pennsylvania
DEPARTMENT OF TRANSPORTATION

Pennsylvania Traffic Records Integration Plan

FINAL REPORT

June 26, 2019

By Vikash V. Gayah, S. Ilgin Guler, and Eric T. Donnell
Pennsylvania State University

COMMONWEALTH OF PENNSYLVANIA
DEPARTMENT OF TRANSPORTATION

CONTRACT # 400015622

WORK ORDER # PSU 008

1. Report No.

FHWA-PA-008-PSU WO 008

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle

Pennsylvania Traffic Records Integration Plan

5. Report Date
June 26, 2019

6. Performing Organization Code

7. Author(s)

Vikash V. Gayah, S. Ilgin Guler, Eric T. Donnell

8. Performing Organization Report No.
2019-07

9. Performing Organization Name and Address

Thomas D. Larson Pennsylvania Transportation Institute
Pennsylvania State University
201 Transportation Research Building
University Park, PA 16802

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

4400015622, PSU WO 008

12. Sponsoring Agency Name and Address

The Pennsylvania Department of Transportation
Bureau of Planning and Research
Commonwealth Keystone Building
400 North Street, 6th Floor
Harrisburg, PA 17120-0064

13. Type of Report and Period Covered

September 4, 2018 – July 4, 2019

14. Sponsoring Agency Code

15. Supplementary Notes

Robert Ranieri and David Kelly of the Bureau of Maintenance and Operations served as the project technical advisors. Heather Sorce was the research project manager.

16. Abstract

The objective of this project was to create a Traffic Records Integration Strategic Plan for Pennsylvania that can be used to support PennDOT's strategic safety plan and traffic safety research. The specific data systems considered in this project included crash, vehicle, driver, roadway, citation and adjudication, and injury surveillance. This strategic plan outlines the current state of Pennsylvania's Traffic Records System, provides a detailed integration plan that outlines the barriers and benefits of integrating the associated components, and provides a suggested order in which they should be performed based on their feasibility, value, estimated cost, benefits, and ability to overcome identified barriers.

17. Key Words

Data integration, traffic record integration, safety data

18. Distribution Statement
No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	**20. Security Classif. (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified	Unclassified	102	$92,761.15

ACKNOWLEDGEMENTS

The authors wish to acknowledge Lesley A. Ross, Associate Professor of Human Development and Family Studies, and Daniel Max Crowley, Assistant Professor of Human Development and Family Studies, for their assistance with some of the interviews with agency representatives.

DISCLAIMER

This work was sponsored by the Pennsylvania Department of Transportation. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Commonwealth of Pennsylvania at the time of publication. This report does not constitute a standard, specification, or regulation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AOPC	Administration Office of Pennsylvania Courts
ARNOLD	All Road Network of Linear Reference Data
CMF	Crash Modification Factor
CRS	Core Crash Database
CDART	Crash Data Analysis and Reporting Tool
CODES	Crash Outcome Data Evaluation System
CRN	Crash Record Number
DUA	Data Use Agreement
DMV	Department of Motor Vehicles
DOT	Department of Transportation
DUI	Driving under the Influence
EMS	Emergency Medical Services
FARS	Fatality Analysis Reporting System
HIPAA	Health Insurance Portability and Accountability Act
HPMS	Highway Performance Monitoring System
HSM	Highway Safety Manual
IRB	Institutional Review Board
LF	Liquid Fuel
MOU	Memorandums of Understanding
MSP	Michigan State Police
NGA	National Governor's Association
NHTSA	National Highway Traffic Safety Administration
PCR	Patient Care Report
PCIT	Pennsylvania Crash Information Tool
PA DOH	Pennsylvania Department of Health
PennDOT	Pennsylvania Department of Transportation
PHC4	Pennsylvania Health Care Cost Containment Council
PJN	Pennsylvania Justice Network
PASDA	Pennsylvania Spatial Data Access
PTOS	Pennsylvania Trauma Outcome Study
PTSF	Pennsylvania Trauma Systems Foundation
PII	Personally Identifying Information
RMS	Roadway Management System
SPF	Safety Performance Functions
SSN	Social Security Number
STA	State Transportation Agencies
SHSP	Strategic Highway Safety Plan

SWOT	Strengths, Weaknesses, Threats and Opportunities
TRSP	Traffic Record Strategic Plan
TRCC	Traffic Records Coordinating Committee
TRPAA	Traffic Records Program Assessment Advisory
UConn	University of Connecticut
VIN	Vehicle Identification Number

CHAPTER 1

Introduction

This report provides a Traffic Records Integration Strategic Plan for Pennsylvania, which can be used to support the Pennsylvania Department of Transportation's (PennDOT) strategic highway safety plan (SHSP) and current and future traffic safety research. The strategic plan summarizes the current state of Pennsylvania's Traffic Records System, provides a detailed integration plan that outlines the barriers and benefits of integrating the component data systems, and provides a suggested order in which data integration should be performed based on the feasibility, value, estimated cost, benefits, and ability to overcome identified barriers of each integration possibility.

This report first summarizes information about existing practices and experiences with traffic records integration from published journal articles, research reports, and other state transportation agencies (STAs). In addition, a review of the existing research literature, survey of individual STAs, and phone interviews with key stakeholders from STAs that have the most mature integration plans is provided.

Next, the report provides an overview of the current state of the core data systems that were considered in the strategic plan. These systems include the following:

- Crash data;
- Vehicle data;
- Driver data;
- Roadway data;
- Citation and adjudication data; and
- Injury surveillance data.

The overview includes a description of each of these systems, potential variables that may be used to link the various data systems, data access policies, and potential barriers to data integration, as well as resources and desires for integration of these data systems across the Commonwealth. It also includes a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis that enumerates the strengths, weaknesses, opportunities, and threats associated with each data system and data integration in Pennsylvania.

The report then provides a proposed data integration plan. This includes a summary of how crash data may be integrated or linked with other safety-related data systems, including the additional knowledge that can be gained through integration (e.g., what questions might be answered that previously could not be), barriers that exist to integrating data system pairs, and potential strategies to overcome these barriers. It also includes an assessment of potential data integration frameworks for Pennsylvania safety data, including a recommended order in which data systems should be integrated and the proposed integration frequency.

CHAPTER 2

Background and Current Practices

High-quality and reliable traffic data can be used by STAs to make more informed decisions on the implementation of safety and operational improvements on their roadways. With regards to safety, the National Highway Traffic Safety Administration (NHTSA) has identified six core data systems that might be helpful to highway safety decision-makers:

- Crash data;
- Vehicle data;
- Driver data;
- Roadway data;
- Citation and adjudication data; and
- Injury surveillance data.

The crash database is generally defined as the repository that stores law enforcement officer crash reports (e.g., record of crashes that occur on state roads, including location, vehicles, people involved, and available injury outcomes). The vehicle database provides an inventory of data that enables the titling and registration of each vehicle under the agency's jurisdiction. Information on vehicle make, model, year of manufacture, and body type are often maintained within this database. The driver database is the repository that stores information on licensed drivers within a state and their driving history. The roadway database is the repository that stores geometric and traffic information on state-owned and/or local roadways (e.g., roadway centerline and geometric data, location reference data, geographical information system data, travel and exposure data, etc.). The citation and adjudication databases, which are often managed by multiple state or local agencies, provide information about citations, arrests, and dispositions. Finally, the injury surveillance database consists of multiple data subsystems from pre-hospital medical services, trauma registry, emergency department, hospital discharge, rehabilitation databases, payer-related databases, and mortality (e.g., death certificates, autopsies, and coroner and medical examiner reports).

Data for each of these core data systems are generally collected and maintained by separate agencies/groups. This separation is a significant barrier to gaining a complete understanding of highway safety performance, which prevents improving upon the status quo by identifying and implementing the most beneficial and effective safety improvements. There are numerous challenges associated with linking these individual data systems, including high costs, legislative restrictions, potential liabilities, lack of well-defined linkage variables, data ownership, accessibility, and data storage/maintenance. Additionally, the quality of each dataset (e.g., accuracy and completeness of each record) can also be a barrier to linking the individual data systems. If the data to be linked are not accurate or complete, the resulting integrated dataset would have less value for analysis (National Highway Traffic Safety Administration, 2018a).

Overcoming these barriers facilitates the linking of individual data elements across the various data systems. Two different types of data linkages are possible: (1) providing a system interface or (2) data integration.

A system interface implies that the linkages are performed in close to real-time, while data integration often takes place at regularly scheduled points in time, such as annually (National Highway Traffic Safety Administration, 2018a). Most of the discussion below pertains to datasets that are integrated at regular intervals. According to NHTSA, an integrated traffic records system would be similar to that shown in Figure 1.

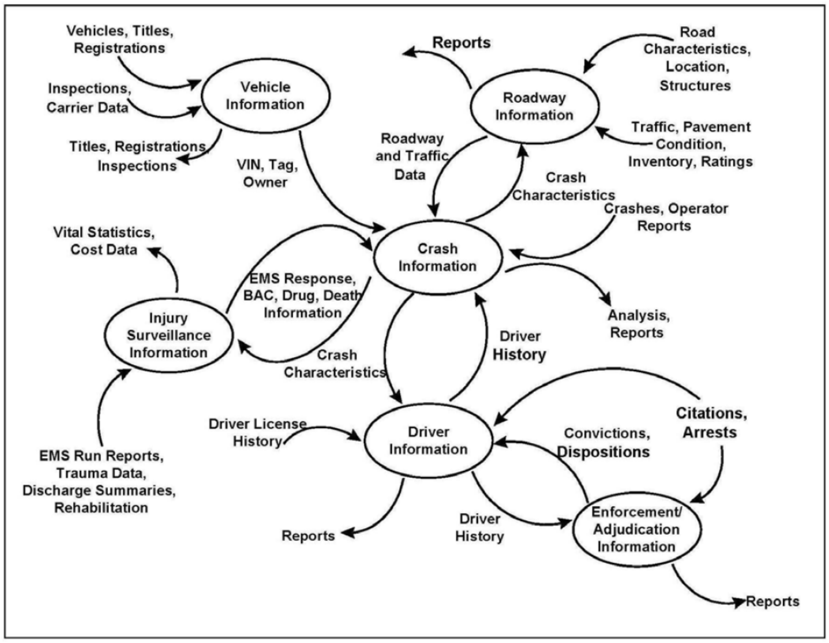

Vehicles, Titles, Registrations
Inspections, Carrier Data
Vehicle Information
Titles, Registrations Inspections
VIN, Tag, Owner
Reports
Roadway Information
Road Characteristics, Location, Structures
Traffic, Pavement Condition, Inventory, Ratings
Roadway and Traffic Data
Crash Characteristics
Crashes, Operator Reports
Vital Statistics, Cost Data
Crash Information
EMS Response, BAC, Drug, Death Information
Injury Surveillance Information
Crash Characteristics
Driver History
Analysis, Reports
Citations, Arrests
Driver License History
Driver Information
Convictions, Dispositions
EMS Run Reports, Trauma Data, Discharge Summaries, Rehabilitation
Reports
Driver History
Enforcement/ Adjudication Information
Reports

Figure 1. Integrated Traffic Records System (Maryland Traffic Records Coordination Committee, 2016)

There are numerous benefits that can be obtained by linking different core data systems. For one, analysis of the data linked across multiple databases can provide a deeper understanding in relation to crash circumstances and injury outcomes (Conderino et al., 2017). Additionally, linked datasets can reduce the redundancy in collecting similar data in different datasets, which can save costs in the long run. The NHTSA *Traffic Records Program Assessment Advisory* lists other general benefits of integrated data as:

1. Lower costs to achieve desired level of data content and availability;
2. Support for multiple perspectives in data analysis and decision-making;
3. Expanded opportunities for data quality validation and error correction;
4. Additional options for exposure data to form raters and ratio-based comparisons;
5. Enhanced accuracy and completeness of data;
6. Increased relevance of information available for legislative and policy analysis; and
7. Increased support for advanced methods of problem identification, countermeasure selection, and evaluation of program effectiveness.

Additionally, more reliable injury outcome information can be obtained through data records integration (Lopez et al., 2000). More precise target areas for countermeasures and interventions can also be identified with more robust databases. Similarly, analyses can be conducted that highlight the cost of injuries, determine more effective allocation of law enforcement resources, and identify specific roadway features that contribute to crash risk (National Highway Traffic Safety Administration, 2018a). Benefits other than those related directly to road countermeasures can also be obtained. Motor vehicle agencies can use linked data to better assess the effectiveness of granting, suspending, and revoking driver's licenses (National Highway Traffic Safety Administration, 2014). Linked datasets can also be used to improve trauma management in an effort to reduce fatalities associated with traffic crashes (Lopez et al., 2000). Having a more complete and comprehensive database can improve decision-making by more accurately identifying problems, selecting more appropriate countermeasures, and more effectively evaluating programs (Connecticut Traffic Records Coordinating Committee, 2018).

To help states better understand the benefits of data integration and to overcome some of these basic challenges, the Moving Ahead for Progress in the 21st Century Act (MAP-21) provides national funding to STAs for State Traffic Safety Information System Improvement through grants administered by the NHTSA (National Highway Traffic Safety Administration, 2018b). To be eligible for these grants, STAs need to meet the following criteria:

- Have a Traffic Records Coordinating Committee (TRCC) that meets at least three times per year;
- Have a designated TRCC leader;
- Have established a state traffic record strategic plan (TRSP) approved by the TRCC;
- Have demonstrated quantitative progress in relation to the significant data program attributes of:
    - Timeliness;
    - Accuracy;
    - Completeness;
    - Uniformity;
    - Accessibility; and
    - Integration of a core highway safety database; and
- Have certified that an assessment of the existing data and traffic records system was conducted or updated during the last 5 years.

The TRCC is formed in every state to coordinate maintenance and integration of these different datasets, along with planning future projects to aid in improving the traffic safety information system. The goal of this committee is to develop and manage the state's traffic records systems. The TRCC chair provides leadership for committee activities. The TRCC is also responsible for approving the TRSP, which aims to identify the current gaps in the integrated traffic records plans and set goals for the following years. In 2018, all STAs met these criteria and received grants for improvement of State Traffic Safety Information Systems, except for the U.S. Virgin Islands.

The NHTSA provides a Traffic Records Program Assessment Advisory (TRPAA) to help STAs develop an effective traffic records system and to quantify performance (National Highway Traffic Safety Administration, 2018a). The TRPAA describes the ideal system that should be used to record traffic data and provides detailed guidelines on how to collect these data, ensure the quality of the data, and manage the six core datasets. As a part of the TRPAA, 391 assessment questions are provided to help each state evaluate their traffic records data. These questions are ranked as very important or somewhat important, and evidence that would suggest if the state meets the described criteria is provided for clarity. These 391 questions seek to assess the effectiveness of:

- The traffic records coordinating committee;
- Strategic planning for traffic records systems;
- Description and contents of different data systems;
- Applicable guidelines for each data system;
- Data dictionary for each data system;

- Procedures and process flows for each data system;
- Interface of each data system with other components; and
- Data quality control programs for each data system

Within the data quality control program for each data system, there are six core performance attributes that are required, including:

- Timeliness: the span of time between the occurrence of an event and the entry of the information into the appropriate database;
- Accuracy: minimizing the errors in the database and removal of duplicate entries;
- Completeness: both the number of missing records from a database and the number of missing information elements from each record;
- Uniformity: the consistency among records in a database which can be measured against some standard, preferably a national standard;
- Accessibility: the ability to obtain desired data; and
- Integration: the ability of a records system to be linked to another database.

Different performance measures are provided in the TRPAA for each of these attributes. These performance attributes are discussed in further detail in a document titled *Model Performance Measures for State Traffic Records Systems* (National Highway Traffic Safety Administration, 2011).

LINKAGE OF DATA SYSTEMS

One of the primary challenges to a well-integrated traffic records data system is linking the individual data systems. Linkage between these systems can be done directly, manually, deterministically, or probabilistically. Direct linkage can be performed when there is a single unique identifier and the databases are merged seamlessly through this information. In this case, the linked database can be viewed as a single database even if it is actually made up of multiple systems. Manual data linkage is done without the aid of a computer by comparing different records to each other using a single unique identifier. Deterministic data linkage is conducted by using several unique identifiers common to the different datasets. Multiple unique identifiers are used so that there are no errors when records are matched. However, if matching rules are strict (e.g., if they require all fields to match or use a weighted system), this can lead to missing data points that should be matched. As an alternative, probabilistic record linkage is applied. This method relies on common identifiers between different datasets, but does not require that these identifiers be unique. A probability weight is assigned to each data point to identify possible matches. Records that have a high probability of being matched are linked, while those with low probabilities are assumed to be independent and not linked together. The primary drawback of probabilistic data linkage is the potential for many false positive matches, but this comes with the benefit of allowing more records to be linked together than deterministic approaches. The TRPAA provides general guidelines on the common links that can be used between the crash database and other databases; see Table 1. In this table, common identifiers that can be used to link each database to the crash database are shaded in grey. Note that several identifiers that were identified by the research team's survey of state agencies were not listed here, including crash date, crash time, and driver gender.

Table 1. Common identifiers between crash data and other core datasets (recreated and redesigned from National Highway Traffic Safety Administration (2018a))

	Name of Database				
Identifier	**Driver**	**Vehicle**	**Roadway**	**Citation and Adjudication**	**Injury Surveillance**
Full name	▨			▨	▨
Date of birth	▨			▨	▨
Address	▨			▨	▨
Driver's license #	▨			▨	
Photo match	▨			▨	
Vehicle make		▨			
Vehicle model		▨			
Vehicle year		▨			
License plate #		▨			
Vehicle Identification Number (VIN)		▨			
Location of crash			▨		▨
EMS run report #					▨
Unique patient ID #					▨

Within the United States, most traffic records data linkage has been performed in a probabilistic manner as part of the Crash Outcome Data Evaluation System (CODES) program. NHTSA created the CODES program in 1992 to develop and coordinate data linkage and to develop a consistent probabilistic linkage algorithm that could be applied nationally. CODES specifically provides a standardized methodology to probabilistically link police–reported crashes to hospital data to provide information about injuries incurred during collision. A sample of variables used by the Nebraska Department of Transportation (DOT) to link Emergency Medical Services (EMS), emergency department, hospital discharge, trauma registry, and Vital Statistics data to crash data using the CODES methodology is shown in Table 2. As shown, the date of birth, crash date and time, and county in which the crash occurred are common variables that exist in different hospital databases that can be linked to crash data. While these data are specific enough that matches will be found, they are not unique and hence require probabilistic linkages.

Table 2. Variables to link CODES datasets (recreated from Nebraska's Traffic Records Coordinating Committee, (2018))

Crash	EMS	Emergency Department	Hospital Discharge	Trauma Registry	Vital Statistics
First Name	X			X	X
Last Name	X			X	X
Date of Birth	X	X	X	X	X
Crash Date	X	X	X	X	X
Crash Time	X			X	X
Crash County	X	X (hospital county)	X (hospital county)	X	X (county of death)

Initially, seven states were awarded grants through CODES to perform data linkage; however, the following states currently participate in the CODES program:

- Connecticut;
- Delaware;
- Georgia;
- Illinois;
- Kentucky;
- Maryland;
- Maine;
- Minnesota;
- Missouri;
- Nebraska;
- New York;
- Ohio;
- South Carolina;
- Utah; and
- Virginia

A review of the academic research literature reveals that data linkage has been used for many different purposes. The most commonly linked datasets are crash databases and hospital/EMS data. The addition of hospital/EMS data significantly enriches datasets in terms of detailing the severity of injuries and providing additional demographic information. Internationally, linking crash and hospital data has been predominantly done in Australia (Boufous et al., 2008; Boufous and Williamson, 2006; Cercarelli et al., 1996; Lopez et al., 2000; Lujic et al., 2008; Mitchell and Newman, 2002; Watson et al., 2015), New Zealand (Alsop and Langley, 2001; J. Langley et al., 2003; J. D. Langley et al., 2003; Wilson et al., 2012), and Europe (Abay, 2015; Amoros et al., 2006; Aptel et al., 1999; Cryer et al., 2001). Most of these studies have focused on the issue of under-reporting injury outcomes in police crash records, often specifically focusing on bicyclists and pedestrians (Abay, 2015; Alsop and Langley, 2001; Amoros et al., 2006; Aptel et al., 1999; Boufous et al., 2008; Cryer et al., 2001; J. D. Langley et al., 2003; Lujic et al., 2008; Rosman and Knuiman, 1994; Watson et al., 2015). Additionally, many studies focused on motorcycle-related crashes and tried to identify their characteristics, often specifically focusing on helmet use (Cook et al., 2009; Daniello and Gabler, 2012). Other studies considered vehicle crashes involving older drivers (Cook et al., 2000), analyzing the accuracy of injury ratings on police crash reports (Burdett et al., 2015), understanding the unique characteristics of rural area traffic crashes (Clark et al., 2013), impacts of seatbelt use (Han et al., 2017, 2015), and impacts of changes in the maximum speed limit (Vernon et al., 2004). Still other studies simply provided methodological insights on how to link data and insights into the current status quo of public health (Burch et al., 2014; Clark, 1993; Conderino et al., 2017; Mcglincy, 2004). The linkage between the datasets in the above-mentioned studies was either done manually (Cryer et al., 2001; Rosman and Knuiman, 1994), in a deterministic manner (Clark, 1993; Watson et al., 2015), or in a probabilistic manner (Boufous and Williamson, 2006; Burch et al., 2014; Burdett et al., 2015; Clark et al., 2013; Conderino et al., 2017; Cook et al., 2009, 2000; Daniello and Gabler, 2012; Han et al., 2017, 2015; Lujic et al., 2008; Rosman, 2001; Watson et al., 2015; Wilson et al., 2012).

All of the above-mentioned studies utilized a crash database based on police crash reports. However, different forms of the medical dataset have been utilized, including: hospital admission data (Cryer et al., 2001; Han et al., 2015; Kuhl et al., 1995), hospital discharge data (Conderino et al., 2017; Han et al., 2015; Lopez et al., 2000; Lujic et al., 2008; Rosman, 2001; Rosman and Knuiman, 1994; Wilson et al., 2012), emergency department information system data (Conderino et al., 2017; Han et al., 2015; Watson et al., 2015), ambulance reports (Clark, 1993), hospital trauma registries (Clark, 1993; Watson et al., 2015), and hospital death records (Lopez et al., 2000; Rosman, 2001). A few studies merged non-medical datasets to crash data records. One study used worker compensation data, which includes all claims for injury or disease that resulted in death for which payments were made, to study work-related traffic crashes (Boufous and Williamson, 2006). Another study utilized Utah Department of Transportation data to analyze the safety impacts associated with the change in the national maximum speed limit law. In this study, traffic volumes obtained from traffic statistics data, along with speed limits obtained from the roadway database, were merged with police crash reports to conduct this analysis (Vernon et al., 2004). Finally, the Fatality Analysis Reporting System (FARS) database, which is an annual census of all traffic crashes that result in a fatality within 30 days of the reported incident, and the National Automotive Sampling System – Crashworthiness Data System database, a sample of traffic crashes resulting in vehicle damage that require towing, were merged to understand the differences between urban and rural crash mortality rates (Clark et al., 2013).

SURVEY OF STAs

In addition to reviewing the academic literature on data linkages, the research team also performed a short survey of all STAs to learn more about their specific practices and to identify additional relevant literature from STAs to review. The survey was organized into the following general sections:

- State represented;
- Organization of six core data systems and identifiers that can be used to link individual data systems;
- Current integration practices, challenges to integration, and best practices to overcome challenges;
- Availability of documentation on data records system and integration plans; and
- Contact information for follow-up interviews.

The survey was hosted through Penn State and Qualtrics, an online survey software. It was distributed to the TRCC chair and coordinator (if these differed) from each STA identified as of January 2017; a complete list of all individual contacted and associated contact information can be found at the following weblink https://www.transportation.gov/sites/dot.gov/files/docs/resources/government/traffic-records/26691/20180602-trcc-chairs-and-coordinators.pdf.

A complete list of the survey questions and summary of all relevant responses can be found in Appendix A. The remainder of this section provides a brief summary of the most pertinent findings.

Responses were received from 36 individuals out of the 95 individuals contacted as a part of this survey effort, which represented the following 30 unique STAs:

- Colorado
- Connecticut
- Delaware
- Guam
- Hawaii
- Louisiana
- Maine
- Maryland
- Massachusetts
- Michigan
- Minnesota
- Missouri
- Montana
- Nevada
- New Jersey
- New York
- North Carolina
- North Dakota
- Ohio
- Oklahoma
- Rhode Island
- South Carolina
- South Dakota
- Tennessee
- Utah
- Vermont
- Virginia
- Washington
- Wisconsin
- Wyoming

All of the responding states indicated that the core data systems are maintained within the state in an electronic database, which greatly facilitates data integration. A summary of the core data systems that have been merged, or have been considered to be merged, with the crash data system for each state that responded to the survey is shown in Table 3. In this table, the grey shaded cells indicate datasets that have been merged or considered to be merged with the crash data. As shown, only the following states have fully integrated or considered fully integrating the six core datasets: Colorado, Maryland, North Carolina, and Wyoming. The most common databases integrated with crash data include the EMS/injury surveillance and roadway databases. Vehicle databases are the least commonly integrated with crash data.

Table 3. Core data system components integrated or considered to be integrated with crash data by state

State	Vehicle	Driver	Roadway	Citation and Adjudication	EMS/Injury Surveillance
Colorado	■	■	■	■	■
Connecticut			■	■	■
Delaware				■	
Hawaii					
Louisiana	■	■	■		■
Maine					■
Maryland	■	■	■	■	■
Massachusetts					
Michigan	■				
Minnesota					
Missouri			■		
Montana			■		■
New York State					
New Jersey			■		■
North Carolina	■	■	■	■	■
Nevada		■	■		■
Ohio					
Oklahoma				■	
Rhode Island		■	■	■	■
South Carolina					
South Dakota	■	■		■	■
Tennessee		■	■		■
Utah			■	■	■
Vermont	■	■			
Virginia		■	■		■
Washington		■	■	■	■
Wyoming	■	■	■	■	■

Additionally, the variables used to link the crash data to the other datasets are provided in Table 4. In this table, the number in each cell is the number of agencies who responded with that identifier. For example, two states have used or considered using the full name when merging driver and crash databases. As shown in Table 4, agencies have different practices in merging their databases, mostly depending on the availability of data. However, merging the roadway database with the crash database is consistently done using the location of the crash for all STAs.

Table 4. Common identifiers between crash data and other core datasets based on survey results

	Name of Database				
Identifier	**Driver**	**Vehicle**	**Roadway**	**Citation and Adjudication**	**Injury Surveillance**
Full name	2	1		1	4
Date of birth	2	2		1	4
Gender					2
Address	1	1		1	
Driver License #	7	3		4	2
License plate #		1			
VIN	1	4			
Date of Crash		1			6
Time of Crash		1			2
Location of Crash			19	2	5
Crash ID #	1			3	
EMS run report #		1			1

Additionally, state representatives identified some challenges for integrating each core dataset with the crash datasets. The overarching themes for all datasets were the lack of communication between agencies and the difficulty in being able to converge on what would be shared and how. The quality or accuracy of data was listed as one of the major challenges in merging different datasets together. The difficulty in determining how to merge different datasets, in terms of merging different database types or lack of unique identifier, were given as some of the major barriers to integration. Another barrier was listed as limited resources. Finally, security of the data and the challenges of personal identifiers were also given as both a challenge and a barrier to merging datasets.

CURRENT STA PRACTICES

The research team reviewed TRSPs and other published documents on data records and integration practices available online or provided by the STAs through the survey. Many TRSPs are updated annually and present performance measures based on the TRPAA. The TRSPs of the following states were available and reviewed as a part of this summary:

- Alaska (Alaska Traffic Records Coordinating Committee, 2015);
- Colorado (Cambridge Systematics, 2018a);
- Connecticut (Connecticut Traffic Records Coordinating Committee, 2018);
- Florida (Cambridge Systematics, 2017);
- Georgia (Georgia Traffic Records Coordinating Committee, 2016);
- Idaho (Idaho Traffic Records Coordinating Committee, 2015);
- Illinois (National Highway Traffic Safety Administration Technical Assessment Team, 2016);
- Kansas (State of Kansas Traffic Records Coordinating Committee, 2015);
- Kentucky (University of Kentucky Transportation Center, 2017);
- Maryland (Maryland Traffic Records Coordination Committee, 2016);

- Michigan (Michigan Traffic Records Coordinating Committee, 2006);
- Montana (Montana Department of Transportation, 2017);
- Nebraska (Nebraska Traffic Records Coordinating Committee, 2018);
- New Mexico (New Mexico Department of Transportation, 2016);
- North Carolina (Tennyson, 2016);
- Oregon (Cambridge Systematics, 2018b);
- Pennsylvania (Pennsylvania Traffic Records Coordinating Committee, 2018);
- Texas (Texas Department of Transportation, 2012);
- Utah (Utah Traffic Records Advisory Committee, 2015); and
- Washington (Washington Traffic Records Committee, 2017).

Many of these agencies conducted a SWOT analysis to understand their current traffic records situation and to identify strategies for improvement, including anticipated benefits. In general, the most common goals of state TRCCs were to improve automated crash reporting and to improve data system linkages. Additional goals were to improve the accuracy, completeness, and uniformity of traffic records data.

Specifically, integrating crash and injury surveillance data is expected to help quantify the severity and overall cost of a crash (including average costs by severity level), as well as the long-term outcomes associated with any resulting injuries. Some states provided statistics on the percentage of criteria for integration mentioned within the TRPAA that are met within their TRSPs. These values are as follows:

- Colorado: 15.4%
- Florida: 46.2%
- Illinois: 15.4%
- Oregon: 15.4%

An overview of other notable findings from existing TRSPs or data integration plans is summarized below:

- Colorado’s TRSP identifies integration of data systems to fully utilize existing datasets as an ideal toward which to strive. This document mentions that most of the data are in individual data systems. The crash and roadway datasets can be linked by manual incorporation of the crash database into the linear referencing system database used for roadway data.
- Connecticut DOT’s TRSP from 2018 specifically focuses on the integration of citation and adjudication data with crash data.
- Florida’s 2016 TRSP mentions that their citation and adjudication data, driver data, vehicle data, and Florida Highway Patrol data have been successfully integrated into a single database. However, roadway and crash data or injury surveillance and crash data have not been linked.
- Idaho’s TRSP indicates that integrating the driver and vehicle databases with the adjudication data is a core priority.
- Illinois’s TRSP indicates that, since 2006, Illinois has been integrating crash data with hospital discharge data as a part of the CODES project despite the lack of a unique identifier between these two datasets. The emergency department data have been included in this since 2009. It is noted that errors in driver license data within the crash file make it difficult to merge the crash and driver databases.
- Michigan’s TRSP identifies two funded projects on integrating the driver and vehicle databases, and integrating pre-hospital, trauma, and crash data.
- Nebraska’s 2016 TRSP includes integrating vehicle records and roadway information with the crash data as projects planned for the future. Nebraska has integrated its crash database with hospital discharge, EMS, and Vital Statistics records for years 2008 through 2014.

- New Mexico's TRSP states that driver and vehicle databases have been integrated since 2016. An extension to this is listed as creating electronic traffic citations to be integrated into the driver records.
- Pennsylvania's TRSP specifically states developing a single statewide traffic records inventory as a consideration. An EMS agency code was added to the police crash reports to aid in the integration of the EMS and Trauma records into the crash database. Additionally, the driver and vehicle systems are being integrated into a single system.

INTERVIEWS WITH OTHER DOTs

As a result of the survey, several STAs were identified as having practices and/or experiences that might be relevant to this project. This included STAs that indicated a high level of safety data integration. Representatives from these STAs were contacted for phone interviews; the individuals from these STAs were those that responded to the survey. Representatives from Colorado, Connecticut, Maryland, Michigan, Nebraska, North Carolina, Utah, and Wyoming were interviewed. All contacts served as chair or co-chair of the TRCC for that state. Three of the interviewees were DOT representatives, three interviewees worked at University Safety Centers and were contracted to perform the data integration on behalf of the state DOT, and one of the interviewees was at the Michigan State Police. Additionally, a short interview with the Nebraska Department of Health and Human Services was also conducted to learn more information about integration of crash data with Vital Statistics. The remainder of this section provides a summary of the phone interviews. This summary includes the current state of data integration for these STAs, as well as the challenges, barriers, and benefits of data integration. Note that one common occurrence from all interviews was that most of the TRCC contacts were not comfortable providing cost estimates since these costs are often experienced by multiple different entities. Only Michigan was able to provide a cost estimate of $2.5 million for its data integration efforts (funded by NHTSA 405c funds, which are for state traffic safety information system improvements).

Colorado

Colorado first considered data integration in 2009 when a project to determine a data integration plan was conducted. The estimated cost for a traffic records data warehouse was determined to be $6.5 million. Due to the large cost, the Colorado DOT did not continue with the proposed approach at the time and instead has considered variations to housing all of the data in one place. Hence, Colorado has since moved on to considering a virtual data sharing approach.

In Colorado, the Department of Revenue officially houses the crash, driver, and vehicle databases. The DOT receives a periodic download of only the crash database. Currently, the Department of Revenue is updating its database system to be able to integrate the crash, driver, and vehicle databases. However, the DOT would not be given direct access to this database due to privacy concerns. Additionally, different software systems often prevent agencies from sharing information with each other.

The Colorado DOT is working to obtain a statewide electronic crash reports and citation system. However, obtaining a statewide citation system is difficult since there are many different court systems within the state.

The Colorado Department of Health conducted a study in 2010 that linked crash and EMS databases using name and dates of birth; however, only 60% of all data were able to be linked due to errors in the databases, and this linkage was only a one-time activity. These two databases have not been integrated since and one of the major barriers is privacy concerns. To overcome this barrier, the Colorado DOT pays the salary of a data analyst working for the Department of Health who can update the health database when needed.
The Colorado DOT has found it useful to sign memorandums of understanding (MOUs) with different stakeholders to share data. They currently have MOUs established with the health department and the State Patrol.

Connecticut

Connecticut DOT was a part of a data integration project sponsored by the National Governors Association (NGA), along with seven other states. This project completed a planning phase early in 2018 and a 6-month implementation phase was concluded in September 2018. As a part of this project they have started integrating, or considered integrating, all six core datasets.

The University of Connecticut (UConn) houses the electronic crash data for the Connecticut DOT in an SQL database (Connecticut Crash Data Repository). As part of the NGA project, UConn has already integrated the toxicology dataset and a portion of the roadway, judicial, and EMS databases with the crash dataset. Additionally, UConn is currently working to integrate the driver and vehicle databases with the crash database. In all cases, the first step was to identify the owners of all of these data systems and make arrangements to access the data. For example, an MOU was established and 10 years' worth of data were obtained to obtain citation information for individuals who pleaded guilty to the citations. To update the crash database in real time, the citation data is received monthly on a CD (in a .csv format) and deterministically linked to the crash database. UConn is still working on obtaining all citation information (i.e., not only those citations for those who pled guilty). The main benefit of obtaining judicial data has been to analyze Driving under the Influence (DUI).

From a public health perspective, 2015 information for trauma center data and EMS data has been obtained as a part of a pilot project. However, no MOU has been established to continue providing access to these data. Additionally, the state toxicology lab has provided the toxicology report for every fatal accident for the last 4 years. This is provided in paper format and has to be scanned and transcribed into the database manually. The main benefit of obtaining this information is to be able to understand the impacts of different drug combinations.

Part of the roadway database has also been linked to the crash database, including traffic volumes, number of lanes, lane width, and horizontal and vertical curvature information. However, Connecticut DOT is in the process of updating its roadway database to a Transportation Enterprise Database. Once this update is complete, all roadway information will be linked to the crash data.

UConn has had limited success working with the Department of Motor Vehicles (DMV) to obtain driver license and vehicle information. Some of the issues identified have been the old data system of the DMV (data are currently housed on a mainframe) and the difficulty of extracting these data in an efficient way. Some other concerns relate to privacy issues. UConn is looking to reengage the DMV by demonstrating the benefits of integrated data.

The main challenges with data integration have been getting agencies to talk to each other and learning details about the different datasets. Stakeholders are sometimes not willing to share their databases, since it could lead to misinterpretation of their data due to how the database is structured. This is specifically true for the judicial system. Hence, it is important to take the time to understand each database and why it is structured the way it is. Additional challenges include different vendors for electronic crash data and getting all of them to conform to the same forms, as well as merging the different data formats together.

To help overcome some of these challenges, it is important to show how the integrated database could be used. The integration of the crash database with other systems is currently being used as an example to convince other agencies to share their data. The most important step has been to show which questions can be answered if an integrated database exists and to highlight that these questions could be answered quickly. Some example questions that Connecticut DOT uses to highlight the importance of data integration include:

- How effective is the current DUI program at reducing and preventing DUI crashes?
- How many DUI crashes involve a person that has multiple DUIs?
- How do injury classifications at the scene of a crash by the office compare to the actual injuries treated at the hospital?
- How effective is the graduate driver's license program in reducing teen crashes and are those drivers who wait until the age of 18 to obtain a driver's license at a greater risk to crash due to reduced driving experience?
- What trends are being seen in drunk or drugged driving in terms of types of drugs and quantities found in an individual's system?

Additionally, it is important to actively include all parties in the TRCC process. To overcome the privacy concerns, UConn has spent significant effort in securing their SQL database using firewalls and multiple layers of security.

Maryland

Maryland data integration began initially as a part of the CODES program. Maryland was one of the first implementers of CODES and started by merging the crash database with the hospital data maintained by Maryland Health Services and the EMS data. One advantage that Maryland has is that there are statewide crash and hospital data available. The initial motivation to join databases was to quantify the benefits of seatbelt and helmet use for NHTSA. The main challenge with merging the crash and hospital/EMS data was the lack of unique identifiers. Hence, they used a probabilistic linkage from the CODES 2000 strategic matching methodology. Currently, these two databases have an interface, which means that they are automatically linked.

Maryland proceeded to merge other databases with its crash data, including the motor vehicle database, the citation database, and the licensing database. These databases are integrated and are linked only once annually. The main challenge with linking these datasets was convincing the various stakeholders and data managers of the benefits of data integration. Additionally, merging of the adjudication databases was challenging due to the differences between the various court systems.

The benefits that Maryland has observed from merging datasets are numerous. Key benefits are:

- Merging the hospital database with crash data has allowed the state to quantify the cost of accidents and to determine the impacts of seatbelt and motorcycle helmet use. The State of Maryland has found that the only way to accurately identify serious injuries is through the hospital data.
- Merging the citation database with the crash database was found to be most useful for highway safety applications. An example is determining whether seatbelts were used. Additionally, this type of data merging allows analysts to identify the types of driver behavior(s) that eventually lead to crashes.
- Merging crash data and licensing data helps to identify communities or locations in which there is a high proportion of drivers involved in crashes. This allows educational efforts to be more focused on the areas in which they can be most beneficial.

Maryland's integrated database is housed at the University of Maryland, which has several benefits. The first benefit that was noted was that the Institutional Review Board (IRB) umbrella and processes within the University helps to more effectively maintain data privacy. The data integration center has more dedicated personnel who are very familiar with all of the datasets and understand how data are collected, maintained, and how they can be used. Maryland's representative noted that it can take up to 1.5 years to become familiar with the structure of various databases. Due to this large startup cost, Maryland employs a dedicated group of personnel to maintain the integrated database. All data analysis is also done in the data integration center, which helps ensure consistent results.

Michigan

The Michigan State Police (MSP) is currently halfway through an integration project that is anticipated to last 3 years. Currently, some of the systems are semi-linked; for example, crash data are linked with the vehicle and driver databases.

Initially, the MSP hired a consultant to do an analysis of how data might be integrated and provide some case studies. The consultant predicted that this would take 4-5 years and would cost approximately $8 million. When Michigan started its data integration plan 1.5 years ago, there was an existing dashboard and set of tools that were being used for other purposes that MSP was able to leverage to streamline the data integration process. Hence, MSP was able to reduce the cost to approximately $2.5 million, including business intelligence tools. A vendor is currently working on this data integration project.

The first step was to create a crash and TRCC data model. In essence, this data model organizes the elements of data systems and standardizes how they are related to each other. The data model not only makes data integration easier, but it can be leveraged for all of the analysis and reporting needs. The basic steps in establishing the data model were to first determine the uses for the databases (i.e., the type of analysis and reporting for which the database would be used). The next step was to establish criteria for the data in terms of which information would be critical and the permissible values for each of those data. Next, the MSP combined this data model with the existing data to determine the requisite integration dataset. The integration of data into this master database is done in different ways depending on the database and the frequency of data collection. For example, the citation data are collected electronically and available in real time. Hence, a change data capture design pattern is used to determine the data that have been changed in the database and update them. A change in the stored data is possible due to the structure of the database. On the other hand, incident data take about a week or the roadway management system takes a day before it can be integrated into the database. These require different levels of the change data capture to be available.

Currently, the MSP is using many different data sources such as citations, crash, breathalyzer, EMS location, roadway condition, and roadway mapping to create the integrated database. These data are all housed in the same place and can be used for different applications (e.g., drunk drivers in a specific location during a specific season). There are more than 100 use cases that MSP has already identified and accommodated for this integrated database. MSP determined these use cases as a result of the consultant's scope analysis and by talking to all stakeholders.

One of the major barriers to sharing data across various agencies and across divisions is data privacy. Hence, it is essential to have a flexible Master Data Management arrangement. MSP uses the Federal Schema for the Master Data Management, which allows MSP to metatag the integrated database to the primary database. This type of communication allows the primary owner of the data to always know where and how their data are being used. The Master Data Management and the data model share the same location, which makes data integration easier and ensures that the database model is flexible.

The main challenges that the MSP has experienced are resistance from several key resource areas with respect to sharing data. For example, the Department of State, which maintains the driver license and vehicle data, has an elected official running the department; for this reason, they are not obliged to and do not share their data. Hence, the MSP has been trying to overcome this challenge by being creative in how they can piece together this information until they can convince the agencies to share the data. For example, for the driver and vehicle data, the MSP has chosen to try to use the data that already exist within the State Police that has been collected for other purposes. Additionally, for missing data (e.g., court data), they have been able to meet the analysis or reporting needs by reconstructing some of the existing datasets. Their objective in using these data is to be able to come up with example uses for integrated data to try to convince the Department of State to share their data. MSP's main approach in trying to get the agencies to cooperate has been to illustrate the benefits of data integration by using the (limited) data they have already integrated. More specifically, they aim to first show how the analysis being currently done can become more efficient, and second show other uses for integrated data. Another challenge that Michigan has experienced has been with the court and citation system, since there are many different types of datasets that are not integrated, which makes it difficult to produce one streamlined database.

The MSP also works with the University of Michigan on maintaining data related to Michigan Traffic Crash Facts. This helps with data analysis since the University of Michigan is not working with the state, has more specific data analysis experience, and can provide unbiased analysis.

Additionally, the MSP is happy to schedule some time to meet with the Pennsylvania TRCC regarding the data integration project and demonstrate some of its integrated data uses. They are motivated to create a regional database that includes both Pennsylvania and Michigan data.

Nebraska

Nebraska began its safety data integration in the late 1990s when the Department of Health and Human Services and Nebraska DOT partnered. Since then, these two entities have continued their data integration efforts. Nebraska was specifically interviewed since it is one of the few states to integrate crash data with Vital Statistics data. Since Vital Statistics, death certificates, and EMS data are all within the Department of Health and Human Services, data sharing is relatively easy in Nebraska. Additionally, hospitals compile injury data and remove personally identifiable information (PII) annually to send to the Department of Health and Human Services. Two major benefits to using Vital Statistics data were noted by Nebraska. First, death certificates contain detailed demographic data, including race and ethnicity, that no other dataset contains. This information is typically provided by the family or the coroner and can be used to improve injury prevention programs. The second benefit is that death certificates provide information about deaths that occur due to crashes on private properties (e.g., tractor crashes on farmland), deaths that occur 30 days after the crash, or a death that happens due to a crash out-of-state. Nebraska uses some of this information for improving its child-death review project.

The linkages between crash data and death certificates, hospital discharge data, and EMS data are done in a probabilistic manner using date of birth, crash date, admission date, and crash location. While some PII is available, probabilistic linkage is preferred due to errors in PII information. The linked data are used by EMS agencies to analyze response times to crashes, and alcohol- and drug-use-related crashes. The linked crash and hospital discharge data are used to accurately estimate the cost of crashes, and used for political purposes. The cost of the whole data share program was estimated to be minimal, and includes the salary of one analyst and minimal charges for the hospital discharge data.

North Carolina

North Carolina started data integration in the mid 1990s. North Carolina noted that most of its data systems were in electronic format at this time, which helped facilitate the data integration process. The first step was to update the driver license history to ensure accuracy and eventual linkage to the crash database. Currently, the crash reports are automatically populated with driver license information at the site when a police officer scans the driver's license.

The second step was to link court data, specifically convictions, to the crash database. The driver license number was used to create this linkage. The citation information is all electronic statewide and is directly integrated into the driver's license database. The next step was to integrate the vehicle database into the crash database using the driver's license information. This is not the most accurate method, since some vehicles are owned by entities other than individuals; however, North Carolina found the resulting database to be sufficient for its purposes. Additionally, this linkage can and is performed automatically. A police officer filing a police report can type in the license plate number, which automatically connects the report to the registration file in terms of owner registration address, etc. The roadway database was initially integrated by triangulating the location using three streets and distances. Currently, some law enforcement have GPS devices and can enter exact location in the crash report. Some others use Google Maps to pull up the coordinate information. Finally, merging the medical data is a work in progress. This was found to be the most challenging piece to integrate. There are many different entities, including trauma registry, EMS, and medical examiner that have to be contacted individually.

As a result of this data integration, North Carolina has been able to quantify the impact of alcohol use, seat belt use, or helmet misuse while riding bicycles on crash outcomes and severities. Being able to quantify these impacts has helped pass laws to prohibit alcohol use, mandate helmet use for motorcyclists, and mandate seat belt use. Other benefits have been to determine the cause(s) of crashes, such as differences between old and young drivers, and better understand and quantify driver distraction. Such information helps focus enforcement efforts and educational efforts to improve roadway safety.

The merged dataset is housed in the state capital, under an agreement with the University of North Carolina for access and data analysis.

The biggest challenge in integrating datasets has been to maintain the personal security on all files. Currently, individuals are not identified in the combined crash and driver license databases. Additionally, the Health Insurance Portability and Accountability Act (HIPAA) has been a major barrier for integrating the medical data. The one solution that North Carolina DOT has been able to identify is to share the crash data with the hospitals and allow them to merge it and return the matches. To motivate this, projects are used to reimburse medical personnel for their time spent integrating the datasets. Some other challenges have been general opposition to data merging and funding such projects.

The best method to overcome data integration challenges is to clearly demonstrate the benefits; for example, quantifying cost savings due to the reduced number of accidents either from small pilot projects or trying to obtain such information from other states. Overall, the main argument that North Carolina has expressed in favor of data integration is that it is improves efficiency. Hence, less staff time is spent on doing everyday tasks and frees up state resources to focus on other areas that can help reduce crashes.

Utah

Utah started its data integration by creating an electronic crash database. The state partnered with the University of Utah and created the Utah Transportation and Public Safety – Crash Data Initiative (UTAPS-CDI). As a part of this initiative, all crash data will automatically go into the UTAPS-CDI database. The university was chosen to house the data to keep consistency in the data analysis, since Utah DOT was experiencing inconsistencies in data analysis between its different units. Additionally, this allows Utah DOT to free-up personnel time to focus on more important tasks such as implementing programs to improve roadway safety. The biggest challenge with this was to ensure that all personnel involved in the project were up to date with their privacy training.

The crash reports and EMS are integrated in the state of Utah through a grant from NHTSA. Name, date of birth, and location of crash are used to merge these two datasets. One of the benefits of having these databases is to determine whether drug use was a factor in crashes. The main barrier to this integration was communication between the vendors and UTAPS. The roadway and crash databases are also merged automatically by determining the location of the crash.

The next step is to merge the court data; however, Utah does not see a need to merge vehicle or driver license data into its databases.

Wyoming

The State of Wyoming created its TRCC in 1998 and has a working group and an executive group. The Highway Safety Department, which is a part of the TRCC, initially started with the creation of a statewide electronic crash database that was implemented in 2008, which is stored in an ORACLE database. Since then, the Wyoming TRCC has been examining how other datasets could be incorporated into the crash data and the challenges of such data integration. They do not currently have an automatic interface with any of the other core data systems; however, they have been able to establish integration at regular intervals.

The data from the motor vehicle and driver license databases are currently stored on an old mainframe. This creates a challenge since mainframes are difficult to synchronize with modern computers, and the security on different systems makes it difficult to pass information between them. Hence, it was determined that the simplest solution for data integration is to periodically copy information from the main frame databases and transfer to the ORACLE database. For the driver database this transfer is done daily. These databases are merged by matching names. The methodology that is used is to take out all the vowels, and collapse some constants (e.g., “ck” becomes a “c”), and then match based on these reduced sets of constants. They have found this to be the easiest way to match data, especially if there are some inaccuracies in the databases. The vehicle database is integrated with the crash database on a monthly basis. The roadway inventory is also linked to the crash database by milepost information.

The next step that Wyoming is trying to achieve is to merge the citation information and other state driver information. The integration of the adjudication and citation databases with the crash database is challenging because even within these systems there is no unified data structure. For example, different municipalities use different citation forms. One of the ways that they are considering overcoming these challenges is to focus on the Supreme Court data and get a snapshot of that information. Another challenge is privacy concerns, since other agencies do not want to share personal data. One of the ways they are considering for overcoming this challenge is to remove personal data before providing snapshots; however, this makes deterministic linkage between the datasets difficult.

Finally, Wyoming is also working toward integrating the EMS database with the crash database. The EMS database is currently being converted to an electronic format. However, linking to this database is challenging due to HIPAA and privacy concerns.

The benefits that Wyoming has observed from merging datasets are numerous. Among the key benefits are:

- By integrating the driver database with the crash database, Wyoming aims to identify the characteristics of drivers that most contribute to crashes. They also aim to identify the kinds of questions in the written driving test that are most associated with long-term safety performance and what kind of driving tests might lead to safer driving practices. Overall, they are interested in improving their licensing system to create a safer roadway system.
- Linkage between the roadway and crash databases has allowed Wyoming to apply *Highway Safety Manual* analysis methods with corrections, to determine how narrow shoulders might contribute to run-off-the-road crashes and make hot-spot analysis (e.g., rank horizontal curves, rank intersections, etc.).
- Linkage between the vehicle and crash databases allows Wyoming to determine the body types that contribute to crashes with more injuries.
- Linking the citation information with the crash database allows Wyoming to categorize driver behavior into six categories (alcohol, drug, speed, seat belt use, distraction, and moving violation) and determine the ratio of citations for each category compared to the number of crashes due to each category. Hence, they can improve how enforcement is deployed in the field and what categories are enforced more carefully.
- Finally, Wyoming expects that the linkage between the EMS database and crash data can allow them to better quantify injuries and obtain better cost estimates for crashes.

Overall, Wyoming found that when working with outside agencies, the issues of privacy, data ownership and reluctance to share data (e.g., due to knowledge of bad database) are the major issues. They have been able to overcome these issues by showing agencies the benefits of linked databases to get over the cost burden. Being able to show the value of data with the statement that some or bad data is better than no data has helped them convince agencies to share data.

CONCLUDING REMARKS

This review examined the academic research literature and current STA practices on integration of traffic records to support safety analysis. The following general findings were identified as a result of the review:

- The primary barriers to data integration are the lack of well-defined linkage variables, privacy concerns, data ownership, and costs. The lack of linkage variables can be overcome through the use of probabilistic linkage methods, which can be used to identify close but not exact matches. In many cases, this means that only a subset of a data system can be integrated with crash data, but this may be sufficient for safety analyses. The other barriers can be overcome by clearly illustrating the benefits of such data integration. Doing so can help convince agencies to more willingly share their data, as well as help justify the costs of doing so. Privacy concerns can be alleviated by removing personally identifying information (PII) from datasets; however, this removes potential linkage variables. Instead, one solution is to have the agency responsible for maintaining the database with PII to perform the data integration first, then remove the PII and provide only a redacted version of the linked database.
- Several states have fully integrated traffic records systems. These agencies include Colorado, Maryland, North Carolina, and Wyoming. Interviews with some of these agencies have identified the following benefits of data integration:
    - Identification of driver characteristics most closely associated with crashes/unsafe driving behavior
    - Identification of questions on the driver exam that are more associated with crashes/unsafe driving in the future
    - Application of *Highway Safety Manual* methods
    - Improved targeted enforcement and educational efforts, especially with respect to driving behaviors such as alcohol and drug use, seat belt use, motorcycle helmet use, and distracted driving
    - Better understanding of crash risk among segments of the population, including younger and older drivers
    - More accurate tracking of crash severity outcomes and estimation of total costs associated with crashes (of varying severity levels)
- Most states integrated only a subset of the data systems. Surprisingly, the most commonly integrated with crash data are the injury surveillance/EMS data, followed by roadway data, driver data, citation and adjudication data, and, lastly, vehicle data.
- Common identifiers or linkage variables used to merge the various data systems included:
    - Driver information
        - Name
        - Date of birth
        - License number
    - Crash information
        - Location
        - Date and time
- Of the previous list, crash information is more conducive to deterministic linkage while the driver information is more conducive to probabilistic linkage.
- Several states highlighted the importance of establishing memoranda of understanding to facilitate data sharing and establishing a secure database to house the data.
- In addition, several interviewed states identified that the data integration is performed by an outside organization that can dedicate the appropriate time and effort to learn the intricacies of the datasets. This was done because the data systems are generally very complex and can take quite some time

to fully understand. Analyses of these data were also performed by these organization to ensure consistency in the approaches used.

CHAPTER 3

Current Situation and SWOT Analysis

This section describes the current state of the core data systems considered for safety data integration. This includes a description of the data structure and interfaces, data sources, data custodians, data elements, current integration with other data systems, and a SWOT analysis that highlights the strengths, weaknesses, opportunities, and threats associated with that data system. Strengths and weaknesses are defined as internal factors that are within the agency's control, while opportunities and threats are factors that are outside the agency's control.

CRASH DATA

This section describes the crash data system for Pennsylvania. This data system provides records of crashes that occur on all roads open to the public, including the location, vehicles, and people involved, as well as injury level suspected by the police at the time of the investigation. The information described here was primarily obtained through an interview with David Kelly (Crash Information Systems and Analysis at PennDOT), who helps to manage this database, as well as the research team's familiarity with this data system from previous projects. This section is organized into several subsections that describe the data structures and interface, sources used to obtain the data, data custodians, and access policies, elements included in the database, and potential linkage variables.

Data structures and interfaces

The crash data system is comprised of multiple individual databases. The first is the working database, which houses the original data. This working database is also known as the crash reporting system, core or input database. The second is an output database called the Crash Data Analysis and Reporting Tool (CDART), which includes additional data elements to describe specific features of interest. The last is a public view of the CDART known as the Pennsylvania Crash Information Tool (PCIT), which excludes any PII from the data.

The core database (CRS) is a DB2 database that includes two instances of each reported crash: a type I record and a type II record. Type I records strictly contain the information obtained from a police crash report. This information is entered either directly using a web-based system used by the police or indirectly using a software that obtains the input information and outputs a *.xml file that can be incorporated into the CRS. Currently, about half of the data in the core database are entered using each of the two methods.[1]

Each Type I record is then copied into a Type II record, which is then reviewed by a data analyst to verify its accuracy and add any missing information.

Every seven days, the CRS records are pulled into the CDART system, which is an Oracle database that allows users to view individual crash record information. In this process, additional data elements are created to help identify or summarize crashes of particular interest. These additional elements are typically flags or summary variables for each crash. Examples include total fatality count, number of pedestrians involved in a crash, etc. The flags allow the database to immediately locate crashes for which these fields apply (e.g., crashes where a pedestrian was involved). This helps make reporting outputs easier, since queries do not have to be built for reporting – the data are already tallied and kept at the crash level in count fields or in flag fields that categorize the individual crashes. Additionally, some data elements are removed and/or merged when incorporated into the CDART. For example, county and municipality codes are merged into one field.

Within the CDART and PCIT databases there are 8 tables that contain usable data, including:

- Crash
- Vehicle
- Comm Vehicle
- Trail Vehicle
- Cycle
- Person
- Roadway
- Flag

The crash table contains crash-level information, including location, time of day, and other factors associated with the crash event. The vehicle table contains information on the vehicles involved in each crash. Separate tables provide information for commercial vehicles (Comm Vehicle), trailers (Trail Vehicle), or motorcycles (Cycle) involved in the crashes. The person table contains information on the drivers, occupants, and pedestrians involved in each crash. The roadway table provides information on the specific roadway on which the crash occurred. The Flag table provides the individual flags developed to make summary reporting more efficient. Each of these tables is linked through a unique Crash Record Number (CRN) associated with each specific crash.

The PCIT system is also an Oracle database that contains many of the same elements included in the CDART system. However, while the CDART system is designed to be used by those within PennDOT, the PCIT is designed to be used by those outside of PennDOT, including police officers or researchers. Thus, all PII is removed from the crash records, such as individuals' names, driver's license numbers, Vehicle Identification Numbers (VINs) of vehicles involved, etc.

[1] Note that there still exist some paper records (approximately 10,000 out of 140,000 crashes) that are entered into the system using intelligent character recognition (ICR). The ICR software also creates an *.xml document that can be incorporated into the CRS. However, this method is being phased out and will be removed by the end of the year 2020.

Data sources

Information initially included in the core database (Type I record) is obtained from police crash reports. As previously mentioned, the police crash reports are either entered into a web-based computer application by a local police department, or paper copies of the police reports are then run through an ICR to create an electronic record of the information. An annotated sample of the first page of a police crash report (AA 500) is shown in Figure 2. Data included in crash records include only reportable crashes, defined as those that occur on a public roadway, involve injury to or death of any person, and/or include damage to any vehicle such that it cannot be moved under its own power.

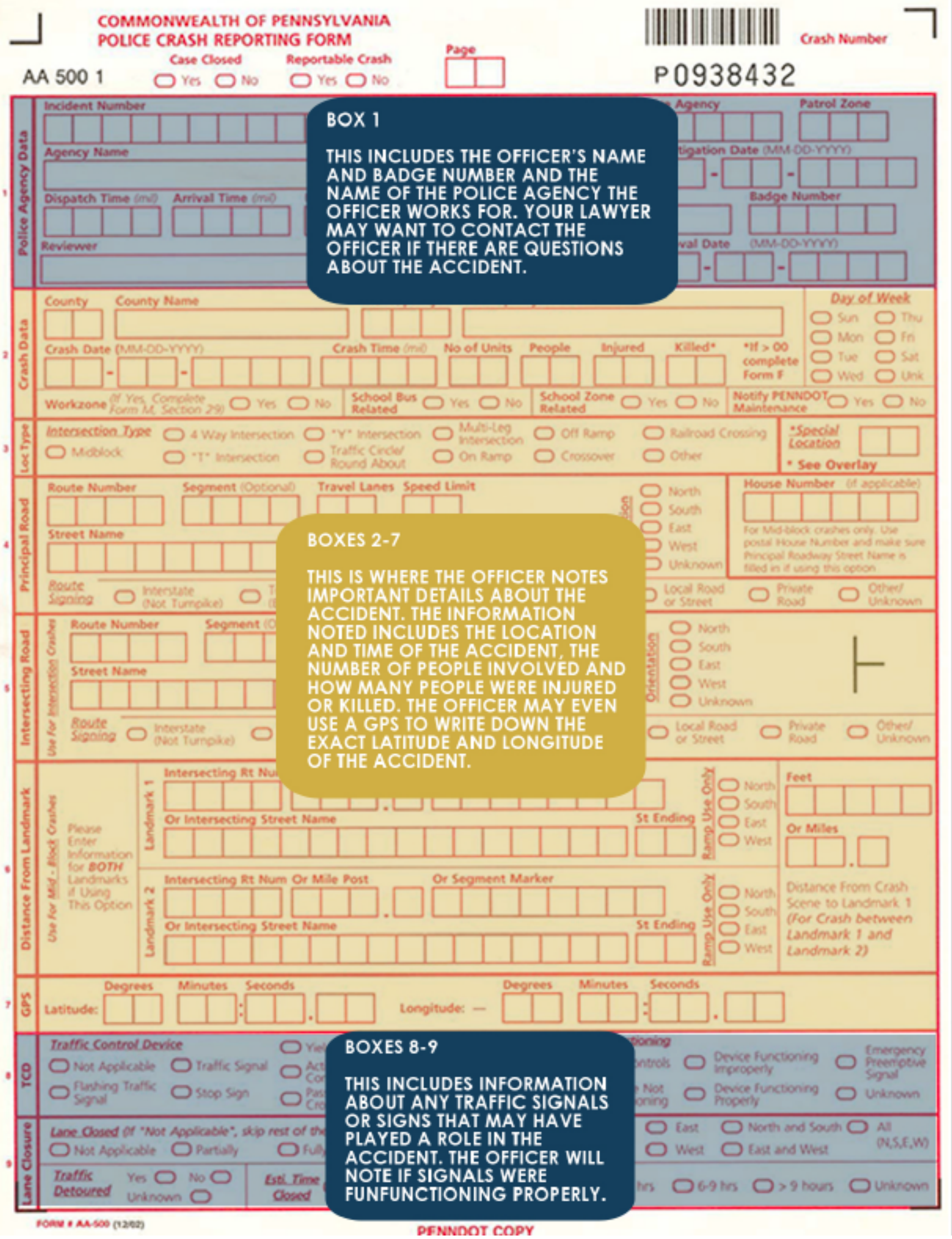
When transitioning a record from a type I to type II record, additional information is obtained from a variety of sources to verify and/or supplement the crash record. The sources include PennDOT's driver's license database (for person-level information), PennDOT Vehicle Registration database (for vehicle-level information), and PennDOT's Roadway Management System database (for infrastructure-level information). These processes are described in the linkages section, as this represents evidence of some amount of one-way data integration already occurring across these systems.

Data custodians and access policies

The crash data are owned by the Crash Information Systems and Analysis Unit within PennDOT. The data—including both the mainframe and Oracle databases—are housed in Harrisburg in a server farm that is operated by PennDOT for the Governor's Office of Administration.

Access to the CDART data is provided only to those with an internal Commonwealth Account who have been included in the CDART user group. Individuals are only included in this group if they have demonstrated a need for accessing these data and received training on accessing and dealing with these data.

Those without an internal Commonwealth Account are considered "public" users and can access the data using one of two methods. First, one-page summaries of the crash data can be obtained from the PCIT. These are known as Public Inquiry/Press Reports and provide only aggregated-level information about crashes in the database by various categories (e.g., see Figure 3, which shows a sample report that presents annual crashes for a specific region for 2008 through 2017). Second, the detailed crash data without any PII can be obtained directly from the PCIT through authenticated access.



COMMONWEALTH OF PENNSYLVANIA
POLICE CRASH REPORTING FORM

AA 500 1

Case Closed ○ Yes ○ No

Reportable Crash ○ Yes ○ No

Page

Crash Number

P0938432

1 Police Agency Data

Incident Number

Agency Name

Dispatch Time (mil) Arrival Time (mil)

Reviewer

Agency Patrol Zone

tigation Date (MM-DD-YYYY)

Badge Number

val Date (MM-DD-YYYY)

BOX 1

THIS INCLUDES THE OFFICER'S NAME AND BADGE NUMBER AND THE NAME OF THE POLICE AGENCY THE OFFICER WORKS FOR. YOUR LAWYER MAY WANT TO CONTACT THE OFFICER IF THERE ARE QUESTIONS ABOUT THE ACCIDENT.

2 Crash Data

County County Name

Crash Date (MM-DD-YYYY) Crash Time (mil) No of Units People Injured Killed* *If > 00 complete Form F

Day of Week
○ Sun ○ Thu
○ Mon ○ Fri
○ Tue ○ Sat
○ Wed ○ Unk

Workzone (If Yes, Complete Form M, Section 29) ○ Yes ○ No

School Bus Related ○ Yes ○ No

School Zone Related ○ Yes ○ No

Notify PENNDOT Maintenance ○ Yes ○ No

3 Loc Type

Intersection Type
○ Midblock
○ 4 Way Intersection ○ "T" Intersection
○ "Y" Intersection ○ Traffic Circle/ Round About
○ Multi-Leg Intersection ○ On Ramp
○ Off Ramp ○ Crossover
○ Railroad Crossing ○ Other

*Special Location
* See Overlay

4 Principal Road

Route Number Segment (Optional) Travel Lanes Speed Limit

Street Name

Route Signing ○ Interstate (Not Turnpike) ○ T

○ North ○ South ○ East ○ West ○ Unknown

House Number (if applicable)

For Mid-block crashes only. Use postal House Number and make sure Principal Roadway Street Name is filled in if using this option

○ Local Road or Street ○ Private Road ○ Other/ Unknown

5 Intersecting Road

Use For Intersection Crashes

Route Number Segment (O

Street Name

Route Signing ○ Interstate (Not Turnpike) ○

Orientation ○ North ○ South ○ East ○ West ○ Unknown

○ Local Road or Street ○ Private Road ○ Other/ Unknown

BOXES 2-7

THIS IS WHERE THE OFFICER NOTES IMPORTANT DETAILS ABOUT THE ACCIDENT. THE INFORMATION NOTED INCLUDES THE LOCATION AND TIME OF THE ACCIDENT, THE NUMBER OF PEOPLE INVOLVED AND HOW MANY PEOPLE WERE INJURED OR KILLED. THE OFFICER MAY EVEN USE A GPS TO WRITE DOWN THE EXACT LATITUDE AND LONGITUDE OF THE ACCIDENT.

6 Distance From Landmark

Use For Mid - Block Crashes

Please Enter Information for BOTH Landmarks if Using This Option

Landmark 1

Intersecting Rt Nu

Or Intersecting Street Name St Ending

Ramp Use Only ○ North ○ South ○ East ○ West

Feet

Or Miles

Landmark 2

Intersecting Rt Num Or Mile Post Or Segment Marker

Or Intersecting Street Name St Ending

Ramp Use Only ○ North ○ South ○ East ○ West

Distance From Crash Scene to Landmark 1 *(For Crash between Landmark 1 and Landmark 2)*

7 GPS

Latitude: Degrees Minutes : Seconds .

Longitude: — Degrees Minutes : Seconds .

8 TCD

Traffic Control Device
○ Not Applicable ○ Flashing Traffic Signal
○ Traffic Signal ○ Stop Sign
○ Yiel ○ Act Cor ○ Pas Cro

tioning
ntrols
Not oning

○ Device Functioning Improperly ○ Device Functioning Properly

○ Emergency Preemptive Signal ○ Unknown

9 Lane Closure

Lane Closed (If "Not Applicable", skip rest of th
○ Not Applicable ○ Partially ○ Fully

○ East ○ West ○ North and South ○ East and West ○ All (N,S,E,W)

Traffic Detoured Yes ○ No ○ Unknown ○

Esti. Time Closed

hrs ○ 6-9 hrs ○ > 9 hours ○ Unknown

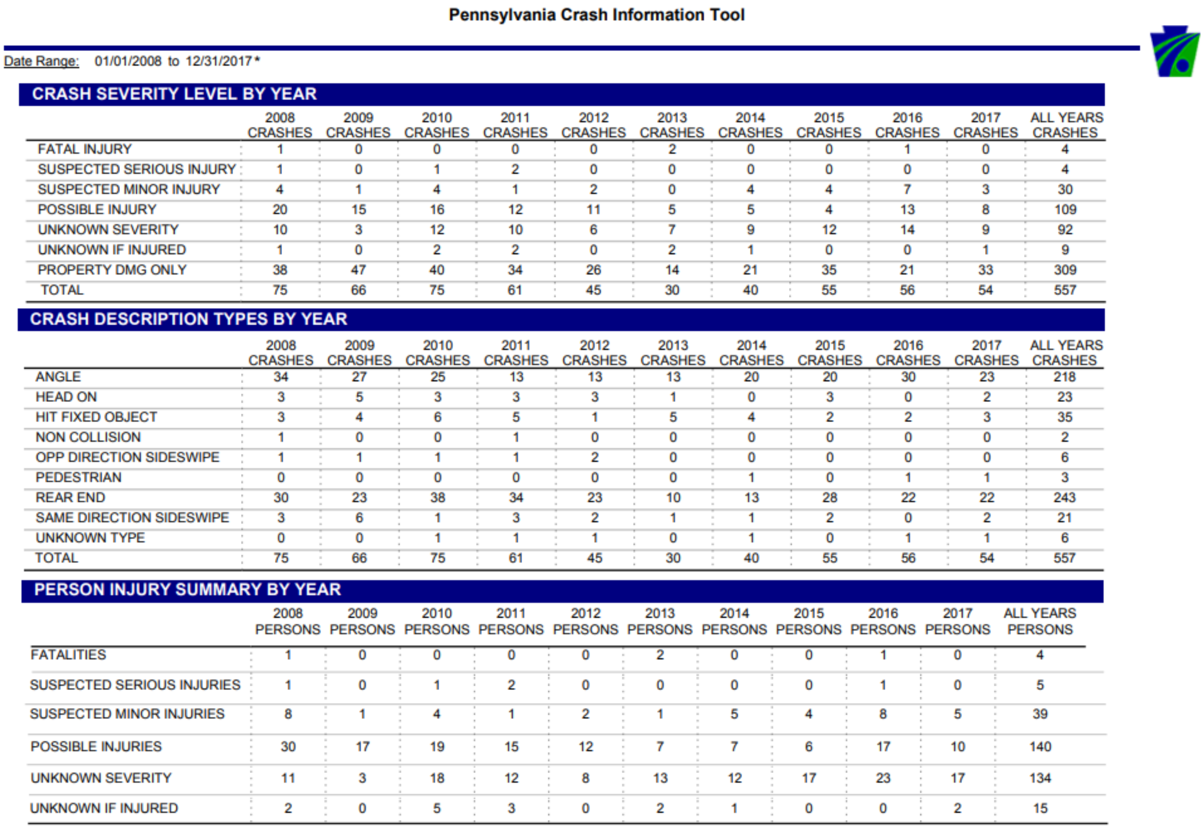
BOXES 8-9

THIS INCLUDES INFORMATION ABOUT ANY TRAFFIC SIGNALS OR SIGNS THAT MAY HAVE PLAYED A ROLE IN THE ACCIDENT. THE OFFICER WILL NOTE IF SIGNALS WERE FUNFUNCTIONING PROPERLY.

FORM # AA-500 (12/02)

PENNDOT COPY

Figure 2. Front page of Pennsylvania crash report



Pennsylvania Crash Information Tool

Date Range: 01/01/2008 to 12/31/2017 *

CRASH SEVERITY LEVEL BY YEAR

	2008 CRASHES	2009 CRASHES	2010 CRASHES	2011 CRASHES	2012 CRASHES	2013 CRASHES	2014 CRASHES	2015 CRASHES	2016 CRASHES	2017 CRASHES	ALL YEARS CRASHES
FATAL INJURY	1	0	0	0	0	2	0	0	1	0	4
SUSPECTED SERIOUS INJURY	1	0	1	2	0	0	0	0	0	0	4
SUSPECTED MINOR INJURY	4	1	4	1	2	0	4	4	7	3	30
POSSIBLE INJURY	20	15	16	12	11	5	5	4	13	8	109
UNKNOWN SEVERITY	10	3	12	10	6	7	9	12	14	9	92
UNKNOWN IF INJURED	1	0	2	2	0	2	1	0	0	1	9
PROPERTY DMG ONLY	38	47	40	34	26	14	21	35	21	33	309
TOTAL	75	66	75	61	45	30	40	55	56	54	557

CRASH DESCRIPTION TYPES BY YEAR

	2008 CRASHES	2009 CRASHES	2010 CRASHES	2011 CRASHES	2012 CRASHES	2013 CRASHES	2014 CRASHES	2015 CRASHES	2016 CRASHES	2017 CRASHES	ALL YEARS CRASHES
ANGLE	34	27	25	13	13	13	20	20	30	23	218
HEAD ON	3	5	3	3	3	1	0	3	0	2	23
HIT FIXED OBJECT	3	4	6	5	1	5	4	2	2	3	35
NON COLLISION	1	0	0	1	0	0	0	0	0	0	2
OPP DIRECTION SIDESWIPE	1	1	1	1	2	0	0	0	0	0	6
PEDESTRIAN	0	0	0	0	0	0	1	0	1	1	3
REAR END	30	23	38	34	23	10	13	28	22	22	243
SAME DIRECTION SIDESWIPE	3	6	1	3	2	1	1	2	0	2	21
UNKNOWN TYPE	0	0	1	1	1	0	1	0	1	1	6
TOTAL	75	66	75	61	45	30	40	55	56	54	557

PERSON INJURY SUMMARY BY YEAR

	2008 PERSONS	2009 PERSONS	2010 PERSONS	2011 PERSONS	2012 PERSONS	2013 PERSONS	2014 PERSONS	2015 PERSONS	2016 PERSONS	2017 PERSONS	ALL YEARS PERSONS
FATALITIES	1	0	0	0	0	2	0	0	1	0	4
SUSPECTED SERIOUS INJURIES	1	0	1	2	0	0	0	0	1	0	5
SUSPECTED MINOR INJURIES	8	1	4	1	2	1	5	4	8	5	39
POSSIBLE INJURIES	30	17	19	15	12	7	7	6	17	10	140
UNKNOWN SEVERITY	11	3	18	12	8	13	12	17	23	17	134
UNKNOWN IF INJURED	2	0	5	3	0	2	1	0	0	2	15

Figure 3. Sample output from PCIT

Data elements

The specific data elements and codes are too numerous to list here but can be obtained from the public data dictionary available on the PCIT website. However, Table 5 provides a summary of some of the data elements included in each of these eight tables.

Table 5. Summary of data elements in crash data tables

Crash table

- Crash record number (CRN)
- Collision category that defines the crash
- Crash date and time
  - Crash year, month, day, and time
- Crash location
  - County, district, municipality code
  - Longitude and latitude
  - Urban or rural indicator
  - Location type (code that defines the crash location)
  - Code that defines any special jurisdiction (fatal crashes only)
- Police
  - Police agency code
  - Police dispatch and arrival times
- Unit involved
  - Total count of all vehicles and pedestrian involved
  - Count of total people involved
  - Young (under 20, inclusive) and old (over 50, inclusive) drivers count
  - Total count of unbelted occupants
  - Count of total pedestrians involved
  - Total amount of automobiles involved
  - Total count of sport utility vehicle (SUV) involved
  - Total amount of commercial vehicles involved
  - Total amount of buses involved
  - Did the crash involve a school bus?
  - Total amount of heavy trucks involved
  - Total amount of small trucks involved
  - Total amount of vans involved
  - Total amount of motorcycles involved
  - Total amount of bicycles involved
- Injury and fatality
  - Number of people that are unknown if injured
  - Number of people with unknown injury severity
  - Total amount of minimal injuries
  - Total amount of moderate injuries
  - Total amount of major injuries
  - Total major injuries of belted occupants
  - Total major injuries of unbelted occupants
  - Total major injuries of bicyclist
  - Total amount of fatalities involved
  - Total deaths of belted occupants
  - Total deaths of unbelted occupants
  - Total deaths of bicyclist
- Weather and illumination
  - Code for the weather type at time of crash
  - Code that defines lighting at crash scene
- Traffic control
  - Traffic control device type and state
  - Intersection type
  - Direction of traffic in closed lane(s)
  - Estimated hours roadway was closed
- Roadway
  - Roadway surface type code (only for fatal crashes)
  - Roadway surface condition code
  - Crash's relativity to the road code
  - PennDOT highway maintenance notified?
- Work zone or school zone or construction zone
  - Did the crash occur in a work zone?
  - Work zone location and type
  - Was traffic rerouted due to work zone?
  - Was there a flagman/patrolman/lane closure?
  - Was there moving work in the zone?
  - Was a median/shoulder in the zone?
  - Did the crash occur in a school zone?

Vehicle table

- CRN
- Unit number (unit number of the vehicle or pedestrian in the crash event)
- Vehicle information
  - VIN
  - License plate
  - Vehicle type (automobile, van, SUV, motorcycle, bicycle, etc.)
  - Vehicle make, model, year, color
  - Vehicle body type code
  - Vehicle registration State
  - Vehicle automation level
  - Vehicle owner address (State, city, ZIP code)
  - Vehicle owner phone number
  - Insurance company name and phone number
- Trailing vehicle count
- Under ride indicator (fatal only)
- Vehicle special usage code
- Vehicle at crash
  - Vehicle position
  - Vehicle movement
  - Vehicle role (0 for N/A, 1 for striking)
  - Vehicle travel direction
  - Estimated travel speed
  - Avoidance maneuver code
  - Driver presence indicator
  - Damage indicator
  - Initial impact point
  - Principle impact point
  - Harmful events (1-4) for this unit
- Total people in this unit
- Commercial vehicle indicator
- Vehicle carrying hazmat indicator
- Vehicle towing indicator and towing company
- Record number (Policy-1, PennDOT-2)
- System date/time the record was posted

Comm Vehicle table

- CRN
- Unit number
- Commercial vehicle information
  - Vehicle configuration code
  - USDOT number
  - PA utility commission number
  - Number of axles on the vehicle
  - Cargo carrier's body type
  - Gross vehicle weight rating
  - Oversize load indicator
- Carrier information
  - Carrier name
  - Interstate commercial carrier number
  - Carrier address
  - Carrier telephone number
- Hazmat
  - Hazmat code for material onboard (one to four)
  - Indicator for hazmat material released

Trail vehicle table

- CRN
- Unit number
- Trailer information
  - Trailer sequence number
  - Trailer registration tag number, and registration year and state
  - Trailer type code

Cycle table

- CRN
- Unit number
- Motorcycle information
  - Motorcycle engine size
  - Motorcycle side bags indicator
  - Motorcycle trailer indicator
- Pedal cycle
  - Pedal cycle head protection indicator
  - Pedal cycle headlight indicator
  - Pedal cycle rear reflector indicator
  - Pedal cycle passenger presence indicator
- Motorcycle driver and passenger
  - Driver safety training indicator
  - Driver head and eye protection indicator (e.g. wear a helmet?)
  - Driver body protection indicator (e.g., wear long sleeves?)
  - Passenger presence indicator
  - Passenger head and eye protection indicator (e.g. wear a helmet?)
  - Passenger body protection indicator (e.g., wear long sleeves?)
- Record type code (Policy-1, PennDOT-2)
- System date/time the record was posted

Person table

- CRN
- Unit number (unit number of the pedestrian assigned to this person)
- Person
  - Person type
  - Person name
  - Person sequential number
  - Person DOB
  - Person address
  - Person age, gender
  - Injury severity code
  - Transported to medical facility indicator
  - Driver/pedestrian physical condition at time of crash
  - Driver/pedestrian telephone number
  - Driver/pedestrian condition code
- Driver information
  - Driver license number
  - Driver license class
  - Driver license State
  - Diver first name
  - Driver middle initial
  - Driver last name
  - Driver's height
  - Driver's address
- Driver license lookup
  - Driver first violation month
  - Driver first violation year
  - Driver last violation month
  - Driver last violation year
  - Previous DUI conviction count
  - Previous moving violation count
  - Previous number of crashes
  - Previous speeding conviction count
  - Previous license suspension count
- Occupants
  - Seat in unit where occupants sat
  - Ejection indicator – only for vehicle occupants
  - Extrication indicator – only for vehicle occupants
- Pedestrian
  - Pedestrian action at time of crash
  - Pedestrian location code
  - Pedestrian signal indicator
- Pedestrian clothing type
- Tests
  - Suspected alcohol/drug use
  - Alcohol test type
  - Alcohol test result
  - Drug test type

  - Age of driver
  - Date of birth from driver’s license
  - Driver actions
  - Driver enforcement compliance
  - Driver license compliance
  - Driver license restrictions
  - Driver license restriction compliance
  - Driver license expired indicator
  - Driver license suspended indicator
  - Non-CDL driver license status
  - CDL driver license restrictions
  - CDL driver license enforcements
  - Drug test result 1 – 4
- Record type code (Policy-1, PennDOT-2)
- System date/time the record was posted

Roadway table

- CRN
- Roadway information
  - Street name
  - Street address of principal roadway
  - County code
  - Roadway sequence number
  - Roadway county code (could differ from county of crash)
  - Route number
  - Route signing
  - Roadway segment
  - Offset within segment
  - Travel lanes (primary direction)
  - Roadway orientation
  - Roadway speed limit
  - Roadway access control code (only for state roads)
  - Roadway ownership (roadway maintained by state, local or private jurisdiction)
  - Roadway information (from RMS lookup)
  - Average daily traffic flow count
  - Average median width
  - Average roadway width
  - Average shoulder width
  - Guide rail height
  - Guide rail type
  - Divisor type
  - Facility type
  - Functional class
  - National highway system (NHS) code
  - Number of lanes
  - Roadway surface type code
  - RMS sequence number
  - Roughness
  - Left/right indicator
  - Truck average daily percentage
  - Urban/rural indicator

Current integration

As described in the previous section, the crash data system is actually comprised of eight tables. Data from each of these can be linked using the unique CRN that is assigned to each crash. However, there are other variables that can be (or in some cases, have been) used to link the individual crashes to other data systems. The Vehicle Registration system is currently used to identify/verify vehicles involved in the crash. This is done by looking up the license plate of vehicles included in the crash report to identify the VIN of PA-registered vehicles involved in the crash. The VIN lookups are usually done within 3 days of the crash. However, if a discrepancy is found, the data analyst verifies and merges the data manually months later from the crash. Between 50% and 60% of the VIN lookups are done manually. The VIN is then used within a separate lookup system (Vintelligence) to verify the make, model, and year of the vehicle. This system can also be used to verify the make, model, and year of vehicles not registered in PA if the VIN is known. Notice that even though this is not a direct linkage, these lookups allow for verifications so that linkages can be made later.

Data from the Driver’s License system is also used to verify the information in the Driver/Pedestrian table within the crash reporting system using an individual’s Driver’s License ID number. This is used to look up or verify that the correct driver license number was entered on the crash report. Similar to VIN lookups, discrepancies are manually corrected months after the crash. Additionally, some of the existing citation information such as previous traffic violations and previous crashes can also be added to the Person table as a part of this verification.

Data from the Roadway Management System (RMS) are also linked to the crash system once the location has been verified using the GIS system. For all state roads in the crash data, RMS data are obtained using Linear Referencing System data (i.e., county, state route number, segment, and offset) determined by the GIS lookup. Approximately 99.6% of all crashes have calculated GPS coordinates identified through the GIS lookup. The remainder of crashes are marked as unknown location.

SWOT Analysis

Strengths

- **Modern data system**: The crash data system consists of a DB2 database (core database) and two Oracle databases (CDART system and PCIT tool). Both are modern databases that facilitate real-time queries and communications features required for integration purposes.
- **Data are well documented**: There exists a comprehensive data dictionary that describes the elements of this database and how the individual data files are linked together. This is required to successfully link or integrate data files.
- **Many potential linkage variables available**: The crash data system contains multiple variables that can be used to link individual crash records with records in other data systems. Potential linkage variables include PII for individuals involved in a crash, such as their name, date of birth, age, gender, driver's license number, and address. For the vehicles involved in the crash, linkage variables include license plate number, VIN, and vehicle year, make and model. The crash location, crash date, and crash time are also available. The EMS agency code was recently added to the police crash report to facilitate data integration with the injury data systems. Note, however, that many of these data elements are only verified and included on individual crash records when the driver involved in the crash has a Pennsylvania license or for vehicles registered within Pennsylvania.
- **Data are owned and maintained within PennDOT**: Although crash data are provided by individual police agencies, the statewide database is owned and maintained by PennDOT. This reduces barriers for integration of crash data with other data systems.

Weaknesses

- **Lag in data entry leads to potential errors**: Some elements of individual crash records are obtained or verified from other data systems. However, for any given record, this is only performed at the time the police submit the electronic crash report to PennDOT. This could lead to errors if any information changes between the time of the crash and the time that the report is submitted. For example, a license plate might move from the vehicle involved in the crash to another vehicle by the time the report is filed, which could lead to the wrong information being pulled from the vehicle registration database. This could lead to incorrect conclusions about vehicle-related factors associated with a crash. Such errors may eventually be corrected but require significant effort from a data analyst.
- **Verification only available for Pennsylvania drivers/vehicles**: The current data exchange can only be used to verify vehicle and driver information for vehicles registered in Pennsylvania or

drivers with a Pennsylvania license. Out-of-state driver and vehicle information cannot generally be validated. Errors for these drivers or vehicles would persist in the crash database.

- **No methods exist to verify information for passengers in crashes**: Passenger information is recorded by police but there is no mechanism in place to validate this information as is done for drivers. Thus, errors in the police report would persist in the crash database.
- **No methods exist to verify information for pedestrians and bicyclists involved in crashes**: Information is recorded by police about pedestrians and bicyclists involved in the crash but there is no mechanism in place to verify this information as is done for registered Pennsylvania drivers. Errors for these individuals would persist in the crash database.

Opportunities

- **Other critical databases maintained by PennDOT**: Three of the five other core data systems are maintained by PennDOT. This significantly reduces potential barriers to integrate crash data with these other core data systems.
- **Existing data exchange**: Data are already being shared between the crash data system and other data systems, although this is mainly used for validation purposes. For example, VINs are obtained for Pennsylvania-registered vehicles involved in a crash from the Pennsylvania vehicle registration database using the vehicle's license plate number. The VIN is then used to verify additional data elements, such as vehicle year, make, and model. The driver's license number of drivers with a Pennsylvania license is also used to obtain data elements such as date of birth from the driver database for validation purposes. Data elements from the roadway database (including the crash location using PennDOT's linear referencing system) are obtained for crashes that are verified using the GIS system. The data exchange is not just one-way, as crash information is also shared with other systems. Specifically, crash records are provided to the driver data system for drivers with a Pennsylvania license to update their driving record with their crash history. This suggests that pathways already exist that could facilitate data integration/sharing among some of the core data systems.
- **Recorded data elements recently updated to improve integration possibilities**: The EMS agency code was added to the police crash report in December 2017. As will be described below, this will be a critical variable that can be used to link crash records with records in some databases that comprise the injury surveillance data systems.
- **Existing projects involve data integration**: Recent and current research projects exist that require crash data to be merged or integrated with other data systems. One example is the set of projects performed to develop or calibrate Pennsylvania-specific safety performance functions, in which roadway data were merged with crash data to identify the number of crashes on state-owned roadways of various functional classifications. Opportunities exist to obtain and store these data from current or future projects. Additionally, a data exchange with the Department of Health and Children's Hospital of Pittsburgh recently occurred, which provided background knowledge in this type of data integration as well as demonstrates some of the potential benefits of data integration.

Threats

- **Little desire for integration from some data systems**: It was noted from representatives from the citation and adjudication system that there is little desire to integrate that data with the crash data system. However, it should be noted that a data exchange currently exists between the citation and adjudication data system and the driver data system. Thus, the citation and adjudication database has an avenue to obtain (and potentially receive) various data elements from PennDOT.

ROADWAY DATA

This section describes the roadway data system, which stores geometric and traffic information on roadways. There are two unique roadway data systems within Pennsylvania: one for state-owned roadways (the Roadway Management System, RMS) and one for non-state-owned local roads (the All Road Network of Linear Reference Data, ARNOLD). ARNOLD is created by combining the Liquid Fuels eligible roadway information with a separately collected ineligible inventory. The information here was primarily obtained through an interview with Frank DeSendi (Planning Division Manager with PennDOT), who manages the spatial RMS database, and Joseph Piper (Transportation Planning Supervisor with PennDOT), who manages the ARNOLD database. This section is organized into several subsections that describe the data structures and interface, sources used to obtain the data, data custodians and access policies, elements included in the database, and potential linkage variables.

Data structures and interfaces

RMS

The main RMS database is stored using an IBM Information Management System. The data are regularly outputted to a DB2 database, which is used for queries, or an Oracle database, which is used for spatial mapping. The latter system allows attributes from the RMS database to be visualized on a map or spatially analyzed. The main RMS data are extensive; however, not all data are outputted into the Oracle database for GIS purposes.

The RMS Linear Referencing Method, or location key, identifies a location on a roadway using the county name, route number, segment number, and finally an offset within the segment. Segments are of varying, field-defined lengths. Segment-level data contain attributes that are considered true for an entire segment, while offset-level data contain information on attributes that can change within a segment or span multiple segments.

The spatial RMS database is comprised of two main tables: RMSSEG and RMSADMIN. The RMSSEG datafile provides the location and physical characteristics of each roadway segment. This includes features such as number of lanes, presence of HOV lanes, pavement condition rating, etc. The RMSADMIN datafile contains information on where attributes change along the network and aggregate network definition attributes into offset-level lengths of identical attribution. These datafiles are constantly updated based on new construction (resulting in new segments being created) or construction on existing roadways (resulting in modification of a segment’s features). Other themed subsystems exist that use the RMS referencing system to define features of interest along the roadway network, such as locations of intersections, traffic volumes, pipes, guiderails, etc. This allows these features to be merged easily with the main RMS data.

ARNOLD

The ARNOLD data system consists of a single datafile stored in an Internal Oracle 12C database, with all maintenance and updates performed within PennDOT. Data are stored according to roadway segments in a similar format as the RMS data. However, in the ARNOLD database, segments generally begin and end at intersections or municipal boundaries and no linear measurements (i.e., offsets) are used within individual segments.

Two sets of roadways are included in the ARNOLD database—Liquid Fuel (LF) local roads and Non-Liquid Fuel roads. Liquid fuel roads are those in townships and municipalities that receive PennDOT funding for maintenance. All of these are linked and included in the ARNOLD database. The Non-LF roads are true local roads that do not receive PennDOT funding. In these cases, maintenance is the responsibility of the local municipality. The Non-LF roads are not yet complete in the database yet, but PennDOT is currently in the process of attributing those geometries with road names, segment numbers, distances, and other information.

Data sources

RMS

The primary component that is used to define the RMS data—the linear referencing system—is defined by trucks equipped with a Distance Measuring Instrument (wheel on the back of the vehicle) that manually drives across all state-owned roads. Traffic engineers use this information to manage locations and measurements for roadway segments and attribution. Other, specific RMS data elements are collected and updated through the STAMPP program[2], through which engineering interns go out into the field and manually check/measure data elements on each segment.

ARNOLD

Data attributes for LF roads are also entered by hand into a tabular database using a cataloging application. These attributes are then linked to the roadway geometry using an Oracle GIS database. The tabular form of the data is updated as changes are received, and these changes are updated and merged nightly with the Oracle database. Typical changes include changing the category of a road from non-LF to LF status or splitting a segment.

Data for non-LF roadway segments are manually collected in the field using hand-held devices with a mobile app. This facilitates the creation of new segments in the field and GPS information is obtained by simply driving along these roadway segments.

Data custodians and access policies

RMS

The RMS data are owned and maintained by the pavement testing and asset management section in the Bureau of Maintenance and Operations within PennDOT. The majority of the data are publicly available through several websites, such as the PennDOT Open Data Portal (https://data-pennshare.opendata.arcgis.com) and the Pennsylvania Spatial Data Access (PASDA) Clearinghouse (http://www.pasda.psu.edu). Only a few data elements—such as bridge condition attributes or pavement

[2] https://www.penndot.gov/ProjectAndPrograms/ResearchandTesting/RoadwayManagementandTesting/Pages/VideoLog-and-STAMPP.aspx

friction data—are restricted and not provided to the public, but are available for use within PennDOT. These data are obtained and managed by those dealing with infrastructure maintenance within PennDOT. They are generally available for use as a part of a PennDOT project or can be obtained through submitting a Right-to-Know[3] request otherwise.

ARNOLD

The ARNOLD data are owned and maintained by the Bureau of Planning and Research within PennDOT. There is an end-user agreement in place to share data with others due to privacy concerns. If used internally within PennDOT, there is no issue with accessing or sharing data. Since this database is not yet complete, there is no information on whether portions of these data will be made available to the public.

Data elements

RMS

Besides RMS, data stored and managed in the spatial database include roadway geometry information, traffic information, pavement and shoulder history, maintenance history, municipal and legislative boundaries, intersections, roadside features, structure locations, railroad crossings information, pavement testing, condition survey information (including guide rail and drainage features), posting/bonding information, location of ITS devices, school and school district boundaries, and roadway projects. These data include any information that are currently included within the Federal Highway Performance Monitoring System (HPMS) database.

These data are often provided in themed data tables. For example, the RMSTRAFFIC table provides traffic information, the RMSINTERSECTION table provides the location of all intersections, RMSRRX provides the location of railroad crossings, and RMSPIPE provides the location and attributes of drainage pipe structures. Often, these data are contextualized using networks defined in the RMSADMIN file. These non-segment boundaries are defined by a segment/offset combination, which indicates the location within each segment where the network definition changes occur. Table 6 and Table 7 provide a summary of the specific data elements in the RMSADMIN and RMSSEG data files, respectively.

[3] **https://www.penndot.gov/ContactUs/Pages/Right-to-Know.aspx**

Table 6. Summary of data elements in RMSADMIN data file

Administrative information	**GIS data**
• PennDOT county code (not the commonwealth's) • PennDOT engineering district number • Maintenance functional class code • Jurisdiction or road ownership code • Road segment posted or bonded indicator • Federal identification code • Federal-aid system code • Federal-aid system status • Federal-aid urban area code • Federal information and processing standards code for urban areas • Federal functional class code • State route number	• Geometry (Oracle spatial data object) • MGE map ID • MGE link ID • Network linear feature (NLF) identifier for use in dynamic segmentation • Distance from start of NLF to segment begin point • Distance from start of NLF to segment end point **Roadway segment data** • Segment number at attribute start point • Segment number at attribute end point • Sequence number for ordering segments • Offset at attribute start point • Offset at attribute end point • Cumulative offset from the beginning of the route in the county to the begin point of the attribute record • Cumulative offset from the beginning of the route in the county to the end point of the attribute record • Roadway segment length in feet • Right/left side indicator • Daytime speed limit

Table 7. Summary of data elements in RMSSEG data file

Segment information

- Segment length
- Segment number
- Segment status (Active or under construction)
- Sequence number for ordering segments
- Beginning segment/offset of RMSADMIN record related to this segment
- Beginning segment/offset of RMSPAVEMENT record related to this segment
- Beginning segment/offset of RMSTRAFFIC record related to this segment
- Number of lanes on a segment (turning lanes, passing lanes exclusively)
- Current annual average daily traffic
- Number of high-occupancy vehicle (HOV) lanes
- Type of HOV lane operation
- Right/left side indicator

Pavement information

- Pavement surface type
- Pavement width
- Pavement survey date
- Pavement friction condition efficient
- Pavement friction survey date
- Pavement friction index number
- International roughness index (IRI) data
- IRI coefficient rating
- Overall pavement index
- Pavement distress survey cycle indicator (a cycle = 2 years)
- Pavement distress survey data switch
- Year of last resurfacing
- Year built
- Treatment type network
- Maintenance responsibility indicator

Location information (GIS information included)

- Mile point
- Begin terminus description
- End terminus description
- Cumulative offset from the beginning of the route in the county to the begin point of the attribute record
- Cumulative offset from the beginning of the route in the county to the end point of the attribute record
- Distance from start of NLF to segment begin point
- Distance from start of NLF to segment end point
- Network linear feature (NLF) identifier for use in dynamic segmentation
- X value begin
- X value end
- Y value begin
- Y value end
- MGE map ID (GIS generated)
- MGE link ID (GIS generated)
- Geometry (oracle spatial data object)
- Urban/rural code

Roadway information (name, functional classification, etc.)

- Street name
- Primary alternate street name
- Secondary alternate street name
- Directional indicator
- One-way indicator
- National Highway Planning Network segment indicator
- National Highway System segment indicator
- Federal-aid primary route segment indicator
- Federal highway performance monitoring system road segment sample site count
- Interstate network segment indicator
- Jurisdiction or road ownership
- Express way network segment indicator
- Toll indicator
- Government agency ownership
- PA scenic byway indicator
- Carriage-way for divided highway segment indicator
- RMS state route number
- Sub route (for sequential numbering of discontinuous sections)
- Traffic route number
- Primary alternate traffic route
- Secondary alternate traffic route
- Traffic route prefix (such as US and PA)
- Primary alternate traffic route prefix
- Secondary alternate traffic route prefix
- Traffic route number suffix
- Primary alternate traffic route suffix
- Secondary alternate traffic route suffix
- Truck parkway network indicator
- Truck route network indicator
- Business plan network
- Special purpose variable (open for district personnel to modify at their discretion)
- Parking lanes indicator
- Route direction
- Access control code

Roadway auxiliary

- Type of barrier or median on divided road segments
- Width of divisor
- Shoulder condition status
- Guardrail survey cycle for a given segment (a cycle = 4 years)
- Guardrail data switch
- Drainage survey cycle for a given segment (a cycle = 4 years)
- Drainage data switch

County and district

- PennDOT county code (not the commonwealth's)
- PennDOT engineering district

ARNOLD

The data elements included in the ARNOLD database are similar to those in the RMS database; however, all data are provided at the individual segment level, since no linear information is available within segments. The general elements are provided in Table 8.

Table 8. Summary of data elements in ARNOLD

Segment information

- Local road segment ID (primary key)
- GIS generated ID
- Road segment number
- Begin latitude of segment
- Begin longitude of segment
- End latitude of segment
- End longitude of segment
- Length of the current segment
- Additional information about segment
- Beginning offset of segment
- Ending offset of segment
- Segment divisor to separate branches or parts
- Segment curb indicator
- Segment sidewalks indicator
- Total number of travel lanes in the segment
- Parking lane width
- Total number of at-grade railroad crossings in the segment
- Width of right-of-way
- Roadway condition indicator

LF vs. non-LF road

- Liquid fuels eligibility indicator
- Liquid fuels indicator
- Length of bituminous surface
- Length of concrete surface
- Length of brick surface
- Length of gravel surface
- Length of seal coated surface
- Length of unimproved surface
- Paved lane length
- The predominant (>50%) type of pavement surface

Area-type

- Area type by population for the roadway segment
- Urbanized area type

Speed limits and traffic information

- Posted speed limit of the roadway segment
- General direction of travel
- Estimated average daily traffic for segment
- Year traffic count was taken
- Traffic pattern of segment
- Daily truck percent of traffic

County and municipality

- PennDOT 2-digit county code
- Municipality ID
- Name of municipality
- PennDOT county code and municipality code concatenation

Median and shoulder information

- Median type indicator
- Shoulder type
- Total shoulder width

Roadway information (type, ownership, and functional classification)

- Functional class code that best describes the roadway
- Road type (such as township or county)
- Beginning intersection road type
- Ending intersection road type
- Beginning intersection road name
- Ending intersection road name
- Township route number of the segment's beginning terminus
- Township route number of the segment's ending terminus
- State route number of the segment's beginning terminus
- State route number of the segment's ending terminus
- State road number of the segment's beginning terminus
- State road number of the segment's ending terminus
- Local road ID
- Local road name
- Local road type
- Roadway posted or bonded indicator
- Federal-aid eligibility indicator
- Federal-aid route number
- Non-state or federal-aid road
- Total number of municipal-owned bridges over 8 feet in length in the segment
- Jurisdiction or ownership of road

Current integration

RMS

As mentioned in the crash data section, RMS data are used to help identify crash locations along the state-owned roadway network using the linear referencing system. Crashes on state roads are located by the combination of county-route-segment (i.e., the specific segment) and offset within the segment, which allows the crash to be directly attributed to a specific segment. This facilitates the identification of specific infrastructure features (e.g., functional classification, geometric properties of the roadway, etc.) for each crash. Such a level of integration has been used in previous studies to develop safety performance functions for different roadway types in Pennsylvania that can predict reported crash frequency based on segment-level characteristics (Donnell et al., 2016, 2014).

ARNOLD

There is no linear referencing information included in the ARNOLD system that could be used to directly link crashes to individual local roadway segments. Instead, crashes would need to be merged with the ARNOLD database using the street name and other identifiers to determine the exact segment in which a crash occurred. Another strategy could be to map both the local roads and crashes on a GIS map and use coordinates to match crashes to the closest local roadway segment. Once the segment is identified in this way, the ARNOLD database could be used to access roadway data associated with that crash.

SWOT analysis

Strengths

- **Modern data system**: The RMS data are stored using an IBM IMS database, and data are outputted to a DB2 or Oracle database for reporting and spatial mapping purposes. The ARNOLD data are stored in a single Oracle database. While having these data stored in the same system would be preferred, these are still modern databases that facilitate the real-time queries and communications features required for integration purposes.
- **Data are well documented**: There exists a comprehensive data dictionary that describes the elements of this database. Furthermore, data elements are stored using a well-defined linear referencing system that is consistent across this and other data systems.
- **Many data elements are publicly available**: A large amount of the RMS data are currently available through the PennDOT Open Data Portal, which reduces barriers for potential integration. There does exist a subset of data elements that are not publicly available, including roadway condition information. However, these are generally available for use internally within PennDOT and thus should be available for safety data integration purposes.
- **Potential linkage variables are available**. The location of a crash can be linked to the roadway data files using the linear referencing system of the RMS database or the GIS mapping of the ARNOLD database. This location information appears to be the most straightforward method for linking records between the roadway and other data systems.

Weaknesses

- **ARNOLD database not fully implemented**: The ARNOLD database is currently under construction and the time schedule for when the elements in the database will be fully populated is unknown.
- **Inconsistency in location information across databases:** The RMS database uses a linear referencing system to identify locations along the roadway. However, there is no linear referencing information included in the ARNOLD database; instead, locations must be identified using street names or other landmarks or GIS mapping. This might serve as a barrier for integration, as two different sets of algorithms or protocols would be needed to merge both state (RMS) and local (ARNOLD) roadway information.
- **Recalibration routines may impact crash locations**: PennDOT's linear referencing system is updated through recalibration routines to maintain its accuracy. However, these recalibration efforts can potentially influence crash locations defined using this linear referencing information. If a segment is recalibrated, the locations of the corresponding crashes on that segment are not automatically updated in the crash database, potentially making that data inaccurate.
- **Historical information is not available:** If data elements are updated in the RMS database, the previous information is usually rewritten. Thus, the information that is pulled from the RMS database might not reflect actual roadway conditions at the time of the crash. Note, however, that RMS datafiles are generally available on an annual basis and that infrastructure-related elements are not likely to change within a small timeframe.
- **Data elements are not updated on a consistent timescale:** Most data elements in the RMS database are updated as needed. However, traffic volumes are updated at various schedules based on the functional classification of the roadway. For example, volumes on freeways are updated annually while volumes on other roadway types are updated less frequently depending on their functional classification. Furthermore, since the ARNOLD database is currently being developed, data are only being updated on a rolling basis.
- **Data are not available for private roadways:** The RMS database includes roadway data for state-owned or federal-aid roads, while the ARNOLD database includes roadway data for local roads. However, roadway information is not available for private roadways, even though records in the crash data system might reflect crashes on private roadways.

Opportunities

- **Roadway data are maintained by PennDOT**: This data system is maintained by PennDOT, which significantly reduces potential barriers to data integration or linkage.
- **Existing data exchange**: Data are already being shared between the roadway and crash data systems as previously described. This data exchange has existed for the last 30+ years. Therefore, well-defined avenues exist to share data between the crash data system and the RMS database.

Threats

- **No date provided for full implementation of ARNOLD:** No date was provided for the implementation of the ARNOLD database. This could become a barrier for integration of crash data with roadway data for local roads.

DRIVER AND VEHICLE DATA

This section describes the driver and vehicle data together, since these two databases share many similarities. The driver data are stored in PennDOT's Driver's License database and the vehicle data are stored in PennDOT's Vehicle Registration database. The information here was primarily obtained through an interview with two PennDOT employees: Amy Thompson (Data Services Section Manager) and Gary Kaskie (Fiscal Services Manager), both PennDOT employees. This section is organized into several subsections that describe the data structures and interface, sources used to obtain the data, data custodians and access policies, elements included in the database, and potential linkage variables.

Data structures and interfaces

Both the driver's license and vehicle registration databases consist of multiple subsystems stored on an IBM mainframe database using COBOL 2. Both are older systems (the vehicle database was implemented in 1986, while the driver's license database was implemented in 1990). PennDOT's Driver and Vehicle Services is currently in the process of building a new system that will accommodate both the driver's license and vehicle registration databases. This process is currently in year 2 of a 6- to 7-year time period. The new system would be a more modern web-based system that could be accessed in Windows (as opposed to only available on a mainframe).

The driver's license database is made up of several subsystems. Each of these is linked to the others using an individual's unique driver's license number. The main subsystems include:

- Customer (root)
- Transaction history
- Exam
- Product
- Traffic safety

The customer subsystem provides personal information for each driver. The transaction history subsystem provides historical information on all transactions that the driver has performed through license services. The exam subsystem provides detailed information on the individual's driver's license exam. The product subsystem provides detailed information on the specific products (e.g., license, photo ID, permit) assigned to an individual. The traffic safety subsystem provides information on license suspensions and violations. Data from each of these systems are extracted daily into a DB2 database for reporting purposes.

The vehicle registration database is also made up of two main subsystems, linked to each other using the vehicle's title number. These are:

- Vehicle (root)
- Suspension

The vehicle subsystem contains information about the vehicle that would be on its title. The suspension subsystem includes other information such as violations and vehicle reconstruction. Data are extracted from this system regularly for reporting purposes. The exact frequency was not known.

Both driver's license and vehicle registration databases were updated mostly in real time.

Data sources

Basic information in the license and vehicle registration databases is primarily obtained from the forms customers fill out when applying for a license or a vehicle title. These are verified and entered directly by PennDOT Driver and Vehicle services employees and contracted business partners (e.g., for driver's license, exams are contracted out to conduct tests and post results). Citation information is obtained electronically from the Administration Office of Pennsylvania Courts (AOPC) on a daily basis for all Title 75 traffic citations. Crash information is provided directly by the crash data system and includes a portion of the crash record number, driver's license number, county where the crash was located, vehicle type, and crash severity.

Data custodians and data access policies

Data are owned by PennDOT Driver and Vehicle Services, which also maintains the data and ensures its security. Access to data is only allowed for PennDOT employees following PA regulation and federal laws for PII. Other governmental organizations can also obtain access to the information by establishing MOUs (if none already exist) among PennDOT and other interested parties.

Data elements

Table 9 and Table 10 provide a summary of the main data elements included in the various subsystems of the driver's license and vehicle registration data, respectively.

Table 9. Summary of data elements in driver’s license database

Customer (root) • Driver name • Driver license number • Driver license class • License expire date • Driver’s address (city/state/zip) • Driver’s date of birth (DOB) • Date of proof • Driver’s social security number (SSN) • Driver’s height, sex, eye color • Organ donation indicator • Driver’s privilege • Driver’s medical restrictions • Record type • Duplicates of license • Endorsements • Commercial driver license (CDL) restrictions • CDL medical self-certificate • CDL medical certificate status (such as non-certified) • Out-of-state (OOS) restriction code • Insurance status, policy number, and expiry date • List of inquiry history **Transaction history** • Driver license number • Driver’s address (city/state/zip) • Driver’s DOB • License status code • Medical restrictions • Total points • Commercial driver license (CDL) restrictions • Habitual offender (HO) indicator	**Exam** • Driver license number • Exam type • Test type • Exam due date • Fail count • Passed date **Product** • Driver license number • Driver’s privilege • Driver’s DOB • Product status • Physical exam date • Medical restrictions • Commercial driver license restrictions • Duplicates • Endorsements • Product class (such as C) • Issue date and expire date • Location where driver license photo was taken • Date when driver license photo was taken • Driver license photo type **Traffic safety** • Driver license number • Driver’s DOB • Driver’s privilege • Interlock control • DUI • Six-point control • Suspension control • Total points • Commercial driver license (CDL) driver qualification (failed) • CDL major control • CDL serious traffic offense (STO) • Violation information (date, type, conviction date, action, etc.)

Table 10. Summary of data elements in vehicle registration database

Vehicle (root)

- Vehicle identification number (VIN)
- Re-issued VIN
- Owner or lessee name and address
- Title
- Title sequence
- Title/registration date
- Title expire date
- Title duplicates
- Non-PA title
- Registration duplicates
- Registration years
- Registration fee
- Registration provider
- Tag
- Tag type
- Primary tag
- Tag color code
- Make
- Model
- Class
- Body
- Vehicle make year
- Purchase date
- Number of seats
- Fuel (such as gasoline)
- Overall tires
- Use
- Privilege use
- Local use fee
- Odometer
- Odometer qualification
- Axle weight rating (AWR)
- Axle transaxle (Axle TX)
- Axles
- Active rollover protection (ARP)
- Automobile air conditioning (A/C)
- Equipment number

Vehicle (root) (cont)

- Unloaded vehicle weight (Unl Wt)
- Gross vehicle weight (GVW)
- Gross vehicle weight rating (GVWR)
- Gross combined weight (GCWT)
- Gross combined weight rating (GCWR)
- Previous renewal date
- Dealership
- Van indicator
- Established work identification number (WID: identifies the document being processed)
- Renewal WID
- Vehicle WID records
- County
- State of origination
- Retired person indicator
- Disable veteran indicator
- Grey market vehicle indicator
- Lemon vehicle indicator
- Vehicle emission inspection and maintenance (I/M) credit
- Vehicle emission I/M required
- List of vehicle emission I/M records (date, test type, station number, odometer reading, etc.)
- List of vehicle lien holder information

Suspension

- VIN
- Unclaimed vehicle indicator
- Stolen vehicle date
- Stolen tag date
- Vehicle purged date
- Junk indicator
- Abandoned indicator
- Suspension/recycle/theft indicator

Current integration

As previously mentioned in the crash data section, the driver's license and vehicle registration databases are used to obtain/verify information from police crash reports, such as VINs for vehicles involved in the crash or information about drivers involved in the crash. Reciprocally, information about crashes involving PA drivers are sent to the driver's license database and added to a driver's record using the driver's license ID number. This information includes the accident record number (last 7 digits of the CRN), county of the crash, vehicle type, injury severity, and crash date. Crash information for individual vehicles is also collected, but not stored in the vehicle registration database. Instead, this information is usually queried from the crash database and provided as reports to private companies (e.g., CarFax or Experian).

SWOT analysis

The driver and vehicle data systems are discussed jointly, since these two databases share many similarities and information was obtained from a single information source that represented both systems.

Strengths

- **Comprehensive data dictionaries**: Both the driver and vehicle databases have a data dictionary/glossary that describes each of the data elements that are included. This is required to successfully link or integrate data files.
- **Many potential linkage variables are available:** The driver and vehicle systems contain multiple variables that can be used to link with the crash databases. For the driver data these include PII such as their name, date of birth, age, gender, driver's license number, and address. For the vehicle database, these include license plate number, VIN, and vehicle year, make, and model.

Weaknesses

- **Outdated databases**: The driver and vehicle data systems are stored on an older IBM mainframe database using COBOL2. These systems were implemented over 30 years ago and thus may not be able to communicate as easily with more modern data systems, which serves as a potential barrier to integration.
- **Data structure is unknown**: The structure of the database—specifically, how data are linked between individual driver and vehicle records—is unclear. No information on this was available from the representative interviewed.
- **Historical information missing for some data elements**: It was specifically noted that the vehicle data system does not contain a historical record for some data elements that are vital for data integration purposes. For example, only the most recent vehicle that is assigned to a specific license plate number is stored in the database; the list of vehicles that might have been assigned this license plate number before the current vehicle is not available. This leads to errors in validating vehicle information if a license plate is moved between the time of a crash and the time when the crash report is entered into the crash data system.
- **Data primarily only available for Pennsylvania:** Both the driver and vehicle databases only have detailed information for drivers with a Pennsylvania license or vehicles registered within Pennsylvania. There is limited data sharing for drivers through the State-to-State Verification Service; however, this is primarily used to ensure that an individual does not have a valid license in more than one state and allows states to verify licenses or identification cards from other states.

Opportunities

- **Driver and vehicle data are maintained by PennDOT**: This data system is maintained by PennDOT's Driver and Vehicle Services, and an existing MOU is in place to share data—including private data—internally within PennDOT, which significantly reduces potential barriers to integration.
- **New database being developed**: A new, modern database is being developed by the Driver and Vehicle Services to store driver and vehicle data. This will replace the outdated system currently being used and is likely to better facilitate data integration.
- **Existing data exchange**: As previously mentioned in the crash data system SWOT analysis, data from the driver and vehicle systems are shared with the crash data system for verification purposes. The driver data system also contains the crash report number and date of any crash associated with

a specific driver. Additionally, citation information is obtained from the court system when a traffic citation case is disposed or when an individual does not respond to a citation and must have their license suspended. This information is stored in a driver's record in the driver data system.

- **Data reports made regularly for outside entities**: Data are regularly shared with non-governmental entities from the vehicle database. This includes reports to private companies on crash history of individual vehicles. This suggests alternate pathways for integration may be possible.

Threats

- **No technical details available on new databases**: Although the new database for driver and vehicle information will be more modern and likely to better facilitate data integration, technical details on this new system were not provided by the driver and vehicle data system representative for this project. This lack of detail means that future integration possibilities are currently unknown.

CITATION AND ADJUDICATION

This section describes the Citation and Adjudication databases. Mark Rothermel, Administration Office of PA Courts, Lisa Polonia (IT Trainer for Magisterial District Judges) and Ami Levin (Data Exchange Manager) provided the majority of the information contained in this section of the report. This section is organized into several subsections that describe the data structures and interface, sources used to obtain the data, data custodians and access policies, elements included in the database, and potential linkage variables.

Data structures and interfaces

The citation and adjudication data are stored in two main systems: one that covers the roughly 550 Magistrate District Judges and another than covers the 67 County Courts of Common Pleas. These systems are proprietary systems that were developed in-house by the AOPC to serve the courts. Each of the two main systems consist of thousands of individual tables that are interrelated and linked using indicators such as individual's name, social security number (SSN), citation number, court docket number, and others. However, the data can be compiled and exported as one singular summary table, if desired. Note that the Philadelphia Courts act independently of the AOPC and thus their data are not included in these data systems.

Data sources

Citation information comes directly from the traffic citation form that is filled out by an officer at the time the citation is issued. Pennsylvania State Police and close to 200 local agencies report this data electronically to the AOPC through an e-filing process and courts access this information through the Pennsylvania Justice Network (PJN) (https://www.pajnet.pa.gov/), which is an online portal that provides a common environment for authorized users to access public safety and criminal justice information. However, the court system is responsible for ensuring the accuracy and completeness of the data. Non-electronic records are provided to the courts, which then put this information into the database. Adjudication information is obtained directly from the court system after a case is complete.

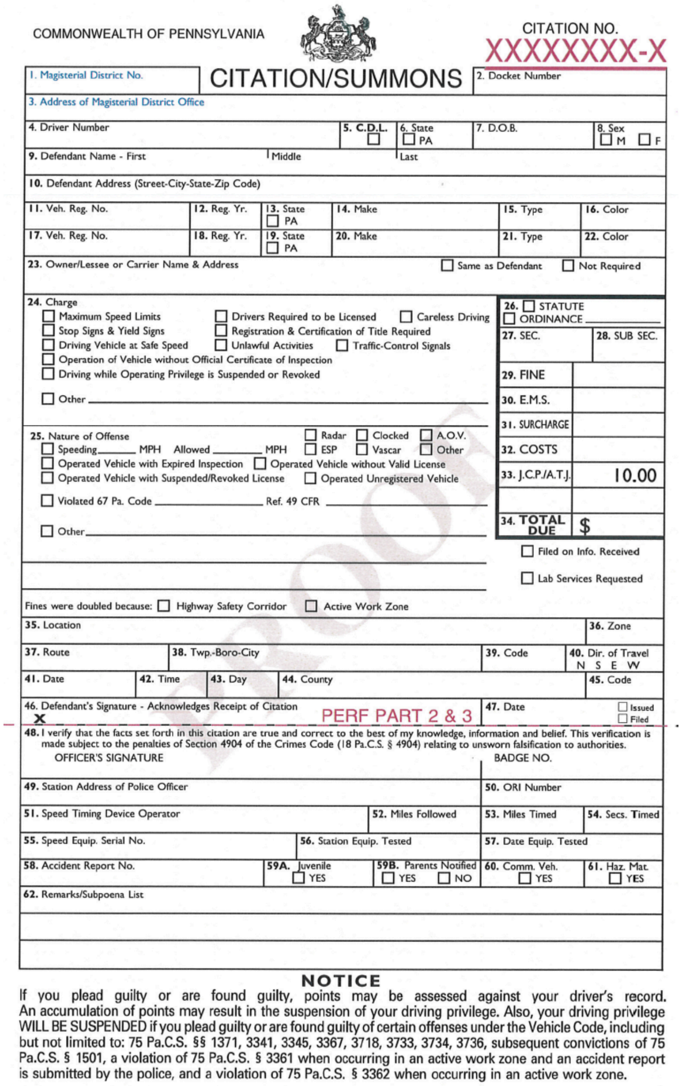
Data custodians and access policies

The Administrative Office of the Pennsylvania Courts essentially “owns” these data and controls which agencies or individuals can access the data. Direct access is restricted to court users, who can perform queries such as searching for all records for a specific individual. However, select citation and adjudication information (including individual court docket sheets for specific offences) can be accessed through the PJN or the Unified Judicial System of Pennsylvania Web Portal[4]. These systems can be queried using information such as the citation number, court docket number, and individual’s name, among other fields. The court system controls what is placed on these “public” pages, as well as what information is redacted. For example, individuals’ names are included on these documents—since these are public data—but addresses are redacted.

Data elements

Data elements of interest that would be stored in these data include anything that is contained on the paper traffic citation; see Figure 4 for an example. A summary of these data elements is provided in Table 11. Note that SSNs are not collected for traffic citations and thus are only available in the citation and adjudication system if the individual was in the system for another offense. In this case, the SSN would be automatically linked to the traffic citation record as well.

[4] https://ujsportal.pacourts.us/



COMMONWEALTH OF PENNSYLVANIA

CITATION NO.
XXXXXXXX-X

CITATION/SUMMONS

1. Magisterial District No.

2. Docket Number

3. Address of Magisterial District Office

4. Driver Number | 5. C.D.L. □ | 6. State □ PA | 7. D.O.B. | 8. Sex □ M □ F

9. Defendant Name - First | Middle | Last

10. Defendant Address (Street-City-State-Zip Code)

11. Veh. Reg. No. | 12. Reg. Yr. | 13. State □ PA | 14. Make | 15. Type | 16. Color

17. Veh. Reg. No. | 18. Reg. Yr. | 19. State □ PA | 20. Make | 21. Type | 22. Color

23. Owner/Lessee or Carrier Name & Address □ Same as Defendant □ Not Required

24. Charge
□ Maximum Speed Limits □ Drivers Required to be Licensed □ Careless Driving
□ Stop Signs & Yield Signs □ Registration & Certification of Title Required
□ Driving Vehicle at Safe Speed □ Unlawful Activities □ Traffic-Control Signals
□ Operation of Vehicle without Official Certificate of Inspection
□ Driving while Operating Privilege is Suspended or Revoked
□ Other ____________________

26. □ STATUTE
□ ORDINANCE ____________

27. SEC. | 28. SUB SEC.

29. FINE	
30. E.M.S.	
31. SURCHARGE	
32. COSTS	
33. J.C.P./A.T.J.	10.00
34. TOTAL DUE	$

25. Nature of Offense □ Radar □ Clocked □ A.O.V.
□ Speeding_______ MPH Allowed ________ MPH □ ESP □ Vascar □ Other
□ Operated Vehicle with Expired Inspection □ Operated Vehicle without Valid License
□ Operated Vehicle with Suspended/Revoked License □ Operated Unregistered Vehicle
□ Violated 67 Pa. Code ______________ Ref. 49 CFR ______________
□ Other____________________

□ Filed on Info. Received

□ Lab Services Requested

Fines were doubled because: □ Highway Safety Corridor □ Active Work Zone

35. Location | 36. Zone

37. Route | 38. Twp.-Boro-City | 39. Code | 40. Dir. of Travel N S E W

41. Date | 42. Time | 43. Day | 44. County | 45. Code

46. Defendant's Signature - Acknowledges Receipt of Citation
X
PERF PART 2 & 3

47. Date □ Issued □ Filed

48. I verify that the facts set forth in this citation are true and correct to the best of my knowledge, information and belief. This verification is made subject to the penalties of Section 4904 of the Crimes Code (18 Pa.C.S. § 4904) relating to unsworn falsification to authorities.
OFFICER'S SIGNATURE
BADGE NO.

49. Station Address of Police Officer | 50. ORI Number

51. Speed Timing Device Operator | 52. Miles Followed | 53. Miles Timed | 54. Secs. Timed

55. Speed Equip. Serial No. | 56. Station Equip. Tested | 57. Date Equip. Tested

58. Accident Report No. | 59A. Juvenile □ YES | 59B. Parents Notified □ YES □ NO | 60. Comm. Veh. □ YES | 61. Haz. Mat. □ YES

62. Remarks/Subpoena List

NOTICE

If you plead guilty or are found guilty, points may be assessed against your driver's record. An accumulation of points may result in the suspension of your driving privilege. Also, your driving privilege WILL BE SUSPENDED if you plead guilty or are found guilty of certain offenses under the Vehicle Code, including but not limited to: 75 Pa.C.S. §§ 1371, 3341, 3345, 3367, 3718, 3733, 3734, 3736, subsequent convictions of 75 Pa.C.S. § 1501, a violation of 75 Pa.C.S. § 3361 when occurring in an active work zone and an accident report is submitted by the police, and a violation of 75 Pa.C.S. § 3362 when occurring in an active work zone.

Figure 4. Example of paper traffic citation

Table 11. Summary of data elements obtained from traffic citations

Citation information • Citation number • Magisterial district number • Docket number • Address of magisterial district office **Driver/defendant information** • Driver (license) number • Commercial driver license indicator • Driver license State (PA or not) • Driver date of birth • Driver sex • Defendant name • Defendant address • Defendant signature and date • Juvenile only (parent notified: Y/N) **Vehicle information** • Vehicle registration number • Vehicle registration year • Vehicle registration state • Make • Type • Color • Owner/Lessee or carrier name & address • Commercial vehicle indicator • Hazardous material **Charge indicators** • Maximum speed limits • Stop sign & yield signs • Driving vehicle at safe speed • Operation of vehicle without official certificate of inspection • Driving while operating privilege is suspended or revoked • Driver required to be licensed • Registration & certification of tide required • Unlawful activities • Careless driving • Traffic-control signals • Other	**Nature of offense** • Speeding (_mph, allowed _mph) • Operated vehicle with expired inspection • Operated vehicle with suspended/revoked license • Operated vehicle without valid license • Violated 67 Pa. code (ref. 49 CFR) • Other **Fine** (fines were doubled because highway safety corridor or active work zone) • Statute • Ordinance • Section • Sub section • Fine • EMS • Surcharge • Costs • Total due **Violation information** • Location • Zone • Route • Township/borough/city and code • Direction of travel • Date • Time • Day • County and code **Police agency** • Officer signature and badge number • Station address of police officer • ORI number • Speed timing device operator • Miles followed • Miles timed • Seconds timed • Speed equipment serial number • Station equipment tested • Date equipment tested • Accident report number

Current integration

Although citation and adjudication data are not directly linked to other data systems, there are avenues for limited integration between law enforcement/the court system and PennDOT available through the PAJNET portal. Specifically, Pennsylvania police officers can use this portal to obtain detailed driver and vehicle information using limited identifiers such as license plate or license number. Officers can also access driver's license images from PennDOT through the JNET photo search. These photos can also be used for facial recognition purposes using the Pennsylvania Chiefs of Police Facial Recognition System (JFRS). Real-time notifications are also available through PAJNET for offenders on watch lists, including change of address information provided to PennDOT. PAJNET also allows users to access PennDOT's list of expired and revoked driver's licenses and vehicle registrations, which are obtained directly from the associated data systems. Vehicle inspection and emissions information are also provided by PennDOT to enforcement officials as a means to identify fraudulent vehicle inspection stickers. Lastly, PennDOT provides PAJNET users with certified vehicle records. While there appears to be ample sharing of PennDOT data with law enforcement/the court system, it is not clear how often the PennDOT data available in the PAJNET are updated and/or if PAJNET users have direct access to the driver's license or vehicle registration data systems.

Citation and adjudication data

Strengths

- **Modern data system**: The citation and adjudication data are stored on proprietary databases developed in-house by the AOPC. Although proprietary, these data systems are modern and facilitate online access and sharing of data electronically and automatically, which is useful for data integration.
- **Well-defined data highway**: Data are already shared electronically to, from, and within the AOPC using the Pennsylvania Justice Network. This online portal provides a common environment for authorized users—with varying levels of access—to access criminal justice information.

Weaknesses

- **Complex and proprietary data tables**: Due to the proprietary nature of the database used to store these data, information on the data structure is not available. It was noted that the data are stored over thousands of individual tables and it was impossible to identify how each of these is linked for individual citation records. A complete data dictionary was also not available.
- **No consistent timeline for data updates:** Data are updated on a rolling horizon based on when a citation is made and entered into the court system, when an individual responds, and when a disposition is obtained. Thus, the time lag for individual records varies. It was noted that individuals have 10 days to respond to the court system for a citation and this occurs for approximately 85% of cases. However, there could be significant lags with some of the records based on individual responses and time for a disposition to be reached.
- **Records may not be kept indefinitely**: Data on individual citations are only required to be kept for three years by law. After that, hard copies may be destroyed and electronic records may be purged. This decision is up to each individual court, so historical records in the database may not be consistent across the state.
- **Database not comprehensive**: While the database contains all (traffic and non-traffic) cases for the approximately 550 Magistrate District Judges and 67 County Courts of Common Pleas that comprise the AOPC, Philadelphia operates independently. Records from cases in the Philadelphia court system are thus not included in the statewide database. Additionally, some traffic-related offenses—like parking citations—are handled locally and are also not included in these data.
- **Little desire for integration**: Representatives of the citation and adjudication database indicated very little desire for integration with crash or PennDOT data (outside what is already being performed). It was specifically mentioned that some pertinent information—such as history of driver or crash records—would not be able to be viewed or used by judges or court personnel. No other information was noted as being especially helpful at this time.

Opportunities

- **Existing data exchange with PennDOT**: As previously mentioned, data sharing already takes place with the driver data system when a disposition is made on a traffic case. Various data elements—including name, date of birth, driver license number, some limited demographic information, the offense and the outcome—is sent to be recorded in the driver's license database. Additionally, when a license suspension is warranted, this information is also shared. The court system also has tools available through the Pennsylvania Justice Network to access driver's license photos directly from PennDOT when a warrant is issued, lists of expired or revoked vehicle registrations or driver's licenses, vehicle inspection/emissions information, and certified vehicle records.

Threats

- **Lack of communication with PennDOT**: Representatives of the citation and adjudication database indicated that there might be a lack of communication within PennDOT about the types of data currently being shared between the Pennsylvania Courts and PennDOT. They indicated that this could serve as a barrier for data integration.
- **Location information is not consistent with PennDOT's linear referencing system:** The citation database includes several fields related to the location of a traffic citation, including route, direction of travel, county, and township/borough/city. In addition, there is another field in which officers can provide more detailed location information (e.g., a specific intersection or location along a route). However, the information in this latter field is not consistent with PennDOT's linear referencing system, which limits the ability to merge the citation and roadway databases.
- **PennDOT data systems**: Representatives of the citation and adjudication database noted that PennDOT uses an antiquated system for its driver's license database, and that this serves as a barrier for data integration.
- **Data only provided when disposition is made**: Only disposition records are sent to PennDOT, which can create time lags if drivers do not respond to or acknowledge citations. Even for the cases with no time lags, the time between the day of citation and the time that PennDOT receives that data can be more than 30 days.

INJURY SURVEILLANCE

This section describes the injury surveillance system, which contains data on injury outcomes from hospitals and associated services. In Pennsylvania, the primary injury surveillance systems are the Emergency Medical Services (EMS), Vital Statistics, Trauma and Health Care. Unfortunately, a contact from Vital Statistics was not available for an interview for this project; however, the other systems were reviewed for this report. The information in this section of the report was primarily obtained through interviews with:

- EMS: Cathy Curley (Program Analyst with the Pennsylvania Bureau of EMS), Kaylen Irwin (Office Operation Specialist with the Emergency Medical Management Cooperative), and Aaron Rhone (EMS Program Manager at the Pennsylvania Department of Health)
- Health Care: JoAnne Nelson (Supervisor of Special Requests with the Pennyslvania Health Care Cost Containment Council (PHC4))
- Trauma: Juliet Altenburg (Executive Director), Lyndsey Diehl (Manager of Data Quality), Stephanie Radzevick (Trauma Data Analyst), Tom Wasser (Research Consultant), Patrick Reilly

(Board Chairman), and Niels Martin (Research Committee Chairman), all with the Pennsylvania Trauma Systems Foundation (PTSF), as well as Kathleen Yetter (Director of Marketing) and Glendene Strickland (Senior Analyst) of Digital Innovation, Inc., which is the vendor that manages the Trauma database.

This section is organized into several subsections that describe the data structures and interface, sources used to obtain the data, data custodians and access policies, elements included in the database, and potential linkage variables.

Data structures and interfaces

EMS

All PA EMS agencies use one of about 11 different vendors to collect and store data into a Patient Care Report (PCR) for all patients served by Emergency Medical Services. The Pennsylvania Department of Health (PA DOH) then collects specific data elements from the PCRs and copies them into one main EMS database. This database is cloud-based and can be accessed from anywhere using the internet. The data are housed through CloudPCR under a contract with EMMCO West. No further information about the specific data protocols or nature of the database itself was available from the interviewees.

Health Care

No information was provided on the specific data structures or interfaces for the Health Care data. However, a data sharing agreement exists which suggests that these data can be exported in a format used for reporting and/or merging with other data systems.

Trauma

The Pennsylvania Trauma Systems Foundation ("The Foundation" or PTSF) is the accrediting body for trauma centers throughout the Commonwealth of Pennsylvania. The foundation was created by the combined efforts of the Pennsylvania Medical Society and the Hospital & Healthsystem Association of Pennsylvania, along with the Pennsylvania State Nurses Association, the Pennsylvania Emergency Health Services Council, and the Pennsylvania Department of Health. The Commonwealth of Pennsylvania first recognized the foundation in December 1984 when Act 209 was signed into law by Governor Thornburgh. Act 209 expired in June 1985 and a comprehensive Emergency Medical Service Act (Act 45) was signed into law in July 1985, again recognizing the Pennsylvania Trauma Systems Foundation as the accrediting body for the trauma centers in Pennsylvania. The PTSF database is focused on collecting data to improve trauma performance.

Trauma data are stored in one of two datasets: a "Full Dataset," meaning all elements in the database, or an "Excluded Dataset." The excluded dataset is smaller because it does not contain EMS information or referring facility data. These Trauma data contribute to the national trauma databank for national statistics. No further information about the specific data protocols or nature of the database itself was available from the interviewees.

Data sources

EMS

EMS personnel complete the PCR throughout the process when EMS services are needed. This information is entered into an agency-specific software (one of 11 vendors) during this time. The PA DOH EMS database pulls a subset of these data elements directly from these individual PCRs provided by the EMS agencies electronically. While the data elements within the PCR differ based on the vendor software, the information pulled into the PA EMS database includes only common variables across the different PCRs, although formatting might not be the same.

Health Care

All inpatient discharge records, except skilled nursing facility, swing bed, transitional care unit, 23-hour observation, and hospice records, are required to be submitted from all Pennsylvania hospitals directly to the PHC4. Edits are applied during hospital data submissions and the hospital will receive an error report identifying records that do not pass the edit criteria.

Trauma

The PTSF represents 40 accredited trauma centers and has data submission expectations. All centers are required to use the same software and data definitions. All accredited trauma centers are required to submit data to the Pennsylvania Trauma Outcome Study (PTOS) in Pennsylvania. The submitted data are de-identified subcomponents of the accredited trauma center's medical record on the trauma patient for that particular patient stay. No data are collected from non-trauma centers in the PTOS. Data collected in the PTOS are comprised of: demographics, pre-hospital information, process of acute care information, clinical data, outcome data, diagnoses, procedures, and payer class information. Approximately 85% of centers reported data within 42 days of discharge. There are additional unaccredited centers that provide data.

Data custodians and access policies

EMS

Individual EMS agencies own the information included in all PCRs. However, the datasets stored by CloudPCR are the property of the Bureau of EMS and the data can only be accessed by authorized representatives of the DOH. Access to this database must be requested directly from the DOH. Access requests must include dates, regions/locations, and reasons for the requests, which are then approved or denied by the DOH after PII is removed. If fewer than five incidents are reported within any one zip code, those specific data are removed from any data sharing due to privacy concerns (specifically, concern that a user could potentially identify a patient using the records), which occurs due to the agreement between the EMS agencies and PA DOH.

Health Care

The PHC4 provides redacted versions of the Health Care data to users when deemed appropriate. These redacted datasets typically have PII and other information that can be used to identify specific patients (e.g., dates). To apply for PHC4 data, applicants must complete PHC4's Data Request Application[5], sign a Confidentiality Data Use Agreement (DUA), and submit these to the PHC4 Special Requests. Agency-level requests are also possible; e.g., the PA DOH has completed an Application – Government document, which contains different language in the DUA. This DUA must be signed by the Applicant and all users who need access to the data PHC4 provides. The Applicant is not permitted to share the data or use it for any other purpose. PHC4 facilitates extensive and regular data integration due to the large amount of data collected.

From the hospital discharge data that are collected, PHC4 produces Health Care public use data files and makes them available for research projects and studies. Hospital discharge records are processed based on the date of discharge and data files are released quarterly. Therefore, a 2017 data file will include all patients who were discharged from a Pennsylvania hospital during 2017, but some of these patients were admitted during 2016.

Trauma

Data access is limited to the PTSF and all accredited trauma centers in the Commonwealth. Outside groups wishing to use the data for non-trauma purposes must undergo a review within PTSF for data use agreement and potentially PTSF access to the penultimate dataset. Access is based on request and applications[6] are reviewed as they are received. The foundation's research committee meets quarterly to review applications. Access to the statewide PTOS data is by application only.

Data elements

EMS

Table 12 provides a summary of some of the data elements obtained from the PCRs of various EMS agencies by the PA DOH. The complete list of data elements and codes is too extensive to present here, but a data dictionary exists for the EMS database.

[5] http://www.phc4.org/services/datarequests/howtosubmit.htm#sr_applications

[6] http://ptsf.org/docs/request/

Table 12. Summary of data elements in the EMS database

Agency data

- EMS agency unique state ID
- EMS agency number
- EMS agency name
- EMS agency state
- EMS agency service area states
- EMS agency service area counties
- EMS agency census tracts
- EMS agency service area zip codes
- Primary type of service
- Level of service
- Organization status
- Organizational type
- EMS agency organizational tax status
- Statistical calendar year
- Total primary service area size
- Total service area population
- 911 EMS call center volume per year
- EMS dispatch volume per year
- EMS patient transport volume per year
- EMS patient contact volume per year
- National provider identifier
- Fire department ID number

Configuration data

- State associated with the certification/licensure levels
- State certification/licensure levels
- Procedures permitted by the State
- Medications permitted by the State
- Protocols permitted by the State
- EMS certification levels permitted to perform each procedure
- EMS agency procedures
- EMS certification levels permitted to administer each medication
- EMS agency medications
- EMS agency protocols
- EMS agency specialty service capability
- Emergency medical dispatch provided to EMS agency service area
- Patient monitoring capability(ies)
- Crew call sign

Vehicle data

- Unit/vehicle number
- VIN
- EMS unit call sign
- Vehicle type
- Crew state certification/licensure levels

Personnel data

- EMS personnel's last name
- EMS personnel's first name
- EMS personnel's middle name/initial
- EMS personnel's State of licensure

Situation data

- Date/time of symptom onset
- Possible injury
- Chief complaint anatomic location
- Chief complaint organ system
- Primary symptom
- Other associated symptoms
- Provider's primary impression
- Provider's secondary impressions
- Initial patient acuity

Injury data

- Cause of injury
- Trauma center criteria
- Vehicular, pedestrian, or other injury risk factor
- Airbag deployment

Arrest data

- Cardiac arrest
- Cardiac arrest etiology
- Resuscitation attempted by EMS
- Arrest witnessed by
- CPR care provided prior to EMS arrival
- Who provided CPR prior to EMS arrival
- AED use prior to EMS arrival
- Who used AED prior to EMS arrival
- Type of CPR provided
- First monitored arrest rhythm of the patient
- Any return of spontaneous circulation
- Date/time of cardiac arrest
- Date/time resuscitation discontinued
- Reason CPR/resuscitation discontinued
- Cardiac rhythm on arrival at destination
- End of EMS cardiac arrest event
- Date/time of initial CPR

History data

- Barriers to patient care
- Advance directives
- Alcohol/drug use indicators

Vital data

- Date/time vital signs taken
- Obtained prior to this unit's EMS care
- Cardiac rhythm/electrocardiography (ECG)
- ECG type
- Method of ECG interpretation
- Systolic blood pressure
- Diastolic blood pressure
- Method of blood pressure measurement
- Heart rate
- Pulse oximetry
- Respiratory rate
- Respiratory effort
- End tidal carbon dioxide (ETCO2)

- EMS personnel's Sate's licensure ID number
- EMS personnel's State EMS certification licensure level

Record data
- Patient care repost number
- Software creator
- Software name
- Software version

Response data
- EMS agency number
- EMS agency name
- Incident number
- EMS response number
- Type of service requested
- Primary role of the unit
- Type of dispatch delay
- Type of response delay
- Type of scene delay
- Type of transport delay
- Type of turn-around delay
- EMS vehicle (unit) number
- EMS unit call sign
- Level of care of this unit
- Repines mode to scene
- Additional response mode descriptors

Dispatch data
- Complaint reported by dispatch
- EMD performed

Crew data
- Crew member ID
- Crew member level
- Crew member response role

Time data
- PSAP call date/time
- Unit notified by dispatch date/time
- Unit enroute date/time
- Unit arrived on scene date/time
- Arrived at patient date/time
- Transfer of EMS patient care date/time
- Unit left scene date/time
- Patient arrived at destination date/time
- Destination patient transfer of care date/time
- Unit back in service date/time
- Unit canceled date/time

Patient data
- Patient's home county
- Patient's home State
- Patient's ZIP code
- Gender
- Race
- Age
- Age units

- Blood glucose level
- Glasgow coma score-eye
- Glasgow coma score-verbal
- Glasgow coma score-motor
- Glasgow coma score-qualifier
- Total Glasgow coma score
- Level of responsiveness (AVPU)
- Pain scale score
- Pain scale type
- Stoke scale score
- Stroke scale type
- Reperfusion checklist
- Revised trauma score

Exam data
- Mental status assessment
- Neurological assessment
- Stroke/CVA symptoms resolved

Protocol data
- Protocol used
- Protocol age category

Medication data
- Date/time medication administered
- Medication administered prior to this unit's EMS care
- Medication given
- Medication administered route
- Medication dosage
- Medication dosage unit
- Response to medication
- Medication complication
- Medication crew (healthcare professionals) ID
- Role/type of person administering medication
- Medication authorization

Procedure data
- Date/time procedure performed
- Procedure performed prior to this unit's EMS care
- Procedure
- Number of procedure attempts
- Procedure successful
- Procedure complication
- Response to procedure
- Procedure crew members ID
- Role/type of person performing the procedure

Disposition data
- Destination/transferred to code
- Destination type
- Destination State
- Destination county
- Destination ZIP code
- Incident/patient disposition
- EMS transport method
- Transport mode from scene
- Additional transport mode descriptors

Payment data • Primary method of payment • CMS service level **Scene data** • First EMS unit on scene • Type of other service at scene • Date/time initial responder arrived on scene • Number of patients at scene • Mass casualty incident • Triage classification for MCI patient • Incident location type • Incident city • Incident State • Incident ZIP code • Incident county	• Final patient acuity • Reason to choosing destination • Type of destination • Hospital in-patient destination • Hospital capability • Destination team pre-arrival alert or activation • Date/time of destination pre-arrival alert or activation **Outcome data** • Emergency department disposition • Hospital disposition **Others** • Potential system of care/specialty/registry patient • Suspected EMS work related exposure, injury, or death • Crew member completing this report

Health Care

Table 13 provides a summary of some of the data elements obtained by the PHC4 for its Health Care database.

Table 13. Summary of data elements in the Health Care database

Inpatient discharge data

- Record identification
  - System assigned unique record sequence number
  - Processing year
  - Processing quarter
- Facility identification
  - Pennsylvania Facility number
  - Facility region code
  - MAID
- Patient data
  - Patient sex code
  - Hispanic/Latino Origin or Descent
  - Race code
  - Pseudo patient identifier
  - Patient age in years
  - Patient age in days
  - Patient zip code
  - Home market share area code
  - Patient home county code
  - Patient state code
  - Accident state
- Admission data
  - Priority (type of visit)
  - Point of origin for admission or visit
  - Admission hour
  - Admitting diagnosis
  - Admission day of week
- Discharge data
  - Patient discharge status
  - Length of stay
  - Discharge hour
  - Discharge day of week
- Diagnosis and procedure
  - Diagnosis codes
  - Procedure codes and day of week
- Physician data (state license and national provider identifier (NPI))
  - Referring physician - state license
  - Attending physician ID (state license, NPI)
  - Operating physician ID (state license, NPI)
  - Code to identify other physician (State license, NPI)
- Payer identification
  - Primary payer
  - Secondary payer
  - Tertiary payer
  - Estimated payer code
  - Payer ID/Health plan ID
- Additional data
  - Type of bill
  - Prospective payment system (PPS) code
  - Procedure coding method used
  - PHC4 diagnosis-related group (DRG)
  - MS-DRG grouper version
  - Cancer code 1 and code 2
  - Major diagnostic category (MDC)
- MediQual data
  - MediQual atlas admission severity (Patient probability of in-hospital mortality)
  - MediQual non-responder
  - Total charges grouper cluster
  - Total charges grouper cell
  - MediQual morbidity
- Summary charges
  - Room & board charges
  - Ancillary charges
  - Drug charges
  - Equipment charges
  - Specialty charges
  - Miscellaneous charges
  - Total charges
  - Non-covered charges
  - Professional fees
- APR grouper data
  - APR MDC
  - APR DRG
  - APR severity of illness subclass
  - APR risk of mortality subclass

Facility profile

- Pennsylvania facility number
- Facility name
- Facility type
- Facility bed count
- National provider identifier (NPI)
- Master provider index (PA medical assistance identifier)
- Unit type and MAID unit code
- Total discharges for facility during quarter
- Total outpatient cases for facility during quarter
- Facility address

Physician profile

- Physician license number/national provider identifier
- Physician name
- Match indicator

Trauma

The PTOS contains more than 300 data elements and more than 500,000 records over 30 years. Definitions for each of the data elements are found in the PTOS manual. Data collected in the PTOS are comprised of: demographics, pre-hospital information, process of acute care information, clinical data, outcome data, diagnoses, procedures, and payer class information. Table 14 contains a summary of data elements summarized by these categories.

Table 14. Summary of data elements in the PTOS database

Demographic data

- Institution number
- Zip code of residence
- Race
- Ethnicity
- Sex
- Date of birth
- Age
- External cause of morbidity/Primary cause of injury E code
- Secondary cause of injury/External cause of morbidity
- Height of fall
- Place of injury of the external cause
- Activity E-code
- Injury date
- Injury time
- County of Injury
- Protective devices
- Primary and secondary type of injury
- Type of burn injury
- Pre-existing conditions

Prehospital data

- Was patient extricated?
- Was scene provider and transport provider the same?
- Are any scene provider data available?
- Scene and/or transport provider
- Scene and/or transport dates and times
- Ambulance scene time
- Ambulance code
- Ambulance unit number
- Was patient care record available?
- Patient care record number
- Life support - highest level of care
- Was a complete set of vital signs taken prior to the patient leaving the scene of injury?
- Prehospital vital signs
    - Paralyzing drugs (removed for 2017)
    - Pulse rate/minute
    - Unassisted respiratory rate/minute
    - Systolic blood pressure
    - GCS-eye opening
    - GCS-verbal response
    - GCS-motor response
    - Matches NTDB initial ED/hospital GCS assessment qualifiers
    - Intubated with artificial airway
    - Is patient's respiratory rate controlled (bagging or ventilator)
    - Controlled respiratory rate
- Referring facility
    - Is this a transfer patient?

Process of acute care (cont'd)

- Attending surgeon specialty
- Was there documentation that the attending anesthesiologist was immediately present in the OR? – if no, specify arrival time
- Admitting service
- Did patient receive a CT scan of the head during the resuscitative phase?
- Did patient require an initial laparotomy/laparoscopy which is not performed within 2 hours of arrival at your facility?
- Was trauma alert called?
- Date and time initial trauma alert called
- Initial level of alert
- Level of alert – specify
- Date and time called
- Provider arrival date and time
- Emergency physician arrival date and time
- Emergency medicine resident arrival date, time, and PGY level
- Attending trauma surgeon arrival date, time, and PGY level
- Senior trauma resident arrival date, time, and PGY level
- Junior trauma resident arrival date, time, and PGY level
- Neurosurgeon arrival date and time
- Neurosurgical resident arrival date and time and PGY level
- Orthopedic surgeon arrival date and time
- Orthopedic resident arrival date, time, and PGY level
- Anesthesiologist surgeon arrival date and time
- Anesthesiologist resident arrival date, time, and PGY level
- Admitting attending trauma surgeon, CRNA date, and time of arrival
- Others called to ED arrival date, time, and PGY level
- Patient monitoring during radiology studies
- Was any CT scan performed at this hospital during resuscitative phase
- 24-hour in-house coverage
- CT study ordered
- CT tech response/arrival time
- Patient monitoring CT studies
- Units of blood hung

Clinical data

- Total prehospital fluids administrated
- Total prehospital units of blood hung
- On admission
    - Paralyzing drugs (removed for 2017)
    - Pulse rate/minute
    - Unassisted respiratory rate/minute
    - Systolic blood pressure

- Is there data/information available from outside facility?
- Date and time of admission at referring facility
- Date and time of discharge from referring facility
- Length of stay
- Diagnostic interventions at referring facility
- Therapeutic interventions at referring facility
- Referral from facility number
- Unresolved occurrences
- Is referral facility clinical data available?
- Paralyzing drugs (removed for 2017)
- Pulse rate/minute
- Unassisted respiratory rate/minute
- Systolic blood pressure
- GCS-eye opening
- GCS-verbal response
- GCS-motor response
- GCS-qualifiers
- Intubated with artificial airway
- Is patient's respiratory rate controlled?
- Controlled respiratory rate
- Temperature
- Temperature route of measurement
- Weight
- Blood alcohol content (deleted for 2017)
- Alcohol screen and results
- Drug screen

Interhospital data

- Provider
- Dates and times
- Ambulance code
- Ambulance unit number
- Patient care record available?
- PCR number
- Life support - highest level of provider
- Life support - highest level of care
- Interhospital vital signs
- Paralyzing drugs (removed for 2017)
- Pulse rate/minute
- Unassisted respiratory rate/minute
- Systolic blood pressure
- GCS-eye opening
- GCS-verbal response
- GCS-motor response
- Matches NTDB initial ED/hospital GCS assessment qualifiers
- Intubated with artificial airway
- Is patient's respiratory rate controlled (bagging or ventilator)
- Controlled respiratory rate

Process of acute care

- Date entered ED
- Time entered ED
- Date transported to post ED destination
- Time transported to post ED destination
  - GCS-eye opening
  - GCS-verbal response
  - GCS-motor response
  - Matches NTDB initial ED/hospital GCS assessment qualifiers
  - Intubated with artificial airway
  - Is patient's respiratory rate controlled (bagging or ventilator)
  - Controlled respiratory rate
  - Temperature
  - Temperature route of measurement
  - Weight
  - Blood alcohol content (deleted for 2017)
  - Alcohol screen and results
  - Drug screen
- Was the first set of vital signs taken within the first 10 minutes of less of patient's arrival to ED?
- When was the initial nutrition assessment performed?
- When was nutrition initially started?
- Type of nutrition
- Date and time of 'order to change vital signs' to greater than one hour
- Is there sequential neurological documentation on ED record of trauma patient with diagnosis of skull fracture, intra-cranial injury, or spinal cord injury?
- Is there hourly documentation beginning with ED arrival?
- Did patient leave ED with a discharge GCS <= 8?

Outcome data

- Discharge status
- Date of death/discharge/transfer
- Time of death/transfer
- Total days in ICU
- Total days in step down unit
- Total hospital days
- Total ventilator days
- Discharge destination
- Discharge to facility number
- Occurrence
- Were there more than 10 occurrences?
- Did patient have discharge diagnosis of cervical spine fracture, subluxation or neuro deficit not addressed on admission?
- Source of final anatomical diagnoses
- Functional status of discharge
- Organs donated
- Discharge weight and unit of measurement
- Burn patient follow-up
- Was burn patient readmitted due to development of an occurrence?
- Burn wound management
- Autopsy requested and results available
- Consults
- Abuses
  - Was the patient being evaluated for abuse?
  - Was a report of suspected abuse made to civil authorities?

• Date administratively discharged from ED • Time administratively discharged from ED • Post ED destination • Interim ED disposition-temporary location • Time for referral • Was operating room available when patient ready to transport from ED to OR? • Was attending surgeon present when the patient arrived in the OR? – if no, specify arrival time	- Was there a police investigation initiated because of this episode? - Was patient discharged to a different caregiver than when they were admitted?

Current integration

No integration with existing safety-related data systems was noted during the interviews.

SWOT analysis

Injury surveillance data are stored in four unique data systems—EMS, Trauma, Health Care, and Vital Statistics. A representative from Vital Statistics was not available to be interviewed for this project and thus these data were not considered as a part of this report.

Strengths

- **Modern data systems:** EMS data are stored in a cloud-based database (CloudPCR), Trauma data are stored on a proprietary system developed by Digital Innovation, Inc., and Health Care data are stored on an SQL server. Each of these databases is modern and facilitates outputting of data in various formats and real-time queries and communications features necessary for data integration.
- **Comprehensive and well-documented databases:** All three databases have well-defined data dictionaries that can be used to identify all data elements stored for any given record.
- **PII available for Health Care:** The Health Care database includes PII that can be used to link to other data systems, including patient name, date of birth, address, and social security number.

Weaknesses

- **PII is not available for all databases**: While PII is collected by individual EMS agencies and hospitals, the statewide EMS and Trauma databases do not contain this information. This limits how individuals can be identified for data integration purposes.
- **PII in Health Care database cannot be directly shared:** Although PII information is stored in the Health Care database, this information cannot be shared outside of the Pennsylvania Health Care Cost Containment Council (PHC4) due to patient privacy regulations. This serves as a barrier to integration, although one that can be overcome as described below.
- **Data added to statewide databases at different intervals**: Each of the databases making up the injury surveillance data system is updated at different intervals. EMS data are supposed to be reported by the individual EMS agencies within 30 days of a patient being transferred and Trauma data are supposed to be reported by individual hospitals within 42 days of the patient being released. There is a 9-month lag for Health Care data being included in the statewide database.
- **Databases are generally updated quarterly:** Related to the previous item, even though the EMS and Trauma data are reported to the state within 1.5 months, the statewide databases are only

updated at regular, quarterly intervals. This means that there may be a significant lag between when an event occurs and when data might be available to analyze this event.

- **Data elements not consistent across three systems:** The individual databases that make up the injury surveillance data system have inconsistent data elements. As previously mentioned, PII is only stored in the Health Care data and not for EMS or Trauma. However, other variables are inconsistent. For example, EMS agency code is available within the EMS and Trauma databases but not within the Health Care database. This presents unique challenges for integration of these three data systems.

Opportunities

- **Well-defined data sharing agreements in place:** Each of the three databases considered that make up the injury surveillance data system have well-defined data sharing agreements in place, particularly for other state agencies (like PennDOT). These can be used to establish the data sharing protocols necessary for data linkage or integration.
- **EMS and Trauma data are directly available:** For the EMS and Trauma databases, which do not store PII, data can be provided directly to the agency or a third party that is performing research.
- **Data-sharing fees generally waived for state agencies:** Although there are significant fees associated with the sharing of injury surveillance data, these fees are generally waived for state agencies and third parties performing research on behalf of state agencies. However, it should be noted that this may not always be the case. It was noted that the specific fees could only be determined when a formal request for data was made.
- **Strong desire for integration**: A strong desire for data integration was noted by representatives for all three databases making up the injury surveillance data system. Specifically, there is a large movement toward a data-driven approach to patient care and the representatives found having access to the cause of an injury—specifically, crash information—might be helpful to improve patient care at the statewide level.
- **Existing data exchange**: It was noted that PennDOT receives death certificates from Vital Statistics. The shared data include name, DOB, race, time and cause of death, and if the death was caused by injury, description, place, and date of injury are also shared. Although further information was not received, this suggests that a pathway exists to obtain information from this portion of the injury surveillance data system.

Threats

Health Care data not directly available: For the Health Care database, which includes PII, data are not shared directly. Instead, the data agreement allows an agency to provide its data to the PHC4, which will merge the data using a list of provided identifiers.

OVERVIEW

Potential linkage variables

After reviewing the data elements included in each of the six core data systems, several potential linkage variables exist that could be used to integrate these different datasets. Figure 5 provides a graphical summary of these linkage variables with respect to how the various data systems can be linked specifically to crash data. The remainder of this section provides a short discussion of the types of variables available to link data across the six core data systems.

Crash data

The crash data system contains many potential deterministic and probabilistic linkage variables. Deterministic variables include driver-related information (e.g., name, license number, address, and date of birth) and vehicle-related information (e.g., license plate, VIN, and registration number). Event-related information (time and location of incident) tends to facilitate probabilistic linkages with other systems.

Roadway data

For state-owned roads, RMS data can generally be linked to crash data using the PennDOT linear referencing system (county, route, segment), which can be used to identify the segment at which a crash occurs. Other variables exist that can be used to verify that the appropriate location is identified, although some of this information is pulled directly from the RMS into the crash database. For local routes, data elements from ARNOLD will be able to be linked to crash data using GPS coordinators (latitude and longitude) as well as street name.

Driver data

Individuals can be directly identified in the driver's license database using their name, date of birth, or driver's license ID number. Additional variables available include insurance-related information, such as company name and policy number.

Vehicle data

Information for specific vehicles in the vehicle registration database can be identified using a vehicle's registration number, license plate number, or VIN for deterministic linkages. Probabilistic variables also exist, including: make, model, and year and color of the vehicle, which can be used to identify a reasonable subset of vehicles that might be of interest. Deterministic linkages can also be made using the owner/lessee name and address, which are recorded in the database.

Citation and adjudication data

This data system includes name, address, and driver's license number of individuals that have received citations. For traffic citations, vehicle information (such as registration number for deterministic linkages or make, model, and color and year for probabilistic linkages) are also available.

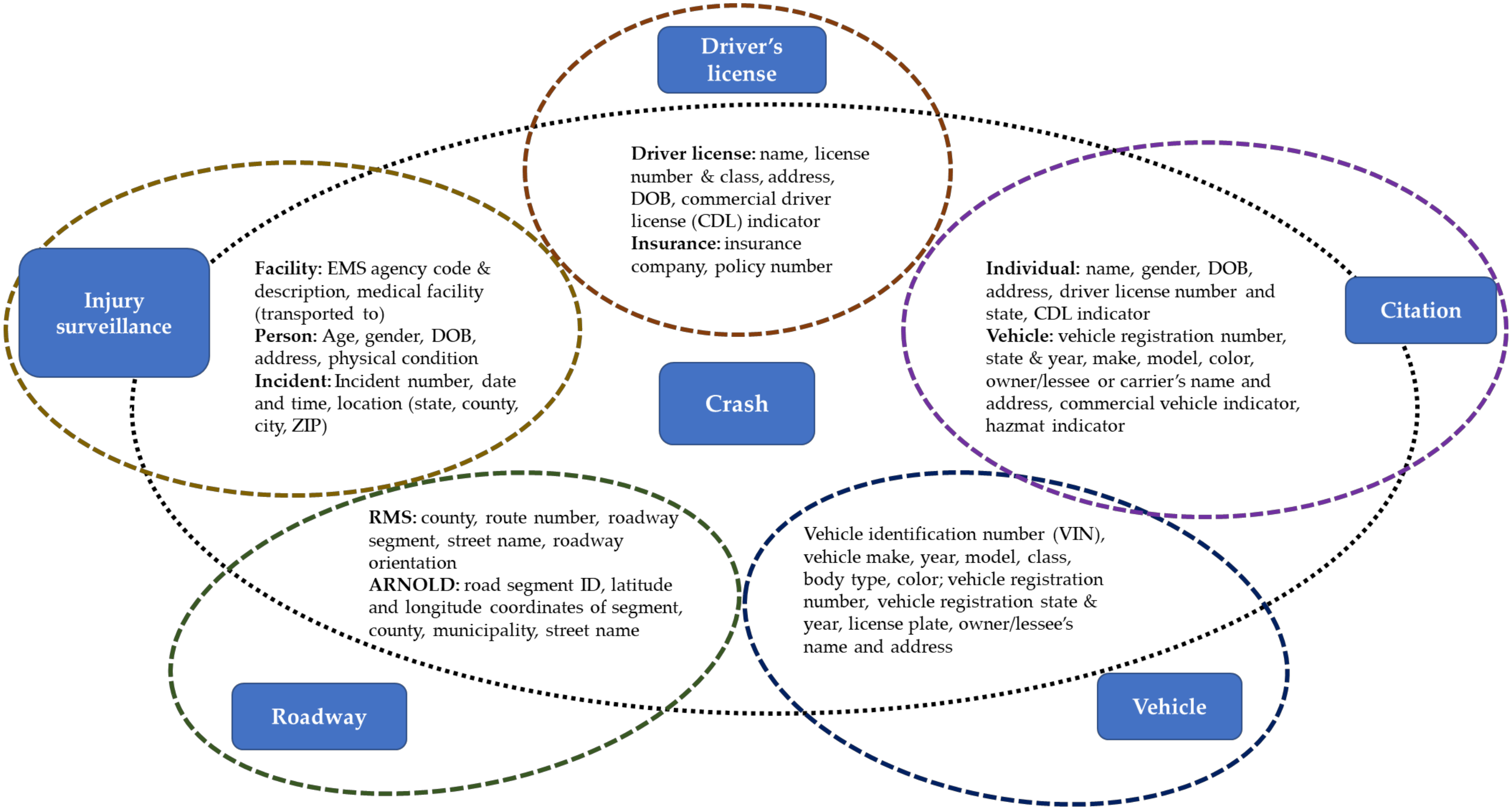

Driver's license
Driver license: name, license number & class, address, DOB, commercial driver license (CDL) indicator
Insurance: insurance company, policy number
Injury surveillance
Facility: EMS agency code & description, medical facility (transported to)
Person: Age, gender, DOB, address, physical condition
Incident: Incident number, date and time, location (state, county, city, ZIP)
Crash
Citation
Individual: name, gender, DOB, address, driver license number and state, CDL indicator
Vehicle: vehicle registration number, state & year, make, model, color, owner/lessee or carrier's name and address, commercial vehicle indicator, hazmat indicator
RMS: county, route number, roadway segment, street name, roadway orientation
ARNOLD: road segment ID, latitude and longitude coordinates of segment, county, municipality, street name
Roadway
Vehicle identification number (VIN), vehicle make, year, model, class, body type, color; vehicle registration number, vehicle registration state & year, license plate, owner/lessee's name and address
Vehicle

Figure 5. Summary of potential linkage variables

Injury surveillance data

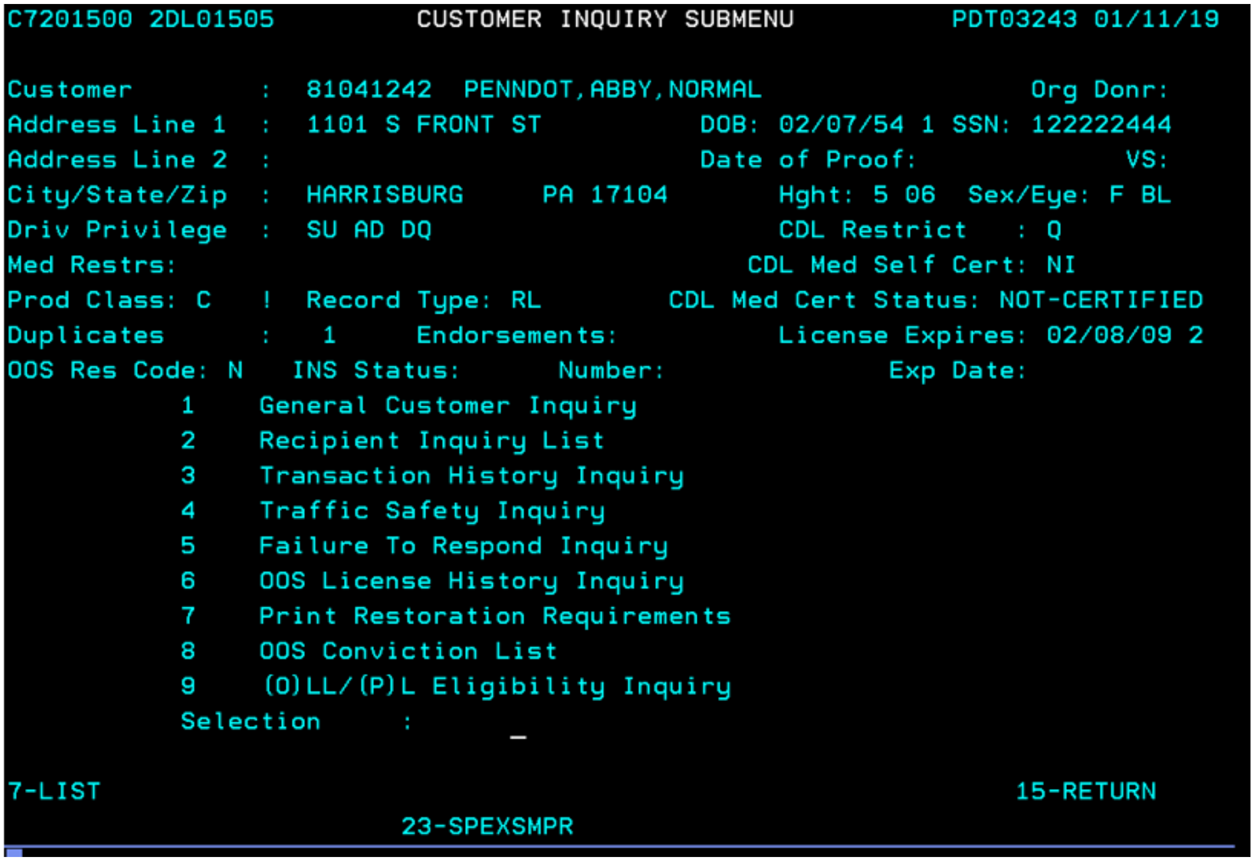
Linkage variables in the injury surveillance data tend to be probabilistic in nature. These include incident-level variables, such as location and time. Facility-level information, such as EMS agency code and nearest medical facility, or individual-level information, such as age, gender, date of birth, can also be used to probabilistically link crashes to specific records in the injury surveillance cases.

Existing barriers and risks to data integration

Multiple barriers to data integration were identified through the interviews with the individuals representing each of the core data systems. These barriers include technical limitations, update data intervals, legal and privacy issues, lack of internal coordination, lack of deterministic linkage variables, and data owner desire and/or resistance to integration. This section describes these barriers in more detail.

Technical limitations

As described in this document, the six core data systems are all stored in various formats and updated at different frequencies, which could make complete data integration difficult. Representatives from citation and adjudication data systems specifically commented that PennDOT uses an antiquated data system that would limit the amount of data integration that could be performed. Two of the systems (the driver's license and vehicle registration databases) are stored on an old IBM mainframe database that is not up to modern standards; see Figure 6 and Figure 7 for an example screenshot of the system. This would complicate the development of a system interface that could integrate these data in real time. Fortunately, the data can be queried and outputted into a more modern DB2 database for reporting purposes, which allows the driver's license and vehicle registration information to be accessed by other systems. Existing examples of this include the availability of driver's license and vehicle registration information to law enforcement officers through the PAJNET portal. This reporting database is better suited for integrating these data with other data systems at discrete points in time.

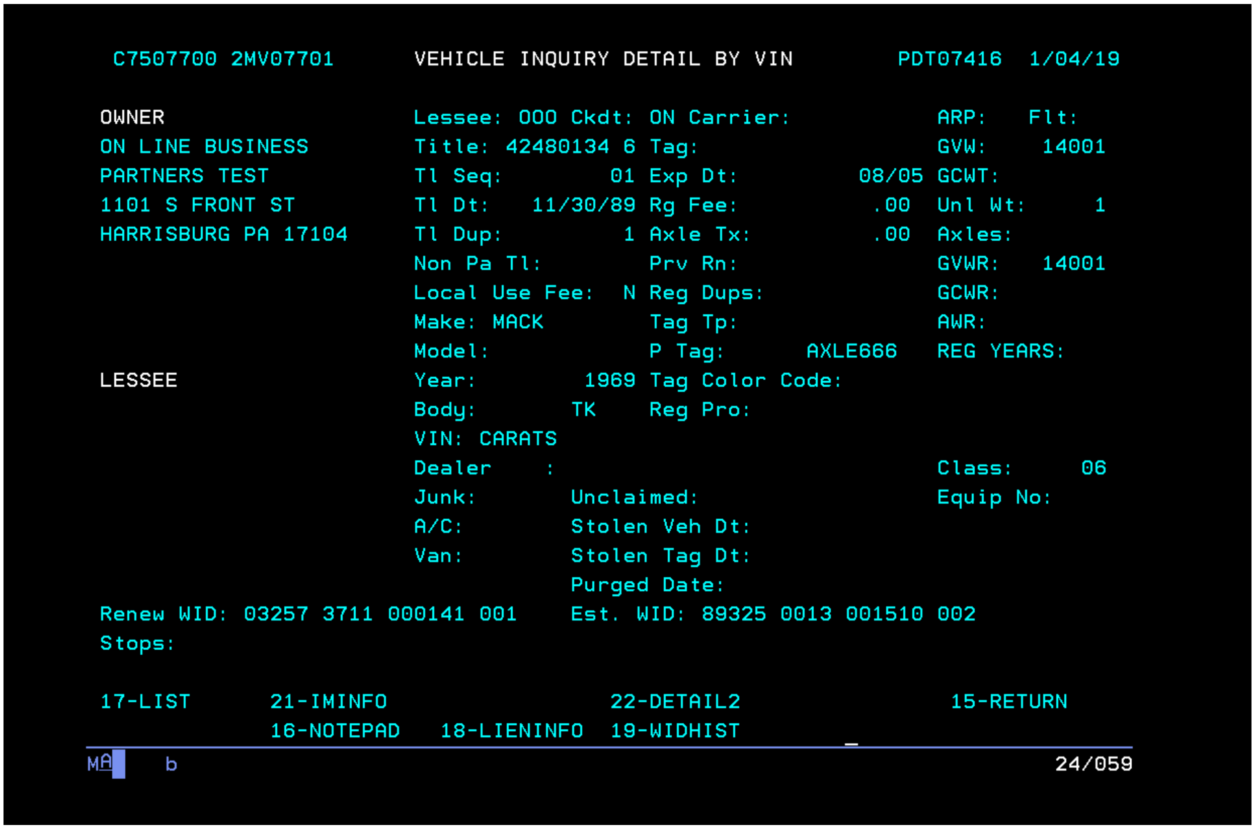


```
C7201500 2DL01505         CUSTOMER INQUIRY SUBMENU          PDT03243 01/11/19

Customer          :  81041242  PENNDOT,ABBY,NORMAL                  Org Donr:
Address Line 1    :  1101 S FRONT ST          DOB: 02/07/54 1 SSN: 122222444
Address Line 2    :                           Date of Proof:            VS:
City/State/Zip    :  HARRISBURG     PA 17104       Hght: 5 06  Sex/Eye: F BL
Driv Privilege    :  SU AD DQ                      CDL Restrict   : Q
Med Restrs:                                      CDL Med Self Cert: NI
Prod Class: C     !  Record Type: RL        CDL Med Cert Status: NOT-CERTIFIED
Duplicates        :   1      Endorsements:         License Expires: 02/08/09 2
OOS Res Code: N   INS Status:      Number:                Exp Date:
           1    General Customer Inquiry
           2    Recipient Inquiry List
           3    Transaction History Inquiry
           4    Traffic Safety Inquiry
           5    Failure To Respond Inquiry
           6    OOS License History Inquiry
           7    Print Restoration Requirements
           8    OOS Conviction List
           9    (O)LL/(P)L Eligibility Inquiry
           Selection     :    _

7-LIST                                                             15-RETURN
                         23-SPEXSMPR
```

Figure 6. Screenshot of driver's license database



```
 C7507700 2MV07701      VEHICLE INQUIRY DETAIL BY VIN       PDT07416  1/04/19

OWNER                  Lessee: 000 Ckdt: ON Carrier:           ARP:   Flt:
ON LINE BUSINESS       Title: 42480134 6 Tag:                  GVW:    14001
PARTNERS TEST          Tl Seq:        01 Exp Dt:         08/05 GCWT:
1101 S FRONT ST        Tl Dt:    11/30/89 Rg Fee:          .00  Unl Wt:      1
HARRISBURG PA 17104    Tl Dup:         1 Axle Tx:          .00  Axles:
                       Non Pa Tl:        Prv Rn:               GVWR:   14001
                       Local Use Fee:  N Reg Dups:             GCWR:
                       Make: MACK        Tag Tp:               AWR:
                       Model:            P Tag:      AXLE666   REG YEARS:
LESSEE                 Year:        1969 Tag Color Code:
                       Body:       TK    Reg Pro:
                       VIN: CARATS
                       Dealer    :                             Class:     06
                       Junk:       Unclaimed:                  Equip No:
                       A/C:        Stolen Veh Dt:
                       Van:        Stolen Tag Dt:
                                   Purged Date:
Renew WID: 03257 3711 000141 001   Est. WID: 89325 0013 001510 002
Stops:

17-LIST      21-IMINFO                22-DETAIL2                15-RETURN
             16-NOTEPAD   18-LIENINFO  19-WIDHIST
MA     b                                                                24/059
```

Figure 7. Screenshot of vehicle registration database

Data update intervals

Another challenge is that these data systems are updated at different frequencies. Driver and vehicle data appear to be updated the most frequently—almost daily and in real-time when driver or vehicle services are performed. Crash data are provided electronically by the various police agencies but this might not be in real time. Furthermore, the crash data are only pulled into the CDART system weekly. These lags can cause errors when validating data elements using other databases (e.g., if the license plate recorded in the crash is changed to a new vehicle, then the vehicle registration will return conflicting information for a vehicle's VIN). It was noted that RMS roadway data are constantly changing due to maintenance, particularly in the spring and summer months. Note that the summer is when the STAMPP program actively updates the RMS data. Linear referencing information also changes constantly and is updated manually. There can be up to a two-week lag between when a change occurs and when the associated RMS data are updated, which could also result in errors between roadway features that exist at the time of a crash and what is reported in the RMS. Similar problems are likely for the ARNOLD database. Citation and adjudication data are updated on a near daily basis as infractions occur and judgments are made; however, there can be a short lag on when these data become available on the public databases. Lastly, the exact frequency of injury surveillance information was not known. However, the availability of these data can have a significant lag, particularly for patients with longer hospital stays.

Legal and privacy

Many of the data systems contain personally identifiable information, which serves as a major limitation to data integration. As described in the Task 1 technical memorandum, PII serves as one of the primary linkage variables that can be used to confidently link records across these systems. Fortunately, four of the six core data systems are owned and maintained by PennDOT; thus, these data systems could be integrated in-house without any significant legal or privacy issues. However, significant privacy issues exist for both the citation and adjudication and injury surveillance data systems, since these datasets contains sensitive PII. Specifically, personal health care data are covered under HIPAA. However, this could potentially be addressed through the careful development of Memoranda of Understanding (MOUs) or Data Use Agreements (DUAs) between PennDOT and the agencies that control these data systems. An MOU could allow PennDOT to provide crash data to agencies with access to EMS, Health Care, or Trauma data so that injury outcomes (alone) could be identified and sent back to PennDOT. This would limit the need to share PII with PennDOT and the potential for any HIPAA violations.

Lack of internal coordination

Representatives of some core data systems specifically indicated that there is a lack of internal coordination that is limiting the potential for data integration within PennDOT. This lack of communication/coordination influenced both the ability to integrate databases owned and maintained by PennDOT (e.g., crash and driver or crash and vehicle data systems) as well as communication with data systems outside of PennDOT. For example, representatives of the citation and adjudication data system indicated that there is some level of data sharing currently going on with PennDOT through the PAJNET portal, but users of some data systems within PennDOT do not seem to be aware of these mechanisms to share data.

Lack of deterministic linkage variables

The majority of the data systems have clear linkage variables that can be used for integration with the other core systems. The crash, vehicle, and driver systems can be linked through deterministic identifiers (driver's license number, driver name, VIN, license plate), although as previously mentioned there might be errors involved due to lack of real-time updates. Crash and roadway systems are also linked using the linear referencing system, since location data are available in the vast majority (99.6%) of the cases. The citation and adjudication data system contains driver's license ID number and social security number, which can be used to link these data to individual drivers in the driver's license database. However, the lack of deterministic linkage variables remains a barrier to integration of crash data with injury surveillance systems. Specifically, the EMS, Health Care, and Trauma databases do not contain any PII or identifying variables that can deterministically link a patient to a specific crash event. Instead, probabilistic methods are needed that rely on other linkage variables, such as relative location or time of incident, which can lead to errors in the integration process. Moreover, as previously mentioned, data are not provided in specific instances where it is very likely that patient confidentiality might be violated (e.g., when fewer than 5 records exist within a single zip code for a specific time period in the EMS data).

Data owner desire for and/or resistance to integration

Although there is much desire for integration from representatives of the crash data systems, representatives of some of the other core data systems did not express the same level of desire for data integration. It was noted in the interviews with EMS representatives that their data needs are being met by the PCR information and they do not see a need to merge their data with other systems (from their perspective). However, they have also indicated that the PA DOH has tried for years to establish links with PennDOT but these could not be accommodated due to regulations within the agencies. The specific regulations that precluded data sharing were not known, however.

Resources and desires for data integration

This section describes some resources that might be possible to leverage when considering integration of transportation safety data within the Commonwealth. This information was obtained from the interviews with representatives of each of the data systems, as well as the research team's existing knowledge and experience.

Existing MOUs and DUAs

Although agreements are not currently in place for PennDOT to acquire data from non-PennDOT-owned data systems (specifically, citation and adjudication and injury surveillance), many of the subsystems associated with injury surveillance have processes in place for data acquisition and sharing. For example, a formal data request process exists in which an agency can specify the dates and purposes of a data request, which is then reviewed by the PA DOH. The PHC4 system also has a well-defined data request application and DUA that potential users must agree to or follow. The DUA prohibits sharing of the data or use for any other purpose than what is agreed upon between the agency and PHC4; however, its existence suggests that mechanisms exist to obtain these data for purposes that might benefit the Commonwealth. Lastly, the PTSF also has an existing data request process in place and well-defined DUA that permits data sharing for research purposes.

PAJNET

Within the citation and adjudication data system, the PAJNET exists as a resource that can be used for data sharing. Although PAJNET does not store any citation and adjudication data, it serves as a secure "highway" for criminal justice data. Thus, PAJNET provides a secure method that approved users or agencies can use to access citation and adjudication systems to obtain data elements for integration. Due to its online nature, the PAJNET portal may not work well with some of the antiquated data systems being used by PennDOT for complete data integration. However, reporting databases can be merged outside of the main databases to link citation data with crash, vehicle, driver, and roadway data at discrete intervals in time.

SWOT analysis summary

Strengths

- **TRCC membership includes key stakeholders from most data systems:** The TRCC includes members from most data systems, which should help facilitate data integration. The exceptions are Trauma, Health Care, and Vital Statistics. However, with the exception of Vital Statistics, which could not be contacted for this report, representatives from the Trauma and Health Care systems generally seem to be responsive and interested in data integration and should be considered for future membership and/or activities.
- **Strong desire for integration of crash data with other systems:** TRCC leadership has intimate knowledge of the crash data system and has expressed a strong desire for integration of crash data with other data systems. This is vital to making data linkage or integration a reality.
- **Many data systems are owned and maintained by PennDOT**: Many potential barriers are alleviated since the crash, roadway, vehicle, and driver data systems are all owned and maintained by PennDOT. This should help expedite data sharing for integration.
- **The TRCC has identified and put into place methods to improve data linkage:** Police crash reports were recently updated to include a variable for EMS agency code. This will provide a direct linkage between the crash data system and EMS and Trauma databases that did not previously exist.
- **Data exchange is already occurring at various levels**: As previously outlined, there are various avenues of data exchange occurring within PennDOT and between PennDOT and other agencies. This provides potential pathways that can be used for future data integration possibilities, or might preclude the need to establish new integration protocols with some data systems (e.g., the citation and adjudication system).
- **Existing projects/activities:** The TRCC has identified ongoing and upcoming projects related to data quality and integration. This includes an audit of police report submissions, which should help improve the quality of data being stored in the crash data system.

Weaknesses

- **Vital Statistics not engaged in the process:** During this project, representatives from the Vital Statistics data were not available for staffing reasons. For this reason, details on this data system were not available.
- **Antiquated driver and vehicle databases:** As previously mentioned, the driver and vehicle data systems are stored on an antiquated database format. While data exchange is possible—and currently occurs—it was noted several times as a barrier for further data integration.

Opportunities

- **Existing data exchange provides pathways for integration**: The fact that data are currently being exchanged between these data systems provides clear pathways that can be used to share more information for data integration. The potential data integration strategies outlined in this report will leverage these existing pathways.
- **Existing guidelines in place to share data:** Both within and outside of PennDOT, existing guidelines and data sharing agreements are already in place to facilitate data exchange. These can be leveraged for data integration purposes in the future.
- **New driver and vehicle data systems:** The upgrade of the existing driver's license and vehicle registration databases provides an opportunity to streamline existing data exchange and integration.

Although technical details were not available for this report, the antiquated nature of the existing systems were noted by several parties as a barrier for data integration.

Threats

- **Systems currently being built/upgraded:** The lack of detailed information about the ARNOLD database (no "live" date) and driver and vehicle databases prohibits a deeper analysis of data integration possibilities.
- **Some data systems are incomplete:** The citation data system does not currently include data from Philadelphia. While this does not prohibit integration, it does mean that any integration that occurs with current systems will be incomplete.
- **Key data elements not included in some databases**: The lack of PII in some of the statewide injury surveillance databases complicates integration. Alternative linkage variables will need to be used or stochastic linkage methods will have to be used instead.
- **Data systems are not updated at the same time scale**: Each database is updated at different temporal intervals. This will complicate data integration, as some portion of the data will always be incomplete at any point in time.

CHAPTER 4

Proposed Data Integration Plan

INTEGRATION PAIR ANALYSIS

This section examines specific issues associated with merging crash data with data from each of the other five core systems. This includes the additional knowledge that can be gained through data integration, the barriers that exist to preclude data integration, and potential pathways to integration that might exist.

Crash and roadway data systems

Purpose of integration and anticipated results

Integration of crash and roadway data is a critical step in implementing methods outlined in the *Highway Safety Manual (HSM)*. This includes applying HSM safety performance functions (SPFs) and crash modification factors, calibrating HSM SPFs or developing local SPFs for a state or local agency, performing network screening activities, identifying appropriate safety countermeasures, prioritizing safety projects, and evaluating the effectiveness of these countermeasures. In order to perform these activities, detailed information on roadway geometry and characteristics are needed for each roadway segment, along with the history of reported crashes on those roadway segments.

More generally, integration of roadway and crash data can be used to identify common physical and functional characteristics of roadway segments that are associated with high crash frequencies and more severe crash outcomes, which can help improve safety management on state-owned roadways.

Barriers

There are very few significant barriers that prevent integration of crash and RMS data. All data are owned and maintained by PennDOT. Several recent projects have successfully merged crash and RMS data to develop Pennsylvania-specific SPFs for various roadway types (Donnell et al., 2016, 2014). However, there are several challenges that should be noted. As previously mentioned, PennDOT's linear referencing system is used to identify crash locations within individual roadway segments. This linear referencing system is also used to locate infrastructure features such as guiderail locations, roadway signage, and traffic control devices. Recalibration of the linear referencing system can lead to errors in how crashes or these infrastructure elements are assigned to individual roadway segments. The accuracy of data elements in the RMS database is also a concern, as recent projects have identified inaccuracies in data elements like roadway direction (one-way vs. two-way roads) and cross-sectional elements (e.g., paved roadway width and shoulder widths). Inaccurate data could lead to incorrect conclusions about which features are associated with more frequent or severe crash outcomes. Thus, the accuracy of the roadway data should be carefully studied before integration is performed.

Crash data have not yet been merged with local road data from the ARNOLD database. However, it is anticipated that significant barriers will not exist here, since a well-defined GIS-based mapping system exists that can be used to link crash locations with specific roadway segments. Spatial methods can be easily developed using GIS software to perform this type of data merging. Doing so will facilitate the development of data-driven, HSM-type methods to safety management on local roadways.

Pathways to integration

Although the RMS and ARNOLD databases both use a different database structure than the crash database, a direct communication link between these databases should be possible. However, due to the different database architecture, establishing a direct link may require development of a specialized algorithm to facilitate communication between these databases. And, due to the differences in how locations are identified in the RMS and ARNOLD databases, linking the two will require different algorithms, which would increase the resources required for algorithm development.

As previously mentioned, the RMS database uses PennDOT's linear referencing system. Each roadway is split into different segments and these are identified using the county code (CTY_CODE), state route number (ST_RT_NO), and segment number (SEG_NO). The variables used in the RMS database to specify the location of a roadway are shown in Table 1. Crash locations are provided based on the county, state route, segment ID, and offset (specific location within the segment). Other roadway features—such as guiderails—are also identified using this county-state route-segment-offset method. The variables that are used to specify the location of a crash in the Roadway table of the crash database are provided in Table 15 and include COUNTY_NAME (county), ROUTE (route number), SEGMENT (segment within a state route), and OFFSET (offset in feet within segment). These variables can be used as common identifiers to deterministically link the RMS and crash databases. Merging the data in this way can readily be performed in real-time (for data integration) or at discrete points in time (for data linkage).

Table 15. Potential linkage variables for merging crash and RMS databases

Crash Data		RMS	
Field name	**Data description**	**Field name**	**Data description**
RDWY_CNTY_CD	County code	CTY_CODE	County of crash
RTE_NUM	State route number (if applicable)	ST_RT_NO	State route number
SEGMENT	Segment along state route (if applicable)	SEG_NO	Segment along state route
OFFSET	Offset within segment along state route (if applicable)		

There is no linear referencing information included in the ARNOLD system that could be used to directly link crashes to individual local roadway segments. Instead, a crash can be mapped to a local road spatially using GIS coordinates of the crash and the GIS mapped local roads by identifying the segment nearest to the crash location. Additionally, common identifiers between crash and ARNOLD databases can be used to validate the accuracy of the segment identification using these spatial methods. These common variables are listed in Table 16.

Table 16. Potential linkage variables for merging crash and ARNOLD databases

Crash Data		ARNOLD	
Field name	Data description	Field name	Data description
SPD_LIM	Speed limit	SPEED_LIMIT	Speed limit
ST_NM	Street name	LR_NAME	Local road name
MUN_CD	3-digit municipality code	MUNICPAL_CODE	5-digit code made up of 2-digit PennDOT county code and 3-digit PennDOT municipality code

Estimated timeline

As mentioned previously, there is an existing linkage between RMS and crash databases. Thus, it can be assumed that the format of RMS data is well known, and the algorithms needed to merge these databases are already developed. The only work needed to be done is to extend the capabilities of the existing merging algorithms. Therefore, the time needed to establish a working merging framework for these databases is expected to be short.

In order to merge the ARNOLD database, an algorithm needs to be developed to identify the nearest road segment to the crash location using its GPS location information. This algorithm should also be able to check and report the potential discrepancies between the merged databases for the segment properties listed in Table 15. The reported cases that need additional verification can be manually checked and merged. Although some work for algorithm development is needed, development of such algorithm is expected to be simple because of the existence of well-defined data dictionaries and the existence of linkage variables that are descriptive enough to facilitate deterministic linkage. However, depending on the errors on both datasets, the manual check and merge process for reported discrepancies can take a significant amount of time.

Resource requirements and cost estimate

The research team estimates the resources and cost required to integrate crash and roadway data to be fairly minimal. Integration between the crash and RMS databases could be done using existing algorithms and thus only some basic labor is required to run the algorithm at regular intervals (approximately every year since this is the frequency at which the RMS data are updated) and spot check for errors in the merged database. Integration between crash and ARNOLD databases would require the development of new GIS-based algorithms, which would require a moderate amount of upfront labor. Minimal labor would then be needed to run the algorithm as the ARNOLD database is updated (again, likely annually) and check for errors.

Metrics and likelihood for success

There are two metrics that can be used to measure the success of merging:

- Percent of non-matched data: this type of error occurs if there are errors (i.e., wrong value, missing data) in linkage variables in either dataset, which results in missing (non-matched) data in the resulting merged dataset. Discrepancies between the verification variables can be solved through manual checking and merging.

- Percent error in matched data: this type of error occurs when the linkage variables are correct, so the datasets can be merged successfully, but the data other than linkage variables in either dataset have errors. Although this is not directly related to the success of the merging process, merging erroneous data can affect the accuracy of any analysis done with the merged datasets.

In terms of percent of non-matched data, merging crash with both the RMS and ARNOLD databases is very likely to be successful. Although the proposed merging process relies on one main linkage variable for each database (linear referencing and GPS coordinates), these variables can be expected to be reliable and available for all records.

As mentioned in the barriers section, there are inaccuracies in the RMS database. The frequency of these inaccuracies is not known.

The likelihood for success of merging the crash and roadway data is likely to be high. Merging the crash and RMS databases has already been done for various projects and thus can easily be performed at any time. Merging the crash and ARNOLD data will require some additional algorithm development, but this is likely to be successful based on the data elements that are available within the two databases and the use of GIS-based mapping for crashes and local roadway segments.

Crash and vehicle data systems

Purpose of integration and anticipated results

Integration between the crash and vehicle data systems can be used as a means to verify information on police crash reports, such as vehicle identification number and features of the vehicle (year, make, and model). It can also be used to study the relative safety performance of various vehicle types (sedan, SUV, light truck, etc.) and vehicle ages (in terms of time, miles traveled, and inspection history). This can help guide new vehicle-related policies, such as inspection costs or crash frequencies based on vehicle age or type.

Barriers

There are no institutional barriers for merging crash and vehicle data and all data are owned and maintained by PennDOT. However, the primary barrier to integration is the outdated nature of the vehicle data system. This is an antiquated system that may require specialized software to create an interface that can communicate with the crash data system for real-time integration purposes. Data linkage is less onerous and can be performed at fairly regular intervals using known linkage variables. However, as previously identified, the lack of historical information for variables (e.g., license plate numbers) might lead to incorrect linkage between crash and vehicle records.

Pathways to integration

Vehicle registration data are stored on an IBM mainframe database using COBOL 2, which is a much older system than the crash database. Although it is possible to establish a communication link between crash and vehicle databases, establishing such connection may require development of a specialized code.

The list of codes that can be used to link the crash and vehicle data systems is provided in Table 17. Direct linkage can be performed between crash and vehicle databases by using the VIN as a unique identifier. Other common identifiers between these datasets are vehicle year, make, and model. These variables can also be used along with the VIN to deterministically merge the data. Although direct linkage is vulnerable to errors in one of the databases, there are no other identifiers that are descriptive enough to use with the VIN for probabilistic merging. In other words, probabilistic linkage would likely not be successful if there is an error in the VIN record due to the lack of common identifiers. Note that the owner address is not considered as a common identifier, because the owner of the vehicle can change between the crash date and data merging date. Since information about all the previous owners of a vehicle are not kept in a vehicle database, owner address cannot be used as a common identifier.

Table 17. Potential linkage variables for merging crash and vehicle databases

Crash Data		Vehicle Data	
Field name	Data description	Field name	Data description
VIN	Vehicle identification number	VIN	Vehicle identification number
MODEL_YR	Vehicle year	VEH-MODEL-YR	Year the vehicle was assigned by the manufacturer
MAKE_CD	Code for vehicle manufacturer		
MODEL_CD	Code for vehicle model	VEH-MODEL-CODE	Code for vehicle model

Estimated timeline

Merging the crash and vehicle databases would require the development of an algorithm that is capable of matching VINs in both databases and verifying matched data with vehicle year and make. The developed algorithm can also report the discrepancies between matched data for manual correction. Since there is an existing data linkage between these databases, development of this algorithm is expected to be simple.

Resource requirements and cost estimate

The research team estimates that merging the crash and vehicle database would not require significant resources or labor costs. Since algorithms already exist to query information from the vehicle database to incorporate into the crash data, these algorithms could be readily modified to perform the requisite merging. These algorithms would need to be run at regular intervals to ensure that the merged dataset is up-to-date. The research team suggests that an automated algorithm be used that can perform the requisite merging at the same time data elements are validated in the crash database using the vehicle registration information.

Metrics and likelihood for success

One metric that can be used for this integration pair is the fraction of crash records that can be linked to vehicles in the vehicle registration database. This metric is not likely to be 100%, as only records involving vehicles registered in Pennsylvania can be merged. Thus, the percentage of crash records involving

Pennsylvania-registered vehicles can be used as an upper bound for the fraction of records matched between the two databases. Other metrics previously described can also be used as metrics for success.

In terms of likelihood for success, merging the crash and vehicle datasets is very likely to be successful since there is only a single linkage variable (VIN) that can be used to deterministically match crash and vehicle records. The only records that are not likely to be successfully matched are those involving non-Pennsylvania-registered vehicles or those with missing or erroneous VINs. This latter case is not to be expected based on the interviews with representatives from the crash and vehicle data systems. Therefore, integrating crash and vehicle data is likely to be very successful.

Crash and driver data systems

Purpose of integration and anticipated results

Integration between the crash and driver data systems can be used to identify driver-related features that are associated with an increased risk of crash occurrence. This includes identifying relationships between crash risk and demographics, such as driver age or time with a license, or scores on a driver's license examination. Such information could be vital for understanding crash risk among different segments of the population, including younger and older drivers. The information obtained from merging crash and driver data could also help when developing policies on if and when relicensing should be performed or to improve the driver's license exam. For example, the current practices review revealed that other states have used such integration to identify the kinds of questions on the written driving test that are most associated with long-term safety performance. This could help identify drivers that are more likely to be involved in crashes at the time of licensing, who could be better targeted with behavioral safety countermeasures. This may be especially usefully with respect to driving behaviors such as alcohol and drug use, seat belt use, motorcycle helmet use, and distracted driving. A merged database could also be used to better match the types of education activities that would be most critical to various geographic areas.

Citation data are already incorporated into the driver data system. Thus, integration could help determine if a driver's traffic citation history is a predictor of risk of crashes occurring, including the types of citations that are most associated with various crash types (e.g., a large correlation between speeding citations and roadway departure crashes might reveal the need for better speed enforcement). Such information could help provide citation penalties that are more targeted to improve driving behavior or identify high-risk drivers that might need more proactive or comprehensive intervention measures.

Barriers

There are no institutional barriers for merging crash and vehicle data, and all data are owned and maintained by PennDOT. However, the primary barrier to integration is the outdated nature of the driver data system. Like the vehicle data system, this is an antiquated system that may require specialized software to create an interface that can communicate with the crash data system for real-time integration purposes. Data linkage is less onerous and can be performed at fairly regular intervals using known linkage variables, including PII that are available in the databases.

Pathways to integration

Driver information data are stored on an IBM mainframe database using COBOL 2, which is an older system than the crash database. Although it is possible to establish a direct communication between crash and driver databases, establishing such a connection may require the development of a specialized code. Data are already being shared between the two systems, and this provides a pathway for regular data linkage. Currently, crash records involving drivers with a Pennsylvania driver's license are added to the driver data system using the driver's license ID number as the linkage variable. Additional linkage variables include driver date of birth, name, and address, as shown in Table 18. The information that is shared to the driver data system includes the accident record number (last 7 digits of the CRN), county of the crash, injury severity, and crash date. The same information can be used to share other data elements from the driver data system back to the crash data system. Since accident record number, county of the crash, injury severity, and crash date are identical between crash and driver license data, deterministic data merging can be performed using these variables.

Table 18. Potential linkage variables for merging crash and driver databases

Crash Data		Driver Data	
Field name	Data description	Field name	Data description
DVR_LIC_NUM	Driver license number	olno	Identification number
DL_DVR_DOB	Date of birth of the driver from the driver's license	oper-birth-date	Birth day
DVR_FI	Driver first name	oper-first-name	First name
DVR_MI	Driver middle initial	oper-mid-name	Middle name
DVR_NM_LAST	Driver last time	oper-last-name	Last name
DVR_ADDR_1	Driver's address	oper-address	Current legal address
DVR_ADDR_CITY	Driver's city	oper-city	City
DVR_ADDR_ZIP	Driver's zip code	oper-zip	Zip code

Estimated timeline

Merging the crash and driver databases would require the development of an algorithm that is capable of matching driver's license numbers in both databases and verifying matched data with driver date of birth and address. The developed algorithm can also report the discrepancies between matched data for manual correction. Since there is an existing data linkage between these databases, development of this algorithm is expected to be simple.

Resource requirements and cost estimate

The research team estimates that merging the crash and driver databases would not require significant resources or labor costs. Since algorithms already exist to query information from the driver database to incorporate into the crash data, these algorithms could be readily modified to perform the requisite merging. These algorithms would need to be run at regular intervals to ensure that the merged dataset is up-to-date. The research team suggests that an automated algorithm be used that can perform the requisite merging at the same time data elements are validated in the crash database using the driver license information.

Metrics and likelihood for success

Similar to merging crash and vehicle records, one metric that can be used for merging crash and driver records is the fraction of crash records that can be linked to individuals in the Driver's License database. This is not likely to be 100%, as only records involving individuals with Pennsylvania driver's licenses can be merged. Thus, the percentage of crash records involving drivers with a Pennsylvania license can be used as an upper bound for the fraction of records matched between the two databases. Other metrics previously described can also be used as metrics for success.

In terms of likelihood for success, merging the crash and driver data systems is very likely to be successful, since the selected linkage variable (driver's license ID number) is deterministic and likely to have an exact match. Only crash records that do not involve Pennsylvania drivers or with erroneously entered driver's license numbers are not expected to be matched. For these reasons, integrating crash and vehicle data is likely to be very successful.

Crash and citation and adjudication data systems

Purpose of integration and anticipated results

Integration between the crash and citation and adjudication data systems can facilitate analysis of various driver-level behavioral characteristics and how they might influence crash outcomes. This can help to improve targeted enforcement efforts, especially with respect to driving behaviors such as drug and alcohol use, speeding, reckless driving, seat belt use, motorcycle helmet use, and distracted driving.

Barriers

One barrier to integration of crash and citation data is that the citation data are owned by the Administration Office of Pennsylvania Courts (AOPC), which does not have a significant desire for data integration with PennDOT's crash data. Another barrier is the complexity of the citation data. The citation and adjudication database was developed specifically for the court system and data are stored in a proprietary format. Furthermore, the exact structure of the data and comprehensive list of data elements are not available. Another barrier is the data update intervals. While records are added to the database daily, this is only done for citations that have reached a disposition. Most citations have a disposition within a short time frame (10 days), but there is a large proportion (up to about 15-20%) that do not. While having access to 80% of the data is certainly useful, the remaining 15-20% represent cases in which individuals did not respond to the citation or those that require a lengthy court process, and these cases might represent some of the more interesting citation cases to study. Furthermore, by law, individual records are only required to be stored for 3 years, and after that time period some court systems might eliminate electronic records. Thus, data availability might not be consistent across the state, which could serve as a barrier to integration.

Another barrier to integration is that the location information included in the citation database is not consistent with PennDOT's linear referencing system. Only the route, direction of travel, county, and township/borough/city are recorded consistently. An additional field is available that allows an officer to record more specific details on the location of a citation (e.g., a specific intersection or location along a route). However, this information will not be recorded consistently and will generally not be able to be matched to individual intersections or roadway segments. Thus, while merging the crash and citation data might facilitate some new knowledge at the route, county, or township/borough/city-levels, detailed insights into specific locations along the roadway network will not be possible.

Pathways to integration

The technical information about the citation and adjudication database is not available. Because of this, the possible compatibility issues and real-time data sharing capabilities are unknown. However, data exchange currently occurs between PennDOT and the court system through the PAJNET portal. Specifically, this portal allows the court system to access elements in the driver and vehicle data systems. It also allows the court system to provide citation information to the driver data system when a disposition is achieved. This citation information is stored in the driver data system and attached to individual drivers. This is significant because it means that an indirect pathway may be available between the crash and citation data using the driver data system as an intermediary. If full data integration could be achieved between the driver and crash data systems, traffic citation information could also be obtained. Specifically, violation information (date, type, conviction date, action, etc.), DUIs, and serious traffic offenses for those with commercial driver's licenses are stored in the driver data system. This might alleviate a need for direct data integration between the crash and citation and adjudication data systems.

If additional data elements are needed from the citation and adjudication data system—e.g., knowledge of other non-traffic citation offenses and their outcomes—then it is possible to perform data linkages using the same PII that is used to provide traffic citation information. This includes name, date of birth, and driver license number. Such information is sufficient for deterministic linkages.

Estimated timeline

The timeline to integrate crash and citation information will be the same as integrating the crash and driver's license databases, since the driver's license database contains the requisite citation information.

Resource requirements and cost estimate

The resources or labor costs to integrate crash and citation information will be the same as integrating the crash and driver's license databases, since the driver's license database contains the requisite citation information.

Metrics and likelihood for success

The metrics and likelihood for success of integrating crash and citation information will be the same as integrating the crash and driver's license databases, since the driver's license database contains the requisite citation information.

Crash and injury surveillance data systems

Purpose of integration and anticipated results

Integration of crash and injury surveillance data could provide a more accurate understanding of crash severity outcomes, which are currently only estimated by police when filling out police crash reports. The most commonly cited gain is a more accurate understanding of crash cost by severity. Such information is vital when performing benefit-cost analyses of future safety projects and behavioral or educational programs. This information could also be disaggregated by crash type, location, time of day, or other features to better understand the impacts of these factors on crash severity distributions. Combining crash and injury severity data from hospitals could also help better quantify the impacts of various strategies designed to reduce crash severity, such as seat belt use (automobiles) or helmet use (motorcycles).

From the injury surveillance side, integration with the crash data could help better understand the types of medical treatment needed for those involved in a crash. By studying patients involved in a crash and their long-term outcomes, better protocols could be put into place for future patients that have been involved in a crash. This is in line with recent trends toward data-driven patient care.

Barriers

The injury surveillance data system is made up of four unique databases (EMS, Trauma, Health Care, and Vital Statistics), and to be able to completely understand the severity outcomes of crashes, linkages to all four of these systems would be required. Since each of them has its own formatting and storage systems, the barriers to integration with the crash data system are different. Furthermore, each of these databases is maintained by a different agency:

- EMS – PA DOH
- Trauma – PTSF
- Health Care – PHC4
- Vital Statistics – PA DOH

The main barrier to integration with the EMS database is a lack of unique linkage variable. PII is not maintained in the statewide database, which makes it difficult to obtain deterministic linkages with crash data. As previously mentioned, EMS agency code has recently been added to police crash reports, which can facilitate probabilistic linkage using other variables.

The Trauma database also faces the lack of PII as a barrier to integration. Additionally, this database does not contain all Trauma patient data. Only data for patients that did not survive while admitted to the trauma center, or those that remained in the trauma center for a certain time period (24 hours for heavy injuries, 36 hours otherwise), are included. Because of that, records for those individuals with minor injuries who are transferred to a trauma center cannot be merged with the crash database.

The primary barrier within the Health Care database is patient privacy. PII is collected and stored in this database but cannot be released to any outside agencies to protect patient privacy.

For Vital Statistics, no representative was available to be interviewed for this project. Thus, the specific barriers are unknown.

Pathways to integration

This section provides potential pathways for integration between the crash data system and the EMS, Trauma, and Health Care databases, respectively.

Crash and EMS data

The EMS database is housed in a Microsoft Azure-based cloud database, but other technical details about this database are unknown. Depending on the structure of the database (i.e., SQL or NoSQL), real-time data communication between crash and EMS databases may be possible. However, since the crash database uses a different database structure (Oracle), a real-time communication link between databases may require software development. Because of this, non-real-time data merging, such as manually requesting data from EMS and importing those data to the crash database, is likely to be easier.

There are no common unique identifiers that can be used to link data between the crash and EMS databases. Instead, there are five common identifiers that can potentially be used to match records between the two databases; these are shown in Table 19. However, the values for these data elements may not be identical in each of the two databases. Possible differences between these values are listed below:

- Difference of recorded EMS unit arrival times by police (crash data) and emergency responders (EMS data)
- Spelling differences of the name of the destination medical facility
- The "TOT_PEOPLE" variable in the crash database is used to record the total number of people involved in the crash. However, a similar variable in the EMS database, "eScene.06," is used to record the total number of patients at the crash location. Thus, these values may not match exactly. However, since the total number of patients cannot be larger than the total number of people involved in a crash, it can be used as an identifier to eliminate unrelated data points.

Furthermore, the variable "eDispacth.01" has a specific value for traffic accidents, so it can be used to identify EMS records that are associated with traffic crashes.

For these reasons, deterministic linkage may not be possible. Instead, probabilistic data merging can be performed in which the most likely matching records are identified in one database using elements from the other. Additional variables from the EMS database noted in Table 19 can also be used to help this process. A set of rules that can detect and correct the irregularities between the crash and the imported EMS dataset can be developed to merge records across these two databases.

Table 19. Potential linkage variables for merging crash and EMS databases

Crash Data		EMS data	
Field name	Data description	Field name	Data description
EMS_AGENCY_CD	EMS agency identifier	eResponse.01	EMS agency number
EMS_ARRIVAL_TIME	Arrival time of EMS unit	eTimes.07	Unit arrived on scene date/time
CRS_DT	Crash date/time		
MED_FACILITY	Medical facility	eDisposition.02	Destination/transferred to
TOT_PEOPLE	Total number of people	eScene.06	Number of patients at scene
RDWY_CNTY_CD	County code	eScene.21	Incident county code
		eScene.19	Incident zip code
		eRecord.01	Patient care report (PCR) number
		eDispacth.01	Complaint reported by dispatch

Crash and Trauma data

The technical information about the PTSF database is not available. Because of this, the possible compatibility issues and real-time data sharing capabilities are unknown.

Directly merging crash and Trauma data using PII does not appear to be possible due to a lack of common linkage variables. However, it may be possible to link records in the EMS and Trauma databases using the patient care report number (PCR number). The PCR number is a unique number that is issued by the EMS agency to identify a patient. The importance of the PCR number is that it remains constant for an individual throughout the hospitalization process until all patient care is completed. Thus, it can be used to track patients from one system to another. Unfortunately, it was reported that only 54% of patient records in the Trauma database have a valid PCR number. The remaining patients either did not arrive to the trauma center via EMS or the patient's PCR number was simply not documented. It was also reported that most transfer patient records have no PCR number. Still, for the 54% of the records that do exist, deterministic linkage could be performed with the EMS database. Thus, crash data could be indirectly linked to a portion of the Trauma data using the EMS database as an intermediary.

Crash and Healthcare data

The technical information about the healthcare database is not available. Because of this, the possible compatibility issues and real-time data sharing capabilities are unknown. However, the inability of the PHC4 to share PII in the Health Care database due to privacy concerns likely means that real-time integration will not be possible.

Instead, deterministic data linkage between the crash and Health Care data can be performed at regular intervals. It was noted by representatives of the Health Care database that the PHC4 could obtain crash data from PennDOT with unique identifiers (PII) and merge these records with the Health Care records. This provides a pathway for deterministic data integration using PII as the linkage variables.

Another avenue is probabilistic linkage. For patients admitted to the hospital, the variables described in Table 20 can be used for linkage purposes. The variable "Field 72a-72c" (external cause of morbidity) can be used to identify all records that involved traffic accidents. Crash date/time and medical facility name can be used to probabilistically merge these records by assuming the hospital admission occurs shortly after the crash occurred. Possible issues with this approach include:

- Crash time and admission time are related but may be different. In some cases, the time difference can be assumed or a narrow range can be applied to identify potential records in the Health Care database for a specific crash time.
- If more than one person is injured in the crash and transported to the same hospital, this method will find the data of all individuals injured in the same crash. To differentiate these data, the "Field 72a-72c" variable that describes the external cause of morbidity can be used. This data item uses the codes in ICD-10-CM, which specifies whether the injured person is the driver, or a pedestrian, cyclist, or passenger.

Table 20. Potential linkage variables for merging crash and inpatient databases

Crash Data		Inpatient data	
Field name	Data description	Field name	Data description
CRS_DT	Crash date/time	Field 12	Admission date
CRS_TM	Time that crash occurred	Field 13	Admission time
MED_FACILITY	Medical facility to which individual was transported	Field 2	The name and address of the facility
		Field 72a-72c	External cause of morbidity (ICD-10-CM injury codes)

Note that this method might not be able to identify individuals who check into a hospital much later than when the crash occurred (e.g., went to the hospital the next day with an injury).

Outpatient data records cover individuals who are not admitted to a hospital (i.e., those with minor injuries and illnesses). This would include those who were denied EMS service. It is not possible to merge these data without access to PII due to a lack of other possible linkage variables.

Estimated timeline

There is no existing linkage between the EMS database and the crash database. Also, there is no information about the data quality of the EMS database. Thus, to develop a probabilistic merging framework, data quality assessment for both databases needs to be done, descriptiveness of the linkage variables needs to be determined, and a probabilistic linkage model needs to be developed. All of these steps require extensive analysis and experimentation on sample datasets from both databases. Even though the initial development time is likely to be the longest among the other dataset pairs, once the model is developed the application of the model is likely to require the same level of effort with other database merging operations.

In order to merge Trauma and crash data, a deterministic merging algorithm is needed. Since the only linkage variable is the PCR number and there are no other variables for validation purposes, the development of this algorithm is expected to be simple.

As previously mentioned, merging crash and Health Care data will require that the crash data files are provided to the PHC4, which will perform the merge and provide the merged files back to PennDOT. The timeline for this process is not currently known, although the presence of well-defined linkage variables suggests that it will not take long. However, the process to set up the required agreements between PennDOT and PHC4 may take some time, since it will require negotiating the frequency at which the merge will be performed, the set of variables that can be provided from PHC4 to PennDOT and any costs that might be incurred.

Resource requirements and cost estimate

The research team estimates that merging the crash and injury surveillance databases would be the most cost intensive of all the integration pairs. New algorithms would need to be developed to probabilistically merge the crash and EMS data, since deterministic linkage variables are not available. This would require significant up-front coding costs. A series of verification algorithms would also be needed to alert an analyst of discrepancies or records that could not be matched. These discrepancies would need to be checked manually by an analyst; thus, significant costs would also be required for an analyst to oversee the merge every time it is performed. For the Trauma data, a deterministic matching algorithm will need to be developed to merge the approximately half of the Trauma records that have a valid PCR with the EMS data before the EMS data are merged with the crash data. This will require a one-time cost to develop this algorithm, but no significant analyst costs are expected each time the merging is performed, since the linkage variable (PCR record number) is deterministic. The costs to merge the crash and Health Care data are unknown, since this must be performed directly by the PHC4. However, since deterministic, PII linkage variables are available, the matching is expected to be relatively straightforward. It is anticipated that PennDOT might incur a small cost each time the merge is performed; however, these details would need to be negotiated between PennDOT and PHC4.

Additionally, although well-defined data sharing agreements are in place and data acquisition costs are generally waived for state agencies, there is a chance that some costs to acquire the data may be incurred based on the frequency at which injury surveillance are obtained. These costs would need to be negotiated between PennDOT and the agencies that own and maintain these individual datasets.

Metrics and likelihood for success

According to NEMSIS data quality dashboard[7], 93% of the nationwide EMS data have valid information. Thus, with an accurately calibrated probabilistic merging model, it is possible to successfully merge most of the crash data with EMS data, when applicable.

Since PCR number from the merged crash and EMS data is needed to merge Trauma and crash data, success of this merging primarily depends on the success of the EMS-crash merge. However, there are also other factors that can lead to a low percentage of matched data. These are:

Missing PCR numbers in Trauma data due to missing trip sheet from EMS agency or human error.
Since PCR number is assigned by EMS agency, people who deny EMS service do not have a PCR number.
Data of people with minor injuries are not available in Trauma data.

[7] https://nemsis.org/view-reports/public-reports/version-3-public-dashboards/v3-public-data-quality-dashboard/

According to representatives with knowledge of the Trauma database, approximately 54% of the Trauma records have an associated PCR number. Thus, we can expect about half of the applicable crash records to be matched to the associated Trauma data.

No details are available on the quality of the Health Care data; however, since PII are available to match these data records, it is expected that the vast majority of relevant crash records can be matched to Health Care records.

The likelihood for success of merging the crash and injury surveillance data is strong, but not as high as merging the crash database with the other data systems due to the need for probabilistic merging, significant algorithm development, analyst verification, and reliance on the PHC4 to merge the crash and Health Care data. Nevertheless, potential linkage variables are available and so the success of merging crash and injury surveillance data will be directly related to the amount of resources available to support this effort.

POTENTIAL INTEGRATION FRAMEWORKS

This section describes potential strategies for safety data integration within Pennsylvania. These strategies include:

- Development of a system interface
- Database linkage at regular intervals
- Creation of a single safety database

Regardless of the approach selected, a critical component of successful data integration is the existence of a flexible data sharing agreement between PennDOT and the agencies that own and maintain individual datasets. This is especially relevant for the creation of a single safety database, since the database might be maintained by an outside party. As described in this report, existing MOU frameworks are already in place with these agencies; however, PennDOT should ensure that these existing MOUs meet their needs for data integration. The research team also recommends that PennDOT create an internal data use guide that outlines who may access these integrated datasets and how they may be used, particularly with respect to sharing information with the public.

The remainder of this section describes each of these options and provides an assessment of the feasibility of each approach.

System interface

The most comprehensive method for data integration is the development of a system interface. In this approach, no additional databases are required to be developed or maintained. Instead, data are stored in each of their respective data systems and are accessed only when necessary. The system interface queries records from each of these systems and links them in real time. For this approach to be possible, direct access to each data system is required. This usually requires modern databases that facilitate cloud-based or electronic access, since these data systems are generally maintained by different agencies and stored on different networks. Furthermore, unique deterministic linkage variables are needed to link records across the different data systems.

Based on the research team's review of the core data systems in Pennsylvania, this approach does not appear to be feasible for safety data integration within Pennsylvania for several reasons. First, two of the six core data systems—the driver and vehicle systems—are stored on older mainframes that do not easily facilitate electronic access outside of local PennDOT networks. Additionally, direct access to the citation and adjudication and injury surveillance data systems does not seem to be likely based on discussions with representatives from these systems. Instead, it is more likely that data from these systems be made available at regular intervals for linkage/integration purposes. Moreover, records in the individual data systems are not updated in real-time or on the same schedule. Thus, information obtained from the various data systems using an interface approach may not be up-to-date. Lastly, the availability of unique deterministic linkage variables is not consistent across the data systems. The most likely linkage variables available for a system interface approach are PII, such as name, date of birth, or social security number; however, PII is either not stored (EMS, Trauma) or not available (Health Care) in some of the data systems. Instead, probabilistic linkages will be necessary for the EMS and Trauma databases, which may require additional filters or steps to ensure that the most likely match is achieved, which serves as another barrier to real-time integration, while non-PII linkage variables are simply not available to facilitate real-time integration for Health Care.

Database linkage at regular intervals

Another option for data integration is to link specific pairs (or groups) of data systems at regular intervals. This is typically done in an ad hoc manner as a part of projects with a specific goal in mind (e.g., as previously mentioned, roadway and crash data are linked for specific HSM application purposes). However, this can simply be performed by an agency at regular intervals for the purposes of having integrated data for future use. One advantage of this approach is that the costs for data integration can be covered by the projects that require the merging of specific databases. If performed as part of a project, the integration is typically directly tied to a tangible knowledge gain outcome. One disadvantage of this approach is the maintenance of the merged databases, particularly if the merging is done as a part of a specific project. These databases need to be obtained upon the completion of individual projects for future use and the merging re-performed when new data records are obtained or updated. If resources are not allocated to update the merged database regularly, the data would quickly become outdated and lose their usefulness. Also, if merged in an ad hoc manner through individual projects, multiple entities might be responsible for merging the same database at different points in time. This could lead to duplicate effort in terms of algorithm development or lessons learned with respect to integration. Another drawback is that some data system pairs may never be integrated unless a specific project is performed to do so.

Based on the research team's review of the core data systems in Pennsylvania, performing database linkage in this way appears to be a feasible approach in Pennsylvania. This method is already being performed as a part of individual projects, such as the development of HSM methods and safety management tools for Pennsylvania. Linking crash data with vehicle, driver, and roadway data is relatively straightforward since it can be done using the known linkage variables available in both databases. Linking crash and citation data can also be done using the driver database as an intermediary, as previously described. The most difficult pair would be linking crash data with injury surveillance data. A probabilistic linkage is possible for the EMS and Trauma data that could be performed as a part of a project. Linking crash data with Health Care data would require working directly with the PHC4, as the linkage variables in the Health Care data (PII) could not be shared outside of the PHC4. Thus, crash data would have to be sent to the PHC4, who would perform the merge directly and send the dataset back to PennDOT.

The research team recommends that the databases be merged in the following order based on the likelihood for success, anticipated costs, and potential for knowledge gain:

1. Crash and roadway data
2. Crash and driver/vehicle data (including citation data obtained from driver database)
3. Crash and Health Care data
4. Crash and EMS/Trauma data

If this method is chosen, the research team recommends that these database linkages be performed on an annual basis. This is based on the update frequency and data entry lag associated with the database and the desire to balance between having the most-to-date information and integration costs. Since the integration will have to be performed from scratch each time, merging the datasets at more frequent intervals would likely not be worth the additional costs.

Creation of a single safety database

A final option for data integration is the creation of a single database that contains all related safety data. As outlined in the review of current state agency practices, multiple other states employ this approach for safety data integration, including Connecticut, Maryland, North Carolina, and Utah. More specifically, these states use a third-party—specifically, a university partner—to leverage their existing resources. In this method, the agency responsible for maintaining this single database will be responsible for obtaining data from each of the six core data systems and merging the data into a usable format. One advantage of this approach is that the data systems are complex and require time to understand their intricacies, thus a single organization will be responsible for learning the nuances for each of the data systems for integration purposes. This organization would also be well-suited to lead data analysis activities since one of the main barriers for analysis—learning the data—would be much less significant. Drawbacks of this approach are similar to performing individual linkages at various intervals. Specifically, the merged dataset would generally lag behind the individual datasets in terms of the data records that are included, unless data are shared on a very regular basis.

Based on the research team's review of the core data systems in Pennsylvania, this appears to be a feasible approach for Pennsylvania data. Linkage between the databases would proceed in a similar manner to that described in the previous section. Note that since Health Care data must be merged at the PHC4, the organization in charge of the combined database would need to send a subset of the data to the PHC4 at regular intervals to obtain the Health Care data, which would then have to be incorporated into the combined database.

However, there are some challenges that should be noted. First, the organization in charge of maintaining this combined database would need to obtain the requisite permissions from each of the data owners to access and use these data. Existing MOUs can be leveraged and it appears that data sharing issues for third-parties can be alleviated if the dataset is being compiled on behalf of a state agency (PennDOT). Second, the organization responsible for maintaining the database would need to ensure that the appropriate privacy and security measures are in place to obtain and store such sensitive data, including data that include PII. A well-defined access plan would also have to be developed to control access to these data. Access to an institutional review board could be helpful to ensure that appropriate protocols are in place and followed.

Similar to the database linkage model, the research team recommends that the databases be merged in the following order based on the likelihood for success, anticipated costs, and potential for knowledge gain:

1. Crash and roadway data
2. Crash and driver/vehicle data (including citation data obtained from driver database)
3. Crash and Health Care data
4. Crash and EMS/Trauma data

If this method is chosen, the research team recommends that these database linkages be performed on a quarterly basis. This is based on the update frequency and data entry lag associated with the database, and the desire to balance between having the most-to-date information and integration costs. Since the integration would be done by a single entity that manages the data, the integration can be performed more frequently and thus provide a more up-to-date integrated safety database than individual linkages at regular intervals.

References

Abay, K.A., 2015. Investigating the nature and impact of reporting bias in road crash data. *Transp. Res. Part A Policy Pract.* 71, 31–45. doi:10.1016/J.TRA.2014.11.002

Alaska Traffic Records Coordinating Committee, 2015. *Alaska Traffic Records Strategic Plan*.

Alsop, J., Langley, J., 2001. Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accid. Anal. Prev.* 33 3 , 353–359. doi:10.1016/S0001-4575(00)00049-X

Amoros, E., Martin, J.-L., Laumon, B., 2006. Under-reporting of road crash casualties in France. *Accid. Anal. Prev.* 38 4 , 627–635. doi:10.1016/J.AAP.2005.11.006

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. *Accid. Anal. Prev.* 31 1–2 , 101–108. doi:10.1016/S0001-4575(98)00051-7

Boufous, S., Finch, C., Hayen, A., Williamson, A., 2008. *Data Linkage of Hospital and Police Crash Datasets in NSW*.

Boufous, S., Williamson, A., 2006. Work-related traffic crashes: A record linkage study. *Accid. Anal. Prev.* 38 1, 14–21. doi:10.1016/J.AAP.2005.06.014

Burch, C., Cook, L., Dischinger, P., 2014. A Comparison of KABCO and AIS Injury Severity Metrics Using CODES Linked Data. *Traffic Inj. Prev.* 15 6, 627–630. doi:10.1080/15389588.2013.854348

Burdett, B., Li, Z., Bill, A.R., Noyce, D.A., 2015. Accuracy of Injury Severity Ratings on Police Crash Reports. *Transp. Res. Rec. J. Transp. Res. Board* 2516 1, 58–67. doi:10.3141/2516-09

Cambridge Systematics, 2018a. *Colorado State Traffic Records Advisory Committee Strategic Plan*.

Cambridge Systematics, 2018b. *Oregon Traffic Records Strategic Plan*, Federal Fiscal Year 2018.

Cambridge Systematics, 2017. *Florida Traffic Safety Information System*.

Cercarelli, L.R., Rosman, D.L., Ryan, G.A., 1996. Comparison of accident and emergency with police road injury data. *J. Trauma Acute Care Surg.* 40 5, 805–9.

Clark, D.E., 1993. Development of a statewide trauma registry using multiple linked sources of data. *Proceedings. Symp. Comput. Appl. Med. Care* 654–8.

Clark, D.E., Winchell, R.J., Betensky, R.A., 2013. Estimating the effect of emergency care on early survival after traffic crashes. *Accid. Anal. Prev*. 60, 141–147. doi:10.1016/j.aap.2013.08.019

Conderino, S., Fung, L., Sedlar, S., Norton, J.M., 2017. Linkage of traffic crash and hospitalization records with limited identifiers for enhanced public health surveillance. *Accid. Anal. Prev*. 101, 117–123. doi:10.1016/j.aap.2017.02.011

Connecticut Traffic Records Coordinating Committee, 2018. *State of Connecticut Strategic Plan for Traffic Records*.

Cook, L.J., Kerns, T., Burch, C., Thomas, A., Bell, E., 2009. *Motorcycle Helmet Use and Head and Facial Injuries: Crash Outcomes in CODES-Linked Data*. National Highway Traffic Safety Administration, Report No. DOT HS 811 208.

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. *Ann. Emerg. Med*. 35 6 , 585–91.

Cryer, P.C., Westrup, S., Cook, A.C., Ashwell, V., Bridger, P., Clarke, C., 2001. Investigation of bias after data linkage of hospital admissions data to police road traffic crash reports. *Inj. Prev*. 7 3 , 234–41. doi:10.1136/IP.7.3.234

Daniello, A., Gabler, H.C., 2012. Characteristics of Injuries in Motorcycle-to-Barrier Collisions in Maryland. *Transp. Res. Rec. J. Transp. Res. Board* 2281 1 , 92–98. doi:10.3141/2281-12

Donnell, E.T., Gayah, V.V., Jovanis, P., 2014. *Safety Performance Functions*, Final report for the Pennsylvania Department of Transportation, FHWA-PA-2014-007-PSU WO 1.

Donnell, E.T., Gayah, V.V., Li, L., 2016. *Regionalized Safety Performance Functions*, Final report for the Pennsylvania Department of Transportation, FHWA-PA-2016-001-PSU WO 17.

Georgia Traffic Records Coordinating Committee, 2016. Georgia Traffic Safety Information System Improvement Grant, *2016-2017 Documentation and Strategic Plan*.

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. *Int. Emerg. Nurs*. 31, 2–8. doi:10.1016/j.ienj.2016.04.004

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 Inj. Prev*. 16 6 , 605–610. doi:10.1080/15389588.2014.999856

Idaho Traffic Records Coordinating Committee, 2015. *Idaho Traffic Record Systems Strategic Plan*.

Kuhl, J., Evans, D., Papelis, Y., Romano, R., Watson, G., 1995. The Iowa Driving Simulator: An immersive research environment. *Computer* (Long. Beach. Calif). 28 7, 35–41. doi:10.1109/2.391039

Langley, J., Stephenson, S., Cryer, C., 2003. Measuring Road Traffic Safety Performance: Monitoring Trends in Nonfatal Injury. *Traffic Inj. Prev*. 4 4 , 291–296. doi:10.1080/714040487

Langley, J.D., Dow, N., Stephenson, S., Kypri, K., 2003. Missing cyclists. *Inj. Prev*. 9 4, 376–9. doi:10.1136/IP.9.4.376

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. *Accid. Anal. Prev*. 32 6 , 771–7.

Lujic, S., Finch, C., Boufous, S., Hayen, A., Dunsmuir, W., 2008. How comparable are road traffic crash cases in hospital admissions data and police records? An examination of data linkage rates. *Aust. N. Z. J. Public Health* 32 1, 28–33. doi:10.1111/j.1753-6405.2008.00162.x

Maryland Traffic Records Coordination Committee, 2016. *Traffic Records Strategic Plan FFY 2016-2020*.

Mcglincy, M.H., 2004. A Bayesian Record Linkage Methodology for Multiple Imputation of Missing Links. *ASA Proc. Jt. Stat. Meet*.

Michigan Traffic Records Coordinating Committee, 2006. *Strategjc Plan 2006-2010*.

Mitchell, M., Newman, M., 2002. Complex systems theory and evolution. *Encycl. Evol*. 1–5.

Montana Department of Transportation, 2017. *2015 Traffic Records Strategic Plan Update*.

National Highway Traffic Safety Administration, 2018a. *Traffic Records Program Assessment Advisory*.

National Highway Traffic Safety Administration, 2018b. *FY 2018 State Grant Determinations* [WWW Document]. URL https://www.nhtsa.gov/es/highway-safety-grants-program/fy-2018-grant-funding-table

National Highway Traffic Safety Administration, 2014. *Linking Traffic Records Data Systems*.

National Highway Traffic Safety Administration, 2011. *Model Performance Measures for State Traffic Records Systems*.

National Highway Traffic Safety Administration Technical Assessment Team, 2016. *State of Illinois Traffic Records Assessment*.

Nebraska’s Traffic Records Coordinating Committee, 2018. *Nebraska Traffic Records System Plan FY2015-2019*.

New Mexico Department of Transportation, 2016. *Statewide Traffic Records System Strategic Plan Federal*

Fiscal Years 2017-2019.

Pennsylvania Traffic Records Coordinating Committee, 2018. 2019 *Traffic Records Strategic Plan*.

Rosman, D.L., 2001. The Western Australian Road Injury Database (1987–1996): Ten years of linked police, hospital and death records of road crashes and injuries. *Accid. Anal. Prev*. 33 1, 81–88. doi:10.1016/S0001-4575(00)00018-X

Rosman, D.L., Knuiman, M.W., 1994. A comparison of hospital and police road injury data. *Accid. Anal. Prev*. 26 2 , 215–222. doi:10.1016/0001-4575(94)90091-4

State of Kansas Traffic Records Coordinating Committee, 2015. *Kansas Traffic Records System Performance Measurement Report*.

Tennyson, N.J., 2016. *Strategic Plan and Information Technology Plan FY2017-2019*.

Texas Department of Transportation, 2012. *2012 Update to the Texas Traffic Safety Information System Strategic Plan*.

University of Kentucky Transportation Center, 2017. *Kentucky Traffic Records Strategic Plan 2017-2021*.

Utah Traffic Records Advisory Committee, 2015. *Utah Traffic Records Information Systems Strategic Plan*.

Vernon, D.D., Cook, L.J., Peterson, K.J., Michael Dean, J., 2004. Effect of repeal of the national maximum speed limit law on occurrence of crashes, injury crashes, and fatal crashes on Utah highways. *Accid. Anal. Prev*. 36 2 , 223–229. doi:10.1016/S0001-4575(02)00151-3

Washington Traffic Records Committee, 2017. *Washington Traffic Records Committee Strategic Plan*.

Watson, A., Watson, B., Vallmuur, K., 2015. Estimating under-reporting of road crash injuries to police using multiple linked data collections. *Accid. Anal. Prev*. 83, 18–25. doi:10.1016/J.AAP.2015.06.011

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. *Accid. Anal. Prev*. 49, 30–35. doi:10.1016/j.aap.2011.03.007