8. Conditions (CN) collects additional information about physical and mental health conditions identified through medical events or disability days;
9. Charge Payment (CP) tracks total charges and sources of payment for medical events; obtains specific information regarding total charges, copayments, out-of-pocket payments, etc.;
10. Caregiver Roster (CR) collects detailed information on non-household members identified as a potential caregiver in the Caregiver supplemental section;
11. Child Preventive Health (CS) collects information on health care status and needs, behavioral problems, accessibility to care, preventative care, height, and weight of any child in the family;
12. Disability Days (DD) assesses each year's impact of any physical illness, injury, or mental or emotional problem on household members' attendance at work or school;
13. Dental Care (DN) obtains details on the nature of any dental care visit, type of dental care provider, treatments and services performed, and prescribed medicines;
14. Event Driver (ED) verifies and modifies information entered; provides an opportunity to add new medical events throughout the interview if the respondent recalls an event later;
15. Employment (EM) covers questions about each person's employment or lack of employment status, including type and size of business, employment length, reasons for unemployment, etc.
16. Overall Structure of Employment (EM-O) collects detailed information on jobs held by each person in the household aged 16 or older and the health insurance provided by each employer;
17. Emergency Room (ER) obtains information on the health conditions requiring emergency room care, medical and surgical services provided, prescribed medicines, etc.;
18. Event Roster (EV) asks for additional detail on event dates, type of event, and type of provider, as well as the charge and payment for each event;
19. Employment Wage (EW) collects detailed information about the wage structure for all non-self-employed;
20. Flat Fee (FF) is a subsection of Charge Payment (CP) and captures information on those types of medical payments that charge a grouped amount, or flat fee, for multiple visits or services;
21. Health Status (HE) assesses the physical and mental health status for both children and adults. Specific areas assessed include the loss of adult teeth, limitations in activities of daily living, etc.;
22. Home Health (HH) obtains information on the types of health care workers providing home health services, reasons for and nature of the care, frequency of visits, length per visits, etc.;
23. Private Health Insurance Detail (HP) collects details on each private health insurance policy, including name of insurance company, policyholder identified, household members covered by each policy, etc.;
24. Time Period Covered Detail (HQ) clarifies the timeframe for which each person was covered by each reported health insurance policy;
25. Hospital Stay (HS) obtains details on the length of stay, reasons or conditions requiring hospitalization, surgical procedures performed, medicines prescribed, etc.;
26. Health Insurance (HX) collects detailed information about various insurance programs, household members covered; program coverage; etc. This section creates a link to job characteristics collected in the Employment (EM) section;
27. Income (IN) collects information about the household members' Federal income tax filing status, itemized deductions for health insurance premiums, tax credits, wages, other income, etc.;
28. Long-Term Care (LC) collects information on institutionalized household member who received long-term care due to an impairment or a physical or mental health problem and their care;
29. Managed Care (MC) determines whether household members are covered under a private managed care plan;
30. Medical Provider Visits (MV) obtains details on the nature of visits, type of and time with provider, health conditions requiring services, treatments and procedures, prescriptions, etc.
31. Over-the-Counter Medicine (OC) collects details about purchases of any over-the-counter medicines used, type of health conditions for which they were purchased and cost;
32. Old Employment/ Private Related Insurance (OE) collects information about the continuation of insurance coverage;

33. Other Medical Expenses (OM) collects information in cases where respondents report expenses for glasses or contact lenses or for insulin and other diabetic equipment;
34. Outpatient Department (OP) obtains information on any outpatient visits (e.g., nature of the contact, type of care received, health conditions requiring outpatient services, treatments, etc.);
35. Priority Conditions (PC) collects information about a select group of medical conditions including sore or strep throat, diabetes, asthma, hypertension, coronary heart disease, angina, heart attacks, other heart disorders, strokes, emphysema, joint pain, and arthritis;
36. Provider Directory (PD) compiles a directory of all medical persons and medical facilities reported by MEPS respondents;
37. Priority Conditions Enumeration (PE) obtains a summary assessment of each person's physical and mental health;
38. Pregnancy Detail (PG) collects additional information (duration, complications, etc.) for women identified in the Condition Enumeration section as having been pregnant;
39. Prescribed Medicines (PM) obtains details on prescribed medicines reported in earlier medical events sections, such as date of use, refill status, etc.
40. Provider Probes (PP) collects information required to create a medical event in the database, i.e., the type of event, the person incurring the event, the health care provider, and the date(s) of the event;
41. Old Public Related Insurance (PR) collects information on household members covered by Medicare, Medicaid or other state or local government sponsored programs;
42. Provider Roster (PV) creates a roster to display the name and street address of each provider and/or facility associated with each person's medical events; this information is strictly confidential;
43. Re-numeration-A and B(RE-A/RE-B) identifies each person and family unit living within each household and defines how family members are related to one another and the size of the family unit (i.e., race, ethnicity, educational attainment, and military status);
44. Review of Employment Information (RJ) reviews employment information for any current job including job status, salary, health insurance benefits, size of employment establishment, etc.;
45. Satisfaction with Health Plan (SP) collects satisfaction information (e.g., ease of access, delays incare, etc.) for private insurance, Medigap, Medicare managed care programs, Medicaid, etc.

Insurance Component Layer

The Insurance Component (IC) collects data from a sample of private and public sector employers on the health insurance plans they offer their employees. The data include the number and types of private insurance plans offered (if any), premiums, contributions by employers and employees, eligibility requirements, benefits associated with these plans, and employer characteristics. The survey is also known as the Health Insurance Cost Study. IC data is divided into two sections: Establishment Information and Health Insurance Information.

Establishment Information Layer

Establishment information includes location of business, number of employees, minimum number of hours per week employees had to work in order to be eligible for health insurance, workforce information (e.g., union membership, percentage of female employees, employees aged 50 or older, hourly pay, types of fringe benefits, types of tax-advantaged benefits, etc.), availability and cost of insurance plans for employees, the health insurance plans, employee eligibility and enrollment in various insurance plans, the number of part-time workers, their insurance eligibility and number of workers enrolled in health care plans, and the number of employees who work less than 30 hours per week. Other information includes: the use of Small Business Health Options Program or use of insurance broker or agent to help purchase a plan, the eligibility of employees for optional health plans (e.g., dental, vision, prescription drug, long-term care, etc.), the total amount paid by employer and employee for optional coverage, the imposition of a waiting period for new employees, any financial compensation or incentives to employees if they did not elect to receive health insurance coverage, the eligibility of employee spouses, domestic partners or other family members in health insurance plans, and the availability and eligibility of retirees in health insurance plans.

Health Insurance Information Layer

The Health Insurance Information Layer contains a number of variables. These include the name of the health insurance plan with the largest (or next largest) enrollment; type of health care provider arrangement available through this plan (e.g., single, family, etc.); whether the plan requires the enrollee to see a gatekeeper or primary-care physician; whether the plan is offered through a union or a trade association; if the plan has been purchased from an insurance underwriter or was it self-insured; if the organization employed a third party administrator (TPA) or purchase administrative services only (ASO) from an insurer for this self-insured plan; if the organization purchased a stop-loss coverage for this plan and the specific stop-loss amount per employee; the percentage of medical expenses paid by the plan; the number of employees enrolled in the various coverages (e.g., single coverage, employee plus one coverage and family coverage); the number of employees enrolled through COBRA; the deductibles for each coverage type; the availability and contributions of health savings accounts or health reimbursement arrangements, and the percentage of total hospital, specialist or prescription drug bill paid by employer versus employee.

Medical Provider Component

The Medical Provider Component (MPC) requests data from hospitals, physicians, home health care providers, and pharmacies identified by HC respondents. Its purpose is to supplement and/or replace information received from the respondents about the health care that was provided to sample household members in the course of the survey year. Patients in sampled households were asked to sign permission forms authorizing contact with their healthcare providers. The purpose of the MPC is to supplement, verify, and/or replace information provided by household respondents about the charges, payments, and sources of payment associated with specific healthcare encounters. This is important because people cannot always accurately answer questions about the health services they received and about the cost of those services. The information is used solely for editing and imputation purposes on the Household Component. Therefore, these data will not be released in a stand-alone file. The seven questionnaires are designed to obtain information on both the medical and financial characteristics of medical events. A brief summary of each questionnaire is found below.

Home Care Health Care Provider Event Questionnaire is used to collect data from home health care agencies which provide medical care services to household respondents. Information collected includes type of practitioner providing care, hours of service and visits provided per month, and the charges and payments for services received.

Home Care Provider Event Questionnaire for Non-Health Care Providers. This is used to collect information about services provided in the home by non-health care workers to household respondents because of a medical condition; for example, cleaning or yard work, transportation, shopping, or child care. Charges and payments for services received are also collected.

Institutional Event Questionnaire for Non-Hospital Facilities. This is used to collect information on services and expenditures for persons from the household sample who were admitted to a nursing home, rehabilitation center, or other non-hospital long-term health care facility during the survey year.

Office-based Provider Event Questionnaire. This is used to collect data from the office-based physician sample, including Doctor of Medicine (MDs) and osteopathy (DOs), as well as providers practicing under the direction or supervision of an MD or DO (e.g., physician assistants and nurse practitioners working in clinics). Providers of care in private offices, as well as staff model HMOs, are included. This includes diagnoses, procedure, and inpatient stay codes, charges or charge equivalents (where available) before any contractual adjustments or discounts, sources and amounts of all payments made, and the reasons for any difference between charges and payments.

Separately Billing Doctors Event Questionnaire. Information from physicians identified by hospitals as providing care to sampled persons during the course of inpatient, outpatient department or emergency room care, but who bill separately from the hospital, is collected in these questionnaires.

Hospital Event Questionnaire. This is used to collect information about hospitals events, including inpatient stays, outpatient department, and emergency room visits. Hospital data are collected not only from the billing department, but the medical records and administrative records departments as well. Medical records is contacted to determine the names of all the doctors who treated the patient during a stay or visit. In many cases, the hospital administrative office also has to be contacted to determine whether the doctors identified by medical records billed separately from the hospital itself. HMO hospitals are included in the data collection effort. This includes diagnoses, procedure, and inpatient stay codes, charges or charge equivalents (where available) before any contractual adjustments or discounts, sources and amounts of all payments made, and the reasons for any difference between charges and payments.

Pharmacies and Other Sources of Prescribed Medicines Event Questionnaire. Pharmacy data collection focuses on the request for a patient profile (a computer generated listing of the prescriptions dispensed to a given customer). Date filled, National Drug Code, medicine name, strength of medicine (amount and unit), quantity (package size/amount dispensed), and payments and amounts by source. However, when possible, a patient profile (a computer-generated printout of a person's drug purchases) is requested to make the request less burdensome on the provider. Generally, patient profiles contain all the data elements requested in the questionnaire. Pharmacy data collection includes drug stores, grocery stores, discount stores, mail order, online, clinics, HMOs, and hospitals.

1996 Nursing Home Component
The 1996 Nursing Home Component (NHC) comprises 12 questionnaires with multiple queries. Data were collected from community sources, and a minor amount of data was collected from nursing home sources. This component is only available for 1996. The following is a listing of the 12 Questionnaire types:

- Nursing Home: Person Characteristics Questionnaires
- Nursing Home: Facility Background and Health Insurance Questionnaires
- Nursing Home: Health Status Questionnaire
- Nursing Home: Facility Residence History Questionnaire
- Nursing Home: Prescribed Medicines Questionnaire
- Nursing Home: Expenditures Questionnaire
- Nursing Home: Use of Services Questionnaire
- Community Sources: Person Characteristics Questionnaires
- Community Sources: Community Questionnaire
- Community Sources: Facility Characteristics Questionnaires (from sampled facility sources)
- Community Sources: Facility-Level Questionnaire
- Community Sources: Facility Characteristics Questionnaires (from new facility sources)

Crash Environment Databases

Travel Demand Models

Traffic Analysis Zone (TAZ) data and associated travel demand model data can be provided by any regional entity with an operational travel demand model. Travel demand models describe present and project future patterns of travel demand, which means the origins and destinations of trips, the modes used to reach those destinations, and the routes used for each individual trip.

TAZ data is typically available from Metropolitan Planning Organizations (MPOs). Some states also manage travel demand models and may have such data at the state level. Traffic Analysis Zone data is typically not posted for download on the internet. It must be requested from the appropriate person within an MPO or state departments of transportation (DOTs). Some organizations have standardized processes for requesting such data whereas others do not.

How to Access the Database with PII
Personally identifiable information is not available in TAZs or from travel demand models.

Consistency of the Data across States
There are no consistent standards for TAZ data across MPOs and State DOTs.

Layers of the Database
Travel demand models include data on TAZs and data on flows between TAZs.

Data on TAZs
Most travel demand models include basic demographic information at the TAZ scale, including population counts and employment counts with a variety of population and employment types. Population is often available by income group. In some cases, the pedestrian friendliness of each zone is scored.

TAZs may also have other data associated with the travel demand model such as trip productions, trip attractions or accessibility. Trip productions are statistically modeled estimates of the number of trips of a particular purpose originating from a given TAZ. Trip attractions are statistically modeled estimates of the number of trips of a particular purpose destined for a given TAZ. Accessibility is a generalized measure of the ability to reach a particular type of destination by a particular mode across the region.

Data on TAZ Flows
Travel demand models predict the number of trips between TAZs by time of day and by mode (auto, public transit, walk, bike, taxi, etc.). They also estimate the expected travel time between TAZs by time of day and by mode.

Parcel Data
Parcel-level data are generally available from local jurisdictions such as counties, cities, and townships. They are typically available only for the present year, but in some cases parcel databased are archived and available for each year. One of the primary limitations of parcel data is that the fields available, as well as the quality of the data, can vary substantially by jurisdiction.

Most jurisdictions in the United States have parcel data available in GIS form. The fields that are most typically available include Taxable value for land; taxable value for buildings; land acreage; year structure built; and information about the most recent sale of the property, including sale value. Less common fields include land use codes; zoning codes; and building area.

Based upon these fields, it is sometimes possible to calculate the percentage of land in different land uses within a district, the number of parcels of different land uses within a district, and if building area is available, floor area ratio, which is the ratio of building floor area to parcel land area.

Based on the acreage in different types of land uses, it is also possible to calculate level of mixed-use via an entropy calculation.

Land use types, land use entropy, and floor area ratio have all been used in safety research to identify potential hot spots for different crash types.

How to Access the Database With PII
Parcel data is requested by contacting the GIS administrator of a local jurisdiction. Some jurisdictions post their parcel data and/or land use data online available for free download. Other jurisdictions require payment for access to their parcel and/or land use data.

Since parcel data surrounding the crash location is unrelated to the identification of crash victims, PII data is not available from parcel data.

Consistency of the Data across States

There is no national data standard for parcels. Some states have standardized state formats, such as the State of Washington (https://depts.washington.edu/wagis/projects/parcels/). But it is quite common to have substantial variations across local jurisdictions within states.

Layers of the Database

Parcel-level data has only a single layer, data corresponding to land parcels.

LODES

Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) data provide information on the residential location of workers, the workplace location of workers, and the flows between the two. The US census synthesizes this data from multiple administrative sources including state unemployment records. The geographical unit of analysis for this data is the census block scale.

LODES data are available back to 2002, and most states are covered for most years. Out of 53 states and US territories, 50 were participating as of 2017.

Data are available from US census' Longitudinal Employer-Household Dynamics website: https://lehd.ces.census.gov/data/#lodes

How to access database with PII

Personally identifiable information is not available from the LODES. Statutory requirements prevent LODES from disclosing personally identifiable information. Noise is added to the data in order to ensure that no personally identifiable information is recoverable.

Consistency of the Data across States

LODES data is completely consistent over US States as well as US territories. The same data can be used nationwide, although not all states are covered for every data year.

Layers of the database

US census has data on several different universes. These universes include people, households, families, workers, housing units, and certain subsets of the above (i.e., population 25+). The primary universes of interest for crash data are people and households.

Residential Area Characteristics (RAC)

Residential Area Characteristics data includes total number of workers, workers by age, workers by earnings category, workers by race, workers by industrial category (North American Industry Classification System or NAICS), workers by sex, and workers by educational attainment. Notably, counts are by job, not by worker, so a worker may be counted twice unless only primary jobs are selected for analysis.

Workplace Area Characteristics (WAC)

Workplace Area Characteristics data includes total number of workers, workers by age, workers by earnings category, workers by race, workers by industrial category (NAICS), workers by sex, workers by educational attainment, workers by firm size, and workers by firm age. Notably, counts are by job, not by worker, so a worker may be counted twice unless only primary jobs are selected for analysis.

Origin to Destination Flows (OD)

This contains the total number of jobs flowing from a given residential block to a given workplace block. Counts are also available disaggregated by worker age, worker income category, and major worker industrial sector.

Census Data

The US census gathers data through the decennial census and the American Community Survey every year. The decennial census is an attempt at a complete count of the US residential population, but only covers age, race, household structure, and residence. The American Community Survey (ACS) covers a wide range of demographic and economic information for persons and households but is only a sample and therefore cannot provide accurate counts at small levels of geography. The smaller the geography considered, the larger the margin of error is for ACS data.

US census data are available at many levels of geography, with the primary ones being State, County, Census Tract, census Block Group, and census Block. Lower levels of census geography next within higher levels in the above scheme.

US census data are available from the American Factfinder website, but also from many secondary data providers such as ESRI Business Analyst, Social Explorer, and National Historic GIS. In some cases, secondary data providers may make the data easier to access than American Factfinder.

How to Access the Database With PII

Personally identifiable information is not available from the US census. Statutory requirements prevent the US census from disclosing personally identifiable information.

Consistency of the Data across States

US census data is completely consistent over US States as well as US territories. The same data can be used nationwide.

Layers of the Database

US census has data on several different universes. These universes include people, households, families, workers, housing units, and certain subsets of the above (i.e., population 25+). The primary universes of interest for crash data are people and households.

Data on the Universe of People

Data on the Universe of People includes a wide range of demographic and economic variables including age, race, sex, ethnicity, migration status, household status, educational status, and employment status.

Data on the Universe of Households

Data on the Universe of Households includes housing tenure, housing costs, vehicle availability, household type (i.e., parents, children, marital status, and unrelated persons).

Data on the Universe of Workers

Data on commuting is for the universe of workers aged 16+. Commuting data includes work location relative to residential location, commute mode, commute time of day, and commute length in minutes.

Highway Performance Monitoring System

The Highway Performance Monitoring System (HPMS) is a national program which includes inventory information for all of the Nation's public roads as certified by the States' Governors annually. HPMS contains data from 50 states, DC and Puerto Rico and US territories. Regardless of road ownership, all roads that are open to public travel are reported in HPMS, including Federal, State, county, city, and privately-owned roads such as toll facilities. HPMS data are used to calculate following performance measures (DOT 2017):

- Rate of fatalities in 23 CFR 490.207(a)(2)
- Rate of serious injuries in 23 CFR 490.207(a)(2)
- Percentage of pavements of the Interstate System in Good condition in 23 CFR 490.307(a)(1)
- Percentage of pavements of the Interstate System in Poor condition in 23 CFR 490.307(a)(2)

- Percentage of pavements of the non-Interstate NHS in Good condition in 23 CFR 490.307(a)(3)
- Percentage of pavements of the non-Interstate NHS in Poor condition in 23 CFR 490.307(a)(4)

National, State, and local transportation decision-making may also use HPMS database in transportation planning process for decision making to analyze trade-offs among the different modes of transportation.

How to access database

This database is available online, and it is free of charge. This database is available in shapefile format. In the HPMS website, the data from 2011 to 2015 are available to the public.

Consistency of the Data across States

Each state is required to annually report and furnish all data per the reporting requirements which is specified in this HPMS Field Manual. The District of Columbia and the Commonwealth of Puerto Rico are treated as states for HPMS reporting purposes. Moreover, United States Territories (Guam, the Commonwealth of the Northern Marianas, American Samoa, and the Virgin Islands of the United States) are required to annually report limited HPMS summary data only, in addition to the separate reporting of certified public road mileage (DOT 2017).

Layers of the database

This database has several layers; the layers are inventory, route, traffic, geometric, pavement and special network. For more details about the data elements and their changes during the time, please visit highway performance monitoring system field manual (DOT 2017).

Meteorological Data, Terminal Aviation Routine (METAR) Weather Data

The most globally consistent weather data is associated with aviation activity. METAR (Meteorological Terminal Aviation Routine) weather data refer to a well-defined format for reporting weather information to pilots and meteorologists. Raw METAR data is standardized by the International Civil Aviation Organization (ICAO), which allows it to be understood throughout most of the world (NWS 2018). Three types of weather reporting systems are used to collect data for METAR formatting: Automated Terminal Information Service (ATIS), Automated Airport Weather Observation System (AWOS) and Automated Surface Observing System (ASOS) data.

ATIS data requires a human to monitor weather at a terminal and integrate these data into automatically observed weather data systems provided by AWOS or ASOS. AWOS data collection systems are maintained and operated by state or local governments under the authorization and certification of the Federal Aviation Administration (FAA). On the other hand, ASOS systems are maintained and operated by a joint effort between the National Weather Service and Department of Defense (NOAA 2017). The ASOS systems provide weather information beyond that of aviation.

How to access database

Weather stations from around the world report weather conditions every hour. The Aviation Weather Center saves up to seven days of this data. METAR data can be downloaded from the Aviation Weather Center website at https://aviationweather.gov/. The data can be accessed through an interactive display or their data server.

The METAR interactive display page contains an image of the current surface observations in the United States. This page uses OpenLayers to provide an interactive display of data including zoom and pan, which allows the user to alter the map display to include the world, and allows the user to click on a specific station or airport to obtain METAR information. The background map can be viewed in Light (vector map display), Dark (raster map display) or Simple (line map display). Data Layers are available in Satellite or Radar. In addition, overlays such as Highways, Top Jet routes and Air Route Traffic Control Center/Flight Information Region (ARTCC/FIR) Bounds are available. Each station or airport is automatically viewed in a station model or square plot layout. The station model layout illustrates the temperature in the upper left, dewpoint in the lower left, coded altimeter setting in upper right, visibility far left, weather near left, cloud

cover colored by flight category and wind barb. The Plot Options section of the interactive map allows the user to select or deselect these options, in addition to data density and scale, as well as time frame for data display.

The data server provides a method to query specific datasets in order to access real-time METAR, PIREP, AIREP, TAF and AIRSIGMET data. This service is oriented towards automated parsing and is an ideal mechanism for users needing more flexible access to raw data that is available through the interactive website. Results are downloaded into CSV (Comma Separated Value) or XML files. In the CSV file, statistics are reported first, followed by a header indicating which fields are being reported. All data is ordered by descending observation time. CSV files can be open in Excel or Access. XML files are an Extensible Markup Language file. METAR XML output has an XML schema, which describe the elements in an XML document in a series of XSD files (e.g., aircraftreport1_0.xsd, airsigmet1_1.xsd, gairmet1_0.xsd, metar1_2.xsd, pirep1_2.xsd, station1_0.xsd, taf1_2.xsd, etc.). These files are plain text files that can be opened in Notepad or a web browser.

Layers of the database

METAR raw data contains information that can be grouped into four categories: Data Station Layer, Wind Data Layer, Weather Conditions Layer and Visibility and Sky Conditions Layer.

Data Station Layer

The Data Station Location Layer includes elements pertaining to the following attributes:

- Observation Type can be either METAR for a regularly reported observation (such as an hourly) or SPECI for a special observation;
- Station Identifier consists of four characters that can be alphanumeric (e.g., McGhee Tyson Airport station identifier is KTYS, with the letter K designating a US Station);
- Date/Time of observation in International Organization for Standardization (ISO); ISO8601 date/time format in Zulu/UTC (e.g., 051853Z indicates the day of the month is the 5th and the time of day is 1853 Zulu/UTC (1:53 PM EST));
- Latitude and Longitude (in decimal degrees) of the station that reported the data;
- Elevation of the station that reported the data;
- Both temperature and dew point are calculated in degrees Celsius. If the temperature or dew point falls below 0 degrees there will be an "M" before the variable meaning minus (e.g., 03/M02 represent a temperature of 3 degrees Celsius with a dew point of minus 2 degrees Celsius).
- Altimeter refers to atmospheric pressure. For example, A30.16 stands for 30.16 inches of mercury for the pressure.
- RMK simply means REMARKS and marks the end of the standard METAR observation and the beginning of the remarks that are put in as necessary. There are many remarks, and the FMH-1 (Federal Meteorological Handbook-1) provides a full listing of them.

Wind Data Layer

The wind data layer includes the following:

- Wind Direction consists of the direction of the winds in degrees from 0 to 359 degrees, as 360 degrees is zero (e.g., 19020G26KT indicates wind direction is 190 degrees true north). For winds traveling greater than 6 knots and with a directional variation of at least 60 degrees, the designation may be 18015KT 150V210. This means the winds are from 180 degrees at 15 knots, but the direction is actually variable between 150 degrees and 210 degrees
- Wind Speed reports the speed of the winds in knots (e.g., 19020G26KT indicates wind is traveling at 20 knots). For winds speeds below 7 knots, the designation is VRB, which means the wind direction is variable. This is the idea of "light and variable" found in a forecast.
- Wind Gusts reports the wind gusts in knots (e.g., 19020G26KT indicates gusts of wind are traveling at 26 knots; KT represents all values are reported in knots). Not all reports will contain information regarding wind gusts as there are criteria which must be met in order to have a gust.

Weather Conditions Layer

Weather conditions are described by intensity and/or proximity, event description, precipitation type, obscuration type and other weather phenomena. The specific criteria that must be met for accurate weather designations is found in the Federal Meteorological Handbook (FMH) No. 1 "Surface Observations and Reports," which contains the designations for more than 100 variable combinations. Generally, intensity is presented as a negative, non-existent, or positive value. Precipitation is categorized as either light (-), moderate ( ), or heavy (+) based on the criteria in the handbook. Proximity of a weather event can be VC or in the Vicinity. Frequently used event descriptors include partial (PR), blowing (BL), showers (SH), thunderstorms (TS) or FZ (Freezing). Examples of precipitation events include rain (RA), snow (SN), and hail (GR) among others. Some common obscuration elements include fog (FG), mist (BR), smoke (FU), sand (SA) and haze (HZ). Examples of other weather phenomena include funnel cloud tornado waterspout (+FC), squall (SQ) and sand or dust storm (SS). For example, a weather designation of –SHRA indicates light rain showers.

Visibility and Sky Conditions Layer

Visibility designates the statute miles of visibility (e.g., 6SM-Visibility means 6 Statute Miles of visibility). Occasionally, the designation can reflect visibility up to 20 or 30 SM, but generally, the designation will be from < 1/4 (visibility below 1/4 SM) up to 10 SM. Sky Conditions indicate cloud cover. The cloud cover will either be FEW (1/8 TO 2/8 cloud coverage), SCT for Scattered (3/8 TO 4/8 cloud coverage), BKN for Broken (5/8-7/8 coverage), or OVC for overcast (8/8 Coverage). The numeric value found after the sky conditions represents the cloud elevation. For example, BKN090 indicates broken cloud cover at 9,000 feet (simply add 2 zeroes to get the height). More than one designator can be used in sequence (e.g., SCT035 BKN090 OVC140). An indefinite ceiling caused by fog, rain, snow, etc., will require a designator as VV (Vertical Visibility). VV is the vertical visibility into the indefinite ceiling. Also, significant clouds such as TCU (Towering Cumulus), CB, (Cumulonimbus, or a shower/thunderstorm), or ACC (Altocumulus Castellanos) will be found at the end of a designation sequence (e.g., SCT035TCU means scattered clouds at 3,500 feet with towering cumulus).

Model Inventory of Roadway Elements

Model Inventory of Roadway Elements (MIRE) is a database provided by FHWA that maintains a listing of roadway inventory and traffic elements which are critical to safety management. The intention behind MIRE was to provide a guideline to help transportation agencies improve their roadway and traffic data inventories. Mire provides a basis for a standard of what can be considered a good/robust data inventory and helps agencies move toward the use of performance measures to assess data quality (FHWA 2018).

Consistency of the data over the states

Consistency of the recorded data elements in the MIRE is based on their priority rating. The priority ratings are broken down into two major categories: critical and value added. Critical elements are those that are necessary for states to collect in order to conduct basic safety management and/or are contained in safety analysis tools such as SafetyAnalyst. On the other hand, value-added elements are those elements whose presence is beneficial but are not crucial to using current versions of safety analysis tools.

Layers of the database

There are a total of 202 elements that comprise MIRE Version 1.0. The MIRE elements are divided among three broad categories: roadway segments, roadway alignment, and roadway junctions. A breakdown of categories and subcategories is shown below.

- I. Roadway Segment Descriptors
- I.a. Segment Location/Linkage Elements

- I.b. Segment Roadway Classification
- I.c. Segment Cross Section
- I.c.1. Surface Descriptors
- I.c.2. Lane Descriptors
- I.c.3. Shoulder Descriptors
- I.c.4. Median Descriptors
- I.d. Roadside Descriptors
- I.e. Other Segment Descriptors
- I.f. Segment Traffic Flow Data
- I.g. Segment Traffic Operations/Control Data
- I.h. Other Supplemental Segment Descriptors
- II. Roadway Alignment Descriptors
- II.a. Horizontal Curve Data
- II.b. Vertical Grade Data
- III. Roadway Junction Descriptors
- III.a. At-Grade Intersection/Junctions
- III.a.1. At-Grade Intersection/Junction General Descriptors
- III.a.2. At-Grade Intersection/Junction Descriptors (Each Approach)
- III.b. Interchange and Ramp Descriptors
- III.b.1. General Interchange Descriptors
- III.b.2. Interchange Ramp Descriptors

Pre-Crash Databases

Department of Motor Vehicle Data

Every state requires that citizens provide personal data (name, address, date of birth, phone number, Social Security Number and, in some cases, medical or disability information) to obtain a driver's license or identification card and to register a vehicle. The information is stored in state-controlled databases along with records of drunken-driving arrests, traffic offenses, and accidents. All state departments of motor vehicles (DMV) maintain the information about driver's license and non-driver ID applicants in electronic databases that share a core set of data elements essential to voter registration, and that have the demonstrated capability to share information with other government databases (BCJ 2009).

How to Access the Database

While pre-authorized federal, state and local agencies (e.g., Police, Immigration, etc.) can freely access DMV data, federal and state laws specify the circumstances in which access is provided to entities (government and others) that have not been authorized, and these rules vary by state. There are numerous laws that govern how DMV information is disclosed. For example, the Federal Drivers Privacy Protection Act requires all States to protect the privacy of personal information contained in a person's motor vehicle record with certain exceptions defined in Title 18, United States Code Section 2721 (BCJ 2009). However, some states sell DMV data access to private-sector users who qualify for the privilege under federal and state laws. The list includes law firms; insurance companies; auto sales, service and towing companies; private investigators and security firms; and companies and nonprofit organizations that employ drivers. Each State Department of Motor Vehicle should be contacted in order to obtain DMV data.

Consistency of the Data across States

Every state's department of motor vehicles database is different, but share similar information, such as name, address, date of birth, etc. These databases are designed to be shared with pre-authorized users on demand (e.g., police, etc.) and can link the identification dataset with other databases (e.g., voter registration, automobile titles, etc.).

Dataset Layers

All state department of motor vehicles drivers' license/non-driver ID databases have the following data elements in common: (1) full name; (2) date of birth; (3) address; (4) a unique driver's license/ID number; and (5) Social Security number (BCJ 2009). Some databases contain additional information including license expiration date, and physical descriptions (e.g., height, weight, hair color). Additionally, when linked with other databases, the DMV database can extract and download other information such as insurance policy number, name of provider, type of coverage (full or liability), any citation numbers, speeding tickets, parking tickets, as well as vehicle registration indemnification, vehicle type, vehicle tag number, and vehicle manufacturer (BCJ 2009).

Crowdsourced Data

Crowdsourcing occurs when data from a network of collaborators is gathered for a particular purpose. This includes gathering data that would otherwise be cost or labor-intensive or for which the available expertise within a defined organization is unavailable or insufficient (Misra *et al.* 2014). Transportation crowdsourcing involves leveraging the combined data gathered by a group of people utilizing multiple types of internet-connected devices, such as smartphones, and various applications to identify an issue related to transportation. The term "crowdsourcing" is applied in common usage to several dissimilar processes (Dennis 2015). A few methods for collecting crowdsourced data used in transportation include:

- **Commercial Providers:** Transportation agencies already obtain aggregated crowdsourced data through contracted third-party commercial providers, most often for traffic speed and vehicle-count information (e.g., HERE Traffic Analytics), to assess travel reliability, congestion, and emissions (e.g., INRIX, Telenav), and define origin and destination sequences (e.g., Cellint Systems, TravelCast).
- **Social Media:** Some transportation agencies have experimented with leveraging internet-based social networks (e.g., predominantly Facebook and Twitter) to obtain public feedback regarding the condition of the transportation system and performance of the agency. The Michigan Department of Transportation and City of Austin, Texas, have used social media to crowdsource information about transportation networks.
- **Internet:** Many travelers use real-time transportation data integrated from the Internet. Examples include the use of open traffic data to obtain network traffic speed estimations or predict the impact of special events on the transportation system. In 2005 Google launched an application programming interface for its online map to provide travel time estimation. Other examples include Waze, Bing Maps, Tom-Tom, etc. These platforms are primarily used to receive turn-by-turn navigation directions and live traffic information. Crowdsourced data gathering is a secondary activity often used to report traffic hazards.
- **Dedicated Platforms:** Transportation agencies use crowdsourcing platforms, where the data generated by the platform is created specifically for transportation system management. These applications can utilize the phone's built-in GPS, accelerometers, camera, and other sensors to allow the agency to collect a wide variety of information. Dedicated crowdsourcing apps include automated vehicle location for public transit (e.g., Moovit, RideScout), pavement condition data collection (e.g., City of Boston Street Bump), bicycle travel and infrastructure data (e.g., Minnesota DOT's Cyclopath), parking management, origin-destination studies, environmental data collection and planning and project prioritization.

Commercial providers of crowdsourced data related to transportation require the purchase of data. Much of the value provided by contracted third-party data providers is the data fusion, analysis, and information packaging necessary to turn raw data into traffic information useful to a transportation agency (Dennis 2015). Social media platforms may require some expertise in their creation, but the data should be open to the public and freely obtained. Though the raw data may be easily accessed the methods of systematically assessing the raw data may prove problematic. The maps

and data associated with open source internet interfaces are proprietary, and through shared, the raw data is owned by the company providing the platform (e.g., Google, Waze, etc.). Dedicated platform data is also proprietary and owned by the platform developer. Each unique company should be contacted for data access.

Crowdsourced data comes in a large number of formats, which presents a challenge for developing systems that can integrate and analyze multiple types of crowdsourced data, and well as integrate the data into legacy systems. Agencies wishing to leverage crowdsourced data must establish data-intake processes that interpret and distribute the data appropriately (Dennis 2015).

Given the types of crowdsourced data available, there does not appear to be any consistency in the attributes of the dataset layers other than some form of geo-referencing found in data from commercial providers, internet platforms, and some dedicated platforms. Some demographic information may be able to be extracted from social media crowdsourcing as well as some geo-referenced information. Due to massive amount and longevity of the data, commercial and internet providers may not be able to link data to demographics. On the other hand, researchers may be able to investigate evolving patterns of mobility trends in an area. Social media demographic information may need to be manually mined and linked adding costs to the use of social media platforms for crowdsourcing data. Furthermore, social media networks are most often used by those demographically younger, skewing the data. Finally, any data obtained that can identify individuals should consider the privacy rights of users, and those knowingly or unknowingly providing data.

Examples of Crowdsourced Data in Transportation

What follows is a list of sources of data that uses crowdsourcing to obtain information relevant to transportation applications. It is not a comprehensive list, nor are the researchers recommending any of the products listed. This list provides a general description and contact information for crowdsourced data sets that are used for mapping and navigation; detailed traffic performance data, ride-sharing and public transit data, pedestrian and bicycling data, and roadway infrastructure conditions.

	Name	Description	Contact
Mapping & Navigation	GOOGLE MAPS	Google Maps provide real-time traffic and incid updates that modify navigation to less congest routes. Google has access to and incorporates location data collected by smartphones contai Google App in order to provide real-time crowdsourced data that provides real-time traf updates, estimates current traffic speeds and pinpoints road diversions. The underlying Goog maps database will adjust directions according based on the real-time input.	Real-time traffic and navigation Inform is viewed by downloading a phone app by using a computer. This aggregated displayed real-time information is free. GOOGLE CONTACT: 1-877-355-5787 www.google.com/contact/

	Name	Description	Contact
	TELENAV Open Street Map (OSM)	TELENAV is partnered with OpenStreetMap, a crowdsourced community of 1.3 million people have gathered GPS data while driving, biking ar walking the streets of the world to mark roadwa and objects in the environment (e.g., trees, park benches, potholes, structure footprint, etc.). Th static map layers include standard, cycle, publi transport and humanitarian. Other data sets inc map notes, map data, and public GPS traces.	Maps are free. Data is free to downloa gpx files. OPENSTREETMAP: https://www.openstreetmap.org/
	PELMOREX Beat the Traffic	The crowdsource phone App provides real-time on estimated travel time, traffic flow, incidents, construction, road closures, and traffic camera locations and weather. Users report traffic inci by shaking their device to alert fellow users of issues on the road. Additionally, real-time traffi reports are broadcast live.	Pelmorex Corp. Headquarters & The Weather Network Media Centre 2655 Bristol Circle, Oakville, Ontario L6H 7W1 https://www.pelmorex.com/en/
	WAZE	WAZE is a phone App and real-time website tha allows commuters to modify maps and add rea traffic data. This includes reports on gas prices arrival time coordination to facilitate car-poolin real-time alerts for police, accidents, road hazar traffic jams. Community forums are also availa	Waze Office Google West Campus 3, 1505 Salado Dr (MTV-GWC3) https://www.waze.com/
Detailed Traffic Performance Data	HERE Technology	By collaborating with vehicle manufacturers (e. BMW, Audi, Daimler, Volkswagen, etc.), HERE is to use passively collected crowdsourced traffic driving data that can help steer people around backups and road hazards due to inclement we conditions and construction slow-downs. The d collected also includes speed, breaking info, windshield wiper activation, headlight use, loca information and data from other sensors. HEREWeGo is also offered. It is a navigation ap	Datasets are proprietary. HERE Industry website: www.here.com

	Name	Description	Contact
	INRIX Analytics	INRIX Analytics provides a number of products rely in part on crowdsourced data. These datas and tools include historic speed and travel time data; accurate origin-destination data; volume a other location-based and market-based analytic parking matrix data and traffic information, inci and road closures for major roads.	INRIX offers a variety of ways to acces INRIX XD Traffic information including traffic tiles, a monitoring site, flexible A and TPEG Connect. INRIX Americas Headquarters 10210 NE Points Dr., Suite 400 Kirkland, WA 98033 (425) 284-3800 info@inrix.com http://inrix.com/
	CELLINT Traffic Solutions	Cellint uses the cell tower signals provided by c phones to assign GPS coordinates to each data point in real time. The system then matches the crowdsourced live signaling data from the netw with cellular signaling pattern maps to determi origin and destination pairs. Thus, the system t various active mobile phones on the roads anonymously to provide the traffic information.	Datasets are proprietary. USA Address: 175 Bellechasse Drive, Chesterfield, MO 63017 Tel: +1-314-728-6255 www.cellint.com/contact-us/contact-u
Ride Sharing and Public Transit	UBER Movement	Uber Movement provides anonymized data fron over two billion trips for use by city officials or researchers, not by commercial entities. The historical data is available to analyze patterns a impacts of events, rush hours, and road closure etc. All data is anonymized and aggregated to ensure no personally identifiable information o behavior can be surfaced through the Moveme tool. Data is available under a Creative Commo Attribution Non-Commercial license.	The Movement Tool can be downloade data is downloadable in machine-read format. Where permitted, shapefile exp are available for use of Movement data geospatial applications. UBER MOVEMENT https://movement.uber.com movement-research@uber.com
	NYC Taxi & Limousine Commission	The New York City Taxi and Limousine Commis provides crowdsourced data on taxi trips includ pick-up and drop-off dates/times, pick-up and d off locations, trip distances, itemized fares, rate types, payment types, and driver-reported pass counts. The For-Hire Vehicle trip records includ fields capturing the dispatching base license number and the pick-up date, time, and taxi zon location ID (in shapefile format).	Data is downloaded on the website. Da available in CSV; taxi zone shapefiles a also available. Website: http://www.nyc.gov/html/tlc/ html/about/trip_record_data.shtml

	Name	Description	Contact
	MOOVIT	Moovit owns and manages a repository of trans data, generating hundreds of millions of data p a day from its App users and 180,000 local edit More than 100 million users utilize Moovit for t urban mobility in more than 1,500 cities in 77 countries and 43 languages. The App provides information on Trip planning (e.g., next arrivals, service alerts, etc.), Bus data (e.g., Real-Time B Arrival Data, digital departure boards in station voice announcements on vehicles), and an Age Dashboard that provides live bus tracking, plan versus actual schedule arrivals, and advanced analytics. Data sets are used by cities and tran agencies to analyze travel patterns to and from specific regions, on a specific line, or at a speci station.	The data is anonymized and aggregate from Moovit. Nir Bezalel VP Research and Development press@moovitapp.com Phone: 972 8 621 3141 https://www.solutions.moovit.com/
Pedestrian and Bicycling	STRAVA Metro/Labs	STRAVA is an App that tracks and maps locatio real-time and collects other metric data (e.g., sp heart rate, etc.) of bicyclists and runners. Each records his/her activity and then shares it with specific followers. Strava then merges routes frequently utilized to create Running or Bicyclin Route Guides. Strava Metro and Strava Labs al use running and bicycling data for other project	Strava Headquarters 500 3rd Street #110 San Francisco, CA 94107 https://www.strava.com/ https://labs.strava.com/ https://metro.strava.com/ Email:partner@strava.com
	WALKScope	WALKscope is an App developed by WalkDenve PlaceMatters for collecting data related to sidewalks, intersections, and pedestrian counts the Denver metro area. This information create inventory of pedestrian infrastructure and coun identify gaps, and build the case for improveme Maps are available to view the data collected.	Jill Locantore at WalkDenver Email: jill.locantore@walkdenver.org http://www.walkscope.org/

	Name	Description	Contact
	Bikeshare (e.g. Divvy) Bikes	Divvy Bikes is a bike-share program in Chicago, Illinois. Divvy uses crowdsourcing to identify location points for bike station placement. An A available to determine bike accessibility at stat Public use trip data is available at https://www.divvybikes.com/system-data Trip data includes Trip start/end day, time and station; rider type based on membership. If the is a member, the dataset includes gender and y of birth.	Divvy Bikes https://www.divvybikes.com/ Data@DivvyBikes.com
	PROJECT SIDEWALK	In order to improve conditions in Washington, D for those with mobility impairments, a team at t University of Maryland is crowdsourcing a map sidewalk impediments. The Walk Score-inspired map, called Project Sidewalk, allows the public catalog and rate the accessibility of sidewalks curb ramps and any obstacles like fire hydrants crumbling pavement both physically and throug virtual mapping. Currently, there are 64,000 labe and 463 miles of D.C. roads covered.	Jon E. Froehlich Associate Professor Computer Science and Engineering University of Washington https://sidewalk.umiacs.umd.edu/ sidewalk@umiacs.umd.edu
Road Infrastructure Condition	STREET Bump	Street Bump is a crowd-sourcing App project th helps residents collect road condition data whi they drive. The data provides governments with time information. Street Bump produced by the Mayor's Office of New Urban Mechanics in Bos Designed and developed by Connected Bits.	1 City Hall Square 5th Floor Boston, MA 02201 617-635-0044 http://www.streetbump.org/ newurbanmechanics@boston.gov.
	FIX MY STREET Platform	FixMyStreet uses Open Source software to laur website that helps people to report street probl like potholes and broken streetlights. Problem reports are sent to site administrators and government authorities. Open source report-mapping software can be deployed anywhere in world. This software is most commonly used fo reporting street issues to government officials, flexible enough to fit any project that uses geographical points.	http://fixmystreet.org/ international@mysociety.org

	Name	Description	Contact
	SEE CLICK FIX	SEEClickFix is a website and App that connects citizens reporting infrastructure issues to city c county officials. It is used to report illegal dump potholes, graffiti, sidewalk problems and more. is proprietary.	770 Chapel Street, 3F New Haven, CT 06510 https://seeclickfix.com/ Toll-free: 1-(800)-369-9060 Local: 1-(203)-752-0777 Sales: sales@seeclickfix.com Partners: support@seeclickfix.com Citizens: contact@seeclickfix.com

SHRP2 Project Data

The United States Congress created the second Strategic Highway Research Program (SHRP 2) in 2005 to address the challenges of moving people and goods efficiently and safely on the nation's highways. Currently, SHRP 2 is administered by the Transportation Research Board of the National Academies, under a Memorandum of Understanding with the Federal Highway Administration (under the U.S. Department of Transportation) and the American Association of State Highway and Transportation Officials.

SHRP 2 addresses four strategic focus areas: the role of human behavior in highway safety; rapid renewal of aging highway infrastructure; congestion reduction through improved travel time reliability; and transportation planning that better integrates community, economic, and environmental considerations into new highway capacity. A naturalistic driving study investigates ordinary driving under real-world conditions in order to make the driving experience safer (SHRP2 2015).

Between 2010 and 2013, 3,400 drivers participated in the Second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study (NDS), which produced over 4,300 years of naturalistic driving data. Data were collected from six sites around the United States. The largest collection sites were in Seattle, Washington; Tampa, Florida; and Buffalo, New York.

Participant vehicles were instrumented with a data acquisition system (DAS) that collected four video views (driver's face, driver's hands, forward roadway, rear roadway), vehicle network information (e.g., speed, brake, accelerator position), and information from additional sensors included with the DAS (e.g., forward radar, accelerometers) (Hankey *et al.* 2016). Dingus *et al.* (2015) and Antin (2011) provided more information about the data collection sites and how they compare to national data in this project.

How to Access the Database with PII

The Naturalistic Driving Study data are considered human subjects' data and must be protected in accordance with federal law. It also contains sensitive and personally identifiable information (PII) such as face video and GPS traces. However, privacy protections promised to participants in the research protocol and consent forms regarding their data continue even after the study ended. Yet, researchers need to have access to the data in order to maximize the financial and personal investment made by the sponsors and the participants.

The most commonly used PII data elements include driver face video; full trip GPS traces which can be used to identify a person's home, work, and school locations; and unaltered forward video of a crash. There are other less commonly

used yet potentially identifying data elements. Data elements that contain any personally identifying information (PII) is available only within a secure data enclave (SDE) to protect the privacy of study participants. Qualified researchers who wish to view and analyze PII must meet eligibility criteria and agree to the requirements of a data use license (DUL) (TRB 2018). In order to both uphold the privacy of study participants and promote the use of the data in important research, a process was created for researchers to establish a data use license (DUL) under which terms they may work with the data. In many cases, Institutional Review Board (IRB) approval of a qualified researcher's plans for data use will be required due to the involvement of human subjects.

Layers of the Database

There are several types of data in the SHRP 2 project. The variables can be classified into several layers namely, Participant Assessments, Vehicle Information, Continuous Data, Trip Summary Data, Event Data, cell phone data and roadway data. The following is a breakdown of the data elements in each layer.

Participant Assessments

- Demographic Questionnaire,
- Driving History
- Driving Knowledge
- Medical Conditions and Meds
- ADHD Screening
- Risk Perception
- Frequency of Risky Behavior
- Sensation Seeking Behavior
- Sleep Habits
- Visual. Physical, and Cognitive Test Results
- Exit Interview

Vehicle Information

- Make, Model Year, Body Style
- Vehicle Condition (tires. battery. etc.)
- Safety and Entertainment Systems

Continuous Data

- Face. Forward. Rear, and Instrument Panel Video
- Vehicle Network Data
- Accelerometers/Gyros, Forward RADAR. GPS
- Additional Sensor Data

Trip Summary Data

- Characterization of Trip Content
- Start Time and Duration of Trip
- Min, Max, Mean Sensor Data
- Time and Distance Driven at Various Speeds, Headway,
- Vehicle Systems Usage

Event Data

- Crashes, Near Crashes, Baselines
- 30s Events with Classifications
- Post-Crash Interviews
- Other Crash Data

Cell Phone Records

- Subset of participant drivers
- Call time and duration
- Call type (text call. pic. etc.)

Roadway Data

- Matching trip GPS to roadway database
- Roadway classifications
- Other roadway data (Hankey *et al.* 2016)

Moreover, detailed investigations of selected crashes are also available for researchers. Some of the available data are listed below:

- Multiple Videos
- Machine Vision Eyes Forward Monitor
- Machine Vision Lane Tracker
- Machine Vision Driver ID
- Accelerometer Data (3 axis)
- Rate Sensors (3 axis)
- GPS
- Latitude, Longitude, Elevation, Time, Velocity
- Forward Radar
- X and Y positions
- Xdot and Ydot Velocities
- Cell Phone
- ACN, health checks, location notification
- Health checks, remote upgrades
- Illuminance sensor
- Passive alcohol sensor
- Video
- Incident push button
- Audio (only on incident push button)
- Turn signals
- Vehicle network data
- Accelerator
- Brake pedal activation
- ABS
- Gear position
- Steering wheel angle
- Speed
- Horn
- Seat Belt Information
- Airbag deployment
- Many more variables (SHRP2 2015)

Strava Metro Data

Strava is a smartphone app, allowing individuals to record, track, and upload their cycling, walking, running or other fitness activities. Typically, bicyclists, pedestrians, and joggers use this app to keep track of their activities privately or publicly. Strava Metro data are the anonymized and aggregated form of activity data of Strava users, who agree to share their activities through the app publicly. The data are aggregated at road segment levels and are currently provided in GIS shapefile.

Strava Metro Data provide information on distance, times, pace, trail routes, and other geographic information of bicycle rides and walks/runs. The fields that are most typically available include number of unique athletes, number of trips, median of times taken by all trips, and sum of commute activities for each road segment. The data are available per direction of road segment and for different time periods (e.g., times of day, weekday, weekend, season, month, and year).

Safety professionals, including departments of transportation, urban city planners, and other safety organizations can use Strava data for assessing the needs of bicyclists and pedestrians. Strava Metro Data help to understand users' riding and walking behavior, e.g., the routes mostly traversed, peak commute times, trip lengths not only at a point but across the entire network of streets of an area. Such visualization helps in taking some proactive measures for ensuring safety of bicyclists and pedestrians. Similarly, Strava data can help in understanding how a new infrastructure development is impacting bicyclists and pedestrians.

Strava Metro Data are not available for public use. Organizations can purchase the data for specific jurisdictions (e.g., city, metropolitan area, county, district) and specific years. The fee depends on the population of the geographic area, the number of years for which data are requested, and the level of detail present in the dataset.

How to Access the Database With PII

Because Strava Metro data are anonymized and aggregated at road segment levels, no PII information is available in the data.

Layers of the Database

Strava Metro data has a single layer, which corresponds to street segments. While the street segment in Strava Metro data has a unique identifier, it does not match with agency's street identifier. Given licensee's requirements, the data may include an area identifier (e.g., city, county, and/or district) of the street segments.

References

ACS, 2018. American college of surgeons, national trauma data bank. FACS.

AHRQ, 2018a. Apcd to asc x12 relationship mapping. United States Health Information Knowledgebase.

AHRQ, 2018b. Medical expenditures panel survey (meps). In: Quality, U.D.O.H.H.R.a.F.H.R.A. ed., Agency for Healthcare Research and Quality

Antin, J.F., 2011. Design of the in-vehicle driving behavior and crash risk study: In support of the shrp 2 naturalistic driving study Transportation Research Board.

BCJ, 2009. Vrm department of motor vehicles databases. In: Justice, B.C.F. ed. Brennan Center for Justice New York, NY.

Dennis, E.P., 2015. Crowdsourcing transportation systems data. Michigan Department of Transportation (MDOT) Center for Automotive Research (CAR). Michigan Department of Transportation (MDOT) Center for Automotive Research (CAR), Michigan Department of Transportation (MDOT) Center for Automotive Research (CAR).

Dingus, T.A., Hankey, J.M., Antin, J.F., Lee, S.E., Eichelberger, L., Stulce, K.E., McGraw, D., Perez, M., Stowe, L., 2015. Naturalistic driving study: Technical coordination and quality control.

DOT, 2017. Highway performance monitoring system field manual. US Department of Transportation, Federal Highway Administration.

FHWA, 2018. What is mire? Federal Highway Administration, Wasgington DC, USA.

Gilbreath, A., Dinkins, C.R., 2003. Hipaa in daily practice Kerlak Enterprises, Inc.

Hankey, J.M., Perez, M.A., McClafferty, J.A., 2016. Description of the shrp 2 naturalistic database and the crash, near-crash, and baseline data sets. Virginia Tech Transportation Institute.

HIPAA, 2018. De-identification of protected heath information. In: Journal, H. ed.

Misra, A., Gooze, A., Watkins, K., Asad, M., Le Dantec, C., 2014. Crowdsourcing and its application to transportation data collection and management. Transportation Research Record: Journal of the Transportation Research Board (2414), 1-8.

MMUCC, 2012. Model minimum uniform crash criteria. DOT HS 811, 631.

NCSL, 2017. Collecting health data: All-payer claims databases. In: Legislatures, N.C.O.S. ed.

NHTSA, 2016. Fatality analysis reporting system (fars) analytical user's manual 1975-2015. US DOT.

NOAA, 2017. Federal meteorological handbook US Department of Commerce. National Oceanic and Atmospheric Administration

NTDB, 2018. National trauma data bank®ntdb research data set user manual and variable description list admission years 2002-2016. American College of Surgeons, Chicago, IL.

NWS, 2018. Aviation weather center metar data. National oceanic and atmospheric administration. In: Service, N.W. ed. National Weather Service

Porter, J., Love, D., Peters, A., Sachs, J., Costello, A., 2014. The basics of all-payer claims databases: A primer for states. Princeton, NJ: Robert Wood Johnson Foundation.

SHRP2, 2015. Strategic highway research program. strategic highway research program. FOTNETDATA.

TDH, 2013. Ambulatory surgical treatment center data system user manual. Office Of Health Statistics Tennessee Department Of Health, Nashville, TN. USA.

TDH, 2017. Hospital discharge data system user manual. In: Health, O.O.H.S.T.D.O. ed. Office Of Health Statistics Tennessee Department Of Health, Nashville, TN. USA.

TnDH, 2014. Death statistical file user manual. Division of Health Statistics Tennessee Department of Health.

TRB, 2018. How do i get access to the safety data?

Zehtabchi, S., Nishijima, D.K., McKay, M.P., Clay Mann, N., 2011. Trauma registries: History, logistics, limitations, and contributions to emergency medicine research. Academic emergency medicine 18 (6), 637-643.

Appendix B

Five Case Studies on Data Integration

Case Study 1

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

Sarah Doggett[a]*, David Ragland[a], Grace Felschundneff

[a] Safe Transportation Research and Education Center (SafeTREC). University of California, Berkeley

*Corresponding Author: doggett_sarah@berkeley.edu

Abstract

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

Introduction

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

According to the Federal Highway Administration (FHWA), traditional crash data sources are insufficient because they exclude both crashes that take place in non-roadway locations (e.g., parking lots, driveways, and sidewalks) and bicycle crashes and pedestrian injuries that do not involve motor vehicles (Stutts and Hunter 1999).

Ideally, the degree of bias in police collision reports could be measured and accounted for in analyses (Elvik and Mysen 1999). This process requires both the reporting level and the uncertainty in the estimate of the reporting level to be known (Hauer and Hakkert 1988). Although researchers have estimated reporting level by linking police collision reports with hospital data, few have determined the level of uncertainty surrounding their estimates.

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

Methodology and Definitions

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

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

Through this process, articles were identified, published between 1999 and 2017. Five of the articles explored data linkage in general but mentioned findings specific to pedestrians and/or bicyclists. Two focused exclusively on bicyclists, two exclusively on pedestrians, and one on both pedestrians and bicyclists.

Most of the studies included in the literature review defined underreporting as hospital records without counterparts in police crash reports. However, Janstrup et al. (2016) found that there are also police crash reports that do not have counterparts in hospital records. This expands the definition of underreporting to crashes that exist only in one dataset. This concept is illustrated in Figure 1.

Some crashes may be reported in both datasets but cannot be linked due to errors in data recording—these crashes are *misreported* but cannot not be distinguished from underreported crashes using current linkage techniques. There are also crashes that are *not reported* to either the police or to hospitals (D in Figure 1), due to lack of serious injury or an unwillingness on the part of those involved in the crash to contact the authorities. To identify these unreported crashes, other datasets such as those collected by insurance companies can be analyzed. One of the studies addressed this and found that the use of only police and hospital injury data can result in a significant underestimation in the actual number and severity of crashes (Short and Caulfield 2016).

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

- A: Cases only appearing in police crash reports
- B: Cases appearing in both police crash reports and hospital records
- C: Cases only appearing in hospital records

- D: Cases that do not appear in either police crash reports or hospital records

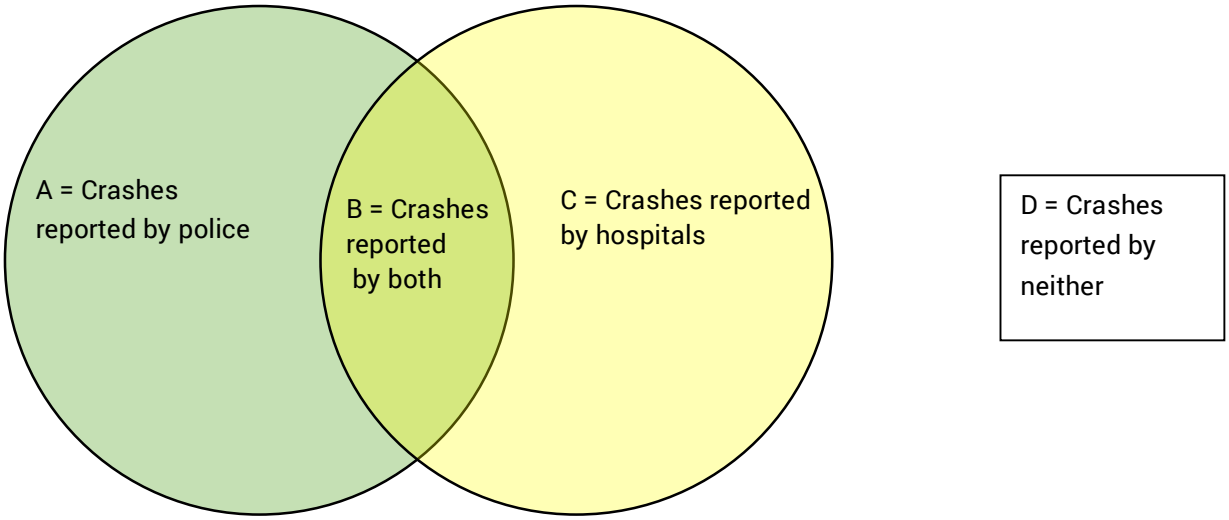

Figure 1 Patterns of Reporting Pedestrian and Bicyclist Crashes by Source

Several studies used the capture-recapture method to determine reporting level, which is not derived from those shown in Table 1. Originally, this method was used to estimate animal populations based on how many animals were captured and then later recaptured in a least two random samples (Tin Tin, et al. 2013). However, researchers have noted that this method may not be appropriate for estimating reporting levels because some of its underlying assumptions, such as a closed study population and an equal probability for each individual to be captured in each sample, may be violated (Tin Tin, et al. 2013, Janstrup, et al. 2016).

The definition used can significantly affect the estimated reporting level (Elvik and Mysen 1999). As an example, assume that 50 cases only appear in police reports (A), 20 cases appear in both police and hospital reports (B), and 30 cases only appear in hospital reports (C). In this scenario, hospital link rate would be 0.4, police link rate would be 0.29, police ascertainment rate would be 0.7, and hospital ascertainment would be 0.5. While definitions 3 and 4 are theoretically the most correct definitions of reporting level, few studies use it when estimating reporting level.

Table 1 Definitions of Reporting Level

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

Literature Review

Due to the different definitions of reporting level, different periods of study, and different study locations, it is difficult to directly compare the ten studies. As shown in Table 2, six of the ten studies used hospital link rate as their definition of reporting level, while two used capture-recapture, two used police link rate, two used police ascertainment rate, and

one used hospital ascertainment rate (this adds up to more than ten because the article by Short and Caufield reported on four different definitions). Even among the six studies using the hospital link definition, estimates of reporting levels for pedestrian crashes range from 44 to 75 percent; for bicycle crashes, estimates range from 7 to 46 percent.

However, the articles all agreed that police collision reports have limitations and underreport certain types of crashes, especially those involving pedestrians and bicyclists. For example, Stutts and Hunter (1999) found that police were unlikely to be contacted about pedestrian incidents not involving motor vehicles, although more than two-thirds of serious injuries fell into this category. Sciortino (2005) found that pedestrian injury victims who were African-American, male, or sustained minor injuries were more likely to be underreported in police crash records. The author suggested that such bias was likely due to the reluctance on the part of pedestrians to summon the police if the police were not initially present at the crash scene. Tarko and Azam (2011) found that pedestrians were less likely to be included in the linked database if they were struck by vehicles on state roads, at Y intersections, or on divided roadways. However, they were more likely to be included in the database if they were struck while crossing a road instead of walking along or standing outside of a roadway.

Drivers are more likely to be included in crash reports than are cyclists (Conderino et al. 2017). Langley (2003) found that only 22 percent of bicyclist crashes on public roads could be linked to a crash report—the percentage increased to 54 percent when it included crashes that also involved motor vehicles. Because crash reports are mainly focused on incidents that involve motor vehicles and which occur on public roadways, they likely capture fewer than one-third of bicyclist injury cases serious enough to require medical treatment (Stutts and Hunter 1999). According to Janstrup et al. (2016), the underreporting rate for crashes involving cyclists in Denmark was 14 percent for those resulting in serious injuries, and 7 percent for those resulting in slight injuries.

The literature identified several possible reasons for underreporting. According to Langley (2003), motor vehicle crashes listed in hospital data are assumed to have occurred on roadways unless another location is specified. Thus hospital derived estimates may overstate the number of motor vehicle crashes on public roads and understate crashes that occur in other locations and those that do not involve motor vehicles. Another possible reason for underreporting is that pedestrian and bicycle crashes are less likely to result in insurance compensation than are motor vehicle crashes. Lujic et al. (2008) found that those entitled to insurance compensation had higher linkage rates than those who were not, presumably because police reports were required as a condition for receiving compensation. According to Watson et al. (2015), it is possible that the severity of injuries resulting from a collision with another vehicle is likely to be more serious, and therefore more likely to be reported.

Several studies have shown that compared with hospital data, police crash reports do not accurately report injury severity. Injury classifications in crash reports are usually based on the KABCO scale, which is less nuanced than the Injury Severity Score that many hospitals use (CDPH 2015). Additionally, KABCO classifications are made at the scene of the crash by officers who typically lack medical training—therefore, less visible but life-threatening injuries such as internal bleeding may be misclassified as non-severe, while more obvious minor injuries such as “minor lacerations with profuse bleeding” may be misclassified as severe *(13)*. Another problem related to estimation of injury severity is that the injury classification in police reports is static and does not necessarily reflect subsequent developments (WABA 2015).

Potential Implications of Study Results

Findings from these ten studies indicate that there is a severe underreporting problem in datasets based only on police collision reports. Underreporting can result in the “projection of false estimates pertinent to crash and fatality rates, and [the] detection of wrong parameters responsible for crash occurrence, thereby making the entire road safety exercise ineffective,” according to Ahmed, Sadullah, & Yahya (2017). When road safety is measured based on the

amount of crashes reported instead of the actual number of crashes that occurred, there is a tendency to mistake trends in crash reporting with trends in traffic safety (Hauer and Hakkert 1988). In addition, the authors found that inaccurate crash data can result in improper prioritization of funding and resources, and under- or overestimation of injury severity risk. For example, Abay (2015) found that estimates based on underreported police-reported crash data minimize the effectiveness of seat belt use in injury severity risk and could have serious policy implications.

In cases of crash underreporting, analysis relying on police data may be biased, according to Janstrup et al. (2016). Reliance on hospital data may also be problematic, as Watson et al. (2015) reported that the level of underreporting varied depending on the data with which the police data was linked. Watson found that when hospital data was examined, approximately two-thirds of the data were not linked to police data. Similarly, when Short and Caulfield (2016) added injury claims data to their analysis, the total number of identifiable injuries was found to be more than three times greater than the number identified by police reports, and five times greater than the number hospitalized.

Because of the inaccuracy of injury classifications in police collision reports, injury severity cannot be used to match datasets. In addition, by only using injury data from police reports, financial costs to crash victims or to the public for health care associated with a crash cannot easily and accurately determined (WABA 2015).

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

Source	Study Period	Study Location	Focus	Definition Used	Findings/Conclusions
Conderino, Fun Sedlar & Norto	2009-2015	New York City	General	Hospital Link Rate	• 50% of hospital reports involving a pedestrian crash lin police report • 45% of hospital reports involving a bicyclist crash linke police report • Sensitivity - 74% • Specificity - 93%
Janstrup et al.,	2003-2007	Funen, Denmark	General	Capture-Recapture	• Compared with car occupants, pedestrians are more lik appear in both police and hospital databases; bicyclists a more likely to appear in either • Only 7% of bicycle crashes resulting in slight injury and 15% of bicycle crashes resulting in severe injury are repo the police
Langley, 2003	1995-1999	New Zealand	Bicyclist	Hospital Link Rate	• Only 22% of bicycle crashes on public roads could be li a crash report • When limited to bicycle crashes on public roads involvi motor vehicles, 54% could be linked to a crash report
Lujic, Finch, Bou Hayen & Dunsm 2008	2000-2001	New South Wale	General	Hospital Link Rate	• 69% of road traffic crashes were linked to police record • Drivers were most likely to have their hospital records to police records (83%) • 46% of bicyclist crashes and 75% of pedestrian crashes linked to police records • Authors hypothesized that underreporting for cyclists i "ambiguity of…laws and regulations" and the fact that cy are "less likely to cause property damage"
Sciortino, Vassa Radetsky & Knu 2005	2000-2001	San Francisco	Pedestrian	Police Ascertainme	• Police reports underestimate the number of pedestrian injuries by 21% (e.g., reporting level is 79%) • African-American pedestrians were less likely than whi pedestrians to be linked to a police report • Women were more likely than men to be linked to a po report
Short & Caulfie	2005-2011 (and Hospita 2010-2011 (Board Data)	Ireland	General	Hospital Link Rate, Link Rate, Hospital Ascertainment Rat Police Ascertainme	• For pedestrian injuries, 28.9% of police records were m with a hospital record; 44.3% of hospital records were m with a police record • For bicyclist injuries, 24.8% of police records were mat with a hospital record; 8.2% of hospital records matched police record •Police Ascertainment Rate was 73.4% for pedestrians a 26.4% for bicyclists •Hospital Ascertainment Rate was 47.7% for pedestrians 80.2% for bicyclists • False Positive Rate – 13% • False Negative Rate – 13%
Stutts & Hunter	1995-1996	Various location California, New Y and North Caroli	Pedestrian and Bicyclist	Hospital Link Rate	• 70% of bicyclist injuries reported by hospitals did not in motor vehicle • 64% of pedestrian injuries reported by hospitals did no involve a motor vehicle • 31% of bicyclist and 53% of pedestrian injuries occurre non-roadway locations (e.g., sidewalks, parking lots, trail • Police crash reports capture less than 33% of serious b injuries
Tarko & Azam,	2003-2008	Indiana	Pedestrian	Police Link Rate	• Pedestrians struck on a state road, at a Y intersection, divided roadway were less likely to be included in both p and hospital databases • Pedestrians struck while crossing a road, as opposed to walking along or standing outside of the roadway were n more likely to be included in both databases • Authors hypothesized that more severe injuries were n likely to appear in both databases

Source	Study Period	Study Location	Focus	Definition Used	Findings/Conclusions
Tin Tin, Woodw Ameratunga, 2(	2006-2011	New Zealand	Bicyclist	Capture-Recapture	• Police reports were linked to insurance, hospital, and mortality records • 13% of hospital reported crashes and 64% of hospital r collisions were linked to police records • 39% of police-reported crashes and 43% of police-repo collisions were linked to hospital records • When compared with self-reported data from the cycli entire linked dataset had a sensitivity of 63.1% and speci 93.5% • The collision-only dataset showed a 40.0% sensitivity a 99.9% specificity
Watson, Watso Vallmuur, 2015	2009	Queensland	General	Hospital Link Rate	• The study used discordance rates between police data hospital records to measure underreporting of crashes t the police • The discordance rate was 44% for pedestrians and 93% bicyclists (e.g., a reporting level of 56% and 7% respectiv • Authors hypothesized that bicyclist injuries are not rep to the police because injuries are generally less serious, a likely to have insurance implications, and are more likely involve young people, who generally have high discorda rates

Limitations of Existing Studies

Hauer and Hakkert (1988) have proposed methods that will account for underreporting in police crash reports. They argue that the variance of the estimated number of crashes that occur can be calculated if the reported number of crashes, the reporting level, and the variance of the reporting level are known. While the studies in this literature review have attempted to establish the reporting level associated with various types of crashes, few have reported on the error surrounding their estimates.

In most real-world cases, true match status is unknown, and link status is used as a proxy. With a perfect matching process, link status and match status would be identical, and link rate would be equivalent to the reporting level. However, there are two sources of error in the process of data linkage—false positives and false negatives. False positives are records that do not belong to the same person/event that are linked (Bohensky 2016). False positives are more likely to occur when identifiers are not discriminative and when files are large (Harron et al. 2016). False negatives, also known as *missed matches*, are records that belong to the same person/event that are not linked (Bohensky 2016). These occur when records have inaccurate or missing data (Harron et al. 2016).

Because linkage and analysis processes are often separated due to privacy concerns, researchers using linked datasets may be unable to determine the extent of bias that linkage errors has introduced into their study (Harron et al 2017).

It is difficult to predict the direction and magnitude of bias resulting from linkage errors due to the "the distribution of errors with respect to variables of interest" which is usually unknown (Harron et al 2017). Missed matches reduce sample size and statistical power and can lead to under-underestimates of exposures or outcomes if the linkage is informative (Harron et al 2017). Because missed matches do not necessarily occur at random, the linked data may not accurately represent the study population, reduce the viability of the research effort and may introduce bias (Harron et al 2017). Selection bias can occur if an individual's presence in the linked dataset is related both to exposure and an outcome of interest (Harron et al 2017).

When data for key variables are missing, cases that are linked are likely to be unrepresentative of the study population as they have less common values such as unusual zip codes (Bohensky 2016). Unfortunately, missing data is common in data linkage. The individual datasets used for linking—police, hospital, insurance, traffic—are generally not developed

for research purposes. The intended use of a dataset largely determines its collection method (Imprialou and Quddus 2017). Therefore, the contents and details of their attributes and the application of various datasets for purposes other than those for which they were designed could result in decreased data quality. Additionally, many variables that would aid incorrectly linking the datasets are often redacted for privacy purposes (i.e., name, address, social security number). Data may be inaccurate due to misspellings or lack of information (i.e., staff might guess the age of an unconscious patient and set the birth month and day to '01') (Christen 2012). Finally, the method of data collection can influence data quality because errors are more likely to occur when data is copied from handwritten forms or transcribed from conversations than when they are entered directly into a database (Harron et al. 2016).

While there are statistical methods for managing data that are missing completely at random or missing at random, it is likely that missing data in crash and hospital data linkages are usually missing not at random, which introduces systemic bias into the analysis of the linked dataset (Bohensky 2016). Studies of linked databases may be significantly impacted by linkage errors, especially if certain types of people or events are more or less likely to have the outcome of interest (Bohensky 2016). Often, the bias introduced by these errors cannot be determined because the errors themselves are unknown.

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

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

Although sensitivity and specificity can be affected by the method of data linkage used, they are much more dependent on the quality of the data that is being linked. Data quality is affected by the accuracy, completeness, consistency, timeliness, accessibility, and believability of the variables used to link the databases (Christen 2012).

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

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

Suggestions for Improvement

Prior to linking data, researchers should carefully examine the individual datasets to ensure data integrity and completeness. Bohensky (2016) recommends that direct, unique identifiers be included within datasets to reduce bias from incorrectly entered data or missing variables. Unfortunately, privacy concerns may prevent this method from being implemented in the United States. In the absence of direct identifiers, Bohensky (2016) suggests the use of financial incentives to encourage data custodians to improve data quality and consistency. Definitions of variables used for linking, such as injury severity, should be standardized at the national or international level so that datasets

from different sources can be reliably linked. Data linkage studies need to develop a clear and systematic method of reporting their methodology so that they can be easily repeatable by other researchers (Bohensky 2016).

Additionally, data linkers must develop indicators to describe the linkage errors present in a linked dataset (Bohensky 2016). Harron et al. (2017) suggests three approaches to evaluating linkage quality. The first approach is to use a gold standard dataset to directly measure missed and false matches—while this is the most accurate approach, it is difficult to apply this methodology in practice as gold standard data is often unavailable. The second approach is to compare the characteristics of linked and unlinked data to determine potential sources of bias—this approach requires a linkage design in which all records in at least one file are expected to link, as well as provision of characteristics of the unlinked records to the researchers. The third approach is to conduct a sensitivity analysis to evaluate how sensitive the results are in response to changes in the linkage method—this approach requires that researchers have access to match weights for each linked record. Match weights do not reveal any sensitive information and thus are more likely to be shared with researchers. However, the sensitivity analysis may be difficult to interpret as the effects of missed matches and false matches cannot be distinguished from each other (Harron et al. 2017).

Further research must be conducted to identify the populations that are most likely to suffer from selection bias during data linkage, in addition to determining the effects that different linkage processes have on reducing such bias (Bohensky 2016).

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

Conclusions

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

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

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

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

References

Ahmed, Ashar, Ahmad Farhan Mohd Sadullah, and Ahmad Shukri Yahya. 2017. Errors in Accident Data, Its Types, Causes and Methods of Rectification-Analysis of the Literature. Accident Analysis & Prevention, July. https://doi.org/10.1016/j.aap.2017.07.018.

Abay, Kibrom A. 2015. Investigating the Nature and Impact of Reporting Bias in Road Crash Data. Transportation Research Part A: Policy and Practice, 71 (January): 31-45. https://doi.org/10.1016/j.tra.2014.11.002

Bohensky, Megan. 2016. Bias in Data Linkage Studies. In Methodological Developments in Data Linkage, First, 63–82. John Wiley and Sons.

CDPH (California Department of Public Health). 2015. Exploratory Analysis of Injury Classification of Crash Victims Using Crash-Medical Linked Data in California. California Department of Public Health. https://archive.cdph.ca.gov/programs/Documents/Exploratory%20Analysis%20of%20Injury%20Classification%20of%20Crash%20Victims%20Using%20Crash-Medical%20Linked%20Data_Final.pdf.

Christen, Peter. 2012. Data Pre-Processing. In Data Matching, by Peter Christen, 39–67. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31164-2_3.

Conderino, Sarah, Lawrence Fung, Slavenka Sedlar, and Jennifer M. Norton. 2017. Linkage of Traffic Crash and Hospitalization Records with Limited Identifiers for Enhanced Public Health Surveillance. Accident Analysis & Prevention 101 (April): 117–23. https://doi.org/10.1016/j.aap.2017.02.011.

Elvik, Rune, and Anne Mysen. 1999. Incomplete Accident Reporting: Meta-Analysis of Studies Made in 13 Countries. Transportation Research Record: Journal of the Transportation Research Board 1665 (January): 133–40. https://doi.org/10.3141/1665-18.

Harron, Katie, Harvey Goldstein, and Chris Dibben. 2016. Introduction. In Methodological Developments in Data Linkage, First, 63–82. John Wiley and Sons.

Harron, Katie, James C Doidge, Hannah E Knight, Ruth E Gilbert, Harvey Goldstein, David A Cromwell, and Jan H van der Meulen. 2017. A Guide to Evaluating Linkage Quality for the Analysis of Linked Data. International Journal of Epidemiology 46 (5): 1699–1710. https://doi.org/10.1093/ije/dyx177.

Hauer, E, and A S Hakkert. 1988. Extent and Some Implications of Incomplete Accident Reporting. Transportation Research Record Methods for Evaluating Highway Improvements (1185): 1–10.

Imprialou, Marianna, and Mohammed Quddus. 2017. Crash Data Quality for Road Safety Research: Current State and Future Directions. Accident Analysis & Prevention. https://doi.org/10.1016/j.aap.2017.02.022.

Janstrup, Kira H., Sigal Kaplan, Tove Hels, Jens Lauritsen, and Carlo G. Prato. 2016. Understanding Traffic Crash Under-Reporting: Linking Police and Medical Records to Individual and Crash Characteristics. Traffic Injury Prevention 17 (6): 580–84. https://doi.org/10.1080/15389588.2015.1128533.

Langley, J D. 2003. Missing Cyclists. Injury Prevention 9 (4): 376–79. https://doi.org/10.1136/ip.9.4.376.

Lujic, Sanja, Caroline Finch, Soufiane Boufous, Andrew Hayen, and William Dunsmuir. 2008. How Comparable Are Road Traffic Crash Cases in Hospital Admissions Data and Police Records? An Examination of Data Linkage Rates. Australian and New Zealand Journal of Public Health 32 (1): 28–33. https://doi.org/10.1111/j.1753-6405.2008.00162.x.

Short, Jack, and Brian Caulfield. 2016. Record Linkage for Road Traffic Injuries in Ireland Using Police Hospital and Injury Claims Data. Journal of Safety Research 58 (Supplement C): 1–14. https://doi.org/10.1016/j.jsr.2016.05.002.

Sciortino, Stanley, Mary Vassar, Michael Radetsky, and M. Margaret Knudson. 2005. San Francisco Pedestrian Injury Surveillance: Mapping, under-Reporting, and Injury Severity in Police and Hospital Records. Accident Analysis & Prevention 37 (6): 1102–13. https://doi.org/10.1016/j.aap.2005.06.010.

Stutts, Jane C., and William W. Hunter. 1999. Injuries to Pedestrians and Bicyclists: An Analysis Based on Hospital Emergency Department Data. Final FHWA-RD-99-078. https://www.fhwa.dot.gov/publications/research/safety/pedbike/99078/index.cfm.

Tarko, Andrew, and Md. Shafiul Azam. 2011. Pedestrian Injury Analysis with Consideration of the Selectivity Bias in Linked Police-Hospital Data. Accident Analysis & Prevention 43 (5): 1689–95. https://doi.org/10.1016/j.aap.2011.03.027.

Tin Tin, Sandar, Alistair Woodward, and Shanthi Ameratunga. 2013. Completeness and Accuracy of Crash Outcome Data in a Cohort of Cyclists: A Validation Study. BMC Public Health 13 (1). https://doi.org/10.1186/1471-2458-13-420.

Watson, Angela, Barry Watson, and Kirsten Vallmuur. 2015. Estimating Under-Reporting of Road Crash Injuries to Police Using Multiple Linked Data Collections. Accident Analysis & Prevention 83 (October): 18–25. https://doi.org/10.1016/j.aap.2015.06.011.

WABA (Washington Area Bicyclist Association). 2015. Modernizing the Collection, Integration and Disclosure of Crash Data: Policy Recommendations for the District of Columbia. Washington Area Bicyclist Association. http://www.waba.org/wp-content/uploads/2016/01/DC-Crash-Data-Policy-Paper-July-2015.pdf.

Case Study 2

Pre-Hospital Response Time and Traumatic Injury—A Review

Sarah Doggett[a]*, David R. Ragland[a], Grace Felschundneff[a]

[a]Safe Transportation Research and Education Center (SafeTREC), University of California, Berkeley

*corresponding author
Email: doggett_sarah@berkeley.edu

Abstract

A significant proportion of fatalities from motor vehicle collisions (MVC) could be prevented through better emergency medical service (EMS) care. Despite a lack of conclusive research, there is a consensus that prehospital time (the time between the MVC and the patient's arrival at the hospital) must be reduced as much as possible. Many studies use response time (the time between EMS dispatch and arrival at the scene) as an indicator of overall prehospital time and a metric of EMS performance. However, there are other components of prehospital time that may be equally important, including the discovery time between the collision and EMS notification, the on-scene time, and the transport time between the scene and the hospital. In rural MVCs, the discovery time can be substantial if there are no witnesses or survivors capable of calling emergency services; technologies that automatically detect MVCs can shorten discovery times in such instances. Response times depend on the distance between the EMS vehicle and the scene; this time can be reduced through deployment strategies that place EMS vehicles close to where motor vehicle collisions are likely to occur. Transport times depend on the distance between the scene and the hospital; this time could be reduced by increasing access to trauma centers, especially in rural areas. On scene time is a component of the total time. However there is a trade-off between minimizing scene time to reduce total time, on one hand, and, on the other hand, providing optimal on-scene care. Increasing capacity of EMS personnel and/or utilizing technology such as telemedicine should be considered as part of this trade-off. Future research is needed to determine the relative benefits and costs of reducing any of these segments of prehospital time

.

Introduction

In 2016, there were 34,439 fatal motor vehicle collisions in the United States (NHTSA, 2018). It is possible that a significant proportion of these fatalities could have been prevented with faster response times and better emergency medical care. For this reason, the Towards Zero Deaths Steering Committee, comprised of representatives from various highway safety stakeholders, identified enhanced emergency medical services as a key area in its national safety strategy (Towards Zero Deaths Steering Committee, 2014).

This paper explores the factors affecting prehospital time, recommends strategies for reducing this time and improving prehospital care, and summarizes areas in which further research is needed.

Importance of Pre-Hospital Time

There is widespread belief in the significance of the 'golden hour' immediately following an injury, during which time resuscitation, stabilization and transport to a medical facility offer the greatest chance of survival for the patient (Harmsen et al., 2015). By reducing prehospital time, more advanced medical care can be provided sooner, resulting in reduced mortality. However, there is a lack of conclusive research on whether the golden hour is important for all types of injuries. In some cases, in which patients need to be treated prior to transfer to a hospital or trauma center, reduced prehospital time may have a negative impact on patients' health outcomes, especially if necessary treatment in deferred for the sake of getting the patient to the hospital as quickly as possible. Additionally, it is likely that reducing prehospital time below a certain threshold will lead to diminishing returns for patient survival. For example, patient outcomes may be similar for prehospital times below 15 minutes, meaning that a reduction in prehospital time from 15 minutes to 10 minutes may not change patient outcomes enough to justify the cost of the reduction.

Many studies use response time, the time between EMS dispatch and arrival at the scene, as an indicator of overall prehospital time and a metric of EMS performance. However, the literature disagrees about whether response time is an important factor in patient survival. For example, one study found that response times only improved survival chances when they were less than five minutes (Blackwell and Kaufman, 2002). A study of severely injured trauma patients in Korea found that longer prehospital times had no impact on mortality and that mortality decreased when scene times were longer than six minutes (Kim et al., 2017). However, another study found that increased EMS prehospital time was likely to be associated with higher mortality rates in rural areas in which the mean EMS response time was greater than 14 minutes (Gonzalez et al., 2009). A study of motor vehicle collisions in Spain found that when response times were reduced from 25 minutes to 15, the probability of fatality decreased by one third (Sánchez-Mangas et al., 2010). In yet another study, the author cautioned that some options for decreasing ambulance service times not only have prohibitive costs, but increase the risk of safety to patients, emergency medical service personnel, and the public because ambulances have a high collision rate. The author further stated that patient outcome should be the main standard of EMS performance (Al-Shaqsi, 2010).

In addition, response time may not be important for all types of patients. Reductions in prehospital time may have a greater impact on less severely injured patients since extremely severely injured patients are likely to die regardless of how soon they are treated (Gonzalez et al., 2009). Response time may also have different impacts depending on the type of emergency—most studies on the impact of response time have been focused on cardiac arrest patients (Wilde, 2013).

According to one study, previous research failed to account for the endogeneity (i.e., EMS personnel are usually aware of how serious an injury is and may adjust the speed of their response accordingly when determining the impact on patient outcomes) which may have resulted in biased estimates of the effect of response times downward toward finding no effect. When Wilde accounted for this endogeneity by using an instrument variable, which affects the

explanatory variable but has no independent effect on the dependent variable, the author found that a one-minute increase in response time led to an 8 to 17 percent increase in mortality (Wilde, 2013).

There are other components of prehospital time that may be of equal, or greater, importance than response time. One study defined total prehospital time as the time between the emergency call and arrival of the patient at the hospital (Harmsen et al., 2015). The author further breaks this time down into activation time (the time between the call and EMS deployment), response time (the time between the call and EMS arrival at the crash scene), on-scene time (the time spent on scene by EMS), and transport time (the time between leaving the crash scene and arriving at the hospital). Other important segments of prehospital time include the discovery time between the crash and the 911 call, and the time between arriving at the scene and arriving at the hospital, which can be substantial if vehicle extraction is necessary.

Factors Affecting Pre-Hospital Time

Incident Detection and Dispatch

Traditionally, EMS agencies are notified of a motor vehicle collision when a witness or someone involved in the collision makes a 911 call. If there are no witnesses and no one involved in the crash is capable of making an emergency call, the discovery time between the accident and EMS notification can be dangerously long. One study found that in the state of Texas this incident detection time is three times as long for rural motor vehicle collisions as it is for urban ones, probably due to the geographic isolation of rural areas (Lu and Davidson, 2017).

There are also in-vehicle systems, known as Advanced Automatic Collision Notification (AACN) technologies, that can detect a crash occurrence and send its location to independent emergency call centers. Personnel at these centers then attempt to make voice contact with vehicle occupants, and if they determine it is not a false alarm, they make a 911 call on the occupants' behalf. These systems can also estimate the likelihood of severe injury in a crash—however the algorithms used for these estimates are not consistent across various technologies (Towards Zero Deaths Steering Committee, 2014).

Even when a 911 call is made directly after a crash occurs, there may still be a significant period of time before an EMS vehicle can be dispatched. Traditional 911 systems can identify a caller's location automatically only when the call comes from a landline (Towards Zero Deaths Steering Committee, 2014). Otherwise, the EMS dispatcher must rely on information from the witness or victim, which may be unreliable or unavailable. Enhanced 911 systems have the ability to accurately locate calls made from cell phones and can identify the nearest emergency call center, however, this system has not yet been fully implemented nationwide. Recently, next generation 911 (NG911) systems have been developed. These systems can receive text messages, pictures, and videos from callers to provide responders with more accurate information about the crash scene (Towards Zero Deaths Steering Committee, 2014). As of 2017, a total of 6 states are completely covered and 13 states are partially covered by NG911 capable services (National 911 Program, 2017). Using NG911, AACN technologies could directly contact EMS agencies, bypassing the need for independently operated emergency call centers. These systems reduce the time between the 911 call and EMS dispatch.

Differences in Rural and Urban Accessibility

While only 19 percent of the United States population resides in rural areas, over half of all traffic fatalities involve rural motor vehicle collisions. In 2011, a total of 75 percent of drivers who were injured in motor vehicle collisions and died during transport to the hospital were in rural areas (Towards Zero Deaths Steering Committee, 2014).

Rural motor vehicle collisions are not intrinsically more deadly—one study found that rural and urban motor vehicle crashes result in similar injury severities (Gonzalez et al., 2009). Mortality rates are similar for severely injured patients

regardless of whether the incident occurs in a urban or a rural setting; this indicates that patients with lower injury severity contribute to the generally higher mortality rate in rural areas (Gonzalez et al., 2009).

This discrepancy could be caused by the relative inaccessibility of trauma centers in rural areas. Although patients who are treated at Level 1 trauma centers within one hour of injury are 25 percent less likely to die as a result, more than 45 million U.S. citizens live over an hour away from a Level 1 or 2 trauma center (Towards Zero Deaths Steering Committee, 2014). In a study of motor vehicle collisions in Texas, Lu and Davidson found that activation time, response time, and transport time were significantly longer for fatal motor vehicle collisions in rural areas than in urban areas (Lu and Davidson, 2017).

Land Use

According to one study, urban sprawl is associated with longer EMS response time (Trowbridge et al., 2009). The authors found that counties with characteristics of sprawl including low density construction, limited street connectivity, and distance between residential development and civic and commercial districts showed greater probability of delayed ambulance arrival than counties with smart growth features. The authors asserted that integrating more comprehensive land-use metrics, including measures of urban sprawl, into EMS dispatch algorithms could improve use of resources and potentially improve response reliability (Trowbridge et al., 2009).

Factors Affecting Pre-Hospital Care

Triage

In the context of prehospital care, triage, the process of prioritizing actions in an emergency, begins when a 911 call is received, at which point dispatchers must decide which EMS crew to send to the scene. Triage continues when EMS responders decide whether the emergency requires the use of sirens and lights as they travel to the scene. Once at the scene, the responders must decide whether to stabilize the patient at the site or to rush the patient to the hospital. Finally, the responders must decide which trauma center is most appropriate and whether use sirens and lights is necessary for the trip.

Without proper triage, EMS agencies cannot effectively prioritize which resources to provide to which patients. Unnecessary use of sirens and lights by EMS vehicles can result in harm to ambulance crews and the public (Al-Shaqsi, 2010). EMS personnel may conclude that the use of such warning signals allow them to disregard stop signs or traffic signals and to drive against traffic. In addition, drivers are generally unclear on how to respond to visual and audible signals from emergency vehicles (Sanddal et al., 2010). EMS responders have fatality rates as high as those for police and firefighters, and 75 percent of these fatalities involve transportation (Towards Zero Deaths Steering Committee, 2014). Ambulances are at least seven times more likely to crash than heavy trucks, and two-thirds of fatalities caused by ambulance collisions are among occupants of other vehicles or pedestrians (Towards Zero Deaths Steering Committee, 2014). When EMS vehicles use sirens and lights, these risks may be exacerbated. However, there is currently no evidence-based model that determines when the risk of using sirens and lights is justified; instead, individual EMS responders are responsible for this decision (Towards Zero Deaths Steering Committee, 2014).

Furthermore, the EMS vehicle that is closest to the scene may lack the necessary equipment to address a particular emergency. Although sending the closest vehicle may result in the shortest response time, comprehensive telephone triage that identifies the tools needed for the situation prior to deciding which vehicle to send could improve patient survival, even while increasing prehospital times (Al-Shaqsi, 2010).

Telemedicine

Telemedicine is the provision of medical services via information and communication technologies to remotely located healthcare workers and patients. It is intended to extend the reach of medical specialty services and is of particular

benefit in the case of pre-hospital care in acute emergencies where treatment delays may negatively impact outcome (Amadi-Obi, Gilligan, Owens, & O'Donnell, 2014).

Amadi-Obi notes that telemedicine has been most extensively studied in the area of stroke management. Telemedicine can provide remote access to a stroke specialist, which the author asserts is a promising solution in locations lacking qualified local medical experts. This is known as the "hub and spoke model," wherein telemedicine links underserved areas (the spoke) with a centrally located stroke expert (the hub). The author further notes that telemedicine provides similar quality of care to in-person medicine, and contends that there is no statistical difference in short-term and long-term mortality between telemedicine and face-to-face consultation. Telemedicine is also more expensive than traditional medicine although there is potential for significant cost savings as a result of reduced length of hospital stays (Amadi-Obi, Gilligan, Owens, & O'Donnell, 2014).

There is less research on the impact of telemedicine on trauma management. According to Amadi-Obi, it is appropriate in addressing major incidents in which a significant deficit of healthcare professionals can be resolved via teleconsultation. For example, teleradiology has improved diagnoses and reduced the expense of transfer of trauma patients. Overall, the author recommends further research incorporating better study designs, larger sample size, and a focus on incorporating smartphone technology (Amadi-Obi, Gilligan, Owens, & O'Donnell, 2014).

Another study also presented concerns about the effectiveness and cost-effectiveness of telemedicine, describing most telemedicine studies as "methodologically weak before-and-after studies that rarely examine patient-centered outcomes," instead focusing on feasibility and convenience for patients (Kahn, 2015). Additionally, the author stated there may be potential unintended consequences related to the complex interpersonal and interprofessional relationships within the healthcare profession because telemedicine compels patients to accept medical advice without the benefit of in-person consultation, compromising patient trust and rapport. The study further suggested that research must examine the crucial issue of context—not only whether telemedicine works—but also how, when, and under which circumstances it works best (Kahn, 2015).

Recommendations

More research must be done to determine whether pre-hospital time is significantly related to patient outcome for motor vehicle collisions. If this research determines that pre-hospital time is important and should be reduced as much as possible, there are several ways to achieve that goal.

The time between the incident and EMS notification can be reduced through the implementation of automatic crash detection technologies in vehicles with the ability to accurately estimate injury severity and directly communicate with EMS dispatchers. These technologies could also improve triage by notifying dispatchers when specialized services such as vehicle extraction are necessary (Towards Zero Deaths Steering Committee 2014). This would prevent dispatchers from sending ill-equipped EMS vehicles to the scene and would reduce overall prehospital time. Triage can also be improved by fully implementing Next Generation 911 throughout the nation so that more detailed information can be provided to EMS responders before they reach the scene. Telemedicine could also improve triage and on-scene patient treatment by allowing more specialized medical professionals to provide input on the proper course of patient care. However, more research is necessary before this can be determined.

A number of strategies to reduce response time, the time between EMS dispatch and the arrival of the EMS crew at the scene, have been developed. There are two main ways of reducing EMS response time: having more EMS vehicles in service at the same time or by positioning the existing vehicles so that they have better access to emergencies (Peyravi et al. 2015). Because the former involves substantial financial costs, research has focused on ways to locate EMS vehicles more efficiently. This research includes dynamic load-responsive ambulance deployment (Peleg &

Pliskin, 2004), discrete event simulations (Wei Lam et al., 2014), decision support systems (Repede & Bernardo, 1994), and geospatial-time analysis of ambulance deployment (Ong et al., 2010). These strategies all involve using retrospective data, GIS information, and various models to determine where and when to locate ambulances. For example, Gonzalez found that when rural ambulance stations locations were moved from the area of highest population concentration to areas with high motor vehicle collision rates and/or major roads, EMS response times to motor vehicle collisions decreased (Gonzalez et al. 2011).

Response time could also be decreased by making it safer for EMS vehicles to travel quickly using sirens and lights. This could be achieved by mandating that EMS vehicles have collision avoidance and other safety systems installed (Towards Zero Deaths Steering Committee 2014). The installation of vehicle-to-infrastructure communication technology, such as road condition warning systems, would allow EMS vehicle drivers to adapt their routes according to real-time traffic and avoid areas where a collision would be likely (Towards Zero Deaths Steering Committee 2014).

Finally, the transport time between the scene and the hospital could be reduced. Making it safer for EMS vehicles to travel quickly could also reduce transport time. Another way of reducing transport time is by locating trauma centers so that there is a Level 1 or 2 trauma center within one hour of anywhere in the nation; this could be done through the regionalization of EMS agencies (Towards Zero Deaths Steering Committee 2014). Finally, transport time may become less important if treatment can be effectively administered during the trip through the use of telemedicine.

Conclusion

Although further research is necessary, reducing pre-hospital time may improve patient outcomes in motor vehicle collisions. Pre-hospital time consists of the time between the incident and EMS notification, the time between EMS dispatch and arrival on scene, the time spent at the scene, and the time between the scene and the hospital. Each of these time periods can be reduced, although the benefit relative to the cost of doing so is not yet determined.

References

Al-Shaqsi, Sultan Zayed Khalifah. 2010. “Response Time as a Sole Performance Indicator in EMS: Pitfalls and Solutions.” Open Access Emergency Medicine : OAEM 2: 1–6.

Gonzalez, Richard P., Glenn R. Cummings, Shanna M. Harlan, Maduri S. Mulekar, and Charles B. Rodning. 2011. “EMS Relocation in a Rural Area Using a Geographic Information System Can Improve Response Time to Motor Vehicle Crashes:” The Journal of Trauma: Injury, Infection, and Critical Care 71 (4): 1023–26. https://doi.org/10.1097/TA.0b013e318230f6f0.

Gonzalez, Richard P., Glenn R. Cummings, Herbert A. Phelan, Madhuri S. Mulekar, and Charles B. Rodning. 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): 30–34. https://doi.org/10.1016/j.amjsurg.2007.11.018.

Harmsen, A.M.K., G.F. Giannakopoulos, P.R. Moerbeek, E.P. Jansma, H.J. Bonjer, and F.W. Bloemers. 2015. “The Influence of Prehospital Time on Trauma Patients Outcome: A Systematic Review.” Injury 46 (4): 602–9. https://doi.org/10.1016/j.injury.2015.01.008.

Kim, Jungeun, Kyoung Jun Song, Sang Do Shin, Young Sun Ro, Ki Jeong Hong, and James F. Holmes. 2017. “Does Prehospital Time Influence Clinical Outcomes in Severe Trauma Patients?: A Cross Sectional Study.” Prehospital Emergency Care 21 (4): 466–75. https://doi.org/10.1080/10903127.2017.1294223.

Lu, Yongmei, and Aja Davidson. 2017. “Fatal Motor Vehicle Crashes in Texas: Needs for and Access to Emergency Medical Services.” Annals of GIS 23 (1): 41–54. https://doi.org/10.1080/19475683.2016.1276102.

National 911 Program. 2017. “2017 National 911 Progress Report.” https://www.911.gov/pdf/National-911-Program-Profile-Database-Progress-Report-2017.pdf.

NHTSA. 2018. “Quick Facts 2016.” US DOT. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812451.

Peyravi, Mahmoudreza, Soheila Khodakarim, Per Örtenwall, and Amir Khorram-Manesh. 2015. “Does Temporary Location of Ambulances (‘Fluid Deployment’) Affect Response Times and Patient Outcome?” International Journal of Emergency Medicine 8 (1). https://doi.org/10.1186/s12245-015-0084-1.

Sánchez-Mangas, Rocío, Antonio García-Ferrrer, Aranzazu de Juan, and Antonio Martín Arroyo. 2010. “The Probability of Death in Road Traffic Accidents. How Important Is a Quick Medical Response?” Accident Analysis & Prevention 42 (4): 1048–56. https://doi.org/10.1016/j.aap.2009.12.012.

Towards Zero Deaths Steering Committee. 2014. “Toward Zero Deaths: A National Strategy on Highway Safety.” http://www.towardzerodeaths.org/strategy/.

Wilde, Elizabeth Ty. 2013. “DO EMERGENCY MEDICAL SYSTEM RESPONSE TIMES MATTER FOR HEALTH OUTCOMES?: RESPONSE TIMES.” Health Economics 22 (7): 790–806. https://doi.org/10.1002/hec.2851.

Case Study 3

An Approach to Assess Residential Neighborhood Accessibility and Safety: A Case Study of Charlotte, North Carolina.

Louis Merlin[a]*, Eric Dumbaugh[a], Amin Mohamadi Hezaveh[b], Christopher R. Cherry[b]

[a] School of Urban and Regional Planning, Florida Atlantic University, Boca Raton FL

[b] Civil and Environmental Engineering, University of Tennessee, Knoxville, TN

* Corresponding Author: lmerlin@fau.edu

Abstract

Understanding relationships between the built environment, urban form, and road safety is an important part of the planning process. Transportation planning aims to increase access to destinations through improved mobility and land use strategies that enhance proximity. This study proposes an approach to link transportation accessibility data with home-based safety data to understand whether accessibility also has implications for safety. This approach relies on two main data elements. First, accessibility data from transportation planning is applied to fine-geographic resolution data in an urban area – in this case Census block groups. Next, police crash data is geocoded to the residential location (corresponding block group) of the crash victim. Crash generation models (e.g., negative binomial count models) can be developed at the block group level to assess the extent accessibility influences crash counts of residents of a neighborhood. This method was applied to data from Charlotte, North Carolina, using work, non-work, and population auto accessibility metrics generated from a gravity model utilized in Charlotte's long-range planning efforts. Similarly, demographic data from the Census was included in the modelling effort. Two predictive negative binomial regression models were estimated for total crash counts and injury crash counts. Controlling for demographics, we found that increasing work and non-work accessibility decreased crash counts (for both total and injury crashes). We found that increasing population accessibility increased crash counts. We speculate that increased population accessibility results in concentrated travel demand and ultimately increased crash exposure, overcoming individual-level reductions in crash exposure from lower VMT. Despite data limitations, this work illustrates the importance of including planning-level data in safety analysis and presents areas for improvement of the approach.

Introduction

The built environment can influence both the likelihood and the severity of a crash occurring at a particular location. There are several studies and literature reviews on the influence of the built environment on crashes (Ewing & Dumbaugh, 2009; Retting, Ferguson, & McCartt, 2003; Rothman, Buliung, Macarthur, To, & Howard, 2014; Stoker et al., 2015). Studies to date have addressed issues such as where hot spots of crashes occur (Wier, Weintraub, Humphreys, Seto, & Bhatia, 2009), the influence of the built environment on the severity of pedestrian and bicycle crashes (Chen & Shen, 2016), and how various roadway design characteristics affect roadway crash rates (Abdel-Aty et al., 2009).

The effect of a traveler's residential location on their crash risk has not been studied as thoroughly. Historically, most crash analysis focuses on crash location because of proximate causes present at the crash site and crash data generally includes details of the crash location. The link between the built environment at a traveler's residential location and the likelihood of traffic crashes may not be obvious at first since the residential location of a person is not necessarily or even frequently where their crashes occur. However, there is extensive literature on how the residential built environment influences travel behavior (Ewing & Cervero, 2010; Stevens, 2016), and travel behavior may in turn influence crash risk. We take a "home based approach" (HBA) to safety analysis in this paper. This approach assigns crashes to the home location of the crash victim, rather than the location of the crash that is usually employed in a crash location-based approach. This is done by geocoding victim address locations and aggregating those crashes to small geographic analysis units, like Census block groups or Traffic Analysis Zones (TAZs). While ours is not the first paper to take a HBA approach, here we point out that a more systematic analysis of the built environment and demographic features of residential locations is possible and is relevant to the study of traffic safety. This particular paper focuses on the implications of auto accessibility to population, to work, and to non-work destinations as an exemplar of this home-based approach.

Ewing and Dumbaugh (Ewing & Dumbaugh, 2009) posit that one of the ways that the residential built environment may influence crash risk is via exposure. Certain residential built environments are associated with increased travel exposure, i.e. higher vehicle miles traveled (VMT), and therefore increased crash risk. Indeed, the divergence in crash fatalities per capita and crash fatalities per mile illustrates that countries with more vehicular travel experience a higher number of crash fatalities in part due to higher rates of vehicular travel (OECD & ITF, 2016). For example, even though the US is only 1.86 times riskier on a per-mile basis than the UK, it is 3.52 times riskier (or about 1.89x more) on a per-capita basis, due to the larger amounts of per-capita vehicular travel in the US. Therefore, greater vehicular travel in the US is associated with higher probability of being in a crash at the national level; it is logical to extrapolate this trend and assume that greater vehicular travel would be associated with higher probability of being in a crash on the subnational level as well, with high-VMT residential locations being associated with more crashes while low-VMT residential locations would be associated with fewer.

Several papers examine the relationship between sprawl or compactness and crash risk. Presumably, those who live in more sprawling environments drive more and are therefore more likely to be subject to a crash. Ewing and Hamidi (Ewing, Hamidi, & Grace, 2016) find that sprawling counties are associated with more fatal crashes, but fewer injury crashes, than more compact counties. Mohamed et al. (Mohamed, vom Hofe, & Mazumder, 2014) find that greater jurisdictional sprawl is inversely related to the number of fatal and injury crashes, though they do not take into account the potential spillover effect, i.e. that residential locations are not necessarily crash locations. That is, crash locations rather than residential locations are their object of analysis. Lucy (Lucy, 2003) conducts an interesting study comparing the risk of death in a traffic crash with the risk of death by murder, examining data at city, county, state, and federal levels. He finds that traffic fatality rates are highest in exurban areas and outer counties as compared with inner counties. Yeo et al. (Yeo, Park, & Jang, 2015) conduct a path analysis of the relationship between sprawl, VMT,

traffic fatalities, income, and fuel cost for 147 urbanized areas in the US. They find that sprawl has both a direct effect on increasing traffic fatalities and an indirect effect on increasing traffic fatalities through VMT. In summary, the literature suggests that built environments associated with higher VMT are associated with more traffic fatalities; however, these environments are not consistently associated with more traffic injuries as well.

While most of the literature on the relationship between residential built environment and traffic risk has focused on the complex and multidimensional construct of sprawl, the literature on the relationship between the built environment and travel behavior distinguishes between several different measurable dimensions of the built environment that each influence travel behavior: density, design, diversity, and destination accessibility (Ewing & Cervero, 2010). The advantage of focusing on these individual dimensions is that they are more readily measurable and less abstract than multi-dimensional sprawl. They can also be measured at finer scales than the county level, such as the Census Block Group or Traffic Analysis Zone. This allows for neighborhood-level scales of spatial analysis.

The literature on travel behavior and the built environment strongly suggests that the single most important variable related to reduced vehicle miles traveled is destination accessibility, hereafter referred to simply as accessibility (Ewing & Cervero, 2010; Stevens, 2016). Therefore, this paper examines the relationship between accessibility (as measured at a person's residential location) and crashes, based on the hypothesis that those who live in areas of high accessibility will engage in less vehicular travel and therefore experience reduced rates of vehicular crashes on a per-capita basis.

In comparison with previous papers on sprawl and crash risk, this paper proposes a method, using new safety data approaches, to examine the role of the built environment on crashes. Our approach has two main contributions. First, the method allows a level of analysis on a much finer scale, the Census block group. Second, we analyze the effect of particular built environment variables, accessibility and density, rather than an aggregate sprawl index. This makes the analysis more transparent and replicable for future research and analysis efforts. Furthermore, this paper examines the relationship between three measures of accessibility: accessibility to the metropolitan population, accessibility to all jobs, and accessibility to non-work related jobs. This paper also controls for the characteristics of the resident population to identify any relationship between the accessibility of a person's residential environment and their vehicular crash risk. We rely on a crash database from Mecklenburg County (Charlotte) North Carolina to illustrate the capability of this approach in identifying the effect of accessibility at the block group level on crash prediction for that geographic unit.

Methodology

The method proposed here requires two key pieces of information. First, to examine whether the accessibility of a crash victim's residential location is correlated with their crash rates, we require accessibility data from each residential location, i.e., how many destinations can a resident access from their home within reasonable travel time. Residential locations are geocoded and then identified by the crash victim's home Census block group, a scale of analysis that preserves the confidentiality of crash victims. Therefore, Census block groups are the unit of analysis for all built environment and demographic variables. Next, we generate accessibility metrics for each Census block group based on its regional location. Destinations often considered in accessibility analysis includes number of jobs, population, or other non-work destinations that attract trips (e.g., retail). The network travel time from every origin and destination is calculated, generally using the travel time matrix generated from regional travel demand modeling efforts. Accessibility can be defined by the number of destinations reached in a certain travel time, or it can be represented through a gravity-based approach, where the attractiveness of destinations from an origin diminished based on travel time impedance. The gravity approach is used here, described later in this paper. Census data also includes a trove of socioeconomic control variables to assist in model generation.

Application

Mecklenburg County is the setting to illustrate this analysis. This geographic area includes urban and suburban areas of the Charlotte, North Carolina, metropolitan area, and includes about one million residents. The county population is densest in the central business district and in the southeast quadrant of the county. Accessibility on many metrics varies between block groups.

University of North Carolina's Highway Safety Research Center (HSRC) provided comprehensive data on crashes that occurred on surface streets in Mecklenburg County in 2013. Since crash data are limited to Mecklenburg County, only residents of Mecklenburg County are included in this analysis (i.e., crash victims who live outside the county, but experienced crashes in the county were excluded). Unfortunately, this dataset does not include crashes that occurred on the Interstate highway system in the county. This omission possibly skews the results of the analysis, but we still are able to demonstrate the method with this dataset. There are 555 Census block groups in the county, and after removing block groups with no households, 550 Census block groups remained. In addition to accessibility, we collected demographic and built environment data for each block group to analyze patterns in total and injury crashes.

HSRC matched crash data to crash victim residential locations, removing identifiers. In 2013, there were 20,899 total crashes in the county, including 241 pedestrian crashes, and 101 bicyclist crashes. HSRC matched each person in control of a crashed vehicle (including pedestrians and cyclists) to their residential block groups (where they live). A total of 29,161 vehicle drivers and 269 pedestrians were matched to the 20,899 crashes (no cyclists remained after the matching process), 23,829 of those drivers and 250 of the pedestrians live in Mecklenburg County (refer to Figure 1). Note that the number of drivers is larger than the number of crashes because more than one vehicle may be associated with each crash. These drivers were associated with 29 fatal crashes and 5,804 injury crashes. The dependent variables analyzed were total crash counts and injury crash counts by Census block group, as there were not enough fatal crashes for statistical analysis.

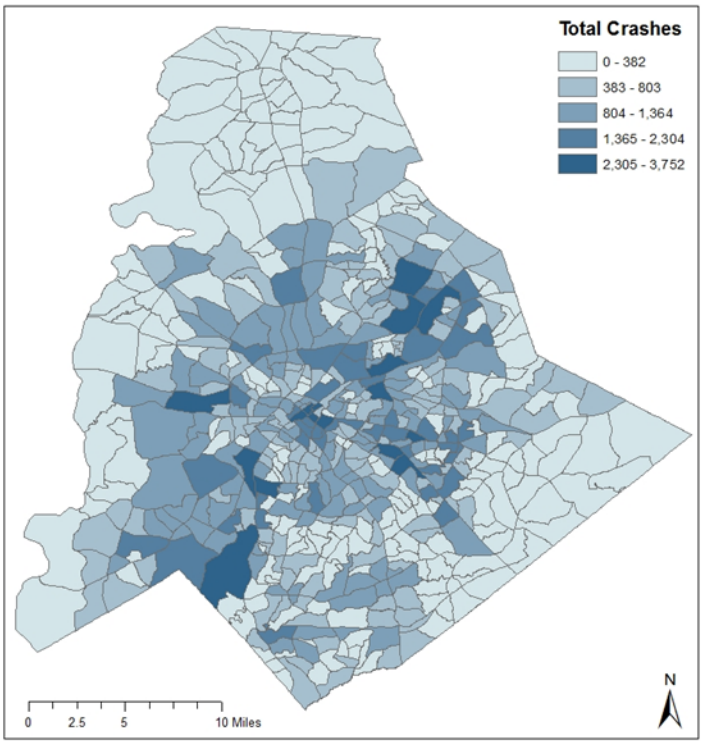

Figure 1. Total crashes by Census block group in Mecklenburg County, North Carolina, 2013.

We also examined related demographic and built environment data for each Census block group. The five-year (2009-2014) American Community Survey provided demographic data. Demographic variables included race/ethnicity (black and Hispanic), age (age 5 -17, and age 65+), educational attainment (no high school degree), and household poverty status. Built environment data included population density, employment density, and accessibility (accessibility to work, accessibility to non-work locations, and accessibility to population). Vehicle Miles Traveled (VMT) in each block group was also examined to control for the amount of driving on freeways and arterials within each Census block group.

Accessibility was calculated using gravity models (see equations below). Three types of destinations were considered: work, nonwork, and population. LEHD Origin-Destination Employment Statistics, or LODES (US Census Bureau, 2016), were used to determine work and nonwork locations for the 12-county Charlotte metropolitan area. Calculations for work accessibility included all jobs as potential destinations, whereas calculations for nonwork trips only included employment sectors associated with nonwork travel: Retail Trade (NAICS 44-45); Health Care and Social Assistance (NAICS 62); Arts, Entertainment, and Recreation (NAICS 71); Accommodation and Food Services (NAICS 72); Other Services (NAICS 81). Population counts for each destination in the 12-county metropolitan area were obtained from the 2010 United States Census. The 2010 Charlotte Regional Transportation Planning Organization (CRTPO, https://www.crtpo.org/) provided zone-to-zone travel time data for auto travel, both peak and off-peak.

The Gamma function was utilized as the impedance function for calculating accessibility, as this function is used within the region's travel demand model for trip distribution (Kimley-Horne and Associates, 2016). Likewise, parameters were adopted from the region's trip distribution model. Specifically, the impedance for non-work trips had a Beta of 2.14 and a Gamma of 0.36, while the impedance function for work trips and population had a Beta of 1.18 and a Gamma of 0.15.

$$A_{work_i} = \sum_j jobs * t_{ij}^{1.18} e^{-0.15*t_{ij}}$$

$$A_{nonwork_i} = \sum_j nonwork\ jobs * t_{ij}^{2.14} e^{-0.36*t_{ij}}$$

$$A_{population_i} = \sum_j population * t_{ij}^{1.18} e^{-0.15*t_{ij}}$$

The above formulae illustrates three equations for accessibility of a given Census block group i. For each accessibility formula, the accessibility of residential zone i is the sum over all other zones j of the number of destinations in zone j (jobs, nonwork jobs, population) times the impedance of travel for traveling from zone i to zone j. The impedance of travel is in turn a function of travel time t_{ij}, where t_{ij} is peak-hour auto travel time for work and population formula and off-peak auto travel time for non-work.

Model Specification

Negative binomial models were used to separately predict total and injury crashes as a function of demographics and built environment variables. All variables were standardized to facilitate the comparison of effect size across variables. The analysis began with a base model of just demographic variables. Subsequently, we add density and accessibility variables to determine if the residential built environment potentially influences the number of crashes residents experienced.

Results

Table 1 lists the results of negative binomial models for total crashes and Table 2 lists the results for injury crashes. In the base model, an increase of total population, percent black, and percent Hispanic populations predicted more total crashes. It is expected that as the population of a given area increases, the number of crashes involving people from that area would also rise. An increase in the percentage of people between the ages of 5 and 17 was associated with fewer crashes. Other variables such as people 65 and up, no high school degree, and poverty status were not statistically significant in the base model.

Density and Accessibility (Model 1 to 6)

Model 1 to 6 offered a comparison between accessibility and density as measures of the residential built environment. In all these models, total population, black, and Hispanic populations were statistically significant and have a positive correlation with the number of crashes. In model 1 and model 2, the percentage of people between the ages of 5 and 17 has a negative relationship that was statistically significant. An increase in the percentage of people with no high school degree was correlated with fewer crashes in model 3. In model 3, 4, 5, and 6 an increase in the percentage of households in poverty predicted fewer crashes. Population density is not statistically significant in model 1, while employment density is associated with fewer crashes in model 2 (β=-0.047). Accessibility to population, work, and non-work locations is significant in model 3 (β=0.196), model 4 (β=0.005), and model 5 (β=0.069), respectively. In terms of effect size, a one standard deviation increase in population was associated with 48.3% more crashes, while a one standard deviation increase in accessibility is associated with 21.6% more crashes, so the effect size of population accessibility is quite large. In model 6, all accessibility measures were examined together in a single regression. In this model, an increase in accessibility to population predicted more crashes (β=0.465), while accessibility to work locations (β=-0.006) and accessibility to non-work locations (β=- 0.172) predicted fewer total crashes, with all accessibility variables being statistically significant.

Model 7 controlled for vehicle miles traveled (VMT) in the resident's home block group as a measure of localized exposure to traffic. There was no change in the statistical significance of demographic variables with the addition of local VMT. Accessibility to population (β=0.213) remained positively correlated with total crashes even after controlling for localized VMT. An increase in VMT in freeways was statistically significantly correlated with fewer total crashes (β=-0.074), while VMT on arterials was not statistically significant.

Injury Crashes

The same patterns found for total crashes generally held for injury crashes as well. In the base model, an increase total population, black, and Hispanic populations was associated with a rise in injury crashes. An increase of people between the ages of 5 and 17 predicted fewer injuries. Other variables such as people 65 and up, no high school degree, and poverty status were not statistically significant.

Model 8 to 13 examine the association of the built environment variables of accessibility and density on injurious crashes. In all these models, total population, black, and Hispanic populations were statistically significant and had a positive correlation. In model 8, 9, 11, and 12 the percentage of people between the ages of 5 and 17 has a negative and statistically significant relationship with crashes. Percent of people aged 65 and up and percent of people without a high school education are not statistically significant in models 8 to 13 (in contrast with the total crash model where no high school degree was negatively associated with crash counts in model 3). Poverty status was correlated with fewer injury crashes in model 10 (similar to total crash models 3, 4, 5, and 6). In model 8, population density is not statistically significant and in model 9 employment density is not statistically significant (this differs from total crash models where employment density was negatively associated). Accessibility to population and to work are associated with an increase in injuries in model 10 (β=0.180) and model 11 (β=0.004), respectively. Accessibility to non-work

locations was not statistically significant in model 12 (differing from total crash models where accessibility to non-work locations was statistically significant in model 5).

In terms of effect size, a one standard deviation increase in population was associated with 48.1% more injury crashes, while a one standard deviation increase in accessibility to population was associated with 19.7% more injury crashes. In model 13, all accessibility measures were tested in conjunction. Accessibility to population predicted an increase in injury crashes (β=0.523). While accessibility to work locations (β=-0.008) and non-work locations (β=-0.201) predicted fewer injury crashes.

Model 14 controlled for vehicle miles traveled (VMT) within the Census block group in the prediction of injury crashes. All demographic variables maintained their statistical associations after controlling for localized VMT. Accessibility to population (β=0.197) remained statistically significant in model 14 even after controlling for local VMT. An increase in VMT in freeways was correlated with fewer injury crashes (β=-0.083), while VMT in arterials was not statistically significant.

	Total Crashes															
	Demographics		Density				Accessibility								VMT Control	
	Base Model		Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	Estimate		Estimate		Estimate		Estimate		Estimate		Estimate		Estimate		Estimate	
Intercept	3.697	***	3.697	***	3.696	***	3.688	***	3.453	***	3.696	***	3.972	***	3.686	***
Demographic Variables																
Population	0.352	***	0.350	***	0.346	***	0.394	***	0.375	***	0.370	***	0.381	***	0.407	***
Black	0.245	***	0.247	***	0.239	***	0.223	***	0.233	***	0.254	***	0.181	***	0.229	***
Hispanic	0.104	***	0.099	***	0.102	***	0.132	***	0.127	***	0.124	***	0.092	***	0.125	***
Age 5 to 17	-0.051	**	-0.048	*	-0.059	**	-0.002		-0.022		-0.032		-0.014		-0.006	
Age 65 and Up	-0.007		-0.003		-0.014		0.017		0.007		0.004		0.006		0.014	
No H.S. Degree	-0.014		-0.014		-0.014		-0.059	*	-0.045		-0.030		-0.044		-0.054	.
Poverty Status	-0.036		-0.041		-0.035		-0.078	**	-0.062	*	-0.056	*	-0.053	*	-0.082	**
Built Environment Variables																
Population Density			0.026													
Employment Density					-0.047	*										
Access to Work									0.005	***			-0.006	*		
Access to Non-Work											0.069	**	-0.172	***		
Access to Population							0.196	***					0.465	***	0.213	***
Vehicle Miles of Travel (VMT)																
VMT Arterials															-0.014	
VMT Freeways															-0.074	***
Estimate indicated with level of significance (<0. 1, *<0.05, **<0.01, ***<0.001) All variables have been standardized																

Table 1. Final negative binomial models for total crashes in Mecklenburg County, North Carolina.

	Injury Crashes															
	Demographics		Density				Accessibility								VMT Control	
	Base Model		Model 8		Model 9		Model 10		Model 11		Model 12		Model 13		Model 14	
Variables	Estimate		Estimate		Estimate		Estimate		Estimate		Estimate		Estimate		Estimate	
Intercept	2.228	***	2.228	***	2.228	***	2.221	***	2.046	***	2.228	***	2.638	***	2.218	***
Demographic Variables																
Population	0.355	***	0.353	***	0.357	***	0.393	***	0.371	***	0.366	***	0.377	***	0.404	***
Black	0.350	***	0.351	***	0.352	***	0.322	***	0.337	***	0.354	***	0.279	***	0.330	***
Hispanic	0.142	***	0.138	***	0.143	***	0.164	***	0.157	***	0.154	***	0.116	***	0.155	***
Age 5 to 17	-0.078	***	-0.075	**	-0.075	**	-0.031		-0.056	*	-0.066	**	-0.045	.	-0.035	
Age 65 and Up	-0.012		-0.009		-0.010		0.011		-0.001		-0.005		-0.004		0.009	
No H.S. Degree	-0.001		0.000		-0.001		-0.043		-0.025		-0.011		-0.017		-0.036	
Poverty Status	-0.028		-0.031		-0.028		-0.063	*	-0.045		-0.039		-0.037		-0.068	*
Built Environment Variables																
Population Density			0.020													***
Employment Density					0.016											
Access to Work									0.004	**			-0.008	**		
Access to Non-Work											0.044		-0.201	***		
Access to Population							0.180	***					0.523	***	0.197	
Vehicle Miles of Travel (VMT)																
VMT Arterials															-0.002	
VMT Freeways															-0.083	***

Estimate indicated with level of significance (<0. 1, *<0.05, **<0.01, ***<0.001)

All variables have been standardized

Table 2. Final negative binomial models for injury crashes in Mecklenburg County, North Carolina.

Discussion

The hypothesis for this study was that higher accessibility areas require less vehicular travel and therefore residents would experience fewer crashes. Our results indicate the opposite; residents of higher accessibility areas, in particular residents of areas with high accessibility to population, experience more crashes. Accessibility to population, work, and non-work locations are each associated with an increase in total crashes (refer to Table 1). The same pattern was observed for injury crashes (refer to Table 2), with the exception of accessibility to non-work locations, which is not statistically significant though positively associated with crashes.

Residents of high accessibility areas are located in concentrations of urban activity. Those living in high population accessibility areas live within the commute sheds of many other urban residents. Therefore, when driving they are exposed to more of other people's driving—high accessibility essentially predicts the overall level of traffic in the surrounding environment. As a result, there is a higher rate of crashes for such residents even if they drive less overall than in areas with lower accessibility (refer to diagram 1). As our findings indicate, accessibility to population had the largest effect on total and injury crashes out of all the accessibility measures (refer to Table 1 and 2). This is intuitive because accessibility to population is the best proxy for exposure to the overall levels of traffic in the surrounding environment.

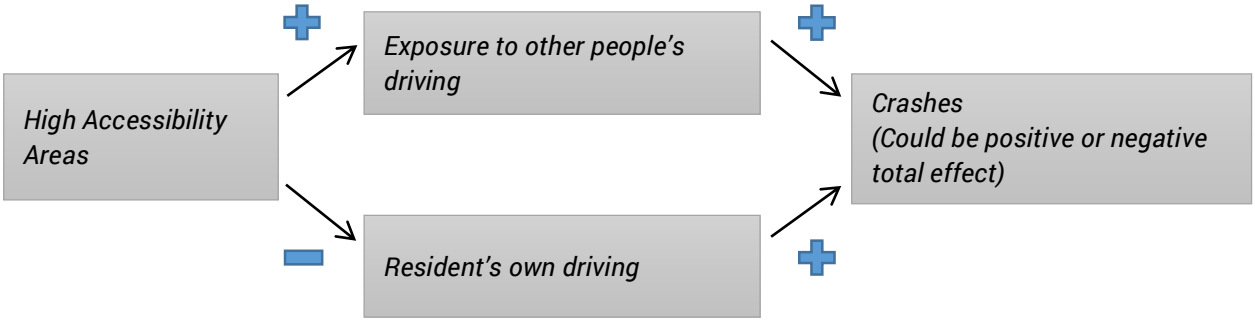

Diagram 1. Conceptual framework: Crashes in high accessibility areas. A positive sign indicates a positive relationship and a negative sign indicates a negative relationship.

The model with all three accessibility variables suggests an even more nuanced interpretation. When accessibility to the population is controlled for, accessibility to work and nonwork destinations has a negative and statistically significant correlation with total crashes and injury crashes. If accessibility to population is a proxy for exposure to other people's driving in the urban environment, then accessibility to work and nonwork may be associated with lower crash risk because they predict lower VMT for the own driving. In other words, the original hypothesis that high accessibility environments are associated with lower crash risk via lower VMT may be true in part, but only after controlling for accessibility to the overall metropolitan population.

Density is the most studied built environment variable, but it may not capture important safety-related effects. Our findings show quite different results for accessibility and density variables. In general, density at a person's residential location showed weak or no relationship with their likelihood of being in a traffic crash. Population density was not statistically significant in total crash and injury crash models. Employment density was only statistically significant in injury crash models, and with a small coefficient (β=-0.047). In contrast, accessibility measures had significant relationships in both total crash and injury crash models. Accessibility to population, as mentioned before, had the largest effect size of all built environment variables examined.

Areas with higher accessibility to population, such as new infill development in close proximity to the central city, are associated with a higher rate of crashes in our results. A one standard deviation increase in accessibility to population is associated with 21.6% more crashes and 19.7% more injury crashes (refer to Table 1 and 2). Although we are not certain of the external validity of these results, it may be the case that when residential development is promoted in high accessibility areas traffic safety may be unintentionally diminished as a result. One implication is that efforts to increase accessibility should be matched with efforts to improve the safety performance of the transportation network.

Limitations and Further Research

The most significant limitation is the lack of data on interstate crashes. It is possible that residents of low accessibility areas travel more and experience more crashes on the interstate system. We undercount crashes that occur on populations that likely generate higher VMT per capita, those that rely on the interstate highway system. If this is the case, then the higher crash counts associated with high population accessibility areas might be overstated in these results.

We only examine the relationship between crashes and residential location for a single urban county. As such, we miss spillover effects where a resident lives towards the edge of the county and experiences a crash in another county. Since all of the crash data obtained is from Charlotte-Mecklenburg County, we are not able to capture these out-of-county crashes, and there may be a systematic undercount of traffic crashes for residents living towards the edge of the county.

Likewise, the range of accessibility analyzed is truncated because the scope of the analysis is limited to a single urban county. The larger metropolitan statistical area (MSA) includes ten counties. Other research finds fatal crash risk increases in exurban and outlying counties (Lucy, 2003). It is possible that accessibility is associated with decreased crash risk and/or decreased injury crash risk when moving from low accessibility locations in exurban counties to inner counties, and is then associated with increased crash risk or injury crash risk within inner urban counties as one moves towards the center of these counties. Since the data here is limited to a single urban county, the full range of accessibility effects may possibly not be captured. However, the positive association of crashes with accessibility is a novel result and should be considered in future analysis concerning the built environment and crashes. Extending this research to a multicounty metropolitan or urbanized area might provide greater insight into the relationship between accessibility and crash rates.

Conclusion

This paper examined the relationship between accessibility, measured at a person's residential location, and their likelihood of being in a crash in Mecklenburg County, North Carolina. The literature review indicated that built environments with higher VMT are associated with more crashes. As a result, our hypothesis was that those who live in areas of high accessibility would engage in less vehicular travel and therefore experience reduced rates of vehicular crashes on a per-capita basis.

We identified the residential Census block group of crash victims with data from the University of North Carolina's Highway Safety Research Center (HSRC). After the residential match was performed no pedestrians and cyclists remained, so the analysis proceeded with only vehicular crash victims. The five-year (2009-2014) American Community Survey provided demographic data (race/ethnicity, age, educational attainment, and household poverty status) to add to our built environment data as controls for the analysis. Negative binomial models separately predicted total and injurious crashes as a function of demographics and built environment variables at the Census block group level.

Our results indicate that residents of higher accessibility areas, in particular residents of areas with high accessibility to population, experience more crashes. This contradicts the original hypothesis that residents of high accessibility areas

would experience fewer crashes. All of the accessibility measures (accessibility to population, work, and non-work locations) were associated with an increase in total and injury crashes (with the exception of accessibility to non-work locations which did not have a significant relationship in injury crash models). One potential explanation for these findings is that people located in high accessibility areas are exposed to other people's driving. Consequently, there is an increase in the rate of crashes for residents even if they drive less than residents of areas with lower accessibility. Although the external validity of the positive association between residential accessibility and crash rates is uncertain, it may be the case that promoting residential development in high accessibility areas diminishes traffic safety as an unintentional byproduct.

The study has significant limitations, however it demonstrates a method to analyze crashes in the context of home-based crash analysis and accessibility. The results are suggestive and more research is required to generate more definitive outcomes. This work does not allow us to investigate a full range of modes or injury severities. Also, the geographically broad accessibility effects may also not have been captured by the single-county analysis which contrasts with previous research that was conducted across multiple counties. Future research should extend the study area to a multicounty metropolitan area with more complete crash data to gain greater insight into the relationship between residential accessibility and crash rates.

References

Abdel-Aty, M., Pande, A., Lee, C., Das, A., Nevarez, A., Darwiche, A., & Devarasetty, P. (2009). Reducing fatalities and severe injuries on Florida's high-speed multi-lane arterial corridors : part I, preliminary severity analysis of driver crash involvements, final report, April 2009. University of Central Florida. Center for Advanced Transportation Systems Simulation. Retrieved from http://ntl.bts.gov/lib/31000/31500/31520/FDOT_BD548-22_rpt_PART_I.pdf

Chen, P., & Shen, Q. (2016). Built environment effects on cyclist injury severity in automobile-involved bicycle crashes. Accident Analysis & Prevention, 86, 239–246. http://doi.org/10.1016/j.aap.2015.11.002

Ewing, R., & Cervero, R. (2010). Travel and the built environment -- A meta-analysis. Journal of the American Planning Association, 76, 265–294. http://doi.org/10.1080/01944361003766766

Ewing, R., & Dumbaugh, E. (2009). The Built Environment and Traffic Safety A Review of Empirical Evidence. Journal of Planning Literature, 23(4), 347–367. http://doi.org/http://dx.doi.org/10.1177%2F0885412209335553

Ewing, R., Hamidi, S., & Grace, J. B. (2016). Urban sprawl as a risk factor in motor vehicle crashes. Urban Studies, 53(2), 247–266. http://doi.org/10.1177/0042098014562331

Kimley-Horne and Associates. (2016). Metrolina Model Users Guide. Charlotte, NC.

Lucy, W. H. (2003). Mortality risk associated with leaving home: recognizing the relevance of the built environment. American Journal of Public Health, 93(9), 1564–1569.

Mohamed, R., vom Hofe, R., & Mazumder, S. (2014). Jurisdictional spillover effects of sprawl on injuries and fatalities. ACCIDENT ANALYSIS AND PREVENTION, 72, 9–16. http://doi.org/10.1016/j.aap.2014.05.028

OECD, & ITF. (2016). Road Safety Annual Report 2016. Paris: International Transport Forum. Retrieved from http://dx.doi.org/10.1787/irtad-2016-en

Retting, R. A., Ferguson, S. A., & McCartt, A. T. (2003). A Review of Evidence-Based Traffic Engineering Measures Designed to Reduce Pedestrian-Motor Vehicle Crashes. American Journal of Public Health, 93(9), 1456–1463. http://doi.org/10.2105/AJPH.93.9.1456

Rothman, L., Buliung, R., Macarthur, C., To, T., & Howard, A. (2014). Walking and child pedestrian injury: a systematic review of built environment correlates of safe walking. Injury Prevention, 20(1), 41–49.

Stevens, M. R. (2016). Does compact development make people drive less? Journal of the American Planning Association, 83(1), 7–18. http://doi.org/10.1080/01944363.2016.1240044

Stoker, P., Garfinkel-Castro, A., Khayesi, M., Odero, W., Mwangi, M. N., Peden, M., & Ewing, R. (2015). Pedestrian Safety and the Built Environment: A Review of the Risk Factors. Journal of Planning Literature, 30(4), 377–392. http://doi.org/10.1177/0885412215595438

US Census Bureau. (2016). LEHD Origin-Destination Employment Statistics. Retrieved December 2, 2016, from http://lehd.ces.census.gov/data/lodes/LODES7/.

Wier, M., Weintraub, J., Humphreys, E. H., Seto, E., & Bhatia, R. (2009). An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. ACCIDENT ANALYSIS AND PREVENTION, 41(1), 137–145. http://doi.org/10.1016/j.aap.2008.10.001

Yeo, J., Park, S., & Jang, K. (2015). Effects of Urban Sprawl and Vehicle Miles Traveled on Traffic Fatalities. TRAFFIC INJURY PREVENTION, 16(4), 397–403. http://doi.org/10.1080/15389588.2014.948616

Case Study 4

Home-Based Approach: A Complementary Definition of Road Safety

Amin Mohamadi Hezaveh[a], Christopher R. Cherry[a*]

[a] Civil and Environmental Engineering, University of Tennessee, Knoxville, TN, United States,

* Corresponding Author: Cherry@utk.edu

Abstract

Current practice of road safety attributes traffic crashes to the location of traffic crashes (i.e., Location-Based Approach –LBA), hence it is challenging to measure individuals' likelihood of involvement in traffic crashes based on where they live. To measure likelihood of involvement in traffic crashes, we introduced Home-Based approach (HBA) as the expected number of crashes that road users who live in a certain geographic area have during a specified period, following epidemiological spatial risk approaches. Census tract and police report crashes were used to extract the location of the traffic crashes (n = 694,276) and home-address of road users (n = 1,113,830) in Tennessee, and accompanying socioeconomics. A Poisson model and a geographically weighted Poisson regression (GWPR) model were used to predict the association between sociodemographic variables and HBA/LBA crash frequency. Results indicate that GWPR models are more suitable than traditional models and all the independent variables have substantial local effects. Findings indicate that residents of areas with higher share of active mode of transportation, higher income and more population of white race are less likely to be involved in the traffic crashes. The HBA approach could be used to identify geographic areas where their residents have a higher likelihood of involvement in traffic crashes and can focus interventions on those populations. The findings are discussed in line with road safety.

Keywords: Macroscopic Crash Prediction Method; Home-Based Approach; Location-Based Approach; Geographically Weighted Regression; Spatial Analysis

Introduction

Traffic crashes are a large negative externalities of the transportation system and are listed as one of the leading cause of early death (Winston Harrington *et al.* 2006, World Health Organization and Organization 2014). World Health Organization reports on road safety demonstrate the variation of the crash fatality rates (i.e., fatality per 100,000 populations) across countries focusing on crash rates per capita. A large body of the road safety literature examined the differences in crash fatality rate and factors influencing it at country level (Koornstra *et al.* 2002, Hermans 2009, Holló *et al.* 2010, Chen *et al.* 2016). This spatial variation reflects the role of exogenous factors such as demographics, economy, infrastructure, culture, safety measures and safety performance indicators (Koornstra *et al.* 2002, Chen *et al.* 2016). Bearing in mind the spatial variation of the crash fatality rate, one may expect that road users' likelihood of involvement in traffic crashes also varies across the space. Knowing about the groups or geographical areas where their residents are more prone to traffic crashes or burden of traffic crashes enable researcher to assign resources or proper countermeasures to improve their safety. Still, less is known about the likelihood of involvement in traffic crashes and its spatial distribution within a country or at fine geographical level such as a neighborhood level or traffic analysis zone.

There are road safety disparities across road user type, income, race, and ethnicities; for example, the crash fatality rate is approximately double in low- and middle-income countries compared to high-income countries (21.5, 19.5, and 10.3 per 100,000 respectively) (WHO 2015). This trend also holds within a country; for example, several studies in the US reported that vulnerable road users (i.e., pedestrians and bicyclists), lower income neighborhoods have higher fatality rates compared to motorized road users and wealthier neighborhoods, respectively (Marshall and Ferenchak 2017). This is also the case for the rural areas where the fatality rate is several times higher than majority of urban areas (Marshall and Ferenchak 2017). In addition, ethnicities such as Hispanic, African-American, and American Indian are more prone to traffic crashes –i.e., higher crash rate (Mayrose and Jehle 2002, Braver 2003, Campos-Outcalt *et al.* 2003, McAndrews *et al.* 2013); this trend also holds for the fatality rate (Schiff and Becker 1996, Baker *et al.* 1998, Harper *et al.* 2000). Bearing in mind that overall road safety burden does not impact population equitably, we may expect the likelihood of involvement in traffic crashes also impacts population inequitably.

Measuring individuals' likelihood of involvement in traffic crashes is challenging. Particularly, due to the current practice of road safety, which attributes traffic crashes to the location of traffic crashes (i.e., Location-Based Approach). The current practice of road safety is best described as "*the number of accidents (crashes) by kind and severity, expected to occur on the entity during a specified period.*" (Hauer 1997 p. 24) This definition enables researchers to estimate the likelihood of occurrence of traffic crashes (kind and severity) for an entity for a specified period. As a results, a large body of the road safety literature is dedicated to the geographic distribution of traffic crashes, identifying the pattern of traffic crash disparities (e.g., hotspots), and evaluating the changes in traffic characteristics, geometry design on road safety outcome –i.e., crash modification factors (HSM 2010). Bearing in mind that human factors contribute in 90% of traffic crashes (Evans 1996, Petridou and Moustaki 2000), we may conclude that this definition overlooks the role of the person involved in the crash compared to other factors.

To accentuate the role of the person involved in the analysis as well as considering the role of environmental factor in the analysis, in this study, we are proposing a new approach to measure road safety at a zonal level that does not rely on locations of traffic crashes, but rather the residential location of the individuals with added information from surrounding demographics. The Home-Based Approach (HBA) is a proposed method to complement the Location-Based Approach (LBA). Additionally, HBA enables researchers and practitioners to identify hotspots where residents have a higher likelihood of involvement in traffic crashes for targeted educational purposes, focusing resources on populations that could benefit most. This approach tends to follow norms associated with spatial epidemiological risk or disease burden studies, where the burden of disease is spatially analyzed and correlated with proximate factors.

The ultimate goal in this study is to scrutinize the road users' likelihood of involvement in traffic crashes at fine geographic level and examine the association between sociodemographic factors surrounding home-address of individuals who were involved in traffic crashes with HBA crash frequency. We use Macroscopic Crash Prediction Models (MCPM) to achieve these goals. We also estimate the LBA crash frequency models for the study area. Above all, we do not have the intention to compare the goodness of fit of LBA and HBA models; our intention is to present these models side by side and discuss the association between sociodemographic variables and crash frequency in each approach.

The remainder of this paper provides a literature review of Macroscopic Crash Prediction Models. In the next sections, we present methodology –including HBA definition, geocoding process, data sources, and the regression model– results, and the discussion of the findings.

Macroscopic Crash Prediction Models

MCPM is one set of methods that explore the relationship between road safety at a macroscopic level with sociodemographic and transportation infrastructure. By using locations of the traffic crashes at the zonal level, researchers identified several factors that associate with crash frequency at the zonal level; including sociodemographic factors such as population density (Huang and Abdel-Aty 2010), age cohorts (Aguero-Valverde and Jovanis 2006, Hadayeghi *et al.* 2010b, Pirdavani *et al.* 2012b, Dong *et al.* 2015), incomes (Pirdavani *et al.* 2012b, Xu and Huang 2015), and employment (Quddus 2008, Hadayeghi *et al.* 2010b). Travel behavior (Naderan and Shahi 2010, Abdel-Aty *et al.* 2011, Dong *et al.* 2014, Dong *et al.* 2015) and road network characteristics such as road lengths with different speed limit (Abdel-Aty *et al.* 2011, Siddiqui *et al.* 2012), road functional classification (Quddus 2008, Hadayeghi *et al.* 2010b), intersections density (Huang and Abdel-Aty 2010, Xu and Huang 2015), and traffic patterns (Quddus 2008, Hadayeghi *et al.* 2010b, Pirdavani *et al.* 2012b) also had association with safety at zonal level.

Heterogeneity is one of the issues that may affect the inferences of the MVPM models where exogenous variables do not vary identically across observations (Xu *et al.* 2017). Heterogeneity could be attributed to the presence of factors (i.e., unknown or known) that are not likely to be available for the analysis and their exclusion from the estimation process may yield biased parameters, which eventually lead to drawing incorrect inferences (Mannering *et al.* 2016). This phenomenon impacts the association among dependent variables and exogenous variables. This issue could be addressed by either spatial or non-spatial methods; each of them has their own advantages. In the non-spatial approach, parameters are drawn from some random distribution and are assumed to vary randomly across observations (Anastasopoulos and Mannering 2011, Chen and Tarko 2014, Xu and Huang 2015). On the other hand, spatial models consider the spatial relationship of the observations to address this issue (Aguero-Valverde and Jovanis 2006, Pirdavani *et al.* 2013b, Xu and Huang 2015).

Methodology

Home-Based Approach Crash Definition

The home-address of the road users who were involved in a traffic crash is one of the data elements that police officer records in crash reports at the crash scene (MMUCC 2012). Using home-address to collect information of the road users to collect data element regarding sociodemographic and travel behavior is a common practice in urban travel demand analysis (Kanafani 1983). In consideration of this practice, we will use the collected home-address of individuals as a basis for further analysis. To tie traffic crashes to the home addresses of the individuals in this study, we define the HBA crash frequency as *the expected number of crashes that road users who live in a certain geographic area have during a specified period*. This definition attributes traffic crashes to individuals and their residential addresses.

To account for individuals who directly contributed to the crashes, we only included drivers, bicyclists, motorcyclists, and pedestrians in this study. We excluded other road users type from our analysis (e.g., vehicle occupants, witnesses) since they did not have a direct role in the traffic crash occurrence.

Data Source and Geocoding Process

The data in this study was provided by Tennessee Integrated Traffic Analysis Network (TITAN), a portal provided by Tennessee Highway Patrol (THP) as a repository for traffic crash and surveillance reports completed by Tennessee law enforcement agencies. For the years 2014, 2015, and 2016, the TITAN records include 694,276 crashes and information of 2,026,506 individuals who were involved in traffic crashes. Each record includes information about the road user type, coordinates of the crashes and addresses of the individual who were involved in traffic crashes.

After obtaining the address of the pedestrians, bicyclists, motorcyclists, and drivers (n= 1,203,887), we used the Bing application program interface (API) services to geocode the addresses. The quality of the geocoding was checked by controlling for the locality of the addresses. Only those records that had an accuracy level of premises (e.g., property name, building name), address level accuracy, or intersection level accuracy was used for the analysis. After controlling for the address quality, 1,113,830, (92.5%) records met the minimum address quality filter. Of those, 1,002,362 had a Tennessee home-address (90.0% of geocoded addresses); there were 111,468 out-of-state addresses.

In this study, one goal was to investigate the relationship between sociodemographic variables and crash frequency at the zonal level. For that reason, we used census tract as the geographic unit. Census data from US survey in 2010 was also used to obtain sociodemographic data elements in each census tract in Tennessee. Table 1 presents the descriptive statistics of the sociodemographic variables. Additionally, we used Highway Performance Monitoring System data for Tennessee in 2015 to obtain Average Annual Daily Traffic for each road segment and calculate total Daily Vehicle Miles Travelled (DVMT) at the census tract level.

Modeling Approach

To evaluate safety at zonal level, traditionally, count data models are commonly utilized owing to the nature of traffic crashes that are usually measured as non-negative integers in a specific period of time (Anastasopoulos and Mannering 2009). Similar to LBA crash frequency, HBA also has non-negative integers. Hence the models that would be used to evaluate HBA crash frequency must follow the nature of counts model. Using the analogy between HBA and LBA crash frequency, and considering the spatial heterogeneity, in this study, we will use Poisson Regression model and the Geographically Weighted Poisson Regression Model (GWPR) to explore the correlation between crash frequency and census tract characteristics.

Poisson Model

In the Poisson regression, the probability that crash frequency at zone *i* equals to n could be written as (Greene 2003):

$$P(n_i) = \frac{\lambda_i^{n_i} \exp(-\lambda_i)}{n_i!} \qquad (1)$$

where λ_i (Poisson parameter) is the expected crash frequency (i.e., HBA or LBA crash frequency) for zone *i* in a three-year period. In order to fit the regression model, the Poisson parameter, λ_i, can be written in a logarithm format (Greene 2003):

$$\ln(\lambda_i) = \beta X_i \qquad (2)$$

where X_i is the vector of the sociodemographic data element extracted from census tract and β is a vector of the estimated coefficients.

Notably, in cases where the mean and the variance are not equal, applying the Poisson regression might lead to inappropriate results. In order to statistically test the existence of over-dispersion in the Poisson model, the Lagrange multiplier method was performed (Greene 2003):

$$LL = \left(\frac{\sum_{i=1}^{N}((y_i - \mu_i)^2 - y_i)}{2\sum_{i=1}^{N}\mu_i^2}\right)^2 \quad (4)$$

where y_i is the observed crash frequency at zone i, μ_i is the predicted crash frequency at zone i, and N is the number of zones.

Geographically Weighted Poisson Regression Model

GWPR can be used to examine whether the association between exogenous variables and crash frequency substantially varies across space (Fotheringham *et al.* 2003). The model can be written as:

$$\ln(\lambda_i) = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)\ln(E_{vi}) + \sum_{k=1}^{K}\beta_k(u_i, v_i)x_{ij} + \epsilon_i \quad (5)$$

where (u_i, v_i) denotes the coordinates of zone i. It should be noted that in the GWPR, $\beta_k(u_i, v_i)$ is a function of the coordinates of the center of census tract i. The following equation can be used to estimate $\beta_k(u_i, v_i)$:

$$\hat{\beta}(\mathrm{u_i}, \mathrm{v_i}) = (\mathrm{X^T W(u_i, v_i) X})^{-1}\mathrm{X^T W(u_i, v_i) Y} \quad (6)$$

where $\hat{\beta}(\mathrm{u_i}, \mathrm{v_i})$ is the vector of estimated coefficients at zone i, X is the matrix of exogenous variables, Y is the $\mathrm{n}\times 1$ vector of the dependent variable (*crash frequency*), and $\mathrm{W(u_i, v_i)}$ is $\mathrm{n}\times\mathrm{n}$ spatial weight matrix. :

$$\mathrm{W(u_i, v_i)} = \begin{bmatrix} \mathrm{w_{i1}} & 0 & \cdots & 0 \\ 0 & \mathrm{w_{i2}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \mathrm{w_{in}} \end{bmatrix} \quad (7)$$

where $\mathrm{w_{ij}}$ is the weight of variable j at location i. In this approach, a regression equation is estimated for each location based on observations at nearby areas. Based on the distance from the regression point each area is weighted (areas that are closer have a higher weight than ones that are farther). There are several methods for estimating the W matrix. In this study, we will use the adaptive bi-square kernel to construe the weight matrix.

$$\mathrm{w_{ij}} = \begin{cases} (1 - \frac{\mathrm{d_{ij}}^2}{\theta_{\mathrm{i(k)}}})^2 & \mathrm{d_{ij}} < \theta_{\mathrm{i(k)}} \\ 0 & \mathrm{d_{ij}} > \theta_{\mathrm{i(k)}} \end{cases} \quad (8)$$

where $\mathrm{d_{ij}}$ is the Euclidean distance between observation i and j and $\theta_{\mathrm{i(k)}}$is an adaptive bandwidth size defined as the $\mathrm{k-th}$ nearest neighbor distance. The best bandwidth is the one with the lowest AICc score (Fotheringham et al. 2003, Hadayeghi et al. 2010a, Nakaya 2014) and we used golden search method to identify the best bandwidth size. For more details about the golden search please see Nakaya (2014). Moran's I (Moran 1950) was also used to test whether the model residuals are spatially correlated. Moran's I values range between -1 to +1. The extreme values are indicators of significant spatial autocorrelation where value close to 0 indicates a random pattern between residuals.

The non-stationarity test was used to evaluate the existence of variation in the estimated coefficients across space (Liu and Khattak 2017). By comparing the upper and lower quartile of the estimated coefficients from GWPR model ($\delta = \beta_{upper} - \beta_{lower}$), we can write:

$$\begin{cases} \delta > 1.96 * SE \text{ and}, \\ 1.96 < \max(|z_i|) \quad Pass\ the\ test\ (local\ coefficient) \\ if\ not \quad failed\ to\ pass\ (global\ coefficient) \end{cases} \quad (9)$$

where SE is the standard error of the coefficient in the global Poisson model, and $|z_i|$ is the absolute value of the significances-score of the GWPR model at census tract i. If the value of δ meet the condition in equation 9, we can conclude there is substantial variations among the estimated coefficients across the space. Otherwise, the coefficient is considered as the global coefficient, which does not have a substantial spatial variation. In order to estimate GWPR model, GWR4.0 software which is developed by Nakaya *et al.* (2012) was used.

Measures of Goodness of Fit

To evaluate and compare the performance of traditional Poisson regression, and GWPR, three statistics were utilized to measure estimation accuracy.

1- *R-squared for the Poisson model*: this statistic assesses the overall goodness of fit based on standardized residuals. Larger values of $R^2_{Poisson}$ (max is 1) indicate better fit. It is defined as (Cameron and Windmeijer 1996):

$$R^2_{Poisson} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2/\widehat{Y}_i}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2/\bar{Y}} \qquad (10)$$

2- where Y_i and $\widehat{Y}_i$ are the observed and crash frequency at location *i* respectively, and $\bar{y}$ is the average number of crashes.

3- *AIC*: a lower AIC represents the better goodness of fit (Bozdogan 1987). A three-point decrease in an AIC value indicates a significant improvement in the goodness of fit (Bozdogan 1987). We can write:

4- $AIC = D + 2k \qquad (11)$

where D denotes the model deviance, and k is the number of parameters. In the GWPR, due to the non-parametric framework of the model, the number of parameters is meaningless. Therefore, an effective number of parameters should be considered, which can be written as (Nakaya *et al.* 2005):

$$K = trace(S) \qquad (12)$$

where S is the hat matrix. For more details, please see (Nakaya *et al.* 2005).

5- *Mean Absolute Deviation:* a smaller value of MAD implies a better model estimation. It can be defined as:

6- $MAD = \frac{\sum_{i=1}^{n}|\hat{Y}_i - Y_i|}{N} \qquad (13)$

We use the abovementioned goodness of fit to compare models within each approach (i.e., Poisson vs. GWPR).

Results

After assigning the crashes to the crash location and individuals' home-addresses and assigning them to corresponding census tracts, we calculated the crash frequency at the zonal level. The average of HBA crash frequency at the census tract level is 243.28 (SD = 164.17). The average of the crashes that occurred at census tract is 134.40 (SD = 171.46). The correlation between crash frequencies at the census tract level of the two approaches was 0.27 which indicated a weak linear relationship. The overall rate is calculated as crashes per 1000 population. Figure 5 and Figure 6 respectively map the LBA crash rate (i.e., number of crashes occurred in a census tract per 1000 population) and HBA crash rate (i.e., number of crashes that residents of a census tract had per 1000 population) for the period 2014-16. The correlation between the two approaches' risk is 0.33 which also indicates a weak linear relationship. Noticeably, as Figure 5 presents, the concentration of the areas with higher crash frequency are along the freeway corridors (e.g., I-40 and I-70) and rural highways. Individuals with a higher likelihood of involvement in traffic crashes are more concentrated around metropolitan areas (e.g., Nashville, Memphis, Knoxville). The comparison of the HBA crash rate over metropolitan areas (Table 2) shows that Middle Tennessee (i.e., Nashville metropolitan area) and Chattanooga respectively have the highest HBA crash rate. The Tri-cities and Greater Knoxville have the lowest HBA crash rate. In addition, Greater Chattanooga and Jackson have the highest LBA crash rate. Alternatively, Knoxville and Tri-cities have the lowest LBA crash rate.

Table 1 Descriptive statistics of the census tract variables

Variable	Mean	Std. Dev.	Max
Total Population	1531.99	784.97	9281.00
Age Cohort Proportion			
16 Years and Younger	0.23	0.08	0.71
16-42 Years Old	0.32	0.11	1.00
43-59 Years Old	0.25	0.07	1.00
60 Years Old and More	0.20	0.10	1.00
White Race Proportion	0.77	0.29	1.00
Means Of Transportation to Work Proportion			
Personal Vehicle	0.92	0.10	1.00
Carpool	0.10	0.08	0.82
Active Mode	0.03	0.07	1.00
Average Travel Time To Work (Minute)	25.18	6.45	65.85
Education Degree Proportion			
High School and Lower	0.52	0.20	1.00
Some College Degree	0.20	0.08	1.00
Bachelors' Degree	0.20	0.12	1.00
Others' Degrees	0.08	0.08	0.73
Median Household Income (1,000 $)	45.85	25.05	249.38
Household Vehicles' Ownership Proportion			
No Vehicle	6.94	9.34	73.35
One Or Two Vehicles	0.70	0.13	1.00
Three Or More Vehicles	0.22	0.13	1.00
Daily VMT (10,000)	5.71	6.88	4.94

Table 2 Crash Frequency and Crash Rate in Metropolitan areas (2014-16 Crashes)

		Crash Frequency		Crash Rate (per 1000 population)	
MPO	Population	HBA	LBA	HBA	LBA
Chattanooga	405,026	76,828	55,195	189.7	136.3
Knoxville	972,667	136,510	81,344	140.3	83.6
Jackson	102,711	15,224	11,620	148.2	113.1
Memphis	1,008,864	171,440	89,082	169.9	88.3
Nashville	1,582,462	312,987	159,049	197.8	100.5
Tri-Cities	346,106	48,579	30,106	140.4	87.0
Non-Metropolitan Areas	1,880,155	240,629	125,474	128.0	66.7
Average				159.1	87.6

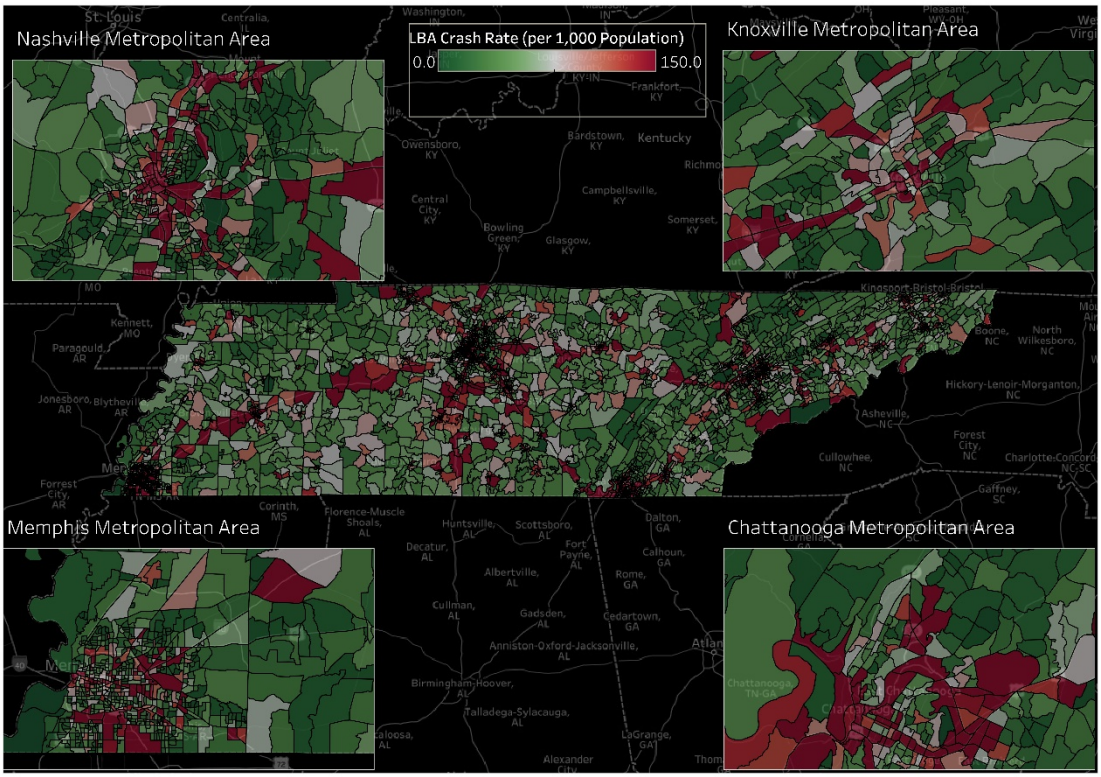

Figure 5 Location-Based Approach Crash Rate

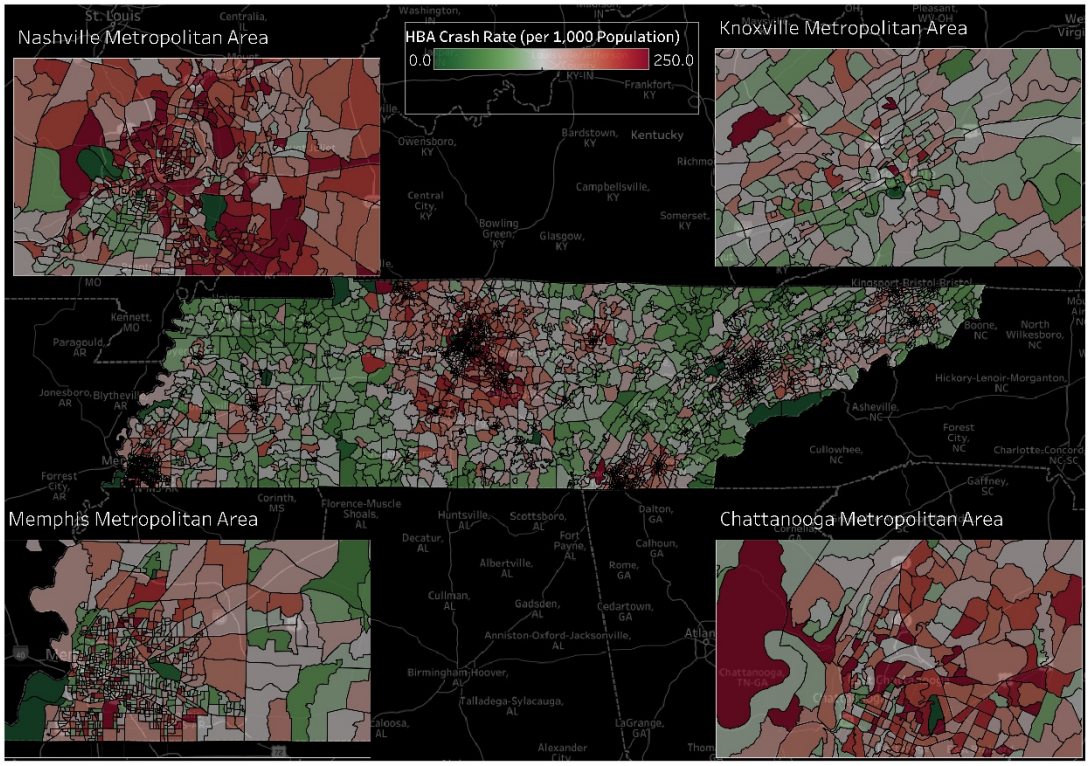

Figure 6 Home-based approach crash rate

Comparison of the Poisson and GWPR models

Table 3 and Table 4 present the results of estimated negative Poisson regression and GWPR model for both HBA and LBA models. Considering the Lagrange multiplier, we can conclude that the over-dispersion is not an issue in both models. Comparison of the GWPR models and Poisson models by using MAD, $R^2_{Poisson}$, and corrected AIC values indicate that GWPR models are more suitable compared to the global models. Moran's I of residuals in both models indicate that spatial autocorrelation is not an issue in the GWPR models.

Table 3 Results of Poisson and GWPR model for LBA results

	Estimate	Standard Error	z(Est/SE)	Mean	STD	Min	Lower Quartile	Median	Upper Quartile	Max
Intercept	3.956	0.025	158.781	3.697	4.038	-18.294	1.083	3.760	6.397	17.028
Population	0.177	0.001	126.602	0.226	0.259	-0.958	0.083	0.213	0.397	1.452
Age cohorts proportion										
Under 16 years	-0.935	0.021	-45.545	-1.225	2.327	-10.305	-2.642	-1.152	0.224	6.550
between 16-42	0.188	0.017	11.119	-0.207	1.943	-8.032	-1.634	-0.192	1.180	6.176
between 43-59	-0.339	0.025	-13.612	-0.862	2.434	-13.003	-2.372	-0.604	0.826	7.871
White Race Proportion	0.200	0.006	32.099	-0.013	1.893	-10.070	-0.826	-0.055	0.716	10.462
Average Travel Time to Work	-0.025	0.000	-99.451	-0.021	0.034	-0.187	-0.043	-0.017	0.000	0.103
Household Income	-2.15E-03	8.60E-05	-24.944	-4.08E-03	1.25E-02	-5.78E-02	-1.21E-02	-2.54E-03	3.97E-03	4.78E-02
Vehicle Ownership										
Household With No-Vehicle	1.709	0.018	96.098	1.278	3.236	-16.932	-0.501	1.519	3.286	12.571
Household With 1 or 2 vehicles	0.858	0.012	69.457	0.690	1.642	-7.321	-0.267	0.592	1.632	6.787
Daily-VMT (10,000 Miles)	0.005	0.000	482.648	0.007	0.004	-0.004	0.005	0.007	0.009	0.025
Travel Model To work										
Personal Vehicle	0.047	0.019	2.539	0.872	2.906	-8.966	-0.812	0.789	2.425	16.626
Active More	1.100	0.028	38.701	1.349	5.889	-29.868	-1.899	1.085	4.488	38.338
Education										
College Degree	0.862	0.018	48.644	0.420	1.733	-6.645	-0.681	0.398	1.461	7.921
Bachelor Degree	0.441	0.016	28.430	0.250	1.889	-5.514	-1.108	0.262	1.609	6.258
Classic AIC:	281805.3			78007.81						
AICc:	281805.5			80603.01						
Percent deviance explained	0.46			0.86						
Deviance:	281775.3			74508.61						
MAD	73.6			35.8						
R Poisson	0.59			0.92						
Lagrange Multiplier	0.28			0.04						
Moran's I of residuals	0.08			-0.01						
Bandwidth	Not applicable			70.00						

Table 4 Results of Poisson and GWPR model for HBA results

	Estimate	Standard Error	z(Est/SE)	Mean	STD	Min	Lower Quartile	Median	Upper Quartile	Max
Intercept	3.044	0.024	127.830	3.871	1.942	-10.303	2.864	4.020	5.027	9.430
Population	0.337	0.001	390.872	0.503	0.125	0.180	0.422	0.501	0.582	0.909
Age cohorts proportion										
Under 16 years	0.218	0.016	13.678	-0.457	0.816	-3.816	-0.999	-0.425	0.091	2.625
between 16-42	0.583	0.014	41.992	-0.179	0.902	-5.505	-0.704	-0.128	0.420	3.776
between 43-59	0.736	0.020	36.364	0.244	0.938	-3.313	-0.361	0.282	0.784	4.155
White Race Proportion	-0.203	0.005	-44.791	-0.010	1.037	-3.758	-0.429	-0.112	0.290	12.610
Average Travel Time to Work	0.010	0.000	56.306	0.005	0.014	-0.052	-0.003	0.005	0.013	0.075
Household Income	0.001	0.000	14.479	0.001	0.005	-0.016	-0.002	0.000	0.003	0.021
Vehicle Ownership										
Household With No-Vehicle	1.56E-04	1.73E-04	0.903	0.001	0.011	-0.059	-0.005	0.001	0.008	0.049
Household With 1 or 2 vehicles	0.278	0.010	27.266	0.287	0.647	-2.177	-0.091	0.285	0.680	3.775
Daily-VMT (10,000 Miles)	5.69E-03	1.30E-04	43.770	0.005	0.011	-0.059	-0.001	0.004	0.010	0.064
Travel Model To work										
Personal Vehicle	0.857	0.020	43.346	0.422	1.245	-4.234	-0.276	0.330	1.072	7.583
Active More	0.018	0.029	0.609	-0.520	2.230	-8.374	-1.933	-0.540	0.801	8.372
Education										
College Degree	0.865	0.014	63.390	0.281	0.802	-2.160	-0.218	0.206	0.725	3.768
Bachelor Degree	0.499	0.012	42.024	0.117	0.713	-3.130	-0.293	0.117	0.537	2.749
Classic AIC:	118441.6			29716.08						
AICc:	118441.7			32681.39						
Percent deviance explained	0.66			0.92						
Deviance:	118411.6			26044.5						
MAD	64.53			29.06						
R Poisson	0.74			0.94						
Lagrange Multiplier	0.07			0.01						
Moran's I of residuals	0.16			-0.005						
Bandwidth	Not applicable			72.00						

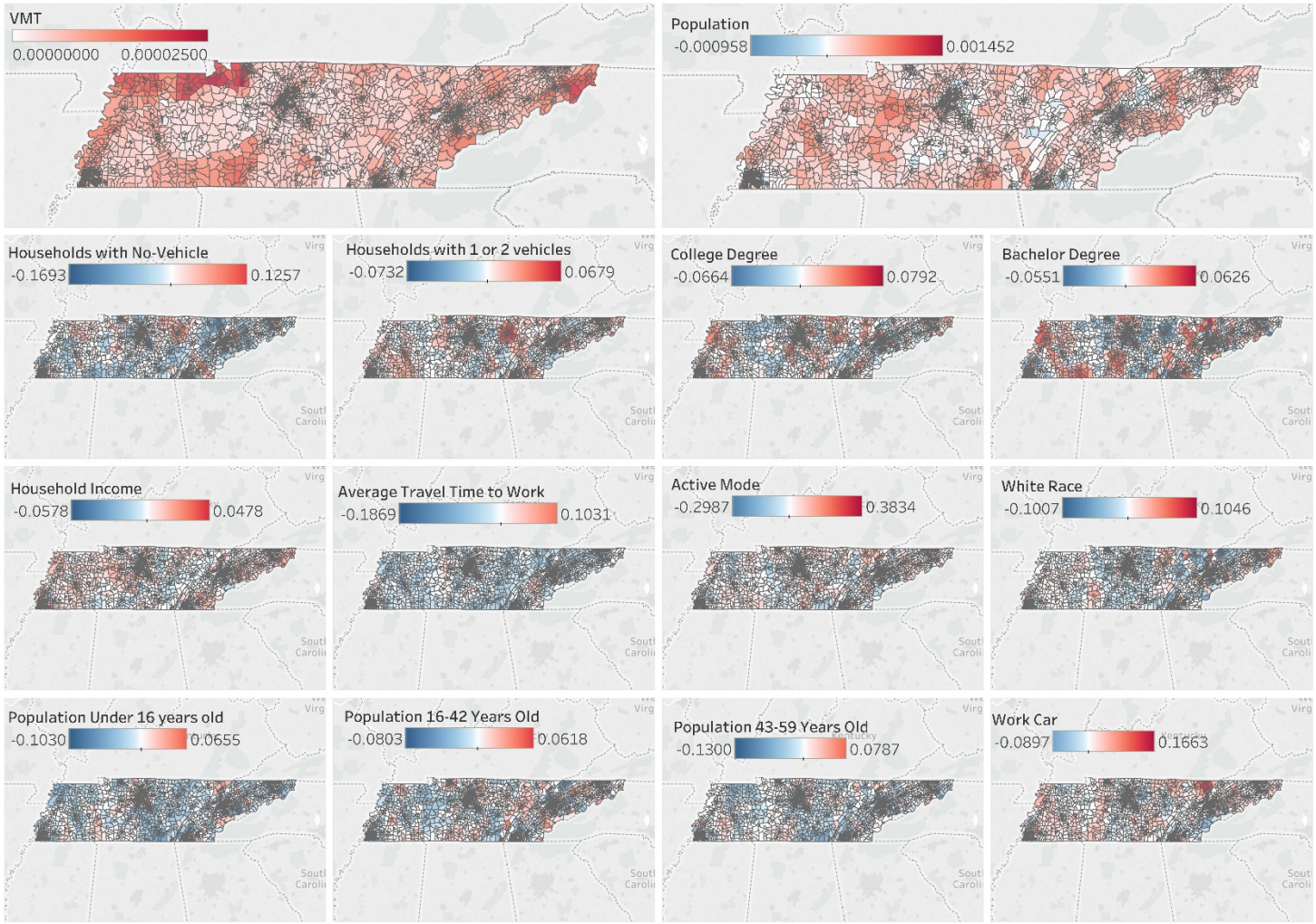

Figure 7 Local visualization of parameter estimates obtained from GWPR model for predicting LBA crash frequency

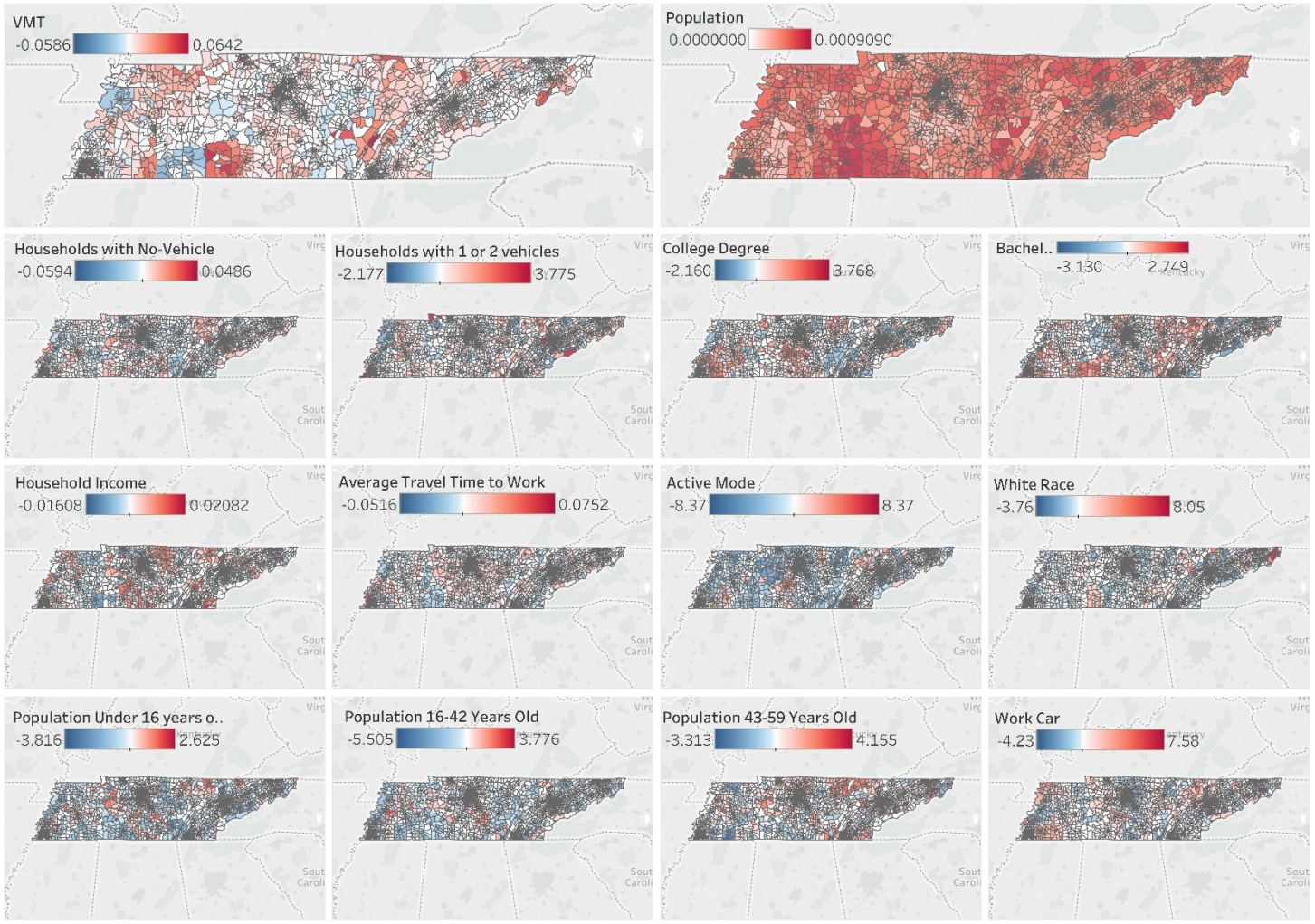

Figure 8 Local visualization of parameter estimates obtained from GWPR model for predicting HBA crash frequency

Figure 7 and Figure 8 present the estimated local coefficients of the GWPR models. The white census tracts present the areas where the local estimated coefficient is not significant. Visual inspection of the figures reveals the presence of the spatial patterns of the estimated coefficients. Likewise, non-stationary test results indicate that all the coefficient in both models has substantial local effects. Considering the range of the estimated coefficients, we learn that all the local estimated coefficient except for population variable in the HBA models includes both negative and positive values. The counterintuitive signs of the local coefficients in the GWPR model is a common issue, and many studies report this issue (Chow *et al.* 2006, Hadayeghi *et al.* 2010b, Pirdavani *et al.* 2013b, Xu and Huang 2015). This issue could arise due to local multicollinearity; however, by controlling for local multicollinearity we realized that the maximum VIF values in the areas with counterintuitive signs is unsubstantial; VIF values range between 1.01 to 3.95. Furthermore, overdispersion of the crash data also could affect the counterintuitive signs. In general, the Poisson regression model may underestimate the variance of the parameters (Lord and Mannering 2010), hence, it produces more significant variables (Xu and Huang 2015).

To discuss the impact of estimated coefficients, we use the mean of the local coefficients of the GWPR models to deliberate the relationship between independent variables and crash frequency at the zonal level. Nevertheless, one needs to consider the magnitude and counterintuitive signs of the local coefficients in the interpretation of the results.

In the LBA model, average travel time to work has a negative sign which indicates that fewer crashes occur in the areas where their residents have to travel longer distances. Areas with longer travel time may be used as a proxy for suburban and rural areas, which usually have a lower crash rate compared to urban areas (Zwerling *et al.* 2005). Cheng *et al.* (2018) explored the association between mean travel time and crash frequency of various road users; their findings indicated the different mean of travel time could have a positive or negative effect on crash frequency for different road users. On the other hand, in the HBA model, as average travel time increases, the HBA crash frequency increases as well. Average travel time to work is a proxy for the exposure to traffic. Considering the sign of average travel time, we may conclude that those who live in the suburban or rural areas (i.e., higher travel time) are engaged in more traffic crashes; however, their crashes may occur in the areas with lower travel time (i.e., urban areas).

In line with previous studies, daily-VMT in the census tract has a positive significant association with both LBA crash frequency (e.g., Lee *et al.* 2015, Cai *et al.* 2017, Cheng *et al.* 2018) and HBA crash frequency. This implies that as traffic exposure increases, the likelihood of involvement in traffic crashes increases; this is also the case for the census population. By setting families with 3 or more vehicles as a base in both models, the proportion of families with no-vehicle and 1 or 2 vehicles have a positive significant association with LBA and HBA crash frequency. The positive sign of vehicle ownership is in agreement with previous research (Lee *et al.* 2015, Xu and Huang 2015).

As expected, travel modes also have a direct impact on both models; the proportion of personal vehicles has a positive impact on both LBA and HBA crash frequency. The positive sign implies that as number of personal vehicle increases, the frequency of HBA and LBA crash frequency increases. Increase in the proportion of personal vehicles has a direct impact on VMT and eventually traffic exposure. Percent of active mode use has varying signs in both models. In the HBA models, active modes have a significant negative sign, which indicates that areas where their residents use active modes more often have better safety records, while, the positive sign of active modes in the LBA model implies that more crashes occur in the areas with active mode users. The positive sign is in agreement with previous studies (Lee *et al.* 2015, Dong *et al.* 2016, Cai *et al.* 2017).

Household income also has varied signs in both models. In the LBA model, the mean of estimated coefficient has a negative sign; this negative sign indicates that as median household income in a census tract increases, the LBA crash frequency decreases. The negative sign also implies that as median income in one area increase, the safety (i.e., LBA crash frequency) improves. This finding is in agreement with the previous literature (Pirdavani *et al.* 2012a, 2013a, Cai *et al.* 2017, Gomes *et al.* 2017). In contrast, in the HBA model, the sign of median household income has a positive association with HBA crash frequency, this sign implies that as median household income in a census tract increases, census residents are more involved in traffic crashes. The positive association could be explained by the trip rate, household travel survey studies in the study area reveal that as household income levels increase, trip rate for a family increases (KRTPO 2008, Lee *et al.* 2013); higher trip rates associated with higher exposure to traffic and eventually increase in the HBA crash frequency.

In line with previous models, in the LBA model, age cohorts have a significant impact on LBA crash frequency. By setting the proportion of population over 60-year-old as the base group, we find that the proportion of other groups have a significant negative impact on LBA crash frequency. This finding is consistent with Dong *et al.* (2016). The estimated coefficients in the GWPR model implies that areas with a higher population of seniors, are more prone to traffic crashes, which is similar to Gomes *et al.* (2017) but contradicts Cai *et al.* (2017). In the HBA model, a positive sign of the group's age between 43-59 years old, and negative sign of groups aged between 16-42 and below 16-year-old implies that the former group has a higher likelihood of involvement in traffic crashes.

Considering race, the parentage of the white population has a negative association with the HBA and LBA crash frequency, which implies that census tracts with higher white population are less likely to be involved in traffic crashes, or have traffic crashes on roadways that intersect their neighborhood. The positive sign agrees with previous research; for example, Gomes *et al.* (2017) reported that an increase in the proportion of the minority races increases crash frequency. Lee *et al.* (2015) also reported that an increase in the proportion of African-American and Hispanic populations increase the crash frequency of vehicle, bicycle, and pedestrian crashes at the zonal level.

Summary and Conclusion

In this study, we introduced a new approach to evaluate road safety that focuses on the home address of individuals (i.e., home-based approach) who were directly involved in traffic crashes instead of the location of the crashes (location-based approach). While LBA explores the geographical distribution of traffic crashes by focusing on the location of traffic crashes, HBA considers the socioeconomics associated with the location of the person involved in the crash. This approach could be used to explore the road safety disparities by considering the factors surrounding home-address of the crash victims.

Comparing of the metropolitan areas exhibited the spatial variation in both approaches' crash rate. The LBA crash rate could be attributed to the activity and infrastructure quality in each metropolitan area. In addition to the activity and infrastructure quality, HBA crash rate also reflects the safety culture. The difference in the safety culture (e.g., drunk driving, seatbelts) of residents of different metropolitans could be investigated in future studies.

In this study, we used GWPR model to address the unobserved heterogeneity in both approaches. The difference between model specifications in HBA and LBA and weak correlation between crash frequency and crash rate in both approaches demonstrated the merits of HBA as a complementary solution in addition to the LBA method as an index to evaluate road safety and identify areas where their residents have a higher likelihood of involvement in traffic crashes.

GWPR model findings indicate that residents who live in the neighborhoods with higher income, higher white race population, more use of an active mode of transportation are less likely to be involved in a traffic crash. Proper safety campaigns could be used to address the safety concerns in the HBA hotspots, particularly by focusing on behavioral interventions that contribute to higher crash risk and injury burden (e.g., speeding). HBA may also be used in road safety campaign design as a solution to prioritize neighborhoods for allocating educational resources.

One shortcoming of this study was that the true value of the exposure of the residents in census tracts is not known. As a result, we used VMT, average travel time to work, and census population as a substitute for the true value of the exposure. These three variables had a significant association with crash frequencies while their magnitude and sign differed in both models. One solution to acquire a better estimate of exposure is to use a travel demand model as an input for the HBA model. Unfortunately, in this study travel demand model was not available. Moreover, in this study we only had access to individuals' traffic crashes in Tennessee. As a result, it is possible that we underestimate individual HBA crash frequency and the crash rate at a geographical level. This issue could be challenging in the areas that individuals have more out-of-state trips.

Future directions

In this study, we only considered sociodemographic variables for the analysis and did not consider the effect of the built environment (e.g., transportation system) in both approaches. Although the causality of the built environment in the LBA is known, we need to scrutinize the association between the built environment, network characteristics and HBA crash frequency. This problem should be addressed in the future studies. Bearing in mind that HBA attributes safety to the home address of the individual, and travel demand model reveals individuals' travel behavior at their origin TAZ that includes home address of travelers, we expect close association between HBA and planning models. As a result, one may imagine HBA integration to travel demand planning frameworks. The HBA could be used in transportation planning process to evaluate the impact of accessibility to activities and transportation modes, as well as trip generation at zonal crash frequencies. Moreover, the HBA could be used in the design of the safety campaigns and prioritizing areas where road users have a higher likelihood of involvement in traffic crashes.

In this study, in order to construe the weighting matrix, we used four different kernel option offered by GWR software namely, fixed Gaussian, fixed bi-square, adaptive bi-square, and adaptive Gaussian. The comparison of four models by using corrected AIC indicated that adaptive bi-square had the lowest value of corrected AIC and therefore is more suitable for presentation purposes. We found that the results were largely consistent with different kernel options and in order to maintain concision, we only included the adaptive bi-square model.

It is also worth mentioning that there are difficulties in accessing the crash data with identifiers and it is not possible to obtain this data in some cases. One possible direction for future research could be to develop a methodology to identify areas where road users have a higher likelihood of involvement in traffic crashes based on the conventional police crash databases that may not have identifiers, or partnering with data owners to assist in matching crashes with spatial datasets.

Acknowledgment

The authors would like to thank the Tennessee Department of Safety and Homeland Security for providing the data for this study. This project was supported by the Collaborative Sciences Center for Road Safety, www.roadsafety.unc.edu, a U.S. Department of Transportation National University Transportation Center

promoting safety. The study design was reviewed and approved by the University of Tennessee Institutional Review Board.

References

Abdel-Aty, M., Siddiqui, C., Huang, H., Wang, X., 2011. Integrating Trip and Roadway Characteristics to Manage Safety in Traffic Analysis Zones. Transportation Research Record: Journal of the Transportation Research Board (2213), 20-28.

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial Analysis of Fatal and Injury Crashes in Pennsylvania. Accident Analysis & Prevention 38 (3), 618-625.

Anastasopoulos, P.C., Mannering, F.L., 2009. A Note on Modeling Vehicle Accident Frequencies with Random-Parameters Count Models. Accident Analysis & Prevention 41 (1), 153-159.

Anastasopoulos, P.C., Mannering, F.L., 2011. An Empirical Assessment of Fixed and Random Parameter Logit Models Using Crash-and Non-Crash-Specific Injury Data. Accident Analysis & Prevention 43 (3), 1140-1147.

Baker, S.P., Braver, E.R., Chen, L.-H., Pantula, J.F., Massie, D., 1998. Motor Vehicle Occupant Deaths among Hispanic and Black Children and Teenagers. Archives of pediatrics & adolescent medicine 152 (12), 1209-1212.

Bozdogan, H., 1987. Model Selection and Akaike's Information Criterion (Aic): The General Theory and Its Analytical Extensions. Psychometrika 52 (3), 345-370.

Braver, E.R., 2003. Race, Hispanic Origin, and Socioeconomic Status in Relation to Motor Vehicle Occupant Death Rates and Risk Factors among Adults. Accident Analysis & Prevention 35 (3), 295-309.

Cai, Q., Abdel-Aty, M., Lee, J., Eluru, N., 2017. Comparative Analysis of Zonal Systems for Macro-Level Crash Modeling. Journal of safety research 61, 157-166.

Cameron, A.C., Windmeijer, F.A., 1996. R-Squared Measures for Count Data Regression Models with Applications to Health-Care Utilization. Journal of Business & Economic Statistics 14 (2), 209-220.

Campos-Outcalt, D., Bay, C., Dellapena, A., Cota, M.K., 2003. Motor Vehicle Crash Fatalities by Race/Ethnicity in Arizona, 1990–96. Injury Prevention 9 (3), 251-256.

Chen, E., Tarko, A.P., 2014. Modeling Safety of Highway Work Zones with Random Parameters and Random Effects Models. Analytic methods in accident research 1, 86-95.

Chen, F., Wu, J., Chen, X., Wang, J., Wang, D., 2016. Benchmarking Road Safety Performance: Identifying a Meaningful Reference (Best-in-Class). Accident Analysis & Prevention 86, 76-89.

Cheng, W., Gill, G.S., Ensch, J.L., Kwong, J., Jia, X., 2018. Multimodal Crash Frequency Modeling: Multivariate Space-Time Models with Alternate Spatiotemporal Interactions. Accident Analysis & Prevention 113, 159-170.

Chow, L.-F., Zhao, F., Liu, X., Li, M.-T., Ubaka, I., 2006. Transit Ridership Model Based on Geographically Weighted Regression. Transportation Research Record: Journal of the Transportation Research Board (1972), 105-114.

Dong, N., Huang, H., Lee, J., Gao, M., Abdel-Aty, M., 2016. Macroscopic Hotspots Identification: A Bayesian Spatio-Temporal Interaction Approach. Accident Analysis & Prevention 92, 256-264.

Dong, N., Huang, H., Xu, P., Ding, Z., Wang, D., 2014. Evaluating Spatial-Proximity Structures in Crash Prediction Models at the Level of Traffic Analysis Zones. Transportation Research Record: Journal of the Transportation Research Board (2432), 46-52.

Dong, N., Huang, H., Zheng, L., 2015. Support Vector Machine in Crash Prediction at the Level of Traffic Analysis Zones: Assessing the Spatial Proximity Effects. Accident Analysis & Prevention 82, 192-198.

Evans, L., 1996. The Dominant Role of Driver Behavior in Traffic Safety. American Journal of Public Health 86 (6), 784-786.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2003. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships John Wiley & Sons.

Gomes, M.J.T.L., Cunto, F., da Silva, A.R., 2017. Geographically Weighted Negative Binomial Regression Applied to Zonal Level Safety Performance Models. Accident Analysis & Prevention 106, 254-261.

Greene, W.H., 2003. Econometric Analysis Pearson Education India.

Hadayeghi, A., Shalaby, A., Persaud, B., 2010a. Development of Planning-Level Transportation Safety Models Using Full Bayesian Semiparametric Additive Techniques. Journal of Transportation Safety & Security 2 (1), 45-68.

Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2010b. Development of Planning Level Transportation Safety Tools Using Geographically Weighted Poisson Regression. Accident Analysis & Prevention 42 (2), 676-688.

Harper, J.S., Marine, W.M., Garrett, C.J., Lezotte, D., Lowenstein, S.R., 2000. Motor Vehicle Crash Fatalities: A Comparison of Hispanic and Non-Hispanic Motorists in Colorado. Annals of emergency medicine 36 (6), 589-596.

Hauer, E., 1997. Observational before-after Studies in Road Safety--Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety.

Hermans, E., 2009. A Methodology for Developing a Composite Road Safety Performance Index for Cross-Country Comparison. UHasselt Diepenbeek.

Holló, P., Eksler, V., Zukowska, J., 2010. Road Safety Performance Indicators and Their Explanatory Value: A Critical View Based on the Experience of Central European Countries. Safety science 48 (9), 1142-1150.

HSM, 2010. Highway Safety Manual. American Association of State Highway and Transportation Officials 2, 10-1.

Huang, H., Abdel-Aty, M., 2010. Multilevel Data and Bayesian Analysis in Traffic Safety. Accident Analysis & Prevention 42 (6), 1556-1565.

Kanafani, A., 1983. Transportation Demand Analysis.

Koornstra, M., Lynam, D., Nilsson, G., 2002. Sunflower: A Comparative Study of the Development of Road. Leidschendam: SWOV Institute for Road Safety Research.

KRTPO, 2008. 2008 East Tennessee Household Travel Survey Final Report. Knoxville Regional Transportation Planning Organization. Knoxville Regional Transportation Planning Organization, 206 Wild Basin Rd., Suite A-300 Austin, Texas 78746.

Lee, J., Abdel-Aty, M., Jiang, X., 2015. Multivariate Crash Modeling for Motor Vehicle and Non-Motorized Modes at the Macroscopic Level. Accident Analysis & Prevention 78, 146-154.

Lee, M., Chan, J., Fucci, A., Jaworski, D., Kline, J., Mccloskey, S., Wilson, L., Fussell, R., Brinkerhoff, P., 2013. Middle Tennessee Transportation and Health Study Final Report. In: Wesatat ed.

Liu, J., Khattak, A.J., 2017. Gate-Violation Behavior at Highway-Rail Grade Crossings and the Consequences: Using Geo-Spatial Modeling Integrated with Path Analysis. Accident Analysis & Prevention 109, 99-112.

Lord, D., Mannering, F., 2010. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. Transportation Research Part A: Policy and Practice 44 (5), 291-305.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved Heterogeneity and the Statistical Analysis of Highway Accident Data. Analytic methods in accident research 11, 1-16.

Marshall, W.E., Ferenchak, N.N., 2017. Assessing Equity and Urban/Rural Road Safety Disparities in the Us. Journal of Urbanism: International Research on Placemaking and Urban Sustainability 10 (4), 422-441.

Mayrose, J., Jehle, D.V., 2002. An Analysis of Race and Demographic Factors among Motor Vehicle Fatalities. Journal of Trauma and Acute Care Surgery 52 (4), 752-755.

McAndrews, C., Beyer, K., Guse, C.E., Layde, P., 2013. Revisiting Exposure: Fatal and Non-Fatal Traffic Injury Risk across Different Populations of Travelers in Wisconsin, 2001–2009. Accident Analysis & Prevention 60, 103-112.

MMUCC, 2012. Model Minimum Uniform Crash Criteria. DOT HS 811, 631.

Moran, P.A., 1950. Notes on Continuous Stochastic Phenomena. Biometrika 37 (1/2), 17-23.

Naderan, A., Shahi, J., 2010. Aggregate Crash Prediction Models: Introducing Crash Generation Concept. Accident Analysis & Prevention 42 (1), 339-346.

Nakaya, T., 2014. Gwr4 User Manual. WWW Document. Available online: http://www. st-andrews. ac. uk/geoinformatics/wp-content/uploads/GWR4manual_201311. pdf (accessed on 4 November 2013).

Nakaya, T., Charlton, M., Lewis, P., Fortheringham, S., Brunsdon, C., 2012. Windows Application for Geographically Weighted Regression Modeling. Ritsumeikan University, Kyoto, Japan.

Nakaya, T., Fotheringham, A.S., Brunsdon, C., Charlton, M., 2005. Geographically Weighted Poisson Regression for Disease Association Mapping. Statistics in medicine 24 (17), 2695-2717.

Petridou, E., Moustaki, M., 2000. Human Factors in the Causation of Road Traffic Crashes. European journal of epidemiology 16 (9), 819-826.

Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., Wets, G., 2012a. Application of Different Exposure Measures in Development of Planning-Level Zonal Crash Prediction Models. Transportation Research Record: Journal of the Transportation Research Board (2280), 145-153.

Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., Wets, G., Year. Developing Zonal Crash Prediction Models with a Focus on Application of Different Exposure Measures.

Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., Wets, G., 2013a. Evaluating the Road Safety Effects of a Fuel Cost Increase Measure by Means of Zonal Crash Prediction Modeling. Accident Analysis & Prevention 50, 186-195.

Pirdavani, A., Brijs, T., Bellemans, T., Wets, G., Year. Spatial Analysis of Fatal and Injury Crashes in Flanders, Belgium: Application of Geographically Weighted Regression Technique.

Quddus, M.A., 2008. Modelling Area-Wide Count Outcomes with Spatial Correlation and Heterogeneity: An Analysis of London Crash Data. Accident Analysis & Prevention 40 (4), 1486-1497.

Schiff, M., Becker, T., 1996. Trends in Motor Vehicle Traffic Fatalities among Hispanics, Non-Hispanic Whites and American Indians in New Mexico, 1958–1990. Ethnicity & health 1 (3), 283-291.

Siddiqui, C., Abdel-Aty, M., Choi, K., 2012. Macroscopic Spatial Analysis of Pedestrian and Bicycle Crashes. Accident Analysis & Prevention 45, 382-391.

WHO, 2015. Global Status Report on Road Safety 2015. World Health Organization. Violence Injury Prevention World Health Organization.

Winston Harrington, Parry, I., Margaret, W., 2006. Automobile Externalities and Policies. Discussion Paper, 45.

World Health Organization, W., Organization, W.H., 2014. The Top 10 Causes of Death.

Xu, P., Huang, H., 2015. Modeling Crash Spatial Heterogeneity: Random Parameter Versus Geographically Weighting. Accident Analysis & Prevention 75, 16-25.

Xu, P., Huang, H., Dong, N., Wong, S., 2017. Revisiting Crash Spatial Heterogeneity: A Bayesian Spatially Varying Coefficients Approach. Accident Analysis & Prevention 98, 330-337.

Zwerling, C., Peek-Asa, C., Whitten, P., Choi, S.-W., Sprince, N., Jones, M.P., 2005. Fatal Motor Vehicle Crashes in Rural and Urban Areas: Decomposing Rates into Contributing Factors. Injury Prevention 11 (1), 24-28.

Case Study 5

Neighborhood-Level Factors Affecting Seat Belt Use

Amin Mohamadi Hezaveh[a], Christopher R. Cherry[a*]

[a] Civil and Environmental Engineering, University of Tennessee, Knoxville, TN, United States,

* Corresponding Author: Cherry@utk.edu

Abstract

Despite well-known safety benefits of seat belt use, some vehicle occupants still choose to not wear them. This is a challenge in Tennessee, which has a lower seat belt use rate compared to the national average. Road side observations and interviews are two main sources for estimating seat belt use rate; however, they have several limitations (e.g., small sample size, social desirability bias). In this study, we used police crash reports to address these limitations. At aggregate levels, 88.0% of persons over 16 years who were involved in traffic crashes in Tennessee in 2016 properly wore their seat belts. In general, young, male individuals, and rear-seat passengers had lower seat belt use rate. A unique aspect of this study is that home addresses of individuals who were involved in a traffic crashes (N = 542,776) were retrieved from police crash reports. Those addresses were geocoded and assigned to their corresponding census tract revealing added information about spatial distribution of seatbelt use. The average seat belt use rate in the metropolitan area was 88% and for the non-metropolitan area it was 87%. A Tobit model was used to evaluate the relationship between seat belt use rate for both drivers and passengers over 16 years old, with neighborhood sociodemographic variables. Population, age cohorts, race, household vehicles' ownership, household size, and education were among the predictors of the seat belt use rate. Results of this analysis could be used in safety campaigns designed to reach geographic area groups with lower seat belt use rates (i.e., seat belt hotspots).

Keywords: Seat Belt Use Rate; Tennessee; Census Data; Tobit Model; Home Address; Seat Belt Hotspots

Introduction

Approximately 1000 individuals die on Tennessee's roads every year most of them vehicle occupants. One known solution that reduces the fatality rate of the vehicles' occupant is proper use of a seat belt. Several studies have reported the merits of wearing a seat belt in reducing crash fatalities and injury rate. Appropriate use of seat belts increases the chance of vehicle occupants surviving potential fatal crashes by 44% - 73% depending on seating position and the type of vehicles involved in traffic crash (Blincoe *et al.* May 2015). Despite the proven effectiveness of these devices, some individuals still do not wear them.

In Tennessee, seat belt use is a primary law and it is mandatory for all the vehicle occupants to be restrained by a seat belt (i.e., secured shoulder and lap belts) when riding in the front seat of a vehicle. Licensed passengers 16 years old or older are responsible for their own conduct. Nevertheless, a ten-year trend of traffic crashes shows that 30% of Tennesseans who died in traffic crashes failed to wear their seat belt properly at the time of the crash. This rate was 54% and 70% for incapacitating injuries and non-incapacitating injuries, respectively (TITAN 2017). NHTSA reports an 88.9% seat belt use rate, based on direct observation, for the front row passengers in 2016 in Tennessee, which was 1.2% lower than the national average (NHTSA 2017). Road side observations of 27,000 vehicles' occupants at 190 sites revealed that females had a seat belt use rate of 93.8% and males had an 85.0% seat belt use rate. Moreover, freeways showed the highest usage rate (91.2%) of all roadway types, while those observed on local roadways had the lowest usage rate (86.1%) (THSO 2016). In addition, another phone interview in Tennessee reported that 90% of respondents always wore their seat belt, females also had higher seat belt use rate than males (Hezaveh *et al.* 2018a).

Not wearing a seat belt could be attributed to personality traits such as forgetfulness, laziness, perceived low risk of injury, and discomfort (Begg and Langley 2001); attitudes, beliefs, and intentions (Fhaner and Hane 1975, Jonah and Dawson 1982, Chliaoutakis *et al.* 2000, Şimşekoğlu and Lajunen 2008); habits (Knapper *et al.* 1976, Chliaoutakis *et al.* 2000, Calisir and Lehto 2002); and lack of enforcement (Jonah *et al.* 1982, Farmer and Williams 2005). Proper countermeasures could target each of these factors through enforcement and education.

Sociodemographic of those who wear their seat belts less frequently is also helpful for identifying and reaching the group(s) with higher risk. Generally, males have lower seat belt use rate in comparison to females (Preusser *et al.* 1991, Reinfurt *et al.* 1997, Nelson *et al.* 1998, Calisir and Lehto 2002, Wells *et al.* 2002, Glassbrenner *et al.* 2004, Gkritza and Mannering 2008, Pickrell and Ye 2009). This is also the case for younger drivers in comparison to older adults (Reinfurt *et al.* 1997, Calisir and Lehto 2002, Glassbrenner *et al.* 2004). Those with higher education and/or income tend to have higher seat belt use rates (Preusser *et al.* 1991, Reinfurt *et al.* 1997, Wells *et al.* 2002, Houston and Richardson 2005). Studies in the United States have also shown that African-Americans compared with Whites or Hispanics are less likely to use a seat belts (Vivoda *et al.* 2004, Gkritza and Mannering 2008, Pickrell and Ye 2009). Several studies have reported that occupants of pickup trucks have the lowest seat belt use rate compared to occupants of other vehicle types [see, e.g., (Boyle and Vanderwolf 2004, Glassbrenner and Ye 2007, Gkritza and Mannering 2008)].

All of these studies tend to focus on direct observation, self-reported survey responses, or investigation of demographics within crash datasets. Although these methods are easy to conduct and can provide information at low cost; they have their limitations that could negatively affect the analysis. A fundamental limitation of self-reported behavioral questionnaires and interviews is that self-report instruments are often vulnerable to socially desirability bias (Lajunen and Summala 2003, Nordfjærn *et al.* 2015, Hezaveh *et al.* 2017, Hezaveh *et al.* 2018b). In the case of road observations, the amount of data that researcher records are very limited; mainly due to the short amount of time that the observers have to record the data and conspicuity challenges. In roadside observations, usually observed data elements are limited to the vehicle type, number of front row occupants, gender, and age group. Also, the number of observations sites are

usually limited, and they usually take place within daylight or in the nighttime in the areas with sufficient lighting to observe inside of the vehicles.

Police crash reports are the main source of evaluating road safety especially for analyzing crash severity and frequency. However, using police crash reports for studying seat belt use has its limitations. The main limitation is possible incorrect assignment of seat belt use or crash severity to individuals by a responding officer. Some studies reported that "some car occupants who survived a crash may falsely claim to police that they were belted in order to avoid a fine." (Cummings 2002). In contrast, several studies of police reports show that reported seat belt use is consistent with roadside observations (Li *et al.* 1999) and National Accident Sampling System Crashworthiness Data System (CDS) (Schiff and Cummings 2004). On the other hand, using police crash reports have several advantages in comparison to the road side observations and interviews. First, it provides a nearly comprehensive dataset of all serious traffic crashes. Second, it covers a vast geographic area with hundreds of thousands of the observations.

While the literature on road safety provides information about the groups with lower seat belt use rate or other factors affecting seat belt use, it does not provide a solution for matching geographical areas where road users with lower compliance rates live. This is one of the main challenges in designing an effective and geographically targeted safety campaign. To date, most safety campaigns provide blanket coverage of regions with lower seatbelt rates, rather than precise and targeted messaging. Targeted education could be more cost effective at increasing overall seatbelt rates.

This study aims to propose a new method to measure seat belt use rate at the neighborhood level and evaluate the relationship between seat belt use rate and socio-demographic variables based on the home address of the individual (i.e., home-based approach) who were involved in traffic crashes at zonal level. Although some studies used police crash reports to evaluate seat belt effectiveness and seat belt use rate, to the best of our knowledge no studies used this dataset for investigating the relationship between sociodemographic data elements and seat belt use rate based on home-address of individuals involved in traffic crashes (i.e., drivers, passengers). Using the home-address of the individuals in a large database of the traffic crashes enables researchers to identify the geographic and surrounding socioeconomic factors that affect seat belt use and neighborhoods where their residents have lower seat belt use rate. Additionally, we will compare the seat belt use rate extracted from police crash reports with other sources of the seat belt use rate in Tennessee. Our findings are not only limited to the front row occupants but include all the vehicle occupants in different times of the day, context, weather, light conditions, and road types. In the next section, we discuss the proposed database, the geocoding process, and the analytical methods. The rest of the paper presents the results and discusses the findings of this study.

Methodology

Database

The data in this study was provided by Tennessee Integrated Traffic Analysis Network (TITAN), a portal provided by Tennessee Highway Patrol (THP) as a repository for traffic crash and surveillance reports completed by Tennessee law enforcement agencies. The traffic crash records from January 1, 2016, through December 31, 2016, were retrieved from TITAN. Each crash record includes information about road user type (e.g., pedestrian), geographic coordinates of the crashes, addresses of the individuals who were involved in traffic crashes, and dozens of other variables related to the crash. The police crash reports database contained 246,777 crashes and information of 580,767 individuals who were involved in traffic crashes in 2016. Data included different road users' classifications; namely driver, motorcyclist, passenger, pedestrian, bicyclist, person in the building, vehicle's owner, witness, and property owner, including 577,131 vehicle occupants (i.e., driver or passenger); 73% of the vehicles occupant were drivers and the rest were passengers.

Geocoding Process

Bing API was used in this study for geocoding the residential addresses of the individuals. Only those addresses with an accuracy level of premise (e.g., property name, building name), address level accuracy, or intersection level accuracy was used in the study. A sample of addresses were verified by manual inspection. After geocoding the home addresses, we were able to retrieve coordinates of 542,776 individuals (94% success rate) which met address-quality-filter criterion. Among geocoded addresses, 62,741 individuals lived out of state. After controlling for age, vehicles' occupants 16 years old and older were selected for the analysis. Census data from 2010 was also used for obtaining sociodemographic data elements. Table 1 provides a summary of the sample characteristics of the variables considered as an input for model estimation for Tennessee.

Tobit Model

Tobin (1958) proposed the Tobit model or censored regression model. In this model, the regression is obtained by making the mean in the preceding correspond to a classical regression model. The general form of the model which is usually given in terms of index function is as following:

$$y_i^* = x_i'\beta + \varepsilon_{i,}$$

Where y_i^* defined as:

$$y_i^* = \begin{cases} y_i \; if \; a < y_i < b \\ a \; if \; y_i \leq a \\ b \; if \; y_i \geq b \end{cases}$$

ε_i assumes that the error term is normally distributed with mean 0 and variance equals to σ^2. In this study, the seat belt use rate is the dependent variables, and β are the variables presented in Table 1. The dependent variable is a proportion confined between 0 and 1.

Model Performance

Veall and Zimmermann (1996) concluded that Maddala pseudo-r-squared is a valid measurement for evaluating the goodness of fit of censored regression. The general form of Maddala pseudo-r-squared displayed below (Maddala 1986):

$$R^2 = 1 - [e^{LL_{Null} - LL_{Full}}]^{2/N}$$

where, LL_{Null} and LL_{Full} are log likelihoods of the null and full model respectively, and *N* is the number of observations. The likelihood function of the Tobit model is:

$$L = \prod_0 \left[1 - \Phi\left(\frac{\beta X}{\sigma}\right)\right] \prod_1 \sigma^{-1} \phi[(Y_i - \frac{\beta X}{\sigma})]$$

where, Φ is the standard normal distribution function, and ϕ is the standard normal density function (Anastasopoulos et al. 2008).

We also used the Akaike Information Criterion (AIC) as a measure of the relative goodness of fit for identification of the models with a better fit in the sample. AIC is a function of the number of parameters in the model (k) and log-likelihood of the model specification $(\ln(L))$; $AIC = 2k - 2\ln(L)$. As a rule of thumb, a three-point change in an AIC value indicates a significant improvement in the goodness of fit (Bozdogan 1987).

Table 1. Sample statistic for state of Tennessee at census tract

	Mean	Std. Err.	[95% Conf.	Interval]
Total Population	1530.02	12.26	1505.98	1554.06
Area (Square Km)	25.89	0.72	0.93	516.94
Age Cohort Proportion				
16 Years And Younger	0.23	0.00	0.22	0.23
16-42 Years Old	0.32	0.00	0.32	0.33

42-60 Years Old	0.25	0.00	0.24	0.25
60 Years Old and More	0.20	0.00	0.20	0.20
Age Median	39.01	0.13	38.75	39.27
Race Proportion				
% Race White	0.77	0.00	0.76	0.78
% Race Black	0.18	0.00	0.18	0.19
% Race Indian	0.00	0.00	0.00	0.00
% Race Asian	0.01	0.00	0.01	0.01
% Race Hawaiian	0.00	0.00	0.00	0.00
Means Of Transportation To Work				
Personal Vehicle	0.92	0.00	0.92	0.93
Carpool	0.10	0.00	0.10	0.11
Bus	0.01	0.00	0.01	0.01
Motorcycle	0.00	0.00	0.00	0.00
Bicycle	0.00	0.00	0.00	0.00
Walk	0.02	0.00	0.01	0.02
Other Means	25.16	0.10	24.96	25.36
Number Of Children	316.08	3.79	308.65	323.50
% Children	0.20	0.00	0.19	0.20
Household Number	597.47	4.58	588.49	606.44
Household Size	2.73	0.08	2.57	2.89
Education Degree Proportion				
% Educated	1021.62	7.99	1005.96	1037.29
High School and Lower	0.52	0.00	0.51	0.53
Some College Education	0.21	0.00	0.20	0.21
Bachelors' Degree	0.20	0.00	0.19	0.20
Others' Degrees	0.08	0.00	0.07	0.08
Median Household Income ($)	45900	389.89	45200.00	46700.00
Number of Housing Unit	679.38	5.02	669.54	689.21
% Occupied Household	0.88	0.00	0.87	0.88
% Vacant Household	0.12	0.00	0.12	0.12
Household Vehicles' Ownership Proportion				
No Vehicle	0.07	0.00	0.07	0.07
One Vehicle	0.33	0.00	0.33	0.33
Two Vehicles	0.37	0.00	0.37	0.38
Three Or More Vehicles	0.22	0.00	0.22	0.23

Results

Seat Belt Use Rate

Tennessee residents had a higher compliance rate (88.2%) than those with out-of-state addresses (86.9%) (t=8.615, P-value = 0.000). The average age of the male occupants (16 years old and older) was 39.4 (SD = 17.5) and for females was 39.2 (SD = 17.7). In addition, the average age of those who wore seat belts properly (i.e., lap and shoulder) was 39.4 (SD 17.6) and those who did not wear seat belt was 38.8 (SD = 17.1). Moreover, females (89.1%) had a higher seat belt use rate in comparison to the males (87.2%). Table 2 shows more details on the gender, age, and seat belt use rate among the vehicles' occupants. In general, the average age of those who wore seat belts was higher than who did not (t=-8.278, P-value = 0.000).

Table 3 also presents the seat belt distributions of the occupants over 16 years old. The highest seat belt use rate was for front passenger (90.0%) followed by driver (88.4%). Seat belt use rate dropped as passengers seating position row number increased (

Table 3).

Table 2. Age and Gender distribution of the vehicles' occupants over 16 years old

	Female			Male			Total*		
	Mean	SD	Obs	Mean	SD	Obs	Mean	SD	Obs
No seat belt	38.70	17.22	25285	38.76	16.97	32178	38.76	17.09	57708
Wear Seat belt	39.24	17.74	205296	39.52	17.54	220700	39.39	17.64	425999
Total	39.18	17.69	230581	39.42	17.47	252878	39.31	17.58	483707

* Including the unknown observations

Table 3. Seat belt use rate (number of observations) among vehicles' occupants over 16 years old regarding seat position

Row	Left	Middle	Right	Other/Unknown
Front	0.88 (395641)	0.55 (912)	0.89 (66464)	0.2 (55)
Second	0.84 (6647)	0.65 (1101)	0.85 (8913)	0.38 (216)
Third	0.74 (424)	0.67 (143)	0.71 (438)	0.12 (54)
Fourth	0.45 (127)	0 (33)	0.50 (166)	0.04 (128)
Other Seats				0.40 (2203)

Table 4 shows the seat belt use rates under different circumstances. Considering the weather condition, occupants seat belt use rate was higher during the harsh weather, and at its lowest rate during clear weather. Regarding the daylight, occupants wore seat belts at higher rates during the daylight and less during night. Seat belt use rates at night was lower when there was no lighting on the road. Regarding the route signing, Interstate and US routes had a higher seat belt use rate than other route types. In addition, seat belt use rate was lowest on frontage and urban roads.

Seat Belt Hot Spot Identification

Using crash data, 480,035 (7.2% of state population) individual addresses in Tennessee were assigned to their corresponding census tracts. The average seat belt use rate at census tracts for the driver's seat was 0.88 (SD = 0.06) and for the passengers' seat was 0.86 (SD = 0.12); the correlation between driver's seat and passengers' seat was 0.32 (P value < 0.000) which indicated a weak positive linear relationship. Figure 1 and Figure 2 presents the number of observations and the seat belt use rate for drivers and passengers over 16 years old at zonal level. Each point represents a seatbelt use rate in a census tract (number of seat belted drivers or passengers divided by total number of participants in crash) and the number of observations reflects the total number of crashes in each census tract. In cases of driver crashes, the range of rates span 60-100% for tracts with reasonably large crash counts (Figure 1). For passenger seatbelt use rate, there are fewer observations in the dataset, and more observations below 60% seatbelt use rates (Figure 2).

Figure 3 and Figure 4 present average seat belt use rate for the drivers and vehicle passengers at zonal levels in Tennessee; the red color in the figures indicates census tracts with low seat belt use rate and green color indicates a high seat belt use rate. Visual Comparison of the Figure 1 and Figure 2 indicated the passengers' seat belt use rate has more variation in Tennessee compared to drivers.

Table 5 presents the average seat belt use rate in metropolitan areas. Average seat belt use rate for vehicle occupants in the metropolitan area was 88% (SD = 0.06), which was slightly higher than the non-metropolitan area (87% SD = 0.06). Moreover, comparison of the six metropolitan areas in Tennessee indicated that drivers' seat belt use rate was the highest among residents of Knoxville, followed by Jackson and Tri-cities. This trend was also the case for passenger's seat belt use rate. The Chattanooga metropolitan area also had the lowest seat belt use rate for both passenger and driver seat. In addition, Chattanooga was the only metropolitan that passenger seat belt use rate was higher than the driver seat belt use rate.

Table 4. Seat belt use rate regarding weather, lighting, and route signage

Variables	Mean	SD	Number of observation
Weather			
Clear	0.868	0.338	395975
Cloudy	0.889	0.314	58743
Fog	0.868	0.339	1377
Smog/Smoke	0.934	0.249	196
Rain	0.884	0.321	54611

Sleet/Hail	0.895	0.307	1181
Snow	0.909	0.287	4749
Blowing Snow	0.912	0.284	272
Severe Cross-Winds	0.902	0.297	123
Blowing Sand/Soil/Dirt	0.922	0.269	51
Other	0.883	0.321	342
Unknown	0.025	0.157	24318
Lighting	Mean	SD	Count2
Daylight	0.879	0.326	389436
Dark-Not Lighted	0.843	0.364	39391
Dark-Lighted	0.860	0.347	69524
Dark-Unknown Lighting	0.787	0.409	1499
Dawn	0.875	0.330	6821
Dusk	0.864	0.343	10632
Other	0.865	0.342	429
Unknown	0.033	0.106	25044
Route Signage	Mean	SD	Count2
Interstate	0.885	0.319	45397
US Route	0.871	0.335	43581
State Route	0.868	0.338	68086
County Route	0.823	0.382	36707
Municipal Route	0.850	0.357	138721
Frontage Road	0.826	0.379	317
Other	0.789	0.408	14054
Unknown	0.796	0.402	195913

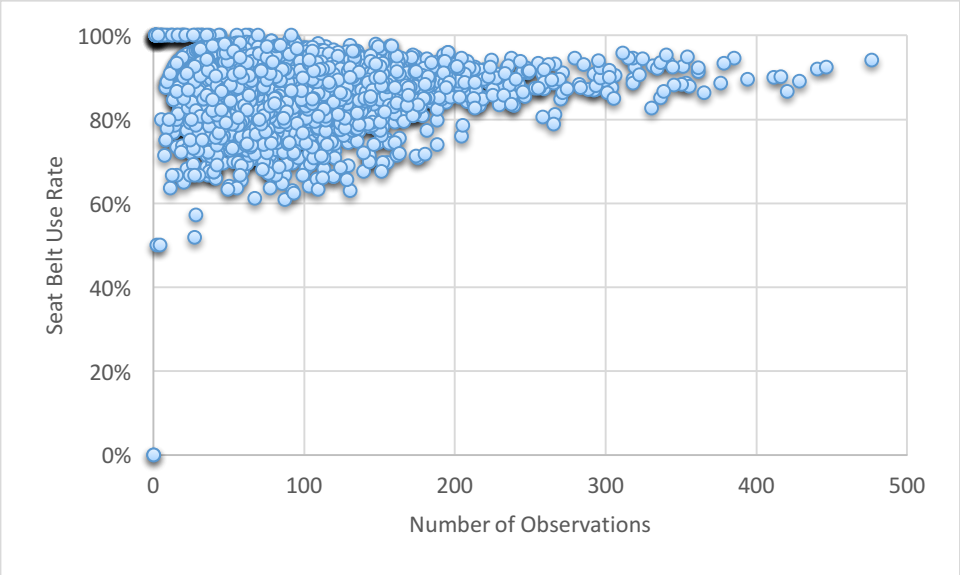

Figure 1. Number of observations and corresponding seat belt use rate at census tract level for drivers

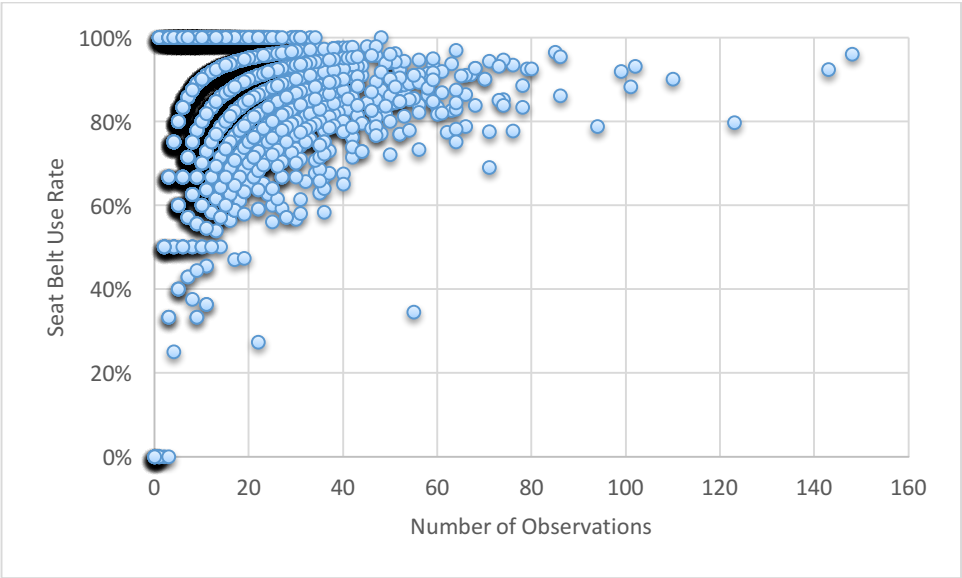

Figure 2. Number of observations and corresponding seat belt use rate at census tract level for passengers (over 16)

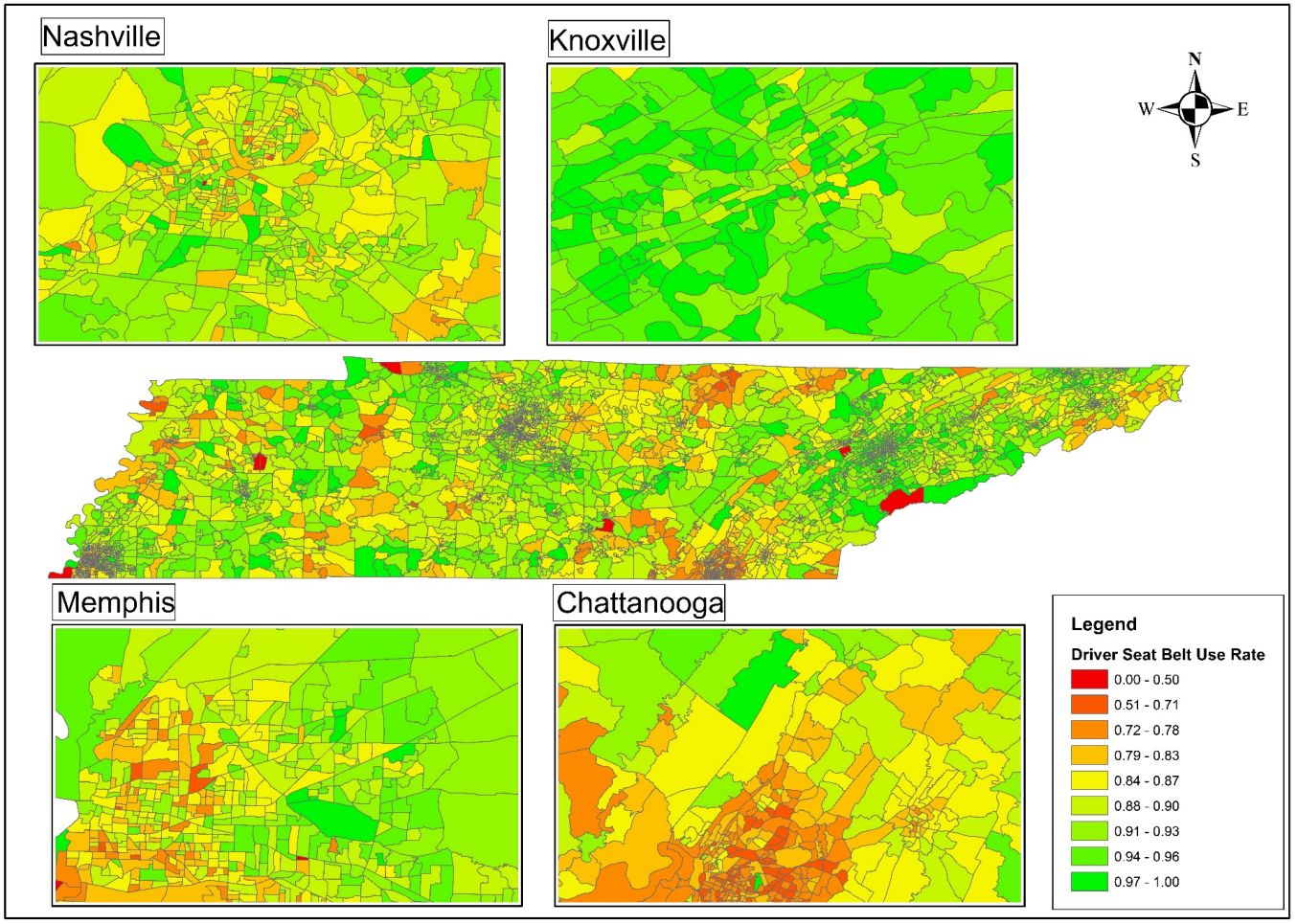

Figure 3 Driver seat belt use rate distribution in Tennessee

Table 5. Mean and Standard Deviation of The Seat Belt Use Rate in Metropolitan Areas

MPO	Driver		Passenger		Overall	
Metropolitan	Mean	SD	Mean	SD	Mean	SD
Knox MPO	0.92	0.04	0.90	0.10	0.91	0.04
Middle TN	0.89	0.05	0.87	0.11	0.88	0.05
Jack	0.90	0.04	0.87	0.11	0.90	0.04
Tri-cities	0.89	0.05	0.88	0.13	0.89	0.05
Chattanooga	0.77	0.07	0.81	0.14	0.77	0.06
Memphis	0.87	0.06	0.83	0.12	0.86	0.06
Non-metropolitan area	0.87	0.06	0.86	0.12	0.87	0.06
Grand Total	0.88	0.06	0.86	0.12	0.87	0.06

Model Estimations

Table 6 presents the results of the Tobit models for predicting seat belt use rate at census tract for drivers' seat belt use rate (DSBUR) model, passengers' seat belt use rate (PSBUR), and their corresponding elasticity values. The chi-square results for all models indicate that both models are significantly different from the null model (DSBUR: $\chi^2 = 328$; PSBUR: $\chi^2 = 233$). The variables that are presented in **Error! Reference source not found.** have a significant correlation with both dependent variables. The mean VIF value for DSBUR model and PSBUR are respectively 1.31 (max = 1.59) and 1.34 (max = 1.71).

Findings of estimated models in Table 6 indicate that population size, the percentage of white race and child percentage at zonal level have a positive association with seat belt use rate in both models. In the DSBUR model, elasticity values indicate that 1% increase in population, child percentage, and portion of white race increase average seat belt use rate by 1.0%, 0.5%, and 0.3% respectively; the corresponding elasticity values for the PSBUR model are higher, 0.8%, 3.7%, and 1.9%, respectively.

Vehicle ownership variables also have a significant association with seat belt use rate; however, the sign of the coefficients are dissimilar in both models. In the DSBUR model, the proportion of household with vehicle (i.e., 0, 1, 2) has a negative association with seat belt use rate. The elasticity values for the proportion of households with one or two vehicles (-2%) is greater than proportion of households with no-vehicles (-0.6%). In the PSBUR model, the proportion of households with one or two vehicles has a positive correlation with passenger seat belt use, whereas proportion of households with no-vehicles has a negative association with passenger seat belt use. The elasticity values for the proportion of families with one or two vehicles is 3% and the corresponding value for households with no-vehicle is -0.3%.

Education-related variable signs are dissimilar in DSBUR model. Percentage of individuals with a college degree has a negative association with drivers' seat belt use rate. On the other hand, the percentage of bachelor degree has a positive association with seat belt use rate in both models. Elasticity values indicate one percent increase in the portion of population with bachelor degree increases seat belt use rate by 0.4% and 1.3%, respectively for DSBUR and PSBUR model.

The metropolitan indicator variable in both models has a positive association with seat belt use rate, which indicates that seat belt use in the metropolitan area is higher than non-metropolitan area. The magnitude and elasticity value of the metropolitan coefficient in the DSBUR model is greater than PSBUR model. Alternatively, the population density variable has a negative correlation with PSBUR variable; the elasticity values indicate that one percent change in population density results in 1.1% reduction in seat belt use rate in the PSBUR model. Household size also has a negative significant association with passenger seat belt use; the elasticity value for this variable is -0.4%.

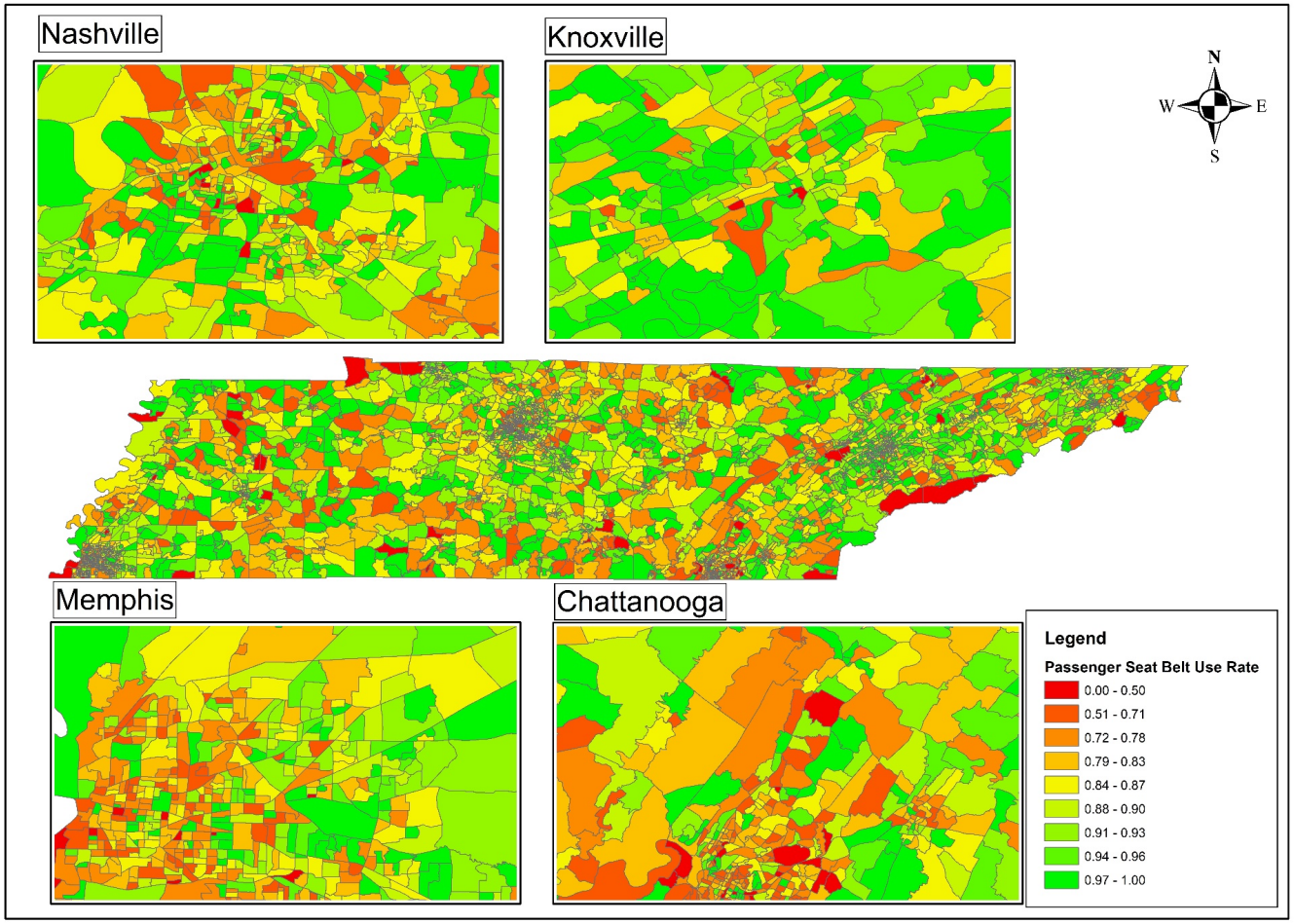

Figure 4 Passengers' seat belt use rate (over 16) distribution in Tennessee

Table 6. Estimated Tobit model for prediction of the seat belt use rate for drivers and passengers

Variable	DSBUR			PSBUR		
	Coef.	Standard Error	Elasticity	Coef.	Standard Error	Elasticity
Population (1,000)	0.006***	0.001	0.010	0.005*	0.003	0.008
% Children	0.023*	0.012	0.005	0.085***	0.024	0.037
% Race White	0.036***	0.004	0.031	0.042***	0.008	0.019
Vehicle Ownership						
% Household with no Vehicle	-0.078***	0.013	-0.006	-0.041*	0.025	-0.003
% Household with One or Two Vehicles	-0.025***	0.008	-0.020	0.036**	0.016	0.029
Education						
% College degree	-0.032**	0.013	-0.007			
% Bachelor Degree	0.016*	0.009	0.004	0.058***	0.018	0.013
Metropolitan Indicator	0.007***	0.002	0.005	0.015***	0.005	0.012
Household Size				-0.001***	0.000	-0.004
Density (1,000 population per square km)				-1.46E-06***	2.24E-07	-0.011
Constant	0.863***	0.008		0.773***	0.016	
Scale parameter	0.004***	7.97E-05		0.014***	3.03E-04	
χ^2	328.37			233.50		
LL_0	5,563.87			2,841.95		
LL_M	5,728.06			2,958.70		
Maddala Pseudo-R^2	0.077			0.056		
N	4,114			4,103		
AIC	-11,436.12			-5,897.41		

* p<.10; ** p<.05; *** p<.01

Source: Authors' analysis of TITAN data and the US Census

Discussion and Conclusion

In this study, we used a police crash database coupled with census data as a source for evaluating the relation between seat belt use rate and sociodemographic variables on a zonal level. Analysis of the vehicles' occupants over 16 years old indicated that seat belt use rate for drivers and the front passenger was approximately 88.2% which was close to roadside observations (88.5%) in Tennessee (CTR 2018). Comparison of the driver and front row passenger seat belt use rate indicated that front row passenger had higher compliance rate which was also in line with the roadside observation in Tennessee (CTR 2018). Generally, the seat belt use rate of the passengers (including back rows) was lower than the driver, which is due to the significantly lower seat belt use rate for passengers in back rows. This lower seat belt use rate for passengers in back rows could be attributed to the current seat belt law in Tennessee which only cover front row passengers.

On average, males and younger individuals were more prone to lower seat belt use rates. Findings also indicated that seat belt use rates were higher in daylight and harsh weather which was consistent with road safety literature (NHTSA 2016). Additionally, the seat belt use rate on interstates was higher than other classes of roads. These findings indicated that police crash reports database is in agreement with roadside observations in Tennessee (CTR 2018) and road safety literature. Variation in seat belt use rate in different circumstances could be attributed to the perception of safety. For instance, those who drive in harsh weather [e.g., rainy or high-speed routes (e.g., interstates)] may perceive more hazard, and as a result, they have a higher seat belt use rate.

Comparison of the driver and passenger seat belt use rate in different metropolitan areas indicated that drivers' seat belt use was higher than other passengers, except for the Chattanooga metropolitan area. Overall, Chattanooga metropolitan area had the lowest seat belt use rate among both metropolitan and non-metropolitan areas. The spatial variation in seat belt use in metropolitan areas reflects different traffic cultures and social psychological factors within Tennessee. Identifying social psychological factors (e.g., attitudes, beliefs, and intentions) that affect seat belt use and using them for educational purposes in safety campaigns would increase seat belt use. Moreover, seat belt use rate distribution map in Tennessee indicated that passenger seat belt use had more spatial variation than driver seat belt use. Spatial variation

in seat belt use could be attributed to both traffic laws in Tennessee and cultural differences which need to be investigated in the future studies.

A Tobit model was used to investigate the association between seat belt use for the drivers and passengers 16 years or older who were involved in traffic crashes and sociodemographic variables of the occupant's home location, at the aggregate level. This is the first time, to the authors' knowledge, that this type of analysis has been conducted. Results indicate that the percentage of white race in the neighborhood had a positive impact on seat belt use rate for both models; this finding parallels previous research (e.g., Gkritza and Mannering 2008, Pickrell and Ye 2009, Bhat et al. 2015). Using a safety campaign in neighborhoods with a high percentage of non-white populations could be used as an effective method for improving seat belt compliance rates. Consistent with other road safety literature, sociodemographic variables have a significant impact on seat belt use rate (Preusser et al. 1991, Reinfurt et al. 1997, Wells et al. 2002, Houston and Richardson 2005). Different neighborhood education levels also have a different effect on seat belt use rate for both models. Percentage of bachelor's degree have a positive impact on the seat belt use rate for both DSBUR and PSBUR models. On the other hand, in the drivers' model, the percentage of a college degree has a negative association with the driver's seat belt use rate. The motive for not wearing a seat belt for lower education driver could be different; perhaps lower seat belt use of drivers with lower-education could be attributed to their subjective norm and attitude toward wearing a seat belt. On the other hand, for higher education and higher income portion of society lower seat belt use rate could be attributed to perceived behavioral control or over-confidence. This may also explain the negative sign of average neighborhood vehicle ownership for the driver seat belt use. Using social psychological tools to investigate how attitudes, beliefs, and values influence seat belt use for different road users would be beneficial for designing a better safety campaign and targeting human factors that predict seat belt use.

Quite the opposite, passengers with higher education and higher vehicle ownership have higher seat belt use rate. This behavior may be attributed to the fact that passengers (particularly front row passengers) have little or no control over the driver's behavior (i.e., perceived behavioral control), and as a result, passengers tend to wear their seat belt more frequently when they are not in the driving position. Psychological factor that affects lower seat belt use rates of the back rows passengers needs to be investigated in the future studies. The home-address environment also has a significant effect on seat belt use rate in both models. Results indicate metropolitan indicator has a positive impact on seat belt use rate for both models. Metropolitan indicator could be used as a surrogate for urban areas, which traditionally have a higher seat belt use rate (NHTSA 2017a).

Future implications

Analysis indicated that using home addresses of the individuals extracted from police crash report was consistent with other source of data for evaluation of seat belt use in road safety literature (e.g., roadside observations). The result of this analysis could be used in the safety campaign design in the programs such as "Click It or Ticket" to efficiently reach neighborhoods with lower seat belt use rate and identify the groups (e.g., lower education, income) that have higher risk. This could be more effective than blanket campaigns that tend to show small population-level effects. There is also a need for developing a methodology that enable researchers and safety practitioners to identify seat belt hotspots. Increase in the enforcement mainly by covering back rows passengers under the primary seat belt use law in Tennessee also could be a practical solution for increasing seat belt use rate of rear seat passengers.

It is also worth mentioning that there are difficulties in accessing the crash data with identifiers and it is not possible to obtain this data in some cases. One possible direction for future researchers could be to develop a methodology to identify seat belt hotspots based on the conventional sources of the data (i.e., temporal and spatial transformation of the models). The sample in this study represents individual who had reported traffic crash in Tennessee in 2016 and careful consideration needed in order to apply the findings to all residents of the state.

Acknowledgments

The authors would like to express gratitude to Tennessee Department of Safety & Homeland Security. This project was supported by the Collaborative Sciences Center for Road Safety (www.roadsafety.unc.edu), a U.S. Department of

Transportation National University Transportation Center promoting safety. The study design was reviewed and approved by the University of Tennessee Institutional Review Board.

References

Begg, D.J., Langley, J.D., 2001. Seat-belt use and related behaviors among young adults. Journal of Safety Research 31 (4), 211-220.

Bhat, G., Beck, L., Bergen, G., Kresnow, M.-J., 2015. Predictors of rear seat belt use among us adults, 2012. Journal of safety research 53, 103-106.

Blincoe, L., Miller, T.R., Zaloshnja, E., Lawrence, B.A., May 2015. The economic and societal impact of motor vehicle crashes, 2010. (revised) (report no. Dot hs 812 013). Washington, dc: National highway traffic safety administration.

Boyle, J., Vanderwolf, P., 2004. 2003 motor vehicle occupant safety survey, volume 2: Safety belt report.

Calisir, F., Lehto, M.R., 2002. Young drivers' decision making and safety belt use. Accident Analysis & Prevention 34 (6), 793-805.

Chliaoutakis, J.E., Gnardellis, C., Drakou, I., Darviri, C., Sboukis, V., 2000. Modelling the factors related to the seatbelt use by the young drivers of athens. Accident Analysis & Prevention 32 (6), 815-825.

CTR, 2018. 2017 survey of safety belt usage in tennessee final report. In: Research, T.U.O.T.C.F.T. ed.

Cummings, P., 2002. Association of seat belt use with death: A comparison of estimates based on data from police and estimates based on data from trained crash investigators. Injury Prevention 8 (4), 338-341.

Farmer, C.M., Williams, A.F., 2005. Effect on fatality risk of changing from secondary to primary seat belt enforcement. Journal of Safety Research 36 (2), 189-194.

Fhaner, G., Hane, M., 1975. Seat belts: Changing usage by changing beliefs. Journal of Applied Psychology 60 (5), 589.

Gkritza, K., Mannering, F.L., 2008. Mixed logit analysis of safety-belt use in single-and multi-occupant vehicles. Accident Analysis & Prevention 40 (2), 443-451.

Glassbrenner, D., Carra, J.S., Nichols, J., 2004. Recent estimates of safety belt use. Journal of safety research 35 (2), 237-244.

Glassbrenner, D., Ye, J., 2007. Seat belt use in 2006—overall results. In: National Highway Traffic Safety Administration, U.S.D.O.T. ed., Washington, DC.

Hezaveh, A.M., Boakye, K., Cherry, C.R., Nambisan, S., 2018a. Factor influencing seat belt use in tennessee: A self-report study. University of Tennessee.

Hezaveh, A.M., Nordfjærn, T., Mamdoohi, A.R., Şimşekoğlu, Ö., 2018b. Predictors of self-reported crashes among iranian drivers: Exploratory analysis of an extended driver behavior questionnaire. PROMET-Traffic&Transportation 30 (1), 35-43.

Hezaveh, A.M., Zavareh, M.F., Cherry, C.R., Nordfjærn, T., 2017. Errors and violations in relation to bicyclists' crash risks: Development of the bicycle rider behavior questionnaire (brbq). Journal of Transport & Health.

Houston, D.J., Richardson, L.E., 2005. Getting americans to buckle up: The efficacy of state seat belt laws. Accident Analysis & Prevention 37 (6), 1114-1120.

Jonah, B.A., Dawson, N.E., 1982. Predicting reported seat belt use from attitudinal and normative factors. Accident Analysis & Prevention 14 (4), 305-310.

Jonah, B.A., Dawson, N.E., Smith, G.A., 1982. Effects of a selective traffic enforcement program on seat belt usage. Journal of Applied Psychology 67 (1), 89.

Knapper, C., Cropley, A., Moore, R., 1976. Attitudinal factors in the non-use of seat belts. Accident Analysis & Prevention 8 (4), 241-246.

Lajunen, T., Summala, H., 2003. Can we trust self-reports of driving? Effects of impression management on driver behaviour questionnaire responses. Transportation Research Part F: Traffic Psychology and Behaviour 6 (2), 97-107.

Li, L., Kim, K., Nitz, L., 1999. Predictors of safety belt use among crash-involved drivers and front seat passengers: Adjusting for over-reporting. Accident Analysis & Prevention 31 (6), 631-638.

Nelson, D.E., Bolen, J., Kresnow, M.-J., 1998. Trends in safety belt use by demographics and by type of state safety belt law, 1987 through 1993. American Journal of Public Health 88 (2), 245-249.

NHTSA, 2016. Traffic safety facts research noteseat belt use in 2015—overall results. In: Analysis, N.S.N.C.F.S.A. ed. NHTSA's National Center for Statistics and Analysis, 1200 New Jersey Avenue SE., Washington, DC 20590.

NHTSA, 2017. Seat belt use in 2018—use rates in the states and territories. In: Analysis, N.S.N.C.F.S.A. ed.

Nordfjærn, T., Hezaveh, A.M., Mamdoohi, A.R., 2015. An analysis of reported driver behaviour in samples of domestic and expatriate iranians. Journal of Risk Research 18 (5), 566-580.

Pickrell, T., Ye, T., 2009. Traffic safety facts (research note): Seat belt use in 2008–demographic results. National Highway Traffic Safety Administration/Department of Transportation, Washington DC.

Preusser, D.F., Williams, A.F., Lund, A.K., 1991. Characteristics of belted and unbelted drivers. Accident Analysis & Prevention 23 (6), 475-482.

Reinfurt, D., Williams, A., Wells, J., Rodgman, E., 1997. Characteristics of drivers not using seat belts in a high belt use state. Journal of Safety Research 27 (4), 209-215.

Schiff, M.A., Cummings, P., 2004. Comparison of reporting of seat belt use by police and crash investigators: Variation in agreement by injury severity. Accident Analysis & Prevention 36 (6), 961-965.

Şimşekoğlu, Ö., Lajunen, T., 2008. Social psychology of seat belt use: A comparison of theory of planned behavior and health belief model. Transportation Research Part F: Traffic Psychology and Behaviour 11 (3), 181-191.

Thso, 2016. Tennessee's average seat belt usage rate soars to 88.95% in 2016. In: Office, T.H.S. ed. Tennessee Highway Safety Office.

Titan, 2017. Tennessee integrated traffic analysis network. Tennessee Highway Patrol

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica: journal of the Econometric Society, 24-36.

Vivoda, J.M., Eby, D.W., Kostyniuk, L.P., 2004. Differences in safety belt use by race. Accident Analysis & Prevention 36 (6), 1105-1109.

Wells, J.K., Williams, A.F., Farmer, C.M., 2002. Seat belt use among african americans, hispanics, and whites. Accident Analysis & Prevention 34 (4), 523-529.

730 Martin Luther King Jr. Blvd.
Suite 300
Chapel Hill, NC 27599-3430
info@roadsafety.unc.edu

www.roadsafety.unc.edu