Development of a Comprehensive Approach for Serious Traffic Crash Injury Measurement and Reporting Systems (2021)

DETAILS

80 pages | 8.5 x 11 | PDF
ISBN 978-0-309-09345-3 | DOI 10.17226/26305

CONTRIBUTORS

Carol A. Flannagan, Jonathan D. Rupp, and N. Clay Mann; National Cooperative Highway Research Program; Transportation Research Board; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2021. *Development of a Comprehensive Approach for Serious Traffic Crash Injury Measurement and Reporting Systems*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26305.

NCHRP

Web-Only Document 302

Development of a Comprehensive Approach for Serious Traffic Crash Injury Measurement and Reporting Systems

Carol A. Flannagan
Jonathan D. Rupp
University of Michigan Transportation Research Institute
Ann Arbor, MI

N. Clay Mann
University of Utah School of Medicine
Salt Lake City, UT

Final Report for NCHRP Project 17-57
Submitted July 2015

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

TRB
TRANSPORTATION RESEARCH BOARD

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, non-governmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

The **Transportation Research Board** is one of seven major programs of the National Academies of Sciences, Engineering, and Medicine. The mission of the Transportation Research Board is to provide leadership in transportation improvements and innovation through trusted, timely, impartial, and evidence-based information exchange, research, and advice regarding all modes of transportation. The Board's varied activities annually engage about 8,000 engineers, scientists, and other transportation researchers and practitioners from the public and private sectors and academia, all of whom contribute their expertise in the public interest. The program is supported by state transportation departments, federal agencies including the component administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation.

Learn more about the Transportation Research Board at **www.TRB.org**.

COOPERATIVE RESEARCH PROGRAMS

CRP STAFF FOR NCHRP WEB-ONLY DOCUMENT 302

Christopher J. Hedges, *Director, Cooperative Research Programs*
Lori L. Sundstrom, *Deputy Director, Cooperative Research Programs*
David Jared, *Senior Program Officer*
Clara Schmetter, *Senior Program Assistant*
Natalie Barnes, *Director of Publications*
Jennifer Correro, *Assistant Editor*

NCHRP PROJECT 17-57 PANEL
Field of Traffic—Area of Safety

John Milton, *Washington State Department of Transportation, Olympia, WA* (Chair)
Leanna Depue, *Governors Highway Safety Association (GHSA), Belton, TX*
Dia Gainor, *National Association of State EMS Officials, Boise, ID*
Helen Holly Hackman, *Boston Medical Center, Boston, MA*
Timothy Kerns, *Maryland Department of Transportation, Glen Burnie, MD*
Jana Simpler, *Delaware Office of Highway Safety, Dover, DE*
David E. Sugerman, *CDC Division of Injury Response, Atlanta, GA*
Ana Maria Eigen, *FHWA Liaison*
John Kindelberger, *NHTSA Liaison*
Noah Smith, *NHTSA Liaison*
Kelly Hardy, *AASHTO Liaison*
Holly Hedegaard, *National Center for Health Statistics Liaison*
Jacob A. Nelson, *AAA National Office Liaison*
Bernardo Kleiner, *TRB Liaison*

Contents

1 Summary

The goal of the National Cooperative Highway Research Program (NCHRP) Project 17-57 is to develop a comprehensive roadmap for states to measure serious injuries in crashes. This goal has been motivated by the Moving Ahead for Progress in the 21st Century Act (MAP-21), which requires a set of performance metrics to include assessment of serious injuries in crashes.

The first task of the NCHRP 17-57 project was to recommend a definition of serious injury for use in these performance metrics. We recommended using a Maximum Abbreviated Injury Scale score of 3 or greater (MAIS 3+) to define serious injury (Flannagan et al., 2012). The key element of this recommendation is to use a diagnosis-based definition of serious injury. However, using a diagnosis-based definition of serious injury for highway performance metrics requires data linkage between crash and medical outcome.

The second task of the project was to recommend near-term solutions for measuring serious injuries in crashes. We recommend two approaches that allow for states to measure serious injuries using a medical-diagnosis-based definition such as MAIS 3+. The first is to use state trauma or hospital discharge databases to count serious injuries in crashes. A majority of states have reasonably comprehensive trauma databases in place and can use them for this purpose while more comprehensive linkage is being put in place. The second near-term approach is to use sampling of hospital records for a subset of crashes. Efficient, stratified sampling can allow states to estimate the number of serious injuries and their association to certain roadway, crash, vehicle, behavioral and occupant characteristics. A third near-term approach discussed was to use regression to “correct” KABCO-based measures. We do not recommend this as a near-term approach except in limited circumstances where other options are not available or older, legacy datasets are being used.

A survey of states indicated that data linkage is a priority for a majority of states. Those that are currently linking are generally doing so using probabilistic linkage methods, typically developed through an existing Crash Outcome Data Evaluation System (CODES) program. Probabilistic linkage is a method of estimating which cases in a pair of datasets refer to the same person, even when the datasets do not contain a unique identifier for those individuals. Probabilistic linkage was the focus of the CODES program and allows states to link datasets after the fact. A variety of alternatives to probabilistic linkage are being considered and tried in a number of states. However, at this time, no state has successfully implemented a non-probabilistic approach to statewide linkage.

This report presents a roadmap for states to develop comprehensive crash-related data linkage systems, with special attention to measuring serious injuries in crashes. The summary below provides a brief background on data and data linkage, and then presents a set of ten steps that states must complete to set up a linkage system at the state level.

Requirements for Linkage

Before presenting the roadmap, it is important to understand the four underlying requirements for linkage to occur. These include: 1) having statewide datasets in place, 2) having common identifiers in the datasets to be linked, 3) having access rules and a mechanism for controlling access, and 4) having a place or mechanism for storing the linked data.

Statewide databases are required for linking data at the state level, and these databases must also be evaluated in certain ways before linkage results can be understood. Characteristics

of databases to be considered include coverage, schema, quality control, and timeliness. Coverage refers to the percent of possible reporting agencies that are participating or percent of possible cases that are included in the database. A schema is a codebook of variables, possible values, and formats. Variables and values should be complete and collected and reported in the same way across reporting units. Quality control (QC) is the process of checking consistency and completeness at the individual data-element level (e.g., percent missing data for different data elements; agreement between related data elements). Database quality issues will multiply through the linkage process, so it is critical to start with high-quality databases. Finally, databases should be quality-checked and available as quickly as possible. In some cases, linkage in near-real-time (e.g., within 24 hrs) is possible. In others, linkage occurs on an annual cycle. Either way, the component databases need to be processed in a timely way to allow linkage to proceed in time for analysis and planning to be completed.

To link people in any pair of datasets, one or more common identifiers must be present in both datasets. The gold standard of identifiers is a single, unique, permanent person-specific, alphanumeric identification code (ID) used in all datasets and assigned to all people in all datasets. However, several less ambitious forms of linking variables can also be used effectively. These include event-specific/person-specific identifiers or collections of non-unique but common identifiers that can be used in combination.

A good statewide data linkage system has rules for access as well as software to allow appropriate access and prevent inappropriate access. Rules for access must comply with the Health Insurance Portability and Accountability Act (HIPAA) and state law, and the level of access may be different for different individuals. It is even possible that state laws that impede linkage may need to be changed.

Finally, the datasets must be stored in some way that allows future use of the linkage. A data warehouse is an integrated and standardized means of storing a variety of different datasets and allowing linkage between subsets of them. Access to only the de-identified linked data can be controlled through individual login and password in the data warehouse analysis software. Although a linked dataset can be stored separately, we recommend the data warehouse model because only one copy of each component dataset is kept and updated. That way, updates to the dataset are reflected in the linked dataset and copies do not proliferate. In addition, the owner of the original dataset can maintain control of the database.

Roadmap to Link Data

The roadmap consists of ten steps:

Step 1: Set up a system for collaboration and communication among all relevant agencies. This is often done through the Traffic Records Coordinating Committee (TRCC), but data linkage projects might also have a focused advisory group and specific buy-in. This group will need to assess whether there are legal hurdles to linkage and address these (possibly through changes to state law) early on.

Step 2: Catalog all available relevant databases. The catalog should include schemas, inclusion criteria, coverage, and quality issues. In some cases, datasets will have to be brought up to a higher level of quality or coverage before linkage can proceed.

Step 3: Determine which databases will be linked. A long-term plan for the order of adding linkages might be mapped out in this step, along with hurdles that need to be overcome for each. The most effective linkage between crash and medical diagnosis data will include EMS (crash-EMS-hospital/trauma).

Step 4: Identify the identifiers. This involves determining what variables the datasets have in common. In some cases, a unique identifier will be present, allowing direct linkage via the identifier. However, in most cases, a unique identifier will not be present and groups of common variables should be identified.

Step 5: Determine the linkage mechanism. If a unique, common identifier is available, linkage should be immediately possible. However, this is unlikely for most state crash databases. Instead, linkage will require the addition of a unique, common identifier to the data systems or probabilistic linkage, which uses a group of common (non-unique) identifiers.

This report provides extensive discussion of potential linkage mechanisms that are used or being piloted by states. We are not aware of any state that is currently using a non-probabilistic approach to linkage involving crash data, though several are piloting methods or planning pilot tests.

Two methods of after-the-fact linkage are probabilistic linkage and hand linkage. Other states are exploring ways to pass a unique identifier between databases to enable future linkage.

Probabilistic Linkage. Probabilistic linkage was the focus of the CODES program. This method uses common variables in two datasets that do not have a unique identifier in common to infer which cases are linked and assign a probability to that linkage.

While probabilistic linkage is an excellent tool for linking datasets that do not share unique identifiers, the quality of the linked dataset needs to be assessed, just as the quality of the original datasets were. Poor quality linkage can result in biased analysis results and requires more complex analytical techniques to overcome this limitation. Linkage quality metrics exist, but are not generally used by states. Although software complexity has been an issue for many states, especially at the beginning, North Carolina (for example) has demonstrated that the method can be used in a timely fashion, producing a linked crash-EMS-trauma dataset for use in planning within two months after the dataset closes (e.g., at the end of the year).

Hand Linkage. Hand linkage is an after-the-fact method that is used to link EMS and trauma data in some states. The process uses software to identify a small set of possible cases in one database (e.g., a crash database) that might link to a single case in another database (e.g., trauma). The choice is then given to a human (e.g., the trauma registrar) who will make a judgment about which case is most likely to be the correct match.

Unique Identifiers. Other linkage methods involve assigning and passing an identifier between agencies (e.g., police and EMS) at the scene. The gold standard of identifiers is the person-specific/global identifier. This is a number that is assigned to an individual and follows them throughout datasets and time, across hospitals and even across different crash events. Two states are working towards this, but it is challenging to implement. A person-specific/event-specific identifier applies to a single person within a specific event. Often, these identifiers do not follow the individual through hospital transfers, though it may be possible to identify transfers after the fact. Finally, an event-specific identifier is assigned to all victims in a single crash event, and then individuals must be separately identified and linked after the fact. For example, Global Positioning System (GPS) location and crash report number are event-specific. Although further identification of individuals is required using probabilistic linkage, the event-specific ID will improve probabilistic linkage quality.

Step 6: Determine a storage mechanism for the linkage. The recommended approach is the data warehouse. The data warehouse typically consists of a front end (reporting or analysis tools), software that can access component databases, and translation software custom-built to standardize each component database. The included databases do not need to be stored in any

specific location, and updates to the original database will be reflected in the data warehouse. The warehouse can link databases that have a common unique identifier. With this approach, a unique identifier, either passed at the scene or added after the fact via probabilistic or hand linkage, is incorporated in the original datasets, wherever they reside.

The data warehouse approach requires some up-front costs to write the translation software and incorporate each database. However, the approach has several advantages. First, it does not proliferate copies of the original datasets, but instead translates the original, allowing for changes to be made in only one place. Second, it allows the owner agency to retain control over the original dataset and its versions and corrections. Third, the data warehouse provides a single system controlling user access on a database-by-database and even variable-by-variable basis. Finally, it allows users to use one software tool for reporting and analysis for all included databases.

An alternative used in some states is to store the linked database separately. The advantage of this approach is reduced cost, but the disadvantage is that the database is static and therefore does not automatically reflect any changes to the original. The database itself may fall under different laws than the original crash dataset because of the inclusion of medical information, and this may create access issues.

Step 7: Harmonize common data elements. For databases to be linked, all common variables must have the same schema. This includes variable formats as well as codes and values. Wherever possible, it is best to standardize to a national schema or data standard, such as the National EMS Information System (NEMSIS) or the Model Minimum Uniform Crash Criteria (MMUCC). This facilitates linkage to other databases that are harmonized with those standards (e.g., National Trauma Data Bank (NTDB)) and allows for re-use of existing Extensible Markup Language (XML), tools, and training materials.

Step 8: Set up a pilot project. A pilot of an assigned-on-scene identifier might occur in a limited geographical area. A pilot of probabilistic linkage should be done on at least one year of whole state datasets because the method's success depends on dataset size and the contents of variables. Piloting allows identification of logistical and dataset issues before a full-scale effort is launched.

Step 9: Set up a sampling program. Step nine is technically optional, but potentially very useful. We recommend setting up a sampling program where medical outcome is obtained for a subset of crash-involved people. This sample provides an estimate of serious injuries in crashes that can be used for performance metrics, without implementing a statewide linkage program. Moreover, it provides a way of evaluating the outcome of a developing linkage system. Early in any linkage process, it is likely that the resulting linked dataset will be biased or of low quality. An independent sample can provide accurate numbers for planning while setting the bar for a more comprehensive approach.

Step 10: Statewide linkage. Finally, the last step is to set up a full-scale statewide linkage program. It should be an expanded version of the pilot program, and any issues that arose during piloting should be resolved.

Recommendations

In carrying out a data linkage program at the state level, there are some challenges that states will have to face. In some cases, finding answers to these challenges at a national level will be efficient. In other cases, state-specific challenges will need to be addressed individually. The

following recommendations revolve around issues that would be most effective to address at a centralized level, so that all states can benefit.

1. Development of a national crash data schema and corresponding XML, based on MMUCC, which would provide the same benefits that NEMSIS has provided to the EMS community. In particular, such a schema should be designed to incorporate MMUCC, additional state-specific variables, and to facilitate linkage to NEMSIS and NTDB schemas.
2. Development of clear methods and criteria for testing quality of linkage systems (probabilistic or otherwise). Levels of linkage quality (in terms of bias, accuracy, and completeness) should also be associated with guidance in how to analyze the data and how to improve linkage quality.
3. Development of a repository for lessons learned, methods used (including those tried and rejected), and contacts in states that can provide advice. This should include (but not be limited to): a) Lists of variables states use for probabilistic linkage (if appropriate) and linkage success; b) Software available and algorithms use for probabilistic linkage, along with the pros and cons of each; c) Non-probabilistic linkage approaches successes and failures; d) Background on the data warehouse model and how to build one over time; e) Lists of vendors used by states for different elements of the data linkage process; and f) Contact information for individuals involved in state data linkage projects to provide assistance or advice.
4. Development of marketing materials that TRCCs can use to advertise the benefits of linkage to all groups that need to be involved. Coordination of a message at the national level would be helpful to gain the involvement of agencies that are not as used to working together (e.g., state health agencies and state departments of transportation).
5. Development and hosting of workshops for state data holders to learn about linkage approaches and discuss challenges with other states.
6. Generate a clear, written interpretation of HIPAA in the context of data linkage that defines clearly what mechanisms must be put in place to link data and still maintain HIPAA compliance. While HIPAA does not prevent data linkage or even including linked (de-identified) data in a state data warehouse, it does put additional security requirements on datasets that include such information.
7. Investigate the potential for vehicle-to-vehicle (V2V) communication to aid in passing identifiers on the scene. This should include assessment of what an application would need to do, potential hurdles in implementation, and estimated short-term (software development) and long-term costs. This project could also investigate the general problem of using event-specific (but not person-specific) identifiers to improve probabilistic linkage among occupants within the event. Such work could be applied to other event-specific linkage approaches (such as passing crash report number to EMS and trauma databases).
8. Develop a more detailed sampling protocol that includes costs of sampling and estimates of sample size needed for a set of target analyses. A pilot sampling project should be included to ensure that logistical challenges and costs are fully identified.

2 Overview

The goal of the NCHRP project 17-57 is to develop a comprehensive roadmap that guides states along the path to measuring the number of serious injuries in crashes. This goal has been motivated by the Moving Ahead for Progress in the 21st Century Act (MAP-21), which requires a set of highway safety performance metrics that includes assessment of serious injuries in crashes.

This report is a comprehensive description of the findings and recommendations from the first three tasks of the NCHRP 17-57 project. Although MAP-21 requires measurement of serious injury, it does not define it. The first task in this project was to recommend a definition of serious injury. As will be explained in the first section, we recommend a medical-diagnosis-based definition of serious injury for use by states.

Choosing a medical-diagnosis-based definition, rather than a police report-based definition, represents a critical point in the entire process of developing state data systems to record serious injury in crashes. If this choice is accepted, as we strongly recommend, then it requires some means of working with medical outcome data. If such analysis is to go beyond simple counts and allow serious injury to be tied to crash, roadway, behavior, vehicle, and occupant characteristics, then linking medical outcome data to crash data is required. The remainder of the project was devoted to exploring options and developing a roadmap for states to achieve this goal (comprehensive linkage of datasets, including crash and medical outcome).

Within the project, there were several tasks designed to gather the necessary information to develop the roadmap. These tasks are listed in Table 1 as they were originally envisioned.

Table 1. Tasks included in NCHRP 17-57

Task	Task Name
1	Outcome identification and short-term solutions.
1a	Perform a literature review to select an injury severity metric based on hospital data.
1b	Analysis and modeling of injury data to develop a near-term method for improved injury severity coding.
2	Roadmap(s) to interim solutions for direct linking crash and hospital data.
2a	Identify best practices for development of crash-EMS-hospital linkage through case studies of states currently performing linking and states working toward linking.
2b	Review strategic highway safety plans and survey state data holders to identify existing procedures and potential barriers to crash/hospital data linkage.
2c	Develop draft roadmap(s) for linking crash, EMS, and hospital data. Solicit feedback from states on these roadmaps. Refine roadmaps based on feedback.
3	Refine roadmap to include methods for linking to other relevant state datasets. Solicit feedback from states on these roadmaps. Refine roadmaps based on feedback.

This report is organized in the following way. First, we start with a review of injury coding systems and injury severity metrics, leading to the recommendation of a specific medical-diagnosis-based metric for use by states. This review will include a comparison of crash-report-based and medical-diagnosis-based metrics using crash data. Second, we present the results of the state survey of data systems and linkage activities, which give context to the current condition of datasets and linkage programs that a roadmap would address. Third, we discuss near-term solutions to measuring serious injuries in crashes, focusing on sampling programs that could be implemented at the state level. Fourth, we present a roadmap to linkage. In this section, we discuss ways in which the comprehensive linkage goal of the project can be facilitated. We also discuss "choice points" where decisions made earlier may facilitate or fail to facilitate future linkages.

3 Measuring Serious Injury

3.1 Injury Classification Systems

To understand the processes used to identify serious injuries, it is necessary to first consider the available injury coding systems. The primary purpose of an injury coding system is to identify specific types of injuries. Ranking injury severity is not necessarily an explicit purpose of injury coding, but ultimately, a ranking system must be imposed in order to identify serious injuries as a distinct class.

The three major injury coding systems relevant to the traffic crash domain are the KABCO scale, the Abbreviated Injury Scale (AIS) (Gennarelli & Wodzin, 2005), and the International Classification of Disease, Clinical Modification (ICD-CM) system (WHO, 1992). Each of these typologies have different scale qualities, may be found in different databases, and have different advantages and disadvantages for use.

KABCO. The KABCO scale is used by police officers on the scene of a crash to judge the general injury severity level of each occupant. The scale was developed by the National Safety Council and is recommended in the MMUCC guidelines for crash data (DOT, 2012). In general, K is Killed, A, B, and C are injuries of decreasing severity, and O is property-damage only. One of the problems with KABCO is that different states use different definitions of the A, B, and C injury codes. The MMUCC 3rd edition (DOT, 2008) and the American National Standards Institute D16.1-2007 (ANSI, 2007) recommended A for Incapacitating Injury, B for Non-Incapacitating Injury, and C for Possible Injury. Most, but not all states have used these definitions, and the lack of universal consistency may create problems in the usage of this scale across different jurisdictions. In 2012, the 4th revision of MMUCC tried to standardize KABCO usage with new definitions: A for Suspected Serious Injury, B for Suspected Minor Injury, and C for Possible Injury.

KABCO does not characterize or "type" injuries. The primary advantage of KABCO is that it is available in the police report database of virtually every state. KABCO is strictly an injury ranking system and not an injury classification system. Each crash-involved party is given a single scale score for his/her apparent overall injury severity level. As such, there is very little that can be done to glean additional information from this scale regarding injury typology.

AIS. The AIS was developed by the Association for the Advancement of Automotive Medicine (Gennarelli & Wodzin, 2005) for the purpose of coding injury types and injury severity, based upon an in-hospital clinical assessment. The AIS system assigns a unique numeric code to each specific injury type and each code is associated with a severity score

ranging from 1 (minor) to 6 (life threatening). Coding is done by coders trained specially on the AIS lexicon using existing medical records.

With AIS, both injury coding and injury severity ranking are embodied in a single system. However, each individual anatomical injury is coded separately. Thus, to identify seriously injured crash victims, it is still necessary to combine a patient's injury severity scores into a single person-level metric. There are a number of systems for doing this that will be described later.

International Classification of Disease, Clinical Modification (ICD-CM). In hospital administrative databases, injuries (as well as diseases, and other aspects of medical typologies) are coded using the International Classification of Disease, Clinical Modification (ICD-CM) system (WHO, 1992). The ICD-CM is a general-purpose classification system for diagnoses of all health conditions and includes codes for both the nature of the injury and causes of injury. Coding is done by trained medical coders who work from hospital records. Medical coders must pass an exam to become a Certified Professional Coder or Registered Health Information Technician(RHIT) and must have background in anatomy and physiology as well as the coding system itself. ICD-CM is widely used in clinical and health research settings, and is commonly found in hospital and trauma databases. The U.S. is currently in transition between using ICD-9-CM and a newer revision, ICD-10-CM.

Unlike AIS, ICD-CM does not include an explicit ranking of injury severity. To be used to identify seriously injured crash victims, ICD-CM must either be mapped to AIS or some other ranking system must be imposed on the coded injuries to assess severity. Both of these approaches will be discussed below.

3.2 Severity Metrics

Each of the three aforementioned coding (or ranking) systems is often manipulated to provide an abridged injury severity metric that suggests the presence of serious vs. non-serious injury in a patient. Severity metrics are typically intended to reflect increasing threat to life, with higher severity scores associated with higher probability of mortality. Severity can also be associated with risk of long-term disability, though the difficulty of obtaining long-term follow-up data has limited studies of this association.

Table 2 summarizes the key injury severity scoring systems based on the three injury coding systems described above. The table includes calculation and cutpoints that have been found in the literature. Details are given in the text that follows.

Table 2 Summary of Injury Severity Metrics and Characteristics

Injury Coding System	Injury Severity Metric	Calculation	Common Cutpoint(s) for Serious Injuries
KABCO	KABCO	All occupants with K or A-injury severity rating on police accident report	KA
AIS	MAIS	Highest AIS severity score of all injuries	3+
	ISS	Sum of squares of highest injury severity in each of three different, most-injured body regions	16+ (9+ is also used)
	NISS	Sum of squares of three highest injury severities regardless of body region	16+ (9+)
ICD	ICISS	Product of Survival Risk Ratios (SRRs) of each individual injury $\text{SRRs} = \frac{num\ patients\ with\ injury\ code\ who\ survive}{num\ patients\ with\ injury\ code}$	<0.90 has been used, but no standard established
	mSRR	Worst (minimum) Survival Risk Ratio among injury diagnosis codes	No standard found
	TMPM	Regression model designed to predict mortality outcome	No standard found
Other	LoS	Length-of-Stay in the hospital, measured in days	4 days has been used; no established standard
	Sentinel Diagnosis	Presence of any of an agreed-upon list of diagnoses	No standard

KABCO-Based. Each of the three aforementioned coding (or ranking) systems are often manipulated to provide an abridged injury severity metrics that suggests the presence of serious vs. non-serious injury. Researchers and practitioners using only police-reported information and KABCO typically use K plus A (KA) to characterize transportation crash injuries resulting IRTAD, 2011]. K, A, B and C (KABC) are also sometimes used to identify all injured occupants [e.g., NHTSA, 2010].

AIS-Based. Because AIS is designed to characterize injury types and has an embedded severity scale for each specific injury, AIS has been used to develop several injury severity metrics that quantify the multiple injuries that may be experienced by the occupant of a transportation crash. The most common AIS-based metric is a single maximum AIS (MAIS) across all body regions that are coded for injuries. MAIS is often used as a measure of overall injury level for an occupant, and MAIS of 3 or greater (MAIS 3+) is commonly used as the cutoff for defining serious injury (e.g., IRTAD, 2011).

Since the maximum AIS severity score does not distinguish between patients with several serious injuries to different body regions and those with more localized injury, Baker et al. (1974) developed the Injury Severity Score (ISS). The ISS is the sum of the squares of the most

severe AIS scores in each of three different body regions. The highest possible ISS is 75. An ISS cutoff of 16+ has been used to define seriously injured occupants (e.g., AACN expert panel). The New Injury Severity Score (NISS) is similar to ISS, but is computed as the sum of squares of the three most severe injuries, regardless of body region (Osler, Baker & Long, 1997).

ICD-CM-Based. Because of its widespread use in hospital settings, there is a great deal of interest in assessing injury severity using ICD-CM. However, the ICD-CM lexicon only characterizes injury and does not assess severity. One approach used to transform injury descriptions to measures of severity is to map ICD-CM to AIS codes and use AIS-based severity scoring. A private software product called ICDMAP™ translates ICD-9-CM codes to AIS 90 (1998 revision).

A more up-to-date application is the ICDPIC mapping procedure developed for use with STATA©, a common statistical package. Injury severity estimates, based upon ICDPIC manipulations of ICD-9-CM, correlate well with hand-calculated AIS provide by trained coders, but these estimates tend to produce a slight but systematic underestimate of true injury severity (Fleischman, Mann, Wang et al., 2012).

Haas et al. (2012) published an evaluation of their mapping from ICD-10 to AIS 1998. The translation produced 57% agreement in overall MAIS and 83% agreement in identifying patients with MAIS 3+ injuries. Agreement in ISS was also evaluated, with promising results (87% of cases resulted in a difference of $\leq$10 points). Although the mapping is to an older version of AIS, the authors discuss the potential to map to the AIS 2005 revision.

Zonfrillo et al. (2015) reported on a mapping between ICD-10-CM (as well as ICD-9-CM) and general AIS categories of AIS 3+, AIS<3, or indeterminable. While less nuanced than a mapping to the full AIS scale, this approach allows mapping to the key cutoff and the definition of serious injury recommended in this report. They report that 27% of ICD-10-CM codes and 17% of ICD-9-CM were rated indeterminable.

An alternative to mapping from ICD-CM to AIS is to use an ICD-CM based severity metric. Because ICD does not include an explicit ranking of injury, ICD-based scoring must introduce severity through another means. One approach relies on computed SRRs for certain classes of ICD-based diagnoses. The SRR is the proportion of patients with the diagnosis code who survive out of all the patients with that code. Thus, the SRR allows categorical injury codes to be ranked with respect to survival (i.e., severity).

The SRR approach presents some challenges for widespread adoption of ICD-based metrics for measuring serious in crashes. Notably, SRRs are database-specific and thus affected by the patient population and treatment protocols of the hospital(s) that contribute to the database (Cryer, 2006). Efforts are being made to standardize SRRs, based on the NTDB (Meredith et al, 2003).

The ICD-9-CM Injury Severity Score (ICISS; Osler, et al., 1996) is the product of all SRRs for a patient's injuries. The calculation is meant to represent the joint probability of survival given the particular group of injuries, and is the most widely used ICD-based injury severity metric.

One of the issues with ICISS is that SRRs calculated from all patients with the diagnosis ("traditional SRR") may not represent the independent probability of survival for each injury. Meredith et al. (2003) reported that ICISS using SRRs calculated only from patients with a single diagnosis ("independent SRR") is preferable to ICISS using traditional SRRs. Another issue with ICISS and mSRR is the lack of a well-established criterion cut point. An ICISS of <0.90 (>10%

chance of mortality) was used by Newgard et al. (2010), but most of the work on ICISS has focus on discrimination performance without consideration of specific cut point choice.

Other approaches use a patient's worst (minimum) SRR rather than the product of all SRRs (Meredith, Kilgo & Osler, 2003). This approach simplifies calculation based on SRRs, relative to ICISS, but has the same problem with a lack of agreed-upon criterion.

An alternative approach that is gaining momentum is the regression-based approach embodied in the Trauma Mortality Prediction Model (TMPM) developed by Glance et al. (2009) using ICD-9-CM codes. The TMPM is based on a regression model developed using mortality as the outcome measure and thus its performance at predicting mortality is, not surprisingly, better than other measures (Haider et al., 2012). An additional advantage is that the model does not depend on SRRs and therefore can be standardized for use in different hospitals.

Other Metrics. In their report on severity metrics, the International Road Traffic Accident Database (IRTAD) (2011) evaluated four candidate metrics, of which two fall outside the categories described above. One was Length-of-Stay in the hospital (LoS), and the other was the presence of a sentinel diagnosis.

Hospital Length-of-Stay is simply the number of days a patient is admitted to the hospital for treatment of crash-related injuries. It is a proxy for injury severity rather than a direct measure, but is used in a number of countries where anatomically-based measures are not available.

Sentinel diagnosis is simply the presence of one or more of a list of specific diagnoses that are selected because of their association with a high probability of hospital admission. These diagnoses can be identified using either AIS codes or ICD-CM codes. "Serious Injury" is then defined as the presence of any of the selected sentinel diagnosis codes and "No Serious Injury" is defined as the absence of all such codes.

There are other prominent injury severity ranking systems that rely on physiologic parameters to calculate (i.e., Revised Trauma Score [RTS] and Trauma and Injury Severity Score [TRISS]). However, these measure are more likely to characterize the physiological stability of a patient and may not correlate well with an anatomical assessment of injury severity.

In addition, there are many other injury severity metrics based on AIS and ICD codes. There is a large literature on measuring injury severity and tying such metrics to survival at hospital discharge. The measures we have selected are the most-studied and most-recognized of available metrics. New developments in measurement of injury severity may bring newer ones to the forefront, but at this time, it makes the most sense to focus on measures that have been studied and used the most. The major advantage of linkage between crash and medical outcome datasets is that if a new metric becomes state-of-the-art, the data will be available to re-compute performance metrics for crashes in the future relatively easily.

3.3 Evaluation Criteria

We selected three criteria on which to evaluate injury severity metrics for use in measuring road safety. Our first criterion is the ability of the metric to predict outcome, specifically survival. The second criterion is availability of data. Given the condition of datasets and data linkages in the U.S. at this time, it is important to discuss the issue of data availability and the impact it will have on states' ability to assess serious injuries using the recommended metric. Finally, the third criterion is ease of use.

Predicting Outcome. A good injury severity metric should be calibrated to an agreed-upon outcome. Survival to the time of hospital discharge is most commonly used for this

purpose. A more serious injury will result in a greater threat to life, and a good injury severity metric should reflect that.

MAIS 3+ has been shown to be a good predictor of in-hospital mortality (Meredith et al., 2002; Kilgo, Osler & Meredith, 2003). In contrast, because of the lack of data linked to in-hospital mortality, "A" injury (from KABCO) has not been included in investigations of this relationship. However, "A" injury has been shown in several studies to be only moderately associated with MAIS 3+ injury (e.g., Farmer, 2003; Compton, 2005). While this is not a direct test of the relationship between "A" and mortality, it does suggest that good performance is not likely.

A larger problem with "A" is the fact that judgments are made by police officers rather than medical professionals, and that "A" is a global measure of injury severity for a single person. Even if correspondence were perfect, "A" would make an effective, straightforward, and readily available measure of overall injury severity, but would leave no further options for analysis of specific injury types (e.g., head injury).

It should be noted that several researchers (e.g., Cryer, 2006) as well as readers of our interim reports have pointed out that attention should also be paid to threat-of-disability as an outcome. In particular, costs to states from injury-related disability can be quite high in the long run, even compared to the costs of treating serious injuries in general. However, in the context of recommending a definition of serious injury to address MAP-21 requirements, the most critical distinction is between medical-outcome-based metrics and police report-based metrics. Using threat-to-life to judge the quality of a serious injury metric will strongly favor a medical-outcome-based metric such as MAIS 3+. Since this requires linkage to medical data, other metrics that are better tied to adverse long-term outcome can be used as well.

Data Availability. Data availability is a significant issue in the U.S. at the state level. In state crash databases, only KABCO is readily available. State hospital discharge datasets, where available (not all states have a statewide trauma registry), include ICD-9-CM (and soon ICD-10-CM) codes, and sometimes AIS codes. ICD-9-CM codes can also be translated into AIS codes as described previously.

Although state trauma registries generally code whether a patient was in a motor vehicle crash (MVC) and include detail on the driver and occupants, the data are not linked to highway variables, crash configuration, and vehicle damage that would help assess details of the highway system in a state with respect to safety. As of this writing, FHWA is proposing to require only a total count of serious injuries in crashes and the ratio of total serious injuries to vehicle miles traveled at the state level (FHWA, 2014) for MAP-21 reporting. These values can be calculated using only a state hospital discharge or trauma database. However, to tie serious injury to implementation of countermeasures, linked information from the state crash database is required.

Current crash data limits analysis of the safety-related performance of roadway, vehicle, and behavior interventions to KABCO-based measures of serious injury based on police reports. However, data linkage between crash and hospital datasets can make AIS- and ICD-based options viable. Detailed discussion of linkage methods will be addressed later in this report. However, efforts to promote good linkage can dramatically change the relative merits of different severity metrics for use in the U.S.. Linkage from crash databases to EMS and hospital databases makes measures of injury severity readily available to analysts and makes AIS or ICD-based metrics viable. These injury codes, in turn, are tied to crash, vehicle, roadway, and environment characteristics that must be accounted for by state DOTs in evaluating the safety performance of their highway systems with respect to serious injuries.

Ease of Use. Ease of use is included here because it facilitates widespread use of a new metric that may be unfamiliar to many highway safety planners. Employees in state agencies who are responsible for compiling performance metrics are typically not trained in statistics. Metrics that require complex calculation are both difficult to compute correctly and difficult to understand. Although calculations can be automated, understanding is still critical to the process.

Ease of use favors MAIS and "A" over others. MAIS is a simple maximum severity of all coded injuries. Both of these are easy to calculate when the appropriate data are available. ICD-based metrics require more sophisticated analysis as well as further work to standardize them on a national level. Some efforts have been made to standardize the calculations (NCHS, 2004), but the general approach will always be more complex than other methods.

3.4 Conclusions

A variety of injury severity metrics, including AIS-based and non-AIS-based metrics were reviewed in Flannagan et al. (2012). In weighing the relative merits of different metrics, much of the crash community has embraced MAIS 3+ as the preferred measure of serious injury. The IRTAD Group recommended using MAIS 3+ as an international standard definition of serious injury in their 2011 Annual Report (IRTAD, 2011). In addition, the European Commission declared in March 2013 that MAIS 3+ should be the metric used in EC countries for this purpose (see European Commission High-Level Group, 2012).

It is clear that the ideal combination of available data and demonstrated predictive value are not currently available in most states. This would require linkage from crash to injury outcome. If serious-injury-based metrics in MAP-21 are to be calculated soon, there must be some attention given to near-term solutions. These might include adaptations of "A" injury (e.g., calibration using state trauma databases) or estimation based on sampling of EMS/hospital records for some crash cases.

Once high-quality linkage between crash and hospital datasets has been achieved, and mapping from ICD-10-CM to AIS 2005 is available (as through Zonfrillo et al., 2015), any ICD- or AIS-based metric becomes viable. At this point, usability combined with predictive value favors MAIS 3+. This metric consistently performs well at predicting survival, and it is easily calculated. Although ICISS may outperform MAIS 3+, its complexity makes it less appealing as the primary metric for measuring serious injuries in crashes. Because linked crash datasets will contain ICD codes, more sophisticated ICISS- and SRR-based analyses can always be conducted by statistical experts. However, until methods are developed and provided to states to automate the calculation of ICISS, MAIS 3+ will be the easiest and most reliable metric for serious injury.

Based on our review, we recommend adopting MAIS 3+ as the primary measure of serious injury for use in MAP-21 performance metrics. Although data linkage needed to use MAIS 3+ to evaluate safety countermeasures has not yet been established in most states, the time has come to move to a medical-outcome-based metric. Diagnosis made by medical professionals is more accurate and reliable than the general injury severity impression of police. Taking the step of defining serious injury in terms of medical diagnosis motivates data linkage, which in turn allows for a much richer and more precise understanding of the relationship between crash, vehicle, and occupant characteristics and injury outcome.

The remainder of this report addresses the many practical implications of requiring a medical-outcome-based measure of serious injury.

4 Data Linkage in States

Since using a medical-diagnosis-based definition of serious injury such as MAIS 3+ calls for linkage between crash and medical outcome data, it is important to first assess the condition of state databases and linkage programs. In collaboration with the NCHRP 20-24 (37K) project, conducted by Cambridge Systematics, we surveyed state TRCCs to learn about their datasets and linkage systems (Cambridge Systematics, 2013).

Fifty-three respondents from 40 states plus the District of Columbia and Puerto Rico responded to the survey including those that responded only after follow-up phone calls were made. States responding are shown in Figure 1.

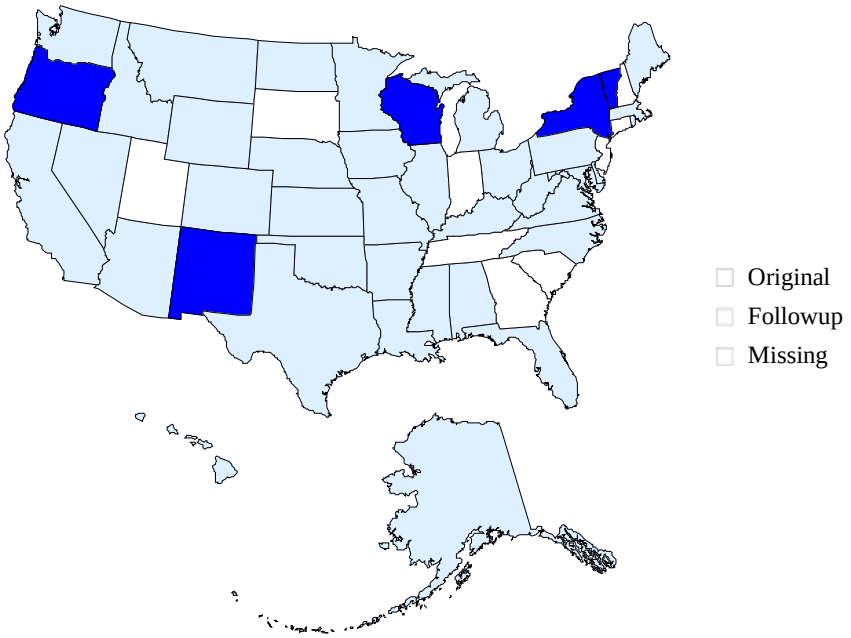

Figure 1. Map of U.S. showing states originally returning surveys (light blue) and states responding to follow-up phone calls (dark blue).

4.1 Definition of Serious Injury

Of the 42 states responding, all except Florida reported that they measure and report on serious injuries as part of their transportation safety improvement efforts. The majority specified four ways in which injury severity is used: research (86%), safety/program planning and management (93%), generating reports (98%), and evaluating and refining existing policy and regulation (83%).

States were also asked to provide the definition of serious injury that they use in reporting. Each state gave a unique answer, and all of the answers are compiled in Appendix A. In reviewing the responses, 33 of the 42 respondents gave definitions that either referenced KABCO “A” injuries or used language from the definition of “A” injury. The remaining states provided unique definitions. Some, such as Texas and Louisiana, use a more inclusive definition of injury in reporting (e.g., KAB).

4.2 Linkage Activities

For each database type, states were asked whether they were linking or planning to link from that database to crash data, and if so, was linkage probabilistic or deterministic, direct or indirect, and whether the linkage was related to a previous CODES program. The CODES program, for which federal funding ended in 2012, promoted probabilistic linkage between state crash and medical outcome databases in participating states. Direct linkage was defined as linkage directly with crash data, whereas indirect linkage was defined as linkage via a different database. Probabilistic and deterministic were defined for the respondents as follows:

> **Deterministic linkage** is based on the number of individual identifiers that can be matched among the combined data sets. When using a deterministic record linkage procedure, two records are considered to match if all or some identifiers above a certain statistical threshold are identical.
>
> **Probabilistic linkage** involves a wider range of potential identifiers and computing weights for each identifier based on its estimated ability to correctly identify a match or a non-match. The weights are used to calculate the probability that two given records refer to the same entity. Pairs with probabilities above a certain statistical threshold are considered matches, pairs with probabilities below another threshold are considered non-matches; and those in between the two thresholds are considered "possible matches".

In the section on data linkage, there were often conflicts between responses when a state sent more than one response. Thus, the percent of state *responses* is not strictly the percent of *states* since it was not possible to determine which respondent most authoritatively spoke for each state. In Tables 3-8, for the questions about details of linkage, we indicate the number of unique states for which at least one respondent indicated that linkage was being done or planned.

Table 3 shows the results of questions related to linkage between crash data and state Emergency Medical Services (EMS) data. About two-thirds of states are linking or planning to link crash-to-EMS databases. About half considered their linkages to be probabilistic and half deterministic. Almost two-thirds were considered direct. One-third of linkages were associated with CODES.

Table 3. Percent of States Engaging in or Planning Linkage Between Crash and State EMS Databases

Question	Response	Percent of States
Data Linkage	Yes	65 %
	No	30 %
	Unknown	5 %
Type of Linkage (among states that are linking) Unique states=24	Probabilistic	50 %
	Deterministic	50 %
	Direct	62 %
	Indirect	38 %
CODES-Related (among states that are linking) Unique states=24	Yes	33 %
	No	60 %
	Unknown	7 %

Table 4 shows the results of questions related to linkage between crash data and state Emergency Department (ED) data. Fewer than half of states are linking or planning to link crash to ED databases, though a substantial portion (14%) did not know. Of those who are linking, most considered their linkages to be probabilistic and about half were considered direct. One-quarter of linkages were associated with CODES.

Table 4. Percent of States Engaging in or Planning Linkage Between Crash and State ED Discharge Databases

Question	Response	Percent of States
Data Linkage	Yes	41 %
	No	46 %
	Unknown	14 %
Type of Linkage (among states that are linking) Unique states=17	Probabilistic	72 %
	Deterministic	28 %
	Direct	50 %
	Indirect	50 %
CODES-Related (among states that are linking) Unique states=17	Yes	26 %
	No	67 %
	Unknown	7 %

Table 5 shows the results of questions related to linkage between crash data and state hospital discharge data. Sixty percent of states are linking or planning to link crash to hospital discharge databases. Of those who are linking, most considered their linkages to be probabilistic and about half were considered direct. Over one-third of linkages were associated with CODES.

Table 5. Percent of States Engaging in or Planning Linkage Between Crash and State Hospital Discharge Databases

Question	Response	Percent of States
Data Linkage	Yes	60 %
	No	32 %
	Unknown	8 %
Type of Linkage (among states that are linking) Unique states=22	Probabilistic	76 %
	Deterministic	24 %
	Direct	48 %
	Indirect	52 %
CODES-Related (among states that are linking) Unique states=22	Yes	38 %
	No	50 %
	Unknown	12 %

Table 6 shows the results of questions related to linkage between crash data and state trauma data. Just over half of states are linking or planning to link crash to state trauma registries. Of those who are linking, 62% considered their linkages to be probabilistic and half were considered direct. About one-quarter of linkages were associated with CODES.

Table 6. Percent of States Engaging in or Planning Linkage Between Crash and State Trauma Registry Databases

Question	Response	Percent of States
Data Linkage	Yes	54 %
	No	43 %
	Unknown	3 %
Type of Linkage (among states that are linking) Unique states=21	Probabilistic	62 %
	Deterministic	38 %
	Direct	50 %
	Indirect	50 %
CODES-Related (among states that are linking) Unique states=21	Yes	28 %
	No	62 %
	Unknown	10 %

Table 7 shows the results of questions related to linkage between crash data and vital records data. About half of states are linking or planning to link crashes to vital records databases, though 10% of respondents did not know the answer to this question. Of those who are linking, half considered their linkages to be probabilistic and just over half were considered direct. About one-quarter of linkages were associated with CODES.

Table 7. Percent of States Engaging in or Planning Linkage Between Crash and Vital Records Databases

Question	Response	Percent of States
Data Linkage	Yes	46 %
	No	43 %
	Unknown	10 %
Type of Linkage (among states that are linking) Unique states=17	Probabilistic	50 %
	Deterministic	50 %
	Direct	53 %
	Indirect	47 %
CODES-Related (among states that are linking) Unique states=17	Yes	22 %
	No	70 %
	Unknown	8 %

Table 8 shows the results of questions related to linkage between crash data and roadway inventory data. Almost 90% of states are linking or planning to link crashes to roadway inventory databases. Of those who are linking, all are deterministic and three-quarters are considered direct. Only 12% of roadway inventory linkages were associated with CODES.

Table 8. Percent of States Engaging in or Planning Linkage Between Crash and Roadway Inventory Databases

Question	Response	Percent of States
Data Linkage	Yes	89 %
	No	8 %
	Unknown	3 %
Type of Linkage (among states that are linking) Unique states=33	Probabilistic	0 %
	Deterministic	100 %
	Direct	76 %
	Indirect	24 %
CODES-Related (among states that are linking) Unique states=33	Yes	12 %
	No	82 %
	Unknown	6 %

States were also asked about state laws related to linkage. Specifically, 75% of states indicated that state law did not set conditions for data linkage, while 11% said the law did set conditions (14% were unknown). The same 11% also indicated that state law required records linkage, while 75% of states did not. Regarding state law that established access to linked records for research, 14% of states had such a law, 64% did not, 17% were unknown, and 3% indicated that the law may allow access with approval.

Finally, respondents were asked to provide a list of identifiers being used for all databases that are being linked to crash. The complete list of identifiers is given in Appendix B.

There is enormous variety in identifiers used across states, even for the same type of database. States using probabilistic linkage to medical outcome databases tend to use date of birth, age, gender, date of crash and date of admission. Several states use name for these linkages and a few use some type of patient identifier or other numeric identifier (e.g., last four digits of social security number).

4.3 Database Coverage

A key issue in implementing statewide data linkage is the coverage of the state databases. States were asked to indicate the percent of all included cases that are captured in each of the various databases in the survey.

Table 9 shows the percent of states with complete, partial or unknown/no coverage for each of seven datasets. In the "complete" coverage category in Table 9, we included all responses of "complete" and any "partial" where the respondent reported 90% or greater coverage. Most states have complete crash datasets. However, closer to half have complete coverage in statewide medical databases, including EMS, hospital, emergency, and trauma databases. Roadway inventory is widely available, but many states do not have complete coverage.

Table 9. Percent of Respondents with Complete and Partial Coverage of State Databases (n=42)

State Database	Complete (90-100%) Coverage	Partial (<90%) Coverage	Unknown
Crash	81% (34)	12%	7%
EMS	45% (19)	17%	32%
Emergency Department	38% (16)	2%	47%
Hospital Discharge	50% (21)	2%	41%
Trauma Registry	40% (17)	7%	47%
Roadway Inventory	48% (20)	24%	25%
Vital Records	57% (24)	2%	38%

In looking at results for medical-outcome databases coverage from our survey, we were concerned at the high percent of unknown responses. The most recent national assessment of state trauma registry attributes was published in 2006 (Mann et al.). Based upon this published information, 32 states maintain some form of a centralized trauma registry. The majority of state data collection efforts *require* hospitals to report data (27 states [84%; 95% CI: 71.8%, 96.9%]). However, variability exists in the *type* of hospital required to submit data to the centralized registry. Thirteen states require data submission from only designated/accredited trauma centers. Another 11 states collect injury data from all acute care facilities. States *requesting* submission of trauma data combine information from a subset of trauma centers with existing registries. Coverage of trauma registries available within individual states vary. It is, however, interesting to note that a significant proportion of the hospitalized trauma occurring in the U.S. is considered to be captured in a trauma registry at the hospital and/or state level. Based upon individual state estimates, Mann et al. (2006) estimated that approximately 66% of registry-eligible trauma occurring in the country is captured in a state, regional or hospital-specific trauma registry.

4.4 Challenges, Priority and Timing

The remaining survey questions cover general issues about the challenges and timing of data linkage. The first of these questions asked which of several challenges states face in implementing data linkage. Respondents could check more than one option, and the percent of respondents choosing each option is shown in Figure 2, ordered from most common to least common answer selected. Funding, confidentiality, data usage issues, and hardware/software issues were identified as a challenge for linkage by more than 50% of respondents.

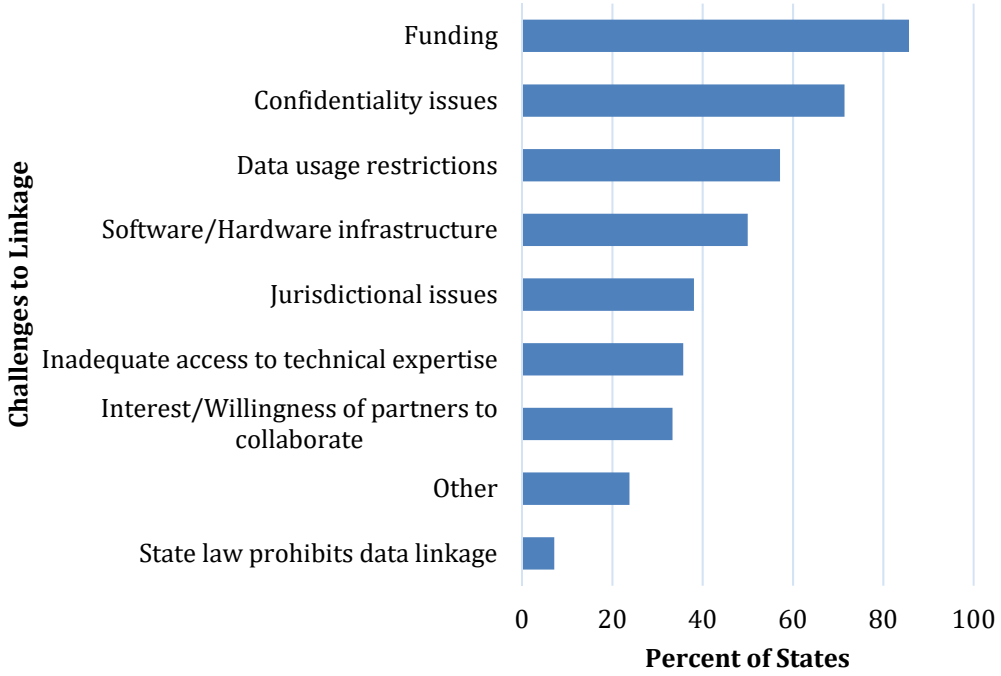

Figure 2. Percent of respondents who indicated challenges they faced in implementing data linkage in their state.

Figure 3 shows the percent of state respondents who indicate each item that would facilitate data linkage in their state. Not surprisingly, over 80% of respondents selected increased funding. Updated equipment/software, willingness of partners to collaborate and enabling legislation were all selected by at least 50% of respondents. These responses mirror the list of challenges selected by respondents and identify the key hurdles to data linkage for most states.

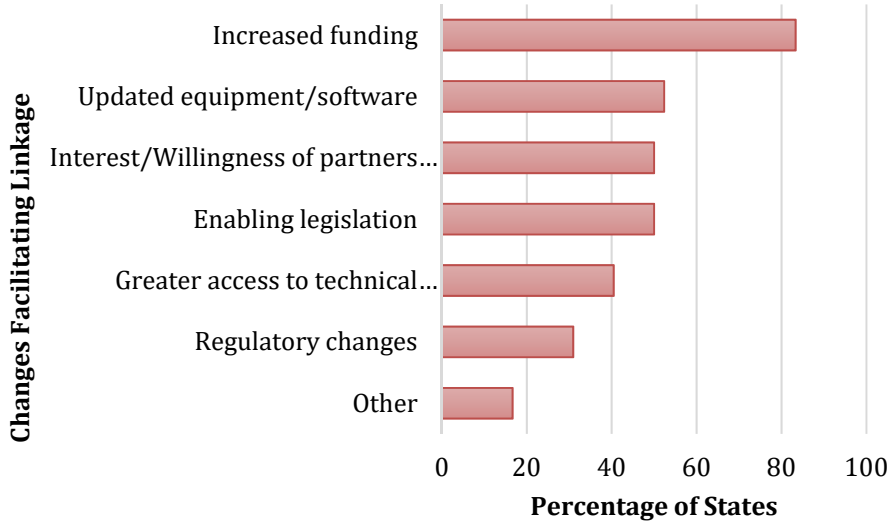

Figure 3. Percent of respondents who indicated changes that would facilitate implementing data linkage in their state.

The final two questions were about the importance and timing of linkage in each state. Of the respondents, 4 (11%) considered linkage to be mission critical, 26 (68%) considered it to be important, and 5 (13%) considered it to be somewhat important. On the issue of timing, ten respondents (27%) estimated that linkage would be implemented within two years, nine (24%) estimated 3-5 years, 11 (30%) estimated 6-10 years, and the remaining seven said linkage would never happen in their state.

In summary, linkage that is being done in states is most often between crash and roadway databases, and such linkage is deterministic. Among medical outcome databases, states using probabilistic linkage often had CODES programs, which were designed to promote probabilistic linkage between state crash and medical outcome databases. EMS is the medical outcome database being linked (or planned to be linked) to crash data in the most unique states. However, almost as many states are linking or planning linkage to hospital discharge and state trauma databases. Probabilistic linkage is common for these databases. The identifiers used for linkage are virtually unique to each state, though name, age, birthdate, gender, and crash time/location are seen in many lists.

One of the challenges of this survey was that questions covered a broad range of topics, and the original survey recipient generally had to find other people to answer specific questions. This was especially true for questions about EMS, hospital, and other medical outcome databases. As a result, the findings of this survey that pertained to these topics seemed to be inconsistent with a previous survey of state trauma databases (Mann et al., 2006). Since the latter survey was likely more accurate on questions about these databases, we present key results of that study here.

For present purposes, the results of the Mann et al. (2006) survey plus our own survey suggest that the majority of states have sufficient coverage and completeness in their trauma registries to facilitate data linkage. Note that most of a decade has passed since the Mann et al. (2006) trauma registry survey, so the current state of trauma registries is expected to be substantially better. Similarly, a majority of states at this time should have state EMS datasets with sufficient completeness and coverage to facilitate data linkage to crash datasets.

5 Near-Term Solutions to Measuring Serious Injury

5.1 Overview of Near-Term Solutions

The time frame for linkage in most states is too slow for the timing of implementation of MAP-21. Indeed, at the time of this writing, the FHWA has released a Notice of Proposed Rulemaking (NPRM) that proposes "A" injury from KABCO as the definition of serious injury for use through 2020. However, the NPRM also recommends putting linkage systems in place by 2020. While linkage efforts are moving forward in states—including the quarter of states that reported a 2-year time frame—meeting the goal of widespread linkage by 2020 will require significant attention.

In Section 3.4, we recommended using MAIS 3+ as the definition of serious injury. However, MAIS 3+ requires medical outcome data from either a state hospital discharge or trauma registry database. In addition, if medical outcome is to be tied to any crash or roadway characteristics, the crash/roadway and medical data must be linked. About half of states surveyed have this linkage process in place or are planning to link. Indeed, the results of the state survey show that while data linkage efforts are important or mission critical in 78% of states surveyed, linkage is still a work-in-progress or in the future for most states.

The challenge is that a definition of serious injury must be chosen and implemented in a short (one year) time frame, yet the appropriate linked datasets will not be in place in time. In contrast, the KABCO scale is captured in the majority of state crash databases, which themselves have complete or high levels of coverage in almost all states.

The temptation, then, is to choose A-injury as the definition of serious injury for MAP-21 purposes to reduce required burden on states and ease compliance with the new reporting requirements. In fact, the NPRM proposes to use "A" injury from the MMUCC 4th Edition (USDOT, 2012) for the next six years (through 2020).

However, choosing A-injury as the definition of serious injury for state reporting is likely to introduce bias in results and may reduce the motivation to implement data linkage. Flannagan & Rupp (2013), Farmer (2003) and Tarko et al. (2010) found that A-injury is biased both with respect to the conditions being considered and the usage of the scale across jurisdictions. For example, Farmer (2003) compared the percentage of A-injuries that were actually MAIS<3 and found differences as a function of geographical region, time of day, manner of collision, driver gender and driver age. Geographic region differences indicate a variability in use of the scale in general, while the other effects indicate differences in the ability or tendency of officers to identify serious injuries based on conditions of the crash or the occupant. Tarko et al. (2010) found similar differences based on vehicle type (car vs. motorcycle) and restraint use. Flannagan & Rupp (2013) detail differences between "A" injury and MAIS 3+ definitions. In addition to substantial overestimation of incidence of serious injuries, "A" injuries tend to over-represent common roadway, crash, vehicle, and occupant categories, such as belted occupants and rear-end crashes. This could lead to suboptimal allocation of countermeasures to prevent serious injuries.

A solution to the short-term need to measure serious injuries and the long-term need to promote linkage between crash and medical outcome data to improve that measurement might be to select MAIS 3+ as the target definition of serious injury and then consider ways in which existing data can be used to *estimate* the number of serious injuries (based on the MAIS 3+ definition). In this way, the target definition of serious injury for reporting purposes will still be MAIS 3+, but the measurement process can make use of data available now, possibly including A-injury rating from crash reports.

The next three sections explore the possibility of using available data in the near term to estimate serious injury as defined by MAIS 3+. We focus this section on solutions whereby states can estimate serious injuries using a medical-outcome-based definition, rather than try to "fix" a non-diagnosis-based definition.

We recommend two potential near-term approaches to estimating serious injuries. First, existing state-level trauma registry databases can be used to count or estimate the total number of people seriously injured in crashes in a state. This approach does not tie injury outcome to crash characteristics, but it can be used to calibrate other counts to an appropriate total. The other approach is to use sampling of a subset of medical records from crash-involved occupants in a state.

Flannagan & Rupp (2013) also explored development and application of a regression equation that corrects for biases in the distribution of A-injuries relative to MAIS 3+. We do not recommend this as a way to estimate serious injuries at the state level because it continues to rely on police-reported information. However, this approach might improve some models that are forced to rely on "A" injury from older databases.

5.2 Using State Trauma Databases

Many states maintain statewide trauma databases that capture MVC-related injuries as well as trauma from other mechanisms. Crash-related trauma can be isolated from other mechanisms, and trauma databases typically include injuries defined by AIS codes directly coded by trained coders using the medical records. The databases often include alcohol involvement, basic vehicle/occupant type (car, motorcycle, pedicyclist) and restraint use, along with patient age and gender.

The available variables in these datasets allow state-level counting of MAIS 3+ serious injuries in crashes as well as a breakdown by alcohol, and sometimes vehicle type and restraint use. However, a challenge to using state trauma databases is that they vary in the extent of coverage within the state. In some cases, only Level 1 and 2 trauma centers are included. In others, not all hospitals that qualify for data capture have systems in place for passing data to the state registry. The not-yet-participating hospitals may or may not be a biased sample (e.g., smaller hospitals in rural areas). Before using state trauma databases for counting crash-related injury totals, statewide coverage and any bias that results from incomplete coverage needs to be assessed. In general, it should be possible to correct for low levels of incomplete coverage without difficulty, but each state must investigate this issue to effectively use state trauma databases to make states comparable.

To the extent that crash descriptors may be addressed within the state trauma database (e.g., alcohol involved), those counts of serious injuries within category will be accurate. However, for data elements only available in crash databases (e.g., road type), state trauma databases can only help make totals across states comparable. Thus, "A" injury would have to remain the definition of injury for planning with respect to specific locations, roadway characteristics, behaviors, and vehicle characteristics.

5.3 Sampling Solution

Sampling of some form of medical outcome information for persons in crashes offers a near-term solution that both avoids the bias and calibration issues associated with KABCO and provides an opportunity to correct for them. More importantly, it represents a potentially cost-effective (though not cost-free) approach to improving measurement of serious injury and tying serious injury to crash, vehicle, behavior, occupant, and roadway characteristics in the near term.

There are multiple approaches to sampling in statistics, each of which has advantages and disadvantages. This section contains a brief background on sampling followed by a recommended approach to sampling medical records associated with crash data. The discussion below focuses on selecting cases from a state crash database for follow-up to obtain medical outcome data from hospital treatment associated with that crash. The approach described would sample only those occupants who are listed on the police report as having been transported by ambulance. Sampling only transported occupants will cover almost all seriously injured (AIS 3+) cases, but may result in missing a larger number of less seriously injured occupants (for more comprehensive analyses of the cost of crashes). Having information about the ambulance service and possibly the destination hospital will make the search for patient records simpler.

Sample vs. Census

A census is a dataset that contains all of the cases in a given population. The Fatality Analysis Reporting System (FARS) is a census of all crashes on public roads in which someone died within 30 days as a result of crash-related injuries (NHTSA, 2013). State crash databases are intended to be censuses of police-reported crashes in a given state.

A sample is a selected subset of a population on which data are collected. “Sampling” means obtaining a probability sample, defined by a) all elements in the population having a non-zero chance of being selected, and b) the selection mechanism being randomized. National Automotive Sampling System datasets (General Estimates System (GES) and Crashworthiness Data System (CDS)) are examples of probability samples of certain crashes. Cochran (1977) describes the advantages of sampling as being: reduced cost, greater speed, greater scope, and greater accuracy. In general, the arguments in favor of sampling over collecting a census revolve around limited resources for gathering information.

States that are able to implement direct linkage between their state crash dataset and state trauma or hospital registry will have a census of police-reported crashes that includes serious injury as a data element. However, for the majority of states without linkage in place, a sampling approach can allow estimation of serious injury incidence under a variety of conditions of interest.

Simple Random Sampling

Simple random sampling represents the most basic approach and is often used as a reference to compare to other sampling approaches. In this context, drawing a simple random sample would involve selecting at random *n* occupants in crashes who were transported by ambulance.

One advantage of simple random sampling is that analysis is fairly simple and estimates of serious injury incidence can be easily generated. The primary disadvantage is that the design can be inefficient (defined as the sample size, *n*, required to obtain estimates with a given precision [confidence interval size]), as well as in terms of cost (due to the large number of hospitals or trauma centers that must be contacted).

Stratified Sample Design

A stratified sample design is one in which a set of mutually exclusive and exhaustive categories are identified and cases are sampled at random from these strata with some known, but possibly unequal, probability. In this case, strata would be based on elements of crashes or occupants such as KABCO injury severity, alcohol involved/not involved, and/or restrained/unrestrained. Kish (1965) suggests that, for outcomes whose variance or cost of data

collection differs dramatically, a stratified sample design can improve efficiencies substantially over a simple random sample.

The best near-term approach to correcting both bias and over-counting is sampling of medical records from crash-involved occupants in a state. Although there is some cost involved, sampling has a number of advantages over other solutions. First, sampling addresses state and local uniqueness in the way KABCO is used by allowing direct assessment of the A-to-MAIS 3+ relationship in the state rather than assuming that models developed from national data can be applied locally. Second, sampling helps build systems, relationships and capabilities that can be leveraged for large-scale direct linkage in the future. The sampling solution does not require that state-level data systems be in place already. Third, sampling allows states that are linking and states that are not to measure serious injury using the same definition. This means that the transition to linked data in states can proceed at different paces without introducing non-comparability across states. Fourth, sampling allows states who are developing linkage systems to evaluate those systems for bias. Finally, sampling is scalable. A larger sample gives greater confidence in the estimates of serious injury, but even smaller samples can be useful for better estimating serious injury incidence.

Details of the sampling approach have been published in Transportation Research Record (Flannagan et al., 2014).

5.4 Regression Solution

This section describes a method that uses regression to adjust for biases in A-injury that were identified in the section on the relationship between KABCO and MAIS. As discussed earlier, we do not recommend this approach for widespread “fixing” of the use of A-injuries to measure serious injury. However, the use of regression-based adjustment might improve analysis of older datasets when better approaches (e.g., sampling) are not available.

The basic approach to the regression solution is to develop a regression equation that uses KABCO and other variables as predictors of the probability that an occupant is seriously injured (MAIS 3+). This estimated probability of injury can then be used instead of observed KABCO in counting serious injuries in crashes. For this analysis, we used the CDS database from 2007-2010 to develop the model and CDS from 2011 to test it.

Model-based mapping between KABCO and MAIS has been developed previously. In 2002, Blincoe et al. used a KABCO-to-MAIS “translator” in an analysis of economic cost of crashes based on the National Automotive Sampling System—GES data. Separate translators were developed for belted occupants, unbelted occupants, unknown belt status occupants, and non-occupants including motorcyclists. Each translator provided the probability of each MAIS level based on the KABCO level within the designated group. Thus, each translator consisted of a 5 (KABCO) by 6 (MAIS) table of probabilities.

Tarko et al. (2010) developed a regression equation based on linked crash and hospital data from the Indiana CODES program. Their model incorporated extra complexities to account for incomplete linkage in the dataset, but the key element for this discussion was an ordered logit model that used KABCO and a large number of other predictors to predict MAIS level.

Although the basic approach seems promising, the Tarko et al. (2010) model used dozens of predictors. Moreover, many predictors were included that did not interact with KABCO. For example, head-on crash was a predictor in the model that, when present, increased the probability of serious injury outcome. However, the inclusion of head-on crash in this way simply accounts for the greater likelihood that someone will be injured in a head-on crash. It does not adjust for any bias in the use of KABCO for head-on crashes vs. other crash types. That is, if KABCO were a

perfect match to MAIS, head-on collisions would still cause more injuries and the head-on variable would still be significant in the ordered logit model. The inclusion of factors like head-on in the Tarko et al. (2010) model serve to permanently encode the relationship between head-on crashes and injury risk for future use of the model. If injury risk in head-on crashes were to decrease over time (e.g., with improvements in occupant protection or collision mitigation), this change would not be reflected in analyses of future data that use the Tarko et al. approach.

Instead, a model designed to adjust for bias in the use of KABCO should only include KABCO and interactions between KABCO and other predictors (like head-on crash). This is, in effect, what Blincoe et al. (2002) did by separating their translators. Each translator is a different relationship between KABCO and MAIS, but none of the translators indicates the overall probability of injury due to belt non-use vs. belt use.

Following this idea, we developed a logistic regression model based on our CDS dataset. The development dataset included data from 2007-2010 and the test dataset included data from 2011. We limited the outcome to MAIS 3+ vs. MAIS 0-2 because we were primarily interested in counting serious injuries. As with previous analyses, we removed fatalities. Our goal in this analysis is not to fully develop a final regression model for use in all states, but to explore the potential value of the regression approach to resolving problems of bias in use of KABCO.

All predictors other than KABCO were implemented as interactions with KABCO. The following predictors were entered (as interactions with KABCO) in addition to KABCO itself in the original model: sex, age, alcohol involvement, restraint use, damage direction, vehicle type, number of vehicles involved, and crash configuration. All interactions with KABCO were significant except number of vehicles involved, which was removed from the final model.

When KABCO is used alone as a predictor of MAIS 3+ injury, the area under the ROC curve (AUC) is 0.88. For the full model, AUC increases to 0.91. The improvement in AUC is fairly small, but significant. However, AUC as a measure of performance is insensitive to differences in total estimated serious injuries and somewhat insensitive to patterns of bias in decision rules. Thus, the real benefit of the regression approach is better seen in comparison of distributions of crash and occupant characteristics for regression approach compared to A-injury alone.

To use the model, we applied the prediction equation to each occupant in the test (2011) dataset. Even occupants with “O” injury severity will have some non-zero predicted probability of having a serious injury. The count of serious injuries will then be the total predicted probability across the condition being evaluated. For example, if we want to estimate serious injuries by crash configuration, we would sum the predicted probability of serious injury for all single-vehicle crashes, then all angle crashes, and so on. Each sum is the estimated total number of serious injuries for that configuration. The regression equation will adjust for both bias and over-counting.

To look at model performance, we built the model on four years of data (2007-2010) and tested it on the most recent year available (2011). Using the test data with non-missing values of predictors, we calculated the predicted total number of serious injuries for each condition of interest. We also calculated the totals for A-injury and MAIS 3+ injuries in the same set of cases. The results for crash configuration are shown in Table 10 and Figure 4.

Table 10 shows that the regression approach partially calibrates the total number of serious injuries, compared to using A-injury alone. In addition, the regression approach partially corrects for bias in A-injury with respect to crash configuration. By definition, the relative distribution of configurations using the regression-based estimate falls somewhere between using exclusively A-injury and using MAIS 3+ (the ideal solution). The regression approach corrects for overestimation

of angle crashes and underestimation of single-vehicle crashes, but in this comparison has over-corrected for head-on risk.

Table 10. Total Serious Injuries by Crash Configuration Based on Three Definitions of Serious Injury (2011 Validation Dataset)

Crash Configuration	Total A-Injuries	Total MAIS 3+ Injuries	Estimated Total MAIS 3+ from Regression Model
Angle	18921	5342	7610
Head-On	2542	681	2444
Rear End	4968	1654	1900
Sideswipe/Opposite Direction	945	775	768
Sideswipe/Same Direction	889	185	373
Single	12968	9115	10577
Grand Total	41233	17751	23672

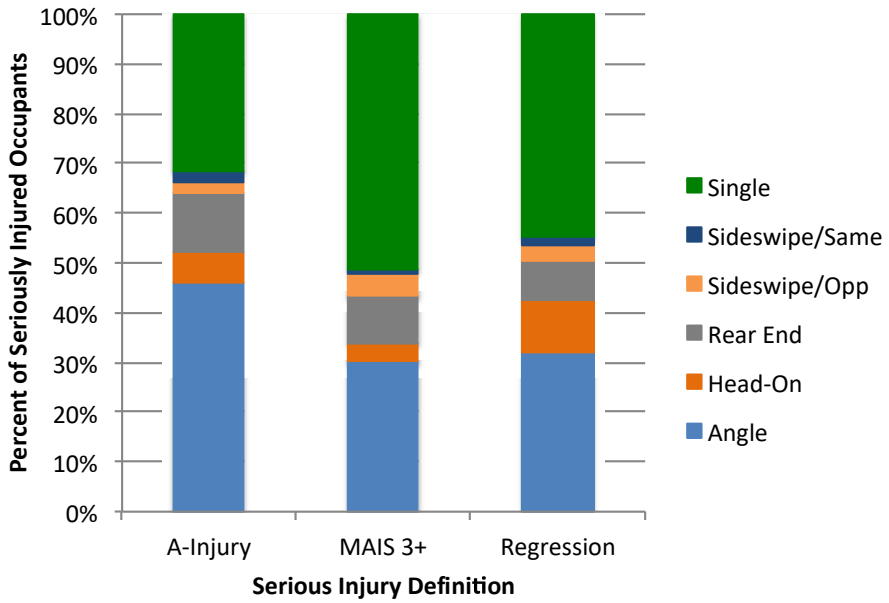

Figure 4. Comparison of relative proportions of different crash configurations for three definitions of serious injury using the 2011 test dataset. Regression approach is described in the text.

Figure 4 illustrates one of the weaknesses of the regression approach. Here, the test dataset from 2011 has an unusually small percentage of head-on collisions and an unusual pattern of relationship between KABCO and MAIS for head-on collisions. The regression approach should adjust for the overall change in percentage, but is insensitive to differences in the *predictive relationship* between KABCO and MAIS. Thus, in the test dataset, the percentage of MAIS 3+ injuries in head-on collisions is overestimated.

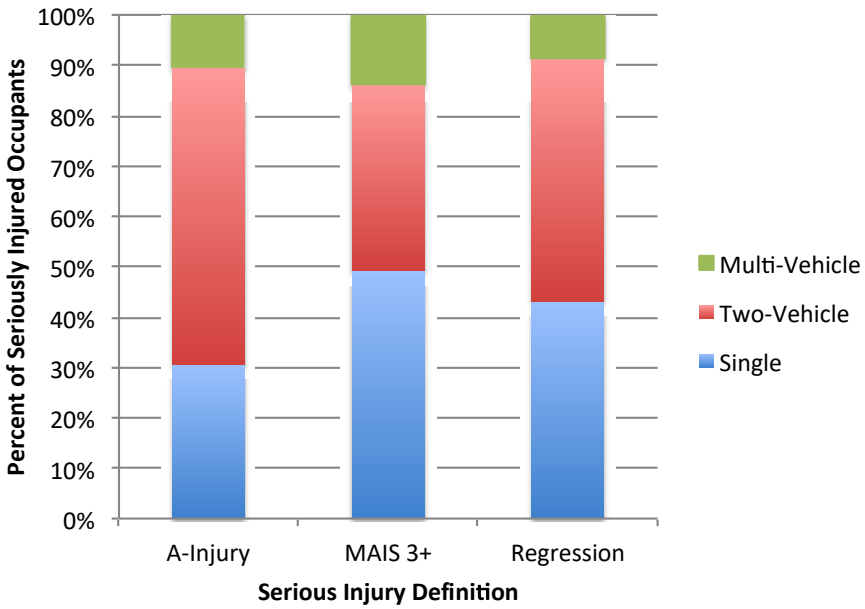

Figure 5. Comparison of relative proportions of number of vehicles involved for three definitions of serious injury (2011 validation dataset). Regression approach is described in the text.

Figure 5 shows the breakdown for number of vehicles involved in a crash. This figure illustrates how bias can be improved even for predictors that are not in the model. Number of vehicles was not a significant predictor, but single-vehicle crashes are included in crash configuration. As a result, bias in use of A-injury that over-counts two-vehicle collisions is influenced by crash configuration and other predictors related to number of vehicle that are included in the regression equation. As with the age analysis, the regression-based distribution falls between that of A-injury and MAIS 3+. The regression helps remove some, but not all bias in the test dataset.

The graphs and tables shown illustrate the potential performance of the regression approach. If the model is carefully developed, estimates can be corrected for factors that both are and are not included in the model itself.

The potential problem with the regression solution is that the data on which it is developed may or may not reflect the patterns in the data on which it is used. Specifically, a development dataset must have both MAIS and KABCO for the same occupants. Since by definition the regression solution is being offered for states that do not have crash data linked to hospital outcome, the development dataset would most likely be CDS. If the patterns of bias relative to predictors seen in CDS hold up across regions, then the regression model developed on CDS data should be appropriate for individual states to use on their data. However, the results for crash configuration shown in Figure 4 illustrate what can happen when the patterns in the state data are not consistent with those in the development dataset.

Additional work is needed to better understand how much patterns vary across states, how robust the regression solution is in the context of the variation seen across states, and how the regression approach itself might be calibrated (e.g., using a Green & Blower (2010) type of approach). Our analysis indicates that the regression approach can help, but it is clearly a limited solution that should only be used in cases where better approaches (e.g. sampling) are not possible.

6 Roadmap to Comprehensive Measurement of Serious Injuries Through Linkage

Although we present two near-term options for measurement of serious injury using a medical-diagnosis-based metric, the most comprehensive, long-term solution is to link crash data to medical outcome data at the state level. In addition, in setting up data linkage systems for this purpose, it is useful to consider linkage of state data systems broadly. Additional linkages will aid in answering additional questions. Thus, this section describes a roadmap to *comprehensive* data linkage at the state level.

We define "comprehensive" as both adding datasets and as including larger swaths of the crash-involved population. At many points in the process, decisions might be made that will facilitate future linkages and increase comprehensiveness. However, some of these paths may result in additional expense, and states will need to decide whether the goal of comprehensiveness is cost-effective at that point. These decision points will be discussed where they appear. It is worth noting that the center of this work is the crash. Thus, we will not consider datasets or linkages that do not involve crash. That said, the principles of linkage and computing infrastructure discussed here would apply for any set of linkages among state datasets (e.g., roadway inventory linked to road repair datasets, etc.).

This report is divided into three major sections. The first presents a set of benefits of linkage beyond just meeting the requirements of Map-21. The second identifies the basic requirements for a linked datasets at the state level. The third describes a series of steps for states to develop such a system. The set of steps, or roadmap, includes alternatives where possible, but in many cases, basic systems (such as complete statewide databases) have to be put in place before linkage can be successful.

6.1 Why Link?

This report began by recommending a medical-outcome-based metric for serious injury. From that, data linkage is logically required. However, it is worth considering the broader set of benefits of data linkage beyond simply being able to count serious injuries for MAP-21 reporting purposes. The following provides a sample list of activities that are enabled by data linkage:

- Calculation of comprehensive costs of crashes
- Cost-benefit analyses around interventions (e.g., infrastructure improvements) that take injury costs into account
- Cost-benefit analyses that take actual state costs into account (fatalities often cost states less than serious injuries in real dollars, even though cost-benefit analyses generally use a large number for each life lost)
- Optimized resource management for EMS and trauma around crashes
- Development of better triage models
- Prediction of high-cost crash locations for better interventions
- Improved data sharing between agencies that need the same datasets
- Identification of crash under-reporting through comparison between reported crashes and trauma/EMS
- Identification of injury hotspots and mechanisms
- Availability/Existence of linked data resources that might aid other entities (e.g., car companies) in understanding their fleet performance; this may become more important as automated vehicles enter the fleet

- Development of hidden-injury prediction models that could aid EMS and hospital treatment of crash-involved patients
- Better validation of EMS practices through the ability to link to diagnosis

general, linked data systems enable richer analyses, and they can make data collection and access more efficient. By collecting each piece of information only once and sharing across data systems, information needed in several datasets can be obtained more efficiently.

The greatest barrier to a fully linked data system is the up-front cost to set up the data collection, linked storage, and access infrastructure. There is a need for a rationale for linkage that goes beyond "MAP-21 says so" and demonstrates real benefit to the state to make the investment.

6.2 Requirements for Linkage

A good performance management framework that incorporates serious injury metrics requires data that has a set of key qualities (Cambridge Systematics, 2013). These include consistency, comparability, and comprehensiveness. Consistency is internal agreement among data elements and databases with overlapping data. Comparability means that performance metrics provided by different states mean the same thing and can be compared directly. Finally, comprehensiveness means that data systems can be used to answer the widest possible range of questions and include the widest possible range of cases. These overarching principles will be considered in the details discussed in Section 7.

To achieve the ideal data qualities described above, a system of data linkage between crash, EMS, hospital, and other datasets at the state level requires certain basic elements. First, there must be statewide databases with good coverage in each area. These datasets must conform to a common schema whenever there are data elements in common. There must be one or more identifiers that specify rows in each dataset that refer to the same individual, and there must be rules for access and a mechanism for secure access. These requirements are discussed in more detail below.

6.3 Dataset Quality

To measure serious injuries in crashes in a state using data linkage, the datasets themselves must be in good enough condition to support the linkage process. Problems in datasets will tend to compound—missing data in any dataset will break a link, so even if each dataset is of reasonable quality, the linked dataset may still be marginal. Thus, attention to the quality of each of the key state datasets is important. Several aspects of dataset quality must be considered.

6.3.1 Coverage

Coverage indicates the percentage of services that are reliably sending data to the state. Within a state, data come from a set of smaller organizations that are providing the on-the-ground service being measured. For example, EMS data are typically collected from a large number of small, independent ambulance and rescue services spread around a state. In many cases, these services are staffed by volunteers in rural areas, and data collection may not be the highest priority for limited time and money. An ideal database will collect from all of the relevant services in the state.

That said, coverage need not be 100% to support good estimates of serious injury incidence. What is more problematic is biased coverage, where certain kinds of services or areas (e.g., urban areas) are more likely to be included in the state database than others. Because

geography affects crash type, it also affects injury risk, so geographical bias in coverage of one database (e.g., EMS) will introduce bias into the linked dataset as well.

6.3.2 Schema Consistency

A database schema is essentially a codebook of variables collected and possible values. It is important to have consistent use of a common schema across all agencies or services providing data. Inconsistencies can occur in different ways. For example, certain data elements may be included in data from some units and not others. Alternatively, different units may use code values differently.

The first issue might arise in hospital data if the hospitals themselves have individual data collection practices and then pull data to satisfy state reporting requirements. One common example of this revolves around the use of the AIS injury coding system.

In the U.S., hospitals are required to use the International Classification of Disease version 9, Clinical Modification (ICD-9-CM) for coding medical records (WHO, 1992). Starting in October 2015, they will be required to use version 10 (ICD-10-CM). The ICD-CM is a general-purpose classification system for diagnoses of all health conditions and includes codes for both the nature of the injury and causes of injury. Coding is done by trained medical coders who work from hospital records. Unlike AIS, ICD-CM does not include an explicit ranking of injury severity. To be used to identify seriously injured crash victims, ICD-CM must either be mapped to AIS or some other ranking system must be imposed on the coded injuries to assess severity.

Only some hospitals (typically larger trauma centers) employ trained AIS coders to recode records into the AIS system. This means that within a state hospital database, there may be some hospitals that consistently provide AIS codes and others that provide only ICD-CM codes. ICD codes can be translated into AIS codes, but the computer-based translation will be different from a human-coder-based translation. Thus, it is important to maximize the consistency and completeness of the incoming data from the set of independent sources across the state and to document inconsistencies in the way data are handled by different sources.

6.3.3 Quality Control

QC is a process of checking consistency and completeness at the individual data-element level. One basic QC issue is missing data. When data are missing in small numbers (e.g., <5% for any one data element), inferences can still be made reliably based on the remaining data, especially if missingness is arguably random. However, as the missingness rate increases, the reliability of inferences based on the remainder decreases. Moreover, the likelihood that data are missing at random also decreases.

QC should also check for consistency among related data elements. For example, if a crash is labeled as single-vehicle, there should only be one vehicle description included on the crash report. These types of redundancies are often built into the crash report form and can be used to test key data elements for consistency.

Quality checking is built into many state data systems. These datasets have typically been developed individually, and in the case of EMS, there are established national schemas with built-in QC rules.

While good quality in component datasets is necessary for a good-quality linked dataset, it is not sufficient. QC rules should be developed for the linked dataset as well.

6.3.4 Timeliness

State crash datasets are used for planning purposes, and highway safety plans must be completed each year in July. Since crash-related fatalities may occur anytime within 30 days of a crash, the annual state crash dataset cannot be finalized until the end of January the following year. However, faster data entry and QC allow a finalized planning dataset to be released as early as possible, giving planners more time to use the data for their process.

Depending on the approach used, the linkage process itself may occur as data are collected or after datasets are finalized. In either case, timeliness in the linked data is dependent on timeliness in the original datasets. Moreover, QC processes using the linked dataset add to the time required to finalize datasets for use in highway safety planning. The linkage approach itself must produce a linked dataset in close to the same timeframe required for the original crash dataset.

6.4 State Datasets

This section provides a brief description of state dataset that might be linked within a comprehensive system. Where possible, each has some indication of available national schemas and the purpose of the linkage. By definition, this can be an ever-expanding list, so the databases discussed below should not be thought of as limiting scope.

6.4.1 Crash

The Federal Highway Administration's (FHWA) Crash Data Improvement Program (CDIP) provides a roadmap for assessing and improving state crash data quality. In addition, the MMUCC, represents a minimum data standard for state crash data. The MMUCC 3rd edition was published in 2008 (DOT, 2008) and is used in most states that use MMUCC. States are encouraged to update to the 4th edition, which notably has changed the definitions of the KABCO data elements in an attempt to better standardize injury reporting. Since MMUCC is a minimum standard, states will generally have a larger number of variables. However, all variables should be collected and reported in the same way from all police units in the state.

6.4.2 Emergency Medical Services (EMS)

The National EMS Information System (NEMSIS), funded through the Office of EMS within the National Highway Traffic Safety Administration (NHTSA), has provided for a common national dataset, database schema and a national EMS registry. Although EMS data do not contain the diagnostic specificity necessary to provide a MAIS-type measure of serious injuries occurring in crashes, the well-developed national schema, close ties between EMS and hospital personnel and records, and the presence of crash location and other key data elements make NEMIS an ideal intermediate dataset for linking crash to hospital outcome. Other valuable measures of injury severity are present in EMS data including the Centers for Disease Control (CDC) Trauma Triage Criteria and the RTS. Moreover, the NEMSIS schema includes standard data elements from the National Trauma Data Standard (NTDS) to enable linkage between EMS and hospital trauma registry datasets (see section 6.4.3). Information about the occupant's condition immediately after the crash as well as the transit time are included in these datasets, providing additional information about patient condition and care.

One aspect of EMS datasets that affects linkage is the fact that each entry in the dataset is a patient run, or a particular point-to-point transport of a particular patient. In most EMS datasets, a patient who is transported a second time (e.g., transferring between hospitals after initial evaluation) will appear in a different record and there will not be a common identifier to

link the two records. Thus, when linkage is made from crash-to-EMS and EMS to medical outcome data, cases with transfers may be less likely to link to final discharge diagnosis (possibly made at the second hospital) and treatment. There are several proven approaches to handling this problem.

6.4.3 Trauma

State trauma databases are collected in a majority of states, though states vary in whether trauma data collection is mandatory or voluntary. They also vary in which hospitals or trauma centers are required to participate. In general, trauma databases include patients whose diagnosis falls into a group of categories defined as "trauma," defined by the American College of Surgeons (ACS) as At least one ICD-9-CM injury diagnostic code in the range: 800–959.9, excluding superficial injuries and at least one of the following:

1. The patient is admitted to the hospital with at least a 24 hour stay;
2. Patient transfer via EMS transport (including air ambulance) from one hospital to another hospital;
3. Death resulting from the traumatic injury.

These diagnoses are generally more severe, on average, than those in a hospital discharge database (see Section 6.4.4) or an ED database (Section 6.4.5).

The ACS has developed the Trauma Quality Improvement Program (TQIP) to assist states and trauma centers in collecting high-quality trauma data. They have also developed the NTDS, which should be used by state trauma databases, but is similar to MMUCC in being a minimum dataset standard that is used with varying levels of compliance. The TQIP program, like CDIP, aims to improve compliance in all states. It should be noted that NTDS and NEMSIS are mutually compliant and integrated, facilitating data linkage between EMS and trauma databases in states.

One advantage of statewide trauma registries is that they typically include AIS codes in addition to ICD codes. Trauma registrars are commonly trained AIS coders and will incorporate coding into their data entry activities. Thus, these datasets are easier to use with an MAIS 3+ definition of serious injury.

6.4.4 Hospital Discharge

Statewide hospital discharge datasets include all patient records at discharge for anyone admitted to a hospital within a state. Compared to state trauma registry systems, state hospital discharge data systems are more inclusive (contain all injured patients treated in all acute care facilities), but provide less detail regarding the patients' injuries (e.g., severity), mechanism of injury and treatment details. Discharge datasets are based upon a universal billing standard (UB-04), while registry systems employ trained abstractors to review medical charts and record specific injury-related information. However, these datasets may be better for injury surveillance in general because they include all injuries that require hospitalization, rather than only the more serious injuries included in the trauma definition.

In 2003, the State and Territorial Injury Prevention Directors Association published a report of the Injury Surveillance Workgroup containing recommendations for using hospital discharge data for injury surveillance (Injury Surveillance Workgroup, 2003). The report not only covered the use of such data for injury surveillance, but made recommendations for standardized reporting and analysis that would facilitate the use of such data. At the time, the group reported that over 40 states collected some data on hospital discharge, though it is unknown how complete the coverage was for these states.

A focus of the Injury Surveillance Workgroup's (2003) report was the use of externalExternal-cause codes in the patient discharge report. External-cause codes are part of ICD coding system and are used to code external causes of injury. The specific codes have been updated in moving from ICD-9-CM to ICD-10-CM. However, the principle of externalExternal-cause codes is that they identify causes, which allow data related to MVC to be identified and separated from the larger dataset. For data linkage, it is important to focus linkage on the cases related to MVC because these cases are a small subset of all hospital data.

Although over 40 states were collecting state hospital discharge data in 2003, many did not reliably include externalExternal-cause codes. The Healthy People 2020 program (see: http://www.healthypeople.gov/2020/topics-objectives/topic/injury-and-violence-prevention/objectives?topicId=24) includes objectives to increase the use of externalExternal-cause codes in both emergency and hospital discharge databases. An evaluation of the use of externalExternal-cause codes for injury surveillance in 2007 (Lawrence et al., 2007) indicated that over half of states either mandated externalExternal-cause codes or obtained over 85% compliance voluntarily. A more recent evaluation (Barrett & Steiner, 2014) indicated 92% compliance for inpatient discharge and 94% for ED data. Thus, a prerequisite to linkage between crash and hospital discharge is wide coverage of hospitals within a state and reliable use of externalExternal-cause codes for all participating hospitals.

The Healthcare Cost and Utilization Project (HCUP), sponsored by the Agency for Healthcare Research and Quality, works with states to standardize and make available state inpatient discharge data (SID) (Barrett & Steiner, 2014). They report that 47 states participate in the HCUP SID, though the specific coverage of hospitals within a state is determined by the state and not clearly reported on the HCUP website. The SID provides a common data standard to facilitate comparison across states.

6.4.5 Emergency Department

ED data represent a further expansion of the sample available for linkage between crash and medical outcome. ED data include everyone seen in the Emergency Department, of whom most are never admitted to the hospital. Relative to trauma and hospital discharge datasets, ED datasets include patients who are much less severely injured, but who are much more numerous.

The HCUP program also works with states to standardize and make available state ED data through the State Emergency Department Databases (SEDD). The SEDD includes only those who were not admitted to the hospital, and as with the SID, hospital inclusion is determined by the state agency submitting data. Thirty-one states participate in the SEDD.

The data and need for E-coding is similar for ED and hospital discharge datasets. These datasets include ICD externalExternal-cause codes, but generally not AIS codes. Consistent use of externalExternal-cause codes is necessary for linkage to the ED dataset as well.

6.4.6 Roadway Databases

Roadway data include a variety of characteristics of the road and traffic in it, which are geographically referenced. These datasets may include physical characteristics (e.g., number of lanes, shoulder width), access control (e.g., public/private, toll), intersections (e.g., traffic control, intersection type), inventory (e.g., signs, pavement within a road segment), traffic characteristics (e.g., volume), structures (e.g., bridges), railroad crossing (e.g., signal type), pavement management (e.g., condition, repair history), and assets (e.g., guardrails, signs) (DeLucia et al., 2012).

The Roadway Data Improvement Program (RDIP) assists states in improving the quality of their roadway data and its management. This includes data elements and linear referencing systems. Safety-related data elements for roadway databases are specified in the Model Inventory of Roadway Elements (MIRE; Council et al., 2007) and the RDIP program encourages states to comply with the MIRE standard.

Linkage to crash records is a key part of roadway safety analysis. Linkage between crash and roadway is done on the basis of crash location (tied to the roadway referencing system) and is relatively straightforward compared to linkage between crash and medical outcome. This linkage, and achieving a certain level of location precision, is essential to planning roadway safety improvements. Moreover, linkage from crash to medical outcome can be carried into the roadway dataset to enhance safety analyses.

6.4.7 Driver Licensing

Driver license databases contain an inventory of all driver's license numbers, along with the demographic and other information about the driver that is contained on the license. This information includes name, address, birthdate, sex, and self-reported height and weight, among other elements. These can be useful in both linkage (via name, address and birthdate) and safety analysis (using age, sex, height, and weight). State driver license files are generally complete, so coverage is not an issue. In addition, police reports routinely include the driver's license number, so linkage is generally very simple. Privacy is usually the only hurdle to overcome in linking crash and license data, though this issue should be considered carefully.

6.4.8 Driver History

State driver history datasets are also indexed by license number but contain citations, arrests, and adjudications. These are critical datasets for understanding safety-related issues such as recidivism among drunk drivers. Like driver license files, the driver history database should be readily linkable to crash via license number. Similarly, privacy is an issue to be considered.

The Governor's Highway Safety Association (GHSA) recommends that states develop a single information system for driver licensing and history data (GHSA, 2014). They also support exchange of such information between states and development of a national driver database. At this time, no such database exists, and driver history is generally not shared between states.

6.5 Common Identifiers

To link people in any pair of datasets, one or more common identifiers must be present in both datasets. The gold standard of identifiers is a single, unique, permanent, person-specific, identification code (ID) used in all datasets and assigned to all people in all datasets. However, several less ambitious forms of linking variables can also be used effectively.

The permanent person-specific ID code allows for analysis of events and treatments that occur outside of the time frame of a single crash event. This is ideal because analysis of follow-up treatment and long-term outcomes require this. However, achieving a statewide person-specific ID is logistically very challenging. A person-specific ID code that is only used for a given event, but is available in all datasets is much easier to implement and will allow effective analysis of serious injuries in crashes.

Using a single ID code across all datasets is also not necessary if each pair of datasets has a common person-specific identifier. In addition, not all people in any one database need the common ID code. For example, only a small number of crash-involved occupants are transported by EMS, so only these people need an ID code for linkage to the EMS dataset. Other occupants

may be assigned a code, but they will not appear in the EMS dataset and their code will not be used. Linkage from crash to hospital would then pass through two stages (crash linked to EMS, and EMS linked to hospital) to get to the subset of people who were transported by EMS and admitted to the hospital.

Finally, identifiers do not have be single or in code. The method of probabilistic linkage will be discussed in detail later in this report, but the key idea is that when a unique common identifier is not available in a pair of datasets, it may be possible to use a set of non-unique common variables to estimate the probability that a person in one dataset is the same as a person in the other dataset. Variables used for this process often include name, birthdate, age, sex, date/time of event, and location of event. Although these identifiers are not unique or single, they must still be common among the datasets to be linked.

6.6 Access Rules & Permissions

Privacy and data protection requirements for any data with personally identifying information (PII) are set by a combination of state-specific laws and the HIPAA. HIPAA applies to health data, so once crash data are linked to health data, the HIPAA rules will apply.

It is important to note that HIPAA does *not* prevent health data from being used for research or public health purposes or from being linked to other datasets. Instead, it sets out the conditions under which such uses may be made. These conditions include definitions of de-identification, requirements for security, and rules for access.

State-specific laws, however, may prevent linkage or use of a linked dataset, or may set additional conditions for permission and access. In some cases, these laws have had to be changed to facilitate data linkage. In any case, they must be known and their requirements addressed.

A good statewide data linkage system requires rules for access and software to allow appropriate access and prevent inappropriate access. Rules for access must comply with HIPAA and state law, and the level of access may be different for different individuals. It is even possible that state laws that impede linkage may need to be changed. A de-identification method that complies with HIPAA allows for a much wider range of individuals to have access (to the de-identified data).

7 Roadmap to Linkage

The previous section described the necessary components of a comprehensive statewide data system with linkages to crash and including medical outcome (among other datasets). However, the process of putting these elements in place is complex and challenging. In this section, we present a process, or roadmap, by which states can reach the goal of comprehensively measuring serious injury in crashes at the state level through linked datasets.

The contents of this roadmap are based on interviews and discussions with staff of state agencies and researchers working with state data. Most of the ideas in the roadmap have been tried and/or are in use in at least one state. In addition, we present reasonable alternatives that allow states to choose what works best given their circumstances.

7.1 Roadmap Overview

Table 11 provides an overview of the steps in the roadmap. Each step is detailed in a subsection below.

Table 11 Summary of Steps to Linkage

Step	Key Goal.
1: Arrange Collaboration Among Relevant Agencies	Facilitate critical communication and decision-making pathways; create a data linkage project group; identify the motivation for and benefits of participation for each group.
2: Catalog Available Databases	Know coverage, contents, schema, inclusion criteria, and potential problems with each database before trying linkage.
Step 3: Determine Databases to Be Linked	Make a plan for the order in which databases will be linked.
4: Identify the Identifiers	Know what is available in existing databases to aid linkage.
5: Determine Linkage Mechanism	For each pair of databases to be linked, choose a mechanism for that linkage. Consider facilitating linkages among more than two databases.
6: Step 6: Determine Database Storage Mechanism	Data must be stored and managed, and access must be protected.
Step 7: Harmonize Common Data Elements	Common data elements in linked files must conform to a common schema.
8: Set Up a Pilot Project	Testing on a small scale helps find problems to fix.
9: Set Up a Sampling Program (Optional but strongly recommended)	Provides pre-linkage ability to measure serious injuries and a way of testing linkage approaches as they are developed.
10: Set Up Statewide Linkage	Pilot project must eventually be launched statewide.

7.2 Step 1: Arrange Collaboration Among Relevant Agencies

In most of the states we interviewed, the TRCC was the group that motivated, initiated, facilitated, and often provided funds to launch data linkage activities. Representation of all relevant agencies, especially the combination of public health and crash/roadway agencies, is necessary for a successful program. In some states, Memoranda of Understanding were signed by agency officials to commit to supporting linkage-related activities specifically. In other states, the agreements to support linkage were less formal, but many of the people we interviewed talked about how important multi-agency participation and advanced buy-in was to the process.

To engage the different agencies that need to participate for a successful program, it is important to start by identifying the motivations and potential benefits of linkage. Some of these may be specific to one agency (e.g., tying diagnosis-based injury outcome to roadway features), while others may serve the great good by enabling research and improving assessment of many different countermeasures. The motivation provided by MAP-21 is not sufficient to sustain an effective multi-agency linkage program. Moreover, knowing what each group hopes to get out of a linked data system will help in prioritizing development of the various components of the system.

At a practical level, the multi-agency discussion around the linkage process should include a number of specific issues that will have to be resolved for a successful program. First, agencies will need to identify what each will put in (e.g., money, staff resources, data), and what each wants to get out (e.g., data access, specific reports, new data elements). Dataset ownership and rules for access must be worked out.

Data access rules must comply with HIPAA, state law, and agency policy. As a result, there should be a legal review early in the process to determine whether any state laws or agency policies need to be changed. This has been necessary in some states and since the process may be slow, it should be addressed early.

7.3 Step 2: Catalog Available Databases

Since the first requirement of a linked data system is statewide datasets in good condition, the second step in the data linkage roadmap is to catalog the available datasets. Although crash linked to medical outcome is necessary to meet the goal of measuring serious injuries in crashes, the cataloguing process should include a much wider variety of datasets. A useful goal might be to measure the *comprehensive* cost of crashes in the state. This would promote inclusion of many datasets that address costs (e.g., Medicare, roadway asset management and repair) in addition to datasets that focus on what happened (e.g., medical outcome and crash data). The items to catalog for each dataset are listed in below.

7.3.1 Data Dictionary

As described earlier, a data dictionary, or schema, is essentially the contents, or codebook, of the dataset. It includes the relational structure of any tables, all variable names, values each variable can take on, and the meaning of numeric codes.

7.3.2 Inclusion Criteria

Inclusion criteria describe how cases are selected to be in the database (or not). Inclusion criteria are critical in linkage because they influence what cases can be in linked datasets and they influence results of any analysis. For example, if the medical outcome dataset is a trauma dataset, then only those occupants with injuries requiring hospital admission and meeting trauma inclusion criteria will be available for matching. This should make it possible to measure MAIS

3+ injuries fairly well, but will not facilitate measurement of all injury costs or of less serious injuries not necessitating hospital admission.

Inclusion criteria may also restrict the dataset to certain reporting units, such as Level 1 and 2 trauma centers. These centers may be more prevalent in urban areas and therefore may bias the state dataset towards events that occur within their catchment areas.

In general, among hospital datasets, trauma registry have the most restrictive inclusion criteria, followed by hospital discharge and then ED. In the long run, linkage to ED data, in addition to hospital and/or trauma data, will be key to broad determination of the injury outcomes of people involved in crashes. The order in which a state chooses to incorporate linkages will depend on the condition of each of the databases.

EMS datasets, by definition, include only cases that are transported by ambulance or other emergency services. Thus, linkage via EMS will not capture those who are transported by private vehicle. As with a trauma registry, EMS linkage will likely capture almost all MAIS 3+ cases, but a future desire to analyze less serious injuries may require direct ED-to-crash linkage in addition to EMS-based linkage.

Crash datasets also have inclusion criteria, and in some states, the criteria include a requirement of injury to one or more persons involved in the crash. A stricter police-reporting requirement will change the nature of the included crashes in both the original and linked crash datasets. This is not likely to be an issue for measuring serious injuries, but might become relevant if injury measurement is eventually broadened to include all injuries.

Finally, other datasets, such as licensing and driver history, will include only those licensed within the state. Out-of-state drivers will not be linkable within a state's databases without some additional agreements with neighboring states.

7.3.3 Coverage

Coverage, described above, reflects the proportion of possible cases in a state that are available in the statewide dataset. For example, some entities may not successfully report to the state, so they will be left out of the state database. Often, this is a small percentage of cases, but they may be systematically biased towards certain types of entities (e.g., small or rural). As a result, the dataset will be biased to some degree, which needs to be assessed and accounted for in analysis.

7.3.4 Quality Control

QC was discussed in the previous section, but in this step, quality issues need to be catalogued. In particular, problems should be evaluated in terms of their specific potential impact on the analysis of serious injuries in crashes. For example, if crash location is used to enable linkage (e.g., to EMS data), then substantial missingness in that data element will hamper the linkage process. In contrast, if specific driver distraction is missing, then only driver-distraction-related analyses will be affected. In other cases, some jurisdictions may apply codes differently than others, biasing results of analysis and indicating the need for better training. Cataloguing quality issues is critical because these problems compound in the linkage process. Only cases that have all necessary identifiers and other key data elements in all datasets can be linked and used in analyses.

One key data element that should be mentioned specifically is the externalExternal-cause code required by hospital datasets. Separating MVC cases from the large number of other cases in trauma, hospital, and ED datasets is critical for linkage. externalExternal-cause code coverage

and consistency should be carefully evaluated at the cataloguing step, and issues in its use should be addressed early in the process.

7.4 Step 3: Determine Databases to Be Linked

The goal of comprehensiveness argues for a data collection and linkage process that maximizes linkages between different datasets as well as successful linkages among cases that are expected to link. This ambitious goal is probably best achieved by a step-by-step process of adding databases and linkages. Thus, this step should involve setting up a long-term plan for adding linkages over time.

Those planning the linkage process should take a number of factors into consideration. First, the condition and coverage of databases need to be good enough to support linkage at the state level. Thus, some databases may not be initial candidates until they are improved. Second, the utility of a linkage should be considered. What questions can be answered through this linkage? Are they high-priority and high-impact? Third, the amenability of datasets to linkage should be considered in planning (though difficult linkages should not necessarily be developed last). Considerations should include the compatibility of data elements, presence/absence of common identifiers, and the amenability of the existing data structure to linkage.

A good starting point for tying injury outcome to crash is linkage from crash-to-EMS to trauma registry. National standardization of EMS databases through NEMSIS, as well as the physical presence of the ambulance at the crash scene and at the hospital location, facilitates linkages between EMS and both crash and trauma registries. Trauma registries commonly include AIS coding and tend to capture nearly all MAIS 3+ cases, making them easiest to work with.

As described earlier, one challenge for EMS linkage is that most EMS databases do not include a patient-specific identifier that tracks transfers. Thus, a patient who is taken by ambulance to a local hospital and is then transferred by ambulance to a trauma center for more specialized treatment will typically have two unlinked entries in the EMS dataset. Crash data will link to the first EMS transport and the first hospital, and trauma registry will link to the second transport. Linking the two is possible, but requires an extra process. Unfortunately, this issue will tend to affect some cases more than others—more seriously injured patients and more rural locations are more likely to be transferred and therefore may fail to link via EMS when the within-EMS patient identifier is not handled.

There are at least two solutions to the patient transfer problem. One is to identify transfer cases in the EMS dataset so that a patient is tracked across runs. This can be accomplished by completing an “internal linkage”. In other words, using common elements within the EMS dataset it is possible to identify (i.e., link) all of the EMS resources reporting data for the same patient. Another approach is to implement two linkages: one with EMS as intermediate step and another direct linkage between crash and trauma registry. This allows checking of the linked cases from two directions and could aid in solving the transfer problem.

To expand the medical outcome linkages, crash-EMS-trauma could be followed by successively expanding the inclusiveness of medical databases (e.g., adding hospital discharge and ED datasets). Each more inclusive hospital dataset will introduce new challenges, including the need to handle ICD codes, but will also expand the breadth of crash-related injuries that can be analyzed.

Roadway linkage to crash is generally more straightforward and has been implemented in most states. Once crash is linked to injury outcome, this information can be brought through to

the roadway database to facilitate analysis of roadway safety as it relates to injury as well as fatality.

Similarly, driver license and history databases are generally already linked and can be tied to injury once injury is available in the crash dataset. It should be noted here that linkage to a state license file will only include in-state drivers. In some states, there is a reasonably large population of out-of-state drivers in crashes, and the inability to link to these drivers might be noted in this step. Solutions include: 1) setting up agreements with neighboring states, and 2) a national driver license and history dataset.

In general, multi-directional linkage among relevant databases is the ideal goal. Linkage from crash-to-EMS, EMS to hospital, and hospital to crash allows for assessment of many aspects of the cost of crashes. Hospital data with good-quality externalExternal-cause codes will represent the universe of crash-related hospital admissions. EMS data often include injury cases that do not have an associated police report. These can help assess how police are using inclusion criteria for police reports and the extent to which police are called in for different types of crashes. For example, pedestrian and bicycle crashes might be underreported when considering only police crash reports.

7.5 Step 4: Identify the Identifiers

State databases that need to be linked were, in most cases, developed independently and prior to thoughts of linkage. As a result, common identifiers may or may not be present in databases to be linked. Once datasets have been catalogued, a high-level linkage schema should be developed to understand how databases are related to each other. This schema will call out the linkages that require a linkage mechanism, and then each linkage can be addressed.

An example of such a linkage schema for Michigan is shown in Figure 6. In Figure 6, each line between data tables must have some means of identifying common cases. The linkages among tables within the crash database are defined in the crash database schema. However, the linkage between the “Person” table in the crash database and the “Medical Record” table in the medical database must be addressed. As a starting point, identifiers in the “Person” table and identifiers in the “Medical Record” must be selected.

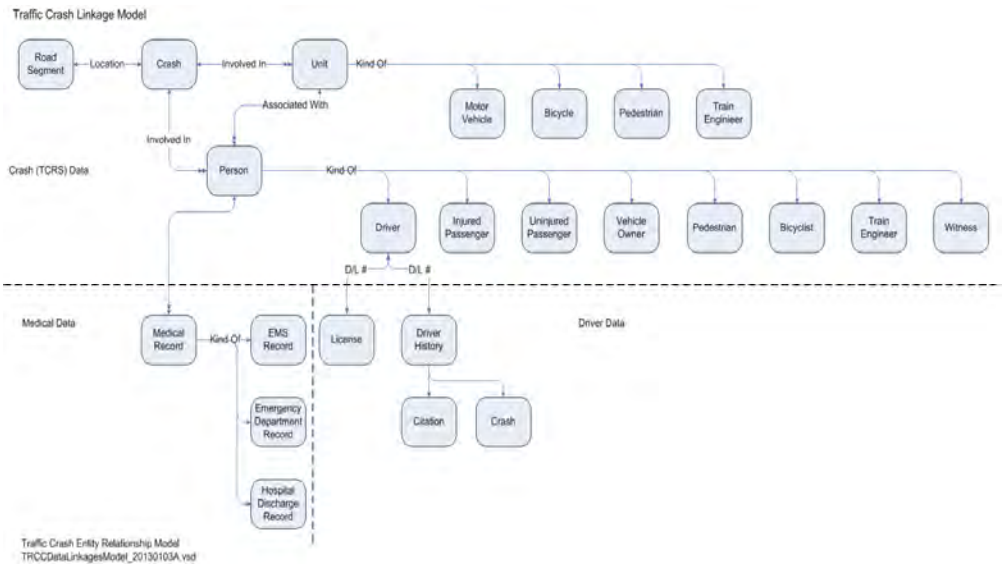

Figure 6. Example linkage schema from Michigan

7.6 Step 5: Determine Linkage Mechanisms

If common, unique identifiers are present in datasets to be linked, then the process is straightforward. However, for linkages across different databases, such as the "Person" to "Medical Record" linkage in Figure 6, common unique identifiers will usually not be present. This section presents a variety of linkage options from which states can choose, along with their pros and cons.

7.6.1 Adding Identifiers After the Fact

The first set of options for identifiers are those that are added to databases after the data are collected. These require datasets to have enough data elements in common that the linkage can be done after the fact. Typically, these approaches are less comprehensive and precise than implementing a process for assigning and passing identifiers in the original records (i.e., at the scene or at/near the time of the event), but they can be easier logistically.

7.6.1.1 Probabilistic Linkage

From 1992-2012, the NHTSA funded the CODES program, which provided both funding and technical support to a set of CODES states to work on linkage of their own data systems. Since 2012, a number of CODES states have continued to self-fund their programs.

One of the key features of the NHTSA-funded CODES program was the CODES Technical Assistance Center, which developed statistical methods and software to support the probabilistic linkage process. In addition to the large volume of content work, the CODES program produced a number of papers focused on improvements to and understanding of probabilistic linkage (e.g., Cook, Olson & Dean, 2001). In addition, the statistical work was implemented in software that is now commercially available as LinkSolv. The papers, analytical reports, and software provide a large storehouse of knowledge and applications of data linkage. However, the loss of centralized technical assistance has left states to implement programs on their own (seeking support individually). This issue will be discussed further later in the report.

Probabilistic linkage is the process of using common (but non-unique) variables in a pair of datasets to compute the likelihood that a case in one dataset refers to the same person as a case in another dataset. A patient in a trauma dataset can, for example, have non-zero probability of matching more than one case in a crash dataset. The technical approach to probabilistic linkage is described elsewhere (e.g., Felligi & Sunter, 1969; Jaro, 1995).

Although it is not necessary to understand all of the technical details of the method, there are certain key concepts that have an effect on how states might go about using this method to link datasets. The basic idea of probabilistic linkage is the assignment of a match weight to each possible pair of cases in each of two datasets. Suppose, for example, that a state has a crash dataset with 100,000 cases and an EMS dataset with 200,000 cases. There are 20,000,000,000 possible matches.

The match weight for a given pair of cases is the sum of the individual match weights for each of the variables that are used in the linkage. If two cases match on the value of a particular variable, then the match weight is shown in Equation 1.

$$w_i = f\left({}^{m_i}/_{u_i}\right) \qquad (1)$$

where m_i is the probability that the values of variable i match given that the cases refer to the same person, u_i is the probability that the values of variable i match given that the cases do not refer to the same person, and f is a monotonic function, typically $\log_2$.

If the two cases do not match on the value of a particular variable, then the contribution to the total weight is given in Equation 2.

$$w_i = f\left({(1 - m_i)}\Big/{(1 - u_i)}\right) \quad (2)$$

In Equations 1 and 2, the value of *m* essentially represents the quality of the variable. In theory, a true match should have the same value of each variable with probability 1. However, data entry errors and missing data generally cause mismatches with some non-zero probability, so *m* can be somewhat smaller than 1.

The denominator component, *u*, reflects the tendency of a variable to match at random, or the discrimination ability of the variable. For example, sex has only two values and will match at random 50% of the time. This creates a large denominator in Equation 1 and results in small match probabilities. By comparison, birthday (without year) has 366 possible values and thus will match approximately 1 out of 366 times at random ($u \approx 1/366$).

In practice, the match weight is specific to the value that does or does not match. This allows for a more nuanced assessment of the information value of each match (or non-match) in helping to determine whether two cases match. For example, if a state has 10 counties, and 80% of crashes occur in one county, then a match on that county is not as informative as a match in another county. The value of *u* for the common county is much greater than for other counties and results in smaller contributions to the total match weight. Similarly, matching on a common last name, such as Smith, is less informative than matching on an uncommon last name.

For computational efficiency, many possible pairs are eliminated from consideration based on a blocking variable such as time of crash (within a certain time window). Nonetheless, once total match weight has been computed for all pairs under consideration, a histogram of total match weight should ideally result in two peaks. One peak, with very low match weights, will contain the clear non-matches, and another with high match weights will contain clear matches. The cases in between represent “possible” matches. It is important to keep in mind that match weight is assigned to each considered pair, so a specific case in one dataset may have a “possible” match to more than one case in the other dataset.

To understand the consequences of the matching process, Figure 7 and
Figure 8 illustrate two different matching scenarios. Figure 7 shows a hypothetical example of distributions of match weights for actual matches (blue) and actual non-matches (pink) in the case where match quality is generally high.
Figure 8 shows a hypothetical example where match quality is generally low. In Figure 7, the matches and non-matches are easily distinguished and there is little uncertainty in the areas between the two distributions. In
Figure 8, there is a large range of uncertainty in which a given match weights could occur from either a true match or a true non-match.

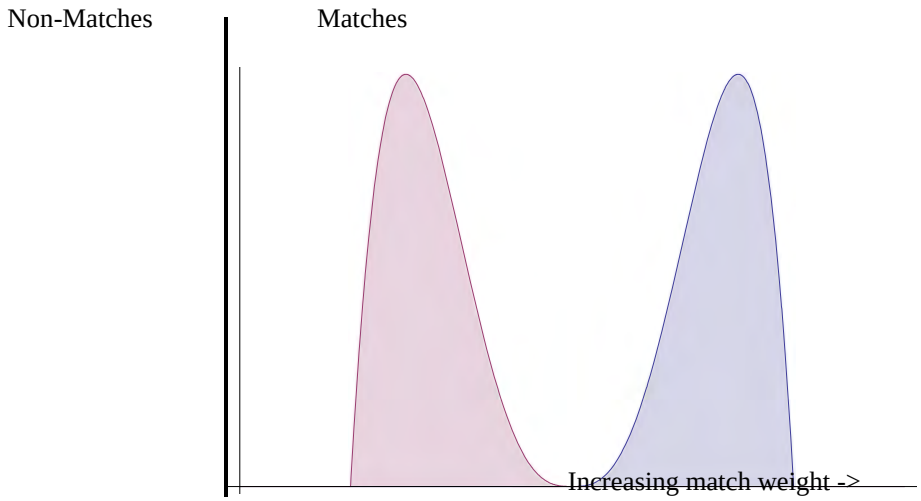

Figure 7 Hypothetical high-quality linkage

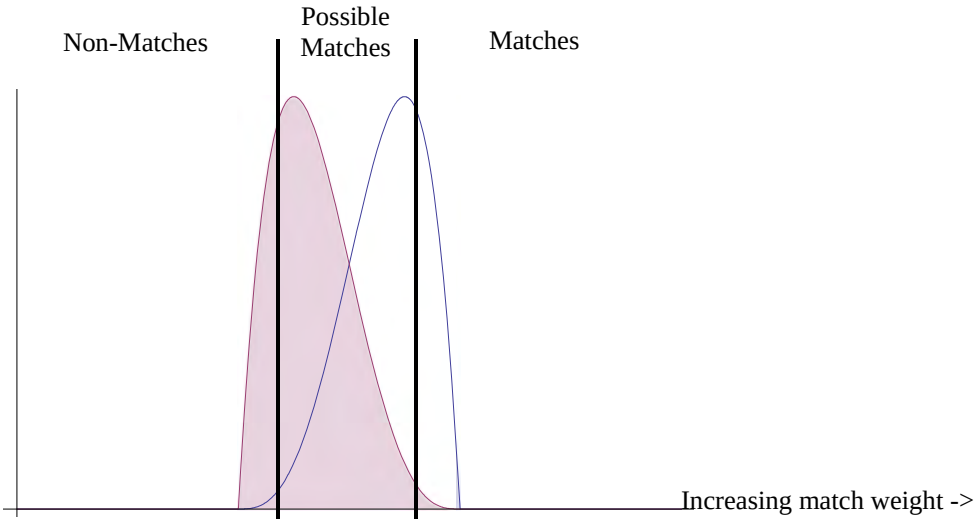

Figure 8. Hypothetical low-quality linkage

When match quality is low enough, some additional process must be used to account for the uncertainty in the "possible linkage" category. Three approaches include: 1) Inspecting each possible match by hand; 2) Selecting all matches of specific quality; or 3) Multiple imputation.

The first approach, hand-inspection of "possible matches" is recommended by Jaro (1995). If there are few matches in this category, the method is potentially feasible. However, in most cases, the number of inspections needed may be too high to be manageable.

The second approach is appealing in its simplicity and because it results in a single linked dataset. To incorporate matches into a data warehouse or other integrated data system, it is necessary to have no more than one match per case. Some states achieve this by requiring a match on each of a complete set of variables. Others keep only matches above a certain cutoff match weight (which can adjust for missing data). Either way, the cutoff approach simplifies the resulting dataset and analysis, but unless the dataset has very high-quality matching variables

with low missingness, the resulting dataset will tend to be biased towards rare events (i.e., unique combinations of match variables).

The tendency for high-weight matches to be biased is a critical issue if states are to use probabilistic linkage as the primary mechanism for measuring serious injury in crashes. Match weights inherently measure the informativeness of a particular case-pair's set of variable values (matches and non-matches). The numerator of a match weight component is a maximum of 1 and is only affected by data quality. However, the denominator represents the probability of a match by chance. Rare values are less likely to match by chance, resulting in smaller denominators and larger contributions to match weight. Thus, these values have a greater likelihood of ending up at the right end of
Figure 8 and being included in a dataset where only high-probability matches are kept.

The consequence to states of retaining a biased linked dataset is that rare events show up in the resulting metrics with greater probability. For example, in the example where one county has 80% of crashes, that county is more common and less likely to show up in the censored dataset (high-weight matches only). When counting serious injuries by county, injuries in the populous county will be undercounted relative to those in less populous counties.

The third, more statistically rigorous, alternative is to use multiple imputation (McGlincy, 2004). Multiple imputation (MI) produces a small set of parallel datasets (3-5) in which a matching row in the second dataset is selected at random for each row in the first dataset (no match may also be selected). Analysis is done in parallel on the datasets and the results are combined. The key benefit of MI is that it accounts for the uncertainty introduced by the matching process and reduces or eliminates bias in that way. In particular, it decreases false negatives—cases that should have matched but did not—which is a key problem in the use of a fixed cutoff.

The software developed for the CODES program handles the MI process. However, a critical disadvantage of the MI approach to handling lower-quality linkage is that it does not lend itself to producing a single linked dataset with one match per case. It is possible to produce a single imputation for use as a linked dataset, but results of analysis will be influenced by unusual random selections of links. In addition, the means of choosing imputed links (McGlincy, 2004) does not guarantee one link per case, since imputation is at the level of a pair of cases and each case can be evaluated as part of several pairs.

If a state chooses to use probabilistic linkage, it will be vitally important to use metrics to assess the overall quality and potential bias in the resulting dataset. Using a single imputed dataset and taking matches above a cutoff have different pros and cons and either might be used to produce one dataset. However, the ideal solution is to have sufficient separation of the matching and non-matching groups. This quality depends on the particulars of the dataset and the variables being used, so a standard set of metrics for linkage quality would be helpful to states trying to determine whether additional identifiers are needed. In particular, some of the identifiers discussed in the next section go hand-in-hand with probabilistic linkage and should improve match quality.

Once probabilistic linkage has produced a matched dataset (of sufficient quality), the linked cases can be assigned a numeric identifier that is inserted in the separate databases. This allows the matches to be recreated on the fly when data are accessed. Thus, the linkage process can be carried out on new data once (updated at regular intervals) and the results can be used by anyone with access to the linked datasets.

In general, probabilistic linkage allows states that have not previously incorporated common identifiers to link datasets going back a number of years. However, the complexity of the process and the potential for producing a biased linked dataset argue for two things: First, there is a strong need for a technical assistance center at the national level that can help states assess the quality of linkage mechanisms, *including but not exclusively probabilistic linkage*, and assess the quality of the linked dataset. Second, probabilistic linkage should ideally be viewed as an intermediate step on a path to incorporating an on-scene identifier (see Section 7.6.2).

7.6.1.2 Hand Linkage

Hand linkage is a form of after-the-fact linkage done by a human, but made logistically feasible by software. This approach, employed in Kansas to link EMS to trauma registry data, uses software to select a small set of EMS runs that are potential matches to a single trauma case. The potential matches are selected based on the timing of the EMS run and the destination hospital, and the trauma registrar selects the correct match by looking at name, address, birthdate, gender, and other identifying information in both records. Once the match is selected, a common identifier can be pushed into both the trauma and EMS databases to enable future linking of de-identified information.

This approach has certain advantages. First, matching can make use of the ability of the human to know what names are likely to be matches even when nicknames or different spellings are used. (Name-based matching is a challenge for probabilistic linkage software.) Second, the workload of hand-matching, which would normally make the process infeasible, is spread among trauma registrars, whose job is data entry and data management. The list of potential matches is made as short as possible through software intelligence, and then the final decision is made by the human. Finally, the trauma registrar is, by definition, allowed to see PII on patients, but once cases are matched, PII can be removed.

There are certain disadvantages to this approach as well. First, it is difficult to assess the matching algorithm or success rate for a human match process. Different registrars may have different criteria for accepting an apparent match, and the overall error rate is unknown without a separate study (which might be warranted). Second, Kansas was able to successfully motivate their trauma registrars to do this because some EMS data are needed in the trauma dataset as well, and these data elements are pushed automatically once the link is made. However, this motivating benefit may not exist for all datasets the state might want to link (e.g., crash), so there must be some consideration for the additional burdens on the trauma registrars (or any other personnel used to hand-link).

7.6.2 Assigning Identifiers At The Time of the Event

The alternative to identifying matches after the fact is to assign some type of identifier at the time of the event, ideally on-scene, and pass that identifier among responding agencies. This group of approaches is more logistically challenging to implement, but once implemented, should allow linkage without introducing bias and without the technical challenges of the after-the-fact approaches. We present several classes of on-scene identifiers below.

7.6.2.1 Event-Specific

An event-specific identifier is one that specifies the crash event, but does not separately identify the people involved in that event. This further identification would have to be done using one of the after-the-fact approaches, but by limiting possible matches to only those people involved in a single crash event, the matching process should be more successful.

The advantage of using an event-specific identifier is logistical simplicity. One approach is to pass either the EMS run number or the crash report number (or both) between agencies, ideally at the scene. In the near future, it may be possible to use vehicle-to-vehicle (V2V), based on dedicated short-range communication (DSRC), to pass the event number automatically to other responding units. Bettisworth et al. (2015) describe some of the data issues, including meeting HIPAA requirements, that need to be addressed to make DSRC data useful for emergency response applications. Solutions to these issues will facilitate the use of V2V to enable data linkage among crash, EMS, and hospital datasets.

GPS location and time can also be used to identify an event. For linkage to roadway datasets, location is already used in states. For linkage between crash and EMS, GPS location and time will not be identical for police and EMS. However, a fairly simple algorithm could choose time-location combinations that match (between police and EMS).

By not having to specify the person, this approach is less time-consuming at the scene and can more easily be automated. The disadvantage is that extra work still has to be done to identify specific occupants for matching to hospital records. The event-specific identifier becomes a good way to limit the field and improve probabilistic (or hand) linkage, but it does not solve the whole problem of finding the same individual in multiple datasets.

7.6.2.2 Person-Specific/Event-Specific

A person-specific/event-specific identifier is one that is assigned to each person involved in a crash, but only for that crash. Trauma bands, which are physical ID wristbands given to anyone seen by rescue personnel, fall into this category. An alternative is for the police to assign numbers to each occupant in a crash and pass along the number to EMS for any occupants who are transported. EMS would then pass the number to the hospital on arrival.

The advantage over event-specific identifiers is obvious. Any linkage method aimed at capturing injury outcome will have to link people—patients in the hospital dataset to occupants/non-motorists in a crash—and a person-specific approach of any kind accomplishes this without additional analysis. The disadvantage is generally logistical burden, particularly at the scene, when other activities (e.g., treating victims) are higher priority. Any on-scene solution must take little time and be very simple, and the person-specific approaches may be difficult to implement in this way.

7.6.2.3 Person-Specific/Global

The gold standard of identifiers is one that is person-specific and permanent, or global. This ID, like social security number (SSN), follows the person throughout all datasets over time and allows for assessment of long-term follow-up, delayed treatment, and repeat visits. It also allows patients to be easily followed when they are transferred between hospitals.

An example of a person-specific/global identifier is the driver's license number. For linkage from crash to driver license/history files, this identifier is ideal. However, young occupants do not have licenses, and crash reports generally do not include the license numbers of non-drivers. Medical outcome datasets also do not tend to include license number. Thus, license number is feasible only for information pertaining specifically to the driver in a crash, but not for linkage to injury outcome.

Alaska and Massachusetts are in the planning stages of implementing two versions of person-specific global identifiers. In Alaska, everyone who interacts with public safety personnel (for crash, crime or other reason) is assigned an Alaska Public Safety Identification Number (APSIN). When police respond to a crash, they look up each occupant in the APSIN system. If a

number has been assigned, it is automatically entered on the crash report. If a number has not been assigned, a new one is generated, and an entry is made in the APSIN database for future use. Alaska is now embarking on a program to allow hospital and trauma registrars to access the APSIN system to put that number in the hospital record.

Massachusetts uses encrypted SSN in a variety of datasets including ten-year death data, trauma registry, hospital discharge and all-care claims data. They use a single encryption algorithm, which translates each SSN into a different number, but one that remains the same across databases. In this way, the patient can give their SSN rather than be assigned a new number, but the SSN does not reside in any of the databases, providing an extra level of security. Massachusetts is exploring incorporating the same identifier into EMS data, and possibly eventually crash data.

One key advantage of these two approaches is that numbers do not need to be passed on-scene between agencies. Common person-specific ID numbers can be accessed separately by each agency that the person encounters. In Alaska, the number can be looked up separately by the rescue or hospital personnel, and in Massachusetts, the patient (or family) can provide the number (to be encrypted). The other obvious advantage is the ability to track the same person beyond the initial EMS run and hospital admission.

The disadvantage of the Alaska approach for other states is the need to implement a statewide ID system first. The APSIN system has been in place in Alaska, thereby facilitating its use in this case. However, setting up a statewide ID system will delay implementation of any data linkage that depends on it. The disadvantage of the Massachusetts approach is that SSN must be provided by the patient or family. Parents do not always have children's SSNs available and some very young children may not have a number at all. Unconscious victims cannot provide SSN, leading to a potential injury-based bias in missing data in that field.

7.6.3 Summary of Linkage Mechanisms Used by States

Table 12 summarizes the linkage mechanisms used by the states we visited and interviewed. Other states, particularly those with CODES programs, reported using probabilistic linkage in the survey. Although some states are trying on-scene identifier methods, most are using probabilistic linkage. This approach can work if the technical challenges are handled and appropriate linkage-quality metrics are used. Although most states have struggled with the linkage process initially, North Carolina has demonstrated that once the system is in place, it can work in a timely way (NC can produce a linked crash-EMS annual state dataset by the end of February—two months after the year ends).

Table 12 Summary of Linkage Mechanisms Used by States Interviewed

Linkage Approach	States
Probabilistic (Linkage Software)	CA, MA (crash-EMS), WA, MT (planned), NC (crash-EMS); (many other CODES states continue to use this method)
Probabilistic (Perfect Matches Only)	UT
Hand Linkage	KS (for EMS to trauma, crash under discussion), NC (for EMS to trauma)
Event-Specific Identifier	FL (pilot)
Person-Specific Identifier (Event-Specific)	AL (pilot)
Person-Specific Identifier (Global)	AK (for crash, under development), MA (for trauma to case-mix (ED, admitted, observation))

As discussed earlier, probabilistic linkage is ideally seen as an intermediate step in the development of a state linkage process. Realistically, event-specific identifiers and even person-specific on-scene identifiers will need to be combined with probabilistic linkage for some time.

Generally, the goal of a linkage system should be to return exactly one linkage for each and every transported or injured case between the crash record and the medical record for that case. This link (per person) should be tied to the crash record so that key information can be further linked to roadway, licensing, and other datasets related to understanding traffic safety. The extent to which this is achieved is a measure of the performance of the linkage process. The choice of linkage mechanism will influence the present and future performance of the linkage system.

7.7 Step 6: Determine Database Storage Mechanism

7.7.1 Data Warehouse

The recommended approach to handling a large number of semi-related databases is to set up a data warehouse. Sometimes called a “hub-and-spoke” system, the data warehouse allows databases to be stored separately, but linked when linkage is possible (i.e., when common identifiers are present). Databases are accessed using software (the hub) that can extract from the component databases and link for analysis and reporting purposes. This approach allows individual agencies to keep control over their databases, including storage location, input, editing, and access, while still allowing access and linkage by other permitted individuals. It also allows databases to be brought into the system one at a time, as resources are available.

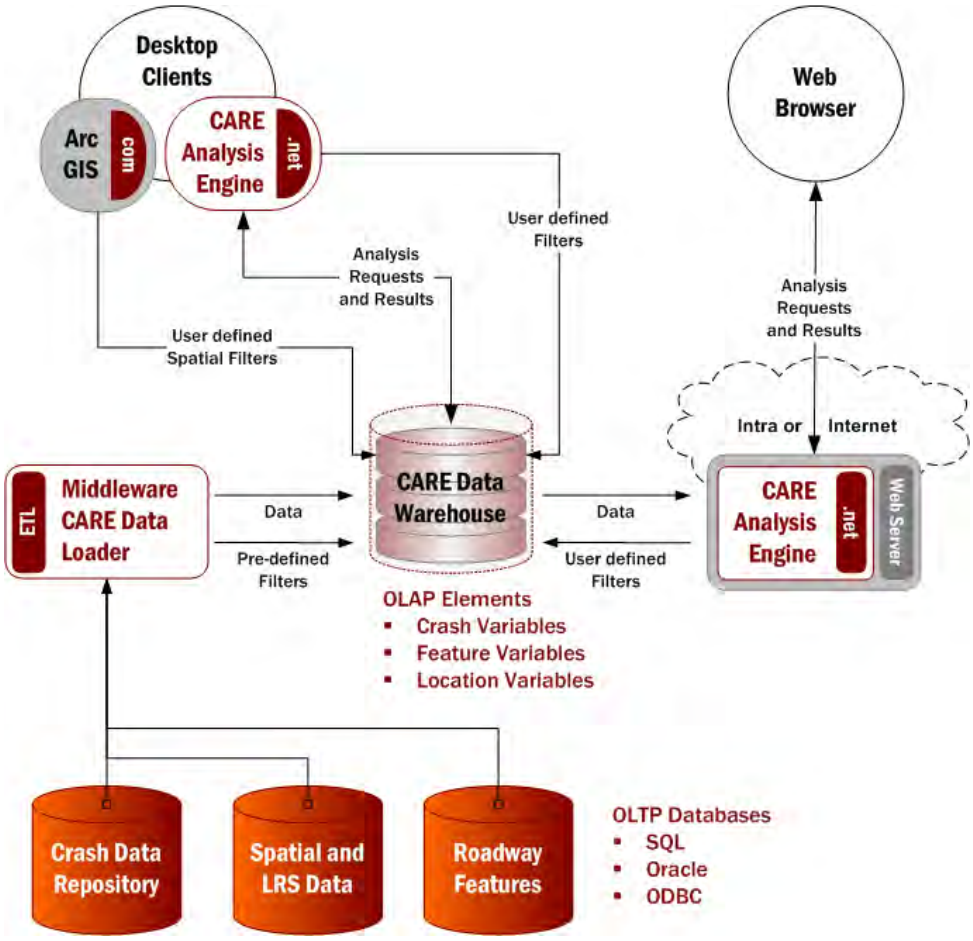

Figure 9. Data warehouse diagram from The University of Alabama Center for Advanced Public Safety's CARE data warehouse (used with permission)

Figure 9 shows an example data warehouse architecture used by the state of Alabama and developed at the University of Alabama Center for Advanced Public Safety. Individual databases represented by the data silo icons (e.g., crash data, linear referencing system (LRS) data, and roadway features data) reside in various locations and formats, generally original to the collecting agency. This allows the originating agency to retain control over the contents and format of the database, and allows them to continue to use existing software for access to that specific database.

The Extract-Translate-Load (ETL) middleware is a dataset-specific translator that changes the format of the original dataset into one that is standardized for use by the CARE data analysis engine. The ETL system may also filter the original dataset, providing a subset of data that is translated. A reasonable goal is to filter as little as possible at the middleware phase.

The analysis engine can be used by web-based software or desktop clients. The client software represents the user's experience and should be designed to facilitate selection of key variables from linkable datasets, filtering of information that is not needed, and development of necessary analyses and reports. There are many commercial software packages available for this purpose, or it can be custom-developed.

The key to using the data warehouse for linkage is in the ETL middleware (and the presence of linking variables). Developing the ETL program for each database is time-consuming and requires an understanding of the standards required by the analysis engine, the contents of the database itself, and the potential uses to be made of it via the analysis engine. For example, all common variable types (e.g., time, date, sex) must have a standard format in all datasets used by the analysis engine (e.g., all time variables must be in 24-hour format, all dates in Julian, and sex must be coded as 1=Male, 2=Female). Knowledge of the needs of the analyst might prompt a recoding of variables (e.g., "Driver Age" is a characteristic of a vehicle that is usually in the person table and not the driver table originally; "EMS run time" might originally only be available by subtracting "Run Start Time" from "Run End Time," but can be computed by the ETL automatically).

Fortunately, this process can be completed one database at a time, and once done, it will not need to be repeated in its entirety. In other words, one-time resources can be used to pull databases such as EMS and trauma into the data warehouse with the expectation that there will not be significant ongoing costs. (Some resources should be set aside for minor updates to the ETL program as new needs are identified by analysts and new variables become available in the original database.)

7.7.2 Separate Linked Dataset

The primary alternative to the more comprehensive data warehouse approach is to handle the linkage separately. In this approach, datasets are linked separately and either the linked dataset is saved on its own or variables are pulled from one dataset into the other permanently.

The state of Washington ran a pilot project to investigate linkage from trauma registry and EMS to crash datasets using probabilistic linkage. The resulting linked dataset includes both medical outcome and crash data. By state law, crash data in Washington are public, but trauma registry data are protected by HIPAA. As a result, the linked variables could not be returned to the original crash dataset. Instead, Washington produced a separate, linked dataset and requires IRB permission to access it. This solution took two years to sort out because of the ambiguous status of a linked dataset with partially public and partially private information.

The advantage of separating the linked dataset from its origins is that it can be done without the overhead of incorporating the component datasets into a data warehouse. In addition, the unique permission issues of the linked dataset can be handled separately. One disadvantage is that changes to the component datasets are not automatically reflected in the linked dataset. In addition, some organization must take responsibility for the linked dataset, even though it is of interest to multiple agencies. Each new dataset must have rules and a process for access, whereas the data warehouse approach centralizes access control (even though access to specific datasets and data elements may still be granted by different agencies).

The single linked dataset approach should be thought of as an intermediate step on the way to incorporation in a data warehouse. Working out linkage issues on a static dataset is helpful as well, since linkage issues can be separated from issues introduced by component dataset updates. However, in the long run, the data warehouse model makes the most sense, given that data linkage is increasingly desirable across a wide variety of state databases (not just crash-EMS-hospital).

7.8 Step 7: Harmonize Common Data Elements

Before linking datasets, it is necessary to have common data elements harmonized. This means that variables are in the same format, and that numeric codes mean the same thing.

An advantage of the data warehouse model is that it provides a partial schema for many common variables (e.g., date formats, sex, age, location, etc.). In addition, it allows the original databases to remain in their original formats, while ensuring that datasets to be linked have common formats.

In selecting harmonized schemas for common data elements, it is important to use an established national schema (e.g., NEMSIS, MMUCC) wherever possible. Using an established schema will enable future linkages to databases with harmonized schemas (e.g., NEMSIS to NTDB). The state also benefits from any existing work done using those schemas, including existing XML, training, manuals, etc.

7.9 Step 8: Set Up a Pilot Project

Almost universally, states that have set up or are setting up linkage programs, reported doing so on a small scale first. Pilot projects may be focused on the logistics of passing identifiers between agencies at the scene, or they may be focused on data issues (e.g., probabilistic linkage, schema standardization across databases).

A good pilot test of on-scene logistics is typically limited geographically. This reduces the number of agencies involved and aids communication, training, and feedback. Logistical problems are the focus of this type of pilot. Missed identifiers, related activities that interfere with patient care or take significant extra time, and confusing processes are among the issues that can be found in trying something out on a small scale. In Florida, the Traffic-Related Injury Prevention task force has proposed a pilot test of passing an identifier at the scene between EMS and police in Orlando for 2014. In Alabama, one EMS service, the fire department, and the department of community health, all in Tuscaloosa, are setting up a pilot project to pass a person-specific identifier between EMS and trauma. Trauma bands are among the options being considered.

A pilot exploration of probabilistic or hand linkage might proceed on a small number of years of data or a small geographic area. Pilot tests can be useful for gathering several types of information. First, a pilot test of probabilistic linkage will identify problems with data structures and data quality that affect all types of linkage, and it can help to prioritize data improvements. Second, a pilot should identify problems with linkage success, both in overall rate and bias. This should feed back to the data collection system by determining additional linkage variables (e.g., name) that should be included to improve the linkage success rate and reduce bias. In Washington, for example, probabilistic linkage was tested on three years of trauma registry and crash data.

A geographically limited pilot is best if statewide datasets are not all in good condition. This is the case in Michigan, where the data linkage committee recommended a pilot project of probabilistic linkage among trauma and crash datasets in one county. However, linkage quality is affected by dataset size, and in particular, the discrimination performance of some variables will change with dataset size (Cook et al., 2001). Linkage testing may start on a smaller dataset, but will need to expand to the statewide dataset before the process is put in place.

For all linkage processes, pilot testing can help to estimate true costs of implementing the linkage program. This is critical for states to be able to plan. However, it is important to account for the likelihood that many costs will go down over time as linkage processes are put in place. For example, initial training and software costs may not repeat often as the process progresses.

7.10 Step 9 (Optional): Set Up a Sampling Program

In the second interim report from this project (Flannagan & Rupp, 2013), we recommended sampling medical outcome data associated with crashes at the state level as an intermediate solution to measuring serious injuries in crashes. In addition to allowing for relatively near-term measurement of serious injuries, sampling also has major advantages in monitoring the progress and success of data linkage approaches. We list this as optional because it is not technically required to complete the process of data linkage, but it is a valuable tool both for achieving the goal of measuring serious injuries early on and for evaluating the quality of linkages as they are developed.

Many, if not most, linkage processes have the potential to bias results, at least in the early phases. Probabilistic linkage, which is the approach most commonly used in states at this time, has the greatest potential for bias. However, even on-the-ground identifier passing has potential for bias. For example, in crashes with multiple victims, passing person-specific identifiers is harder and more likely to fail. This means that single-vehicle crashes are more likely to link, resulting in a bias in the linked dataset towards outcomes typical of single-vehicle crashes. Sampling proceeds independent of these logistical issues, because the logistical problems of sampling revolve around the challenges of getting data from hospitals rather than identifying unique individuals in crashes.

Having a dataset available from a small state sampling program allows for year-to-year assessment of the quality of the linkage process and areas for improvement. In addition, sampling allows serious injury to be measured throughout the development period. Finally, sampling can make use of any source of outcome data without the presence of a statewide database. Thus, sampling may continue to be useful in capturing information about datasets (e.g., ED) that are not yet well covered at the state level, even when other outcome databases are being successfully linked.

7.11 Step 10: Set Up Statewide Linkage

Once pilot linkage is working and the details of the system have been decided, the next step is to set up the linkage process on a statewide basis. Statewide implementation needs to address a number of issues for the long run:

- Who pays for ongoing costs associated with the linkage process?
- Development of training materials for anyone involved in implementation (even probabilistic linkage)
- How is the linkage stored? Numeric codes added to databases? Separate linked dataset?
- Who has access and what is the process for access?
- Establishing a regular process for evaluation of the quality and coverage of the resulting linked database

8 Discussion

The roadmap described in the previous section is divided into steps, which can be thought of as decision points as well. However, the exact path can be different in every state. This section contains additional information and comments on the process that do not fit exactly into the steps in the roadmap, but could be useful in carrying them out.

8.1 Making Progress in Parallel

First, even though steps are numbered, some can be done in parallel. In particular, developing and setting up a data warehouse is generally independent of determining how to link databases. For example, a linked dataset could be stored separately until the component databases are added to a data warehouse. Rows that are linked probabilistically can be assigned permanent numeric identifiers in the component databases sometime in the future. Data harmonization can be done at the same time that identifiers are being selected.

Probabilistic linkage can even be used while on-scene identifiers are being incorporated into the data collection process. A good system for incorporating person-specific numeric identifiers is less subject to bias than probabilistic linkage. However, putting a process in place and establishing its use across all state agencies is time-consuming, and probabilistic linkage can be used in the interim to improve measurement of serious injuries.

Finally, a legal review related to linkage may be time-consuming. This process should start early and can proceed in parallel with the technical development.

8.2 Benefits of a National Standardized Schema and National Datasets

Among those we interviewed who were associated with EMS data management and linkage to crash or trauma, a repeated theme was the benefit of having the national NEMSIS database schema for EMS data. Although states have many challenges in getting all EMS units, including volunteer units, to provide data, all vendors that provide software for EMS data collection output data according to the same schema.

The Highway Performance Monitoring System is a national dataset that provides data on characteristics of the nation’s highways. States submit data, which are standardized at a national level, and the data are compiled at the national level. This system is analogous to NEMSIS in that the schema is standardized, but the coverage of roads and data elements are more limited in the national sample compared to state roadway datasets.

Trauma registries have the NTDS to provide a common structure for their datasets. States are encouraged to move their crash datasets towards the newest MMUCC standard, though compliance is less consistent without a national dataset or XML schema (though one is being developed).

The presence of a national standard, encoded in XML, and a national database tends to speed up the process of standardization of datasets in an area. This, in turn, makes change more efficient because states can share developments. Crash datasets and data linkage to crash would benefit from the kind of coordination that a national dataset and standardized XML schema represents.

Another benefit of national schemas and datasets in all domains is the ability to share across state borders. Smaller states and those with large cities on the border tend to have many crash cases involving drivers from out of state (for linkage to license files) or occupants who are treated in hospitals across state lines. There is no simple within-state mechanism to handle these linkages.

A national schema would mean that if states share data, out-of-state data will look like in-state data and be relatively easy to use. A national dataset would mean that data are available to search for matching records without being arbitrarily restricted by borders. In the meantime, for states that have significant border-related dropout in datasets, the best option is probably to make arrangements with neighboring states to find matching cases. The process for these cases will be different from any in-state system (probabilistic or otherwise), so some level of hand linkage may be necessary.

8.3 Motivators

Another theme of interviews was the benefit of having statutes requiring data reporting. These are state requirements, but it provides the state agency in charge of managing the dataset with some leverage to encourage data reporting. For the most part, these statutes are not enforced, but they do provide motivation.

Interviewees stressed the need to give information back to the entities providing data. This means emphasizing report generation, not only to meet national requirements, but also to give feedback to the individual organizations that provide data to the state. Timely reporting and timely data upload help those groups benefit from the work they put into the data systems.

At the entity level, the motivation for promoting data linkage varies. For departments of transportation (DOTs), MAP-21 requirements motivate the need for crash data to be linked to medical outcome. However, EMS and trauma registries do not generally need crash characteristics for their analyses. For trauma registries, incident location and restraint use are often of greatest interest. In New Jersey, a data linkage project was initiated by a trauma center to capture the incident locations of their crash-related trauma cases. They were interested in targeting educational programs in specific neighborhood schools and centers where certain types of crashes (e.g., pedestrian crashes) are more prevalent. Similarly, knowing when and where crashes happen can help EMS and hospitals better predict future events for resource preparation. There are a number of potential benefits of linking these datasets for all agencies concerned, but it is important to identify and promote these benefits so that agencies are motivated to participate and facilitate.

8.4 Where We Are Now

At this point, state roadway, driver license, driver history, crash, EMS, and trauma databases are either in good condition or are being improved in nearly all states. Statewide hospital and ED databases have more variation in coverage and consistency. The exact condition of a state's databases matters for near-term progress on linkage, but states are generally moving forward on this front (database completeness and quality).

Linkage between EMS and trauma databases is being done in a number of states and tried in others. Linkage between crash and EMS or crash and trauma is planned in many states, but no state that we talked to or surveyed has a fully implemented crash-EMS or crash-hospital linkage that is not probabilistic. The CODES program initiated probabilistic linkage programs in a number of states, and these are being maintained in many cases. Moreover, software developed through CODES, as well as software developed with funds from the CDC are available and help handle some of the technical challenges associated with probabilistic linkage.

In principle, probabilistic linkage is not the preferred approach. However, if sufficiently unique identifiers, such as name, are used, probabilistic linkage comes close to linkage using numeric identifiers in performance. In an effort to avoid the problems of MI, states often select

only high-probability linkages in their resulting database. However, this introduces bias in analysis that will influence results of performance metrics.

Given the strong emphasis in states on probabilistic linkage, it is important to provide states with clear guidance on how to use the approach, how to evaluate the results, and what (if any) versions of a probabilistic linkage system can be considered a final solution to linkage. Otherwise, probabilistic linkage should be considered an intermediate stage on the way to a better approach.

Several states are planning and beginning to implement pilot programs to test crash-to-EMS linkage processes that are not probabilistic. As these develop, it will be helpful to learn from states' different approaches so that a better-tested "toolkit" of approaches is available. Several states have successfully implemented processes to link EMS and hospital data, and these approaches have been described in this document. For the most part, these approaches can be adapted to the crash-EMS linkage process, and indeed, those states are generally considering the possibility.

9 Recommendations

Because data linkage must ultimately occur at the state level, each state will have to work out its state-specific process. In addition, states should have some flexibility in how to implement data linkage while still maintaining comparability of statistics across states. Although many issues that must be addressed, including channels of communication, state law, and agency relationships, are state-specific, many of the problems of data linkage are common across states. In particular, technical challenges follow common themes across states, and technical solutions can be widely applicable. In the context of a national program and a project meant to address this problem across the country, our recommendations focus on steps that can benefit all states and avoid having each state "reinvent the wheel" in its effort to develop a linkage system.

In principle, none of these activities is on the critical path to linkage within states. Many states, particularly those associated with the CODES program, are successfully linking crash, medical outcome, and a variety of other datasets already. However, as the CODES program demonstrated, access to technical assistance and centralized problem-solving facilitates progress. We recommend the following:

A. Many aspects of the linkage process are state-specific and will need to be handled within each state accordingly. However, all states could benefit from a source of technical assistance at the national level. We recommend a broad technical support program that is dedicated to promoting data linkage at the state level through a variety of means, including, but not limited to probabilistic linkage. Specifically, we suggest the following areas of centralized support:

    1. Development of a national crash data schema and corresponding XML, based on MMUCC, which would provide the same benefits that NEMSIS has provided to the EMS community. In particular, such a schema should be designed to incorporate MMUCC, additional state-specific variables, and to facilitate linkage to NEMSIS and NTDB schemas.
    2. Development of clear methods and criteria for testing quality of linkage systems (probabilistic or otherwise). Levels of linkage quality (in terms of bias, accuracy, and completeness) should also be associated with guidance in how to analyze the data and how to improve linkage quality.
    3. Development of a repository for lessons learned, methods used (including those tried and rejected), and contacts in states that can provide advice. This should include (but not be limited to): a) Lists of variables states use for probabilistic linkage (if appropriate) and linkage success; b) Software available and algorithms used for probabilistic linkage, along with the pros and cons of each; c) Non-probabilistic linkage approaches successes and failures; d) Background on the data warehouse model and how to build one over time; e) Lists of vendors used by states for different elements of the data linkage process; and f) Contact information for individuals involved in state data linkage projects to provide assistance or advice.
    4. Development of marketing materials that TRCCs can use to advertise the benefits of linkage to all groups that need to be involved. Coordination of a message at the national level would be helpful to gain the involvement of agencies that are not as used to working together (e.g., state health agencies and state DOTs).

5. Development and hosting of workshops for state data holders to learn about linkage approaches and discuss challenges with other states.

B. We also recommend some additional work that could be done either as part of the work of a national technical assistance program or as separate, smaller efforts:

1. Generate a clear, written interpretation of HIPAA in the context of data linkage that defines clearly what mechanisms must be put in place to link data and still maintain HIPAA compliance. While HIPAA does not prevent data linkage or even including linked (de-identified) data in a state data warehouse, it does put additional security requirements on datasets that include such information.
2. Investigate the potential for vehicle-to-vehicle (V2V) communication to aid in passing identifiers on the scene. This should include assessment of what an application would need to do, potential hurdles in implementation, and estimated short-term (software development) and long-term costs. This project could also investigate the general problem of using event-specific (but not person-specific) identifiers to improve probabilistic linkage among occupants within the event. Such work could be applied to other event-specific linkage approaches (such as passing crash report number to EMS and trauma databases).
3. Develop a more detailed sampling protocol that includes costs of sampling and estimates of sample size needed for a set of target analyses. A pilot sampling project should be included to ensure that logistical challenges and costs are fully identified.

10 References

Barrett M, Steiner C. Healthcare Cost and Utilization Project (HCUP) External Cause of Injury Code (E Code) Evaluation Report (Updated with 2011 HCUP Data). (2014). HCUP Methods Series Report # 2014-01 ONLINE. March 14, 2014. U.S. Agency for Healthcare Research and Quality. Available: http://www.hcup-us.ahrq.gov/reports/methods/methods.jsp.

Bettisworth, C., Hassol, J., Maloney, C., Sheridan, A., and Sloan, S. (2015). Dynamic Mobility Application Policy Analysis: Policy and Institutional Issues for Response, Emergency Staging and Communications, Uniform Management, and Evacuation (R.E.S.C.U.M.E.). USDOT Report No. FHWA-JPO-14-137.

Blincoe, L. J., Seay, A., & Zaloshnja, E. (2002). The economic impact of motor vehicle crashes, 2000 (DOT HS 809 446) Washington, D.C.: US Department of Transportation, National Highway Traffic Safety Administration.

Cambridge Systematics. (2013). Measuring performance among state DOTs: Sharing good practices—serious crash injury. September 2013.

Cochran, W. (1977). *Sampling Techniques*. New York, Wiley and Sons.

Cook, L. J., Olson, L. M., & Dean, J. M. (2001). Probabilistic record linkage: relationships between file sizes, identifiers, and match weights. *Methods of information in medicine*, 40(3), 196-203.

Council, F. M., Harkey, D. L., Carter, D. L., & White, B. (2007). *Model Minimum Inventory of Roadway Elements--MMIRE* (No. FHWA-HRT-07-046).

DeLucia, B.H., Scopatz, R.A. & Lefler, N. (2012). Roadway Data Improvement Program: Informational Resource. Available at http://safety.fhwa.dot.gov/rsdp/downloads /rdip_final061312.pdf.

Department of Transportation. (June, 2008). *MMUCC Guideline Model Minimum Uniform Crash Criteria*, 3rd ed. DOT HS 810 957.

Department of Transportation. (June 2012). *MMUCC Guideline Model Minimum Uniform Crash Criteria*, 4th ed. DOT HS 811 631.

Federal Highway Administration (2014). National Performance Management Measures; Highway Safety Improvement Program, Notice of Proposed Rulemaking. Docket No. FHWA–2013–0020.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

Flannagan, C., Elliott, M., Mann, N., & Rupp, J. (2014). Sampling Serious Injuries in Traffic Crashes at the State Level. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2432, pp. 118–123.

Flannagan, C., Mann, N.C., and Rupp, J. (2012). Interim report 1. National Cooperative Highway Research Program Project 17-57.

Flannagan, C., and Rupp, J. (2013). Interim report 2. National Cooperative Highway Research Program Project 17-57.

Gennarelli, T.A., and Wodzin, E. (2005). Abbreviated injury scale 2005. Association for the Advancement of Automotive Medicine.

Glance L, Osler T, et al. (2009). TMPM–ICD9: A Trauma Mortality Prediction Model Based on ICD-9-CM Codes. *Annals of Surgery*, 249 (6), 1032-1039.

Governor's Highway Safety Association (GHSA). (2014). Traffic Records. Available online: http://www.ghsa.org/html/issues/traffrec.html.

Green, P., and Blower, D. (2010). A new model of crash severities reportable to the MCMIS crash file.

Haider et al. (2012) Should the IDC-9 Trauma Mortality Prediction Model become the new paradigm for benchmarking trauma outcomes? *J Trauma Acute Care Surg,* 72 (6): 1695-1701.

Injury Surveillance Workgroup. (2003). Consensus Recommendations for Using Hospital Discharge Data for Injury Surveillance. Marietta (GA): State and Territorial Injury Prevention Directors Association.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in medicine*, 14(5-7), 491-498.

Kish, Leslie. (1965). *Survey sampling*. New York: J. Wiley & Sons.

Lawrence, B. A., Miller, T. R., Weiss, H. B., & Spicer, R. S. (2007). Issues in using state hospital discharge data in injury control research and surveillance. *Accident Analysis & Prevention*, 39(2), 319-325.

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

National Highway Traffic Safety Administration. (2013). Fatality Analysis Reporting System (FARS) Encyclopedia. www.fars.nhtsa.dot.gov.Main/index.aspx (accessed April 8, 2012).

Newgard, C., Malveau, S., Staudenmayer, K., Wang, N. E., Hsia, R. Y., Mann, N. C., ... & Cook, L. J. (2012). Evaluating the use of existing data sources, probabilistic linkage, and multiple imputation to build population-based injury databases across phases of trauma care. *Academic emergency medicine*, 19(4), 469-480.

Tarko, A., Bar-Gera, H., Thomaz, J., & Issariyanukula, A. (2010). Model-Based Application of Abbreviated Injury Scale to Police-Reported Crash Injuries. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2148, pp. 59–68.

World Health Organization. (1992). International statistical classification of diseases and related health problems (Tenth revision). Geneva, World Health Organization.

Zonfrillo, M., Weaver, A., Gillich, P., Price, J. & Stitzel, J. (2015) New Methodology for an Expert-Designed Map From International Classification of Diseases (ICD) to Abbreviated Injury Scale (AIS) 3+ Severity Injury, *Traffic Injury Prevention*, 16: sup2, S197-S200, DOI: 10.1080/15389588.2015.1054987

11 Appendix A Serious Injury Definitions

Survey State	Definition of serious injury
Alabama	The following defines the injury codes that are applied according to the judgment of the reporting officer: / / 1 Fatal. This code will be entered if a victim is pronounced dead at the scene or before the report is completed. If not, one of the other codes will apply. However, if a victim dies later as a result of the crash this code will need to be updated according to the following di-rections. The Department of Public Safety uses a 30 day counting period for traffic fatalities. If a person dies as a result of injuries received in a traffic crash within 30 days of the date of the crash, that victim is considered to be a traffic fatality, and the victim injury type must be updated to Code 1 in this data item. When it is learned that a victim has died after the crash report has been sent to the Department of Public Safety: (1) Call the FARS representative with this information at 334-242-4427 AND (2) Follow the normal amendment procedure to amend the eCrash given in Item 1.1.1. / 2 Incapacitating. This means that the victim must be carried or otherwise helped from the scene. If the victim needs no help, then either a code 3 or 4 applies even though medical assistance may have been administered at the scene. / 3 Non-incapacitating. If the victim has visible signs of injury, either in a physical or mental sense (e.g., had passed out), but is judged able to walk away from the scene without help, this code applies. The difference between this code and code 4 is strictly in the external evidence of injury. / 4. Not visible but complains of pain. If the victim complains of pain, but there are no visible signs of it, and he or she is able to walk away from the scene of the crash, then this code applies. There is no code for uninjured, in that uninjured occupants are not to be considered in the victim section. There are no codes allowed for 97 or 98 since if a victim is identified some assessment must be made of the severity of the injury according to the classifications given above.
Alaska	Suspected Serious Injury is an injury other than fatal which results in one or more of the following: / * Severe laceration resulting in exposure of underlying tissues/muscle/organs or resulting in significant loss of blood / * Broken or distorted extremity (arm or leg) / * Crush injuries / * Suspected skull, chest or abdominal injury other than bruises or minor lacerations / * Significant burns (second and third degree burns over 10% or more of the body) / * Unconsciousness when taken from the crash scene / * Paralysis
Arizona	Incapacitating (Serious) Injury - Any injury, other than a fatal injury, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred. Often defined as "needing help from the scene." Includes: severe lacerations, broken or distorted limbs, skull or chest injuries, abdominal injuries, unconsciousness when taken from the accident scene.
Arkansas	(the defiition depends on the object of the exercise - however, the injury definitions are below) / / INJURY SEVERITY LEVELS / / / Fatal Injury (code 1) / Any injury that directly results in the death of a living person within 30 days of a MVC. / / Incapacitating Injury (code 2) / Any injury, other than a fatal injury, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred. / Inclusions: / - Severe lacerations / - Broken or distorted limbs / - Skull or chest injuries / - Abdominal injuries / - Unconsciousness at or when taken from the scene / - Unable to leave the scene without assistance / - And others /

Survey State	Definition of serious injury
	Exclusions: / - Momentary unconsciousness / - And others / / Non-Incapacitating Injury (code 3) – / Any injury other than a fatal injury or an incapacitating injury, which is evident to observers at the scene. / Inclusions: / - Lump on head / - Abrasions / - Bruises / - Minor lacerations / - And others / Exclusions: / - Limping (the injury cannot be seen) / - And others /
California	Severe Injury: an injury includes: broken or fractured bones; dislocated or distorted limbs; severe lacerations; skull, spinal, chest or abdominal injuries that go beyond “Other Visible Injuries”; unconsciousness at or when taken from the collision scene; and severe burns. Other Visible Injures: an injury includes: bruises, discoloration, or swelling; minor lacerations or abrasions; minor burns. Complaint of Pain: an injury includes: persons who seem dazed, confused, or incoherent (unless such behavior can be attributed to intoxication, extreme age, illness, or mental infirmities); persons who are limping, or complaining of pain or nausea; any person who may have been unconscious.
Colorado	We classify ALL Injuries according to our crash report, DR 2447: / / 01: Complaint of Injury / 02: Evident - Non-incapacitating / 03: Evident - incapacitating /
Delaware	Incapacitating injury per the MMUCC definition
Florida	
Hawaii	Hawaii doesn't have any attributes in our Motor Vehicle Accident Report for serious injury. Under "Injury Class" on the "All Persons" page, the Injury Classes are: / None / Possible (any non-visible injury reported or claimed that is not fatal, incapacitating or non-incapacitating injury) / Non-Incapacitating (any evident injury, other than fata or incapacitating) / Incapacitating (any injury, other than fatal, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing prior to the accident) / Fatal (an injury received at the scene of the accident that results in death during or after the accident) / Unknown
Idaho	All of the injury definitions in Idaho are straight from ANSI D-16: / / 2.3.4 incapacitating injury: An incapacitating / injury is any injury, other than a fatal injury, which / prevents the injured person from walking, driving / or normally continuing the activities the person was / capable of performing before the injury occurred. / Inclusions: / Severe lacerations / Broken or distorted limbs / Skull or chest injuries / Abdominal injuries / Unconsciousness at or when taken from / the accident scene / — Unable to leave the accident scene without / assistance / — And others / Exclusions: / — Momentary unconsciousness / — And others
Illinois	The serious injury is defined as "A" or "Incapacitated injury" code based on the following severity of injury categories (KABCO) / / K Fatal Injury / A Incapacitating Injury / B Non-Incapacitating Injury / C Reported, not evident / 0 No Indication of Injury /
Iowa	Code 2 Incapacitating Any injury other than a fatal injury which prevents the injured person from walking, driving, or normally continuing the activities the person was capable of performing before the injury occurred. Inclusions: severe lacerations; broken or distorted limbs; skull, chest, or abdominal injuries; unconsciousness; unable to leave the crash scene without assistance.

Survey State	Definition of serious injury
Kansas	As applied by officers in the field, it is very subjective. Our coding manual gives the following instructions: "The person's Injury severity should be listed as the reporting officer observes it to be at the time of the accident: Possible Injury, Non-incapacitating Injury, or disabling injury. If there is uncertainty as to which injury severity code to use, choose P (Possible)." It is defined as: "D - disabling injury (incapacitating): A Disabling injury is any injury, other than a fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities he/she was capable of performing before the injury occurred. Includes severe lacerations, broken or distorted limbs, skull or chest injuries, abdominal injuries, unconsciousness at or when taken from the accident scene, or inability to leave the accident scene without assistance."
Kentucky	Not sure
Louisiana	Incapacitating/Severe, Non-Incapacitating/Moderate
Maine	Title 17-A S. 2 defines serious bodily injury as a bodily injury which creates a substantial risk of death or which causes serious, permanent disfigurement or loos or substantial impairment of the function of any bodily member or organ, or extended convalescence necessary for recovery of physical health. / / Maine's PAR collects injury degree: fatal, incapacitating, non-incapacitating, possible injury and no injury.
Maryland	Maryland uses a 5 point scale. 1 - not injured, 2 - possible injury, 3 - injured, 4 - disabled, 5 - fatal. We define serious injury as 4 (disabled, incapacitating). An incapacitating injury is an injury, other than a fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities he was capable of performing before the injury occurred.
Massachusetts	Fatal, incapacitating, non-incapacitating, possible, no injury, unknown
Michigan	Incapacitating Injury is a synonym for serious injury in Michigan. Injury is any injury, other than fatal, that prevents the injured person from walking, driving, or normally continuing the activities which he or she was capable of performing prior to the motor vehicle traffic crash.
Minnesota	Incapacitating Injury ("A") / An incapacitating injury is an injury, other than a fatal injury, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred. / / Inclusions: Severe lacerations, broken or distorted limbs, skull or chest injuries, abdominal injuries, unconsciousness at or when taken from the accident scene, unable to leave the accident scene without assistance Exclusions: momentary unconsciousness//
Mississippi	We use KABCD where K=Killed, A=Life Threatening, B=Moderate, C=Minor and D=No Injury
Missouri	Missouri's traffic crash report identifies injuries as the following: Fatal, Disabling, Evident-Not Disabling, Probable-Not Apparent, and None Apparent. / Fatal is defined as, "The person was dead or dies within 30 days of the crash from crash-related injuries. / Disabling - The person sustained non-fatal injuries that prevent walking, driving, or continuing activities the person was capable of performing prior to the crash.

Survey State	Definition of serious injury
	Transport by ambulance from the scene does not necessarily indicate the individual sustained disabling injuries. / Evident-Not Disabling - The person sustained visible injuries that were neither fatal nor disabling.
Montana	A serious injury is an incapacitating injury or any injury, other than a fatal injury, which prevents the injured person from walking; driving or normally continuing the activities the person was capable of performing.
Nebraska	Serious injury severity coded as 2 disabling injury - cannot leave scene without assistance (broken bones, severe cuts, prolonged unconsciousness, etc.)
Nevada	Incapacitating Injury or A-injury. Any injury visible, or diagnosed by a physician, that prevents the injured / party from walking, driving, or normally continuing the activities that he/she was capable of / performing prior to the accident. Severe laceration, broken or distorted limbs, unconscious when taken from the accident scene; unable to leave accident scene without assistance.
New Mexico	K Killed A Incapacitated - carried form the scene B Visible Injury C Complaint of Injury - but not visible O No Apparent Injury
New York	Our reports use the KABCO severity score. We typically use K and A as serious injury.
North Carolina	A-injury type (disabling) - Injury obviously serious enough to prevent the person injured from performing his normal activities for at least one day beyond the day of the collision. Massive loss of blood, broken bone, unconsciousness of more than momentary duration are examples. From NC DMV 349 Police Instruction Manual
North Dakota	Incapacitating Injury - Any injury other than a fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities they were capable of performing before the injury occurred.
Ohio	On the Ohio crash report, an incapacitating injury (serious injury) is defined as such: "An injury, other than a fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities the person was capable of performing before the injury occurred. Often defined as "needing help from the scene."
Oklahoma	Oklahoma classifies injuries as follows: / Fatal ─ any injury that directly results in the death of a living person within 30 days of a MVC; / Incapacitating ─ any injury, other than fatal, which prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred; / Non-incapacitating ─ any injury other than fatal or incapacitating that is evident to observers at the scene.

Survey State	Definition of serious injury
Oregon	Code 2 is used for participants who suffer severe injuries. An incapacitating injury is a non-fatal injury which "prevents the injured person from walking, driving or normally continuing the activities the person was capable of performing before the injury occurred". (see ANSI D16.1-1996, page 10, definition 2.3.4) Examples of incapacitating injuries include broken bones, severe bleeding, unconsciousness, etc.
Pennsylvania	Incapacitating injury, including bleeding wounds and distorted members (amputations or broken bones), and requires transport of the patient from the scene
Puerto Rico	Injury resulting of a car accident which involves lacerations, severe hemorrhage or bone fractures and requires hospitalization.
Rhode Island	State of Rhode Island Uniform Crash Report uses a system that is similar to KABCO for injury status: / / 1. Complains of Pain, 2. Non-Incapacitating, 3. Incapacitating, 4. Fatal, 5. No Injury, 6. Unknown / / The definition of "Incapacitating" is not identified, prompting officers to use their own discretion.
Texas	Incapacitating (A) and Non-Incapacitating (B) level severities
Vermont	We use the term "Major Crashes" relating to fatal and incapacitating injuries involved in a crash. This is a standard terms we are now using in the SHSP and w/in the Vermont Highway Safety Alliance group.
Virginia	Visible signs injury, such as bleeding wound, distorted member or had to be carried from the scene
Washington	Police Traffic Collision Report excerpt: / "Serious injury: Any injury which prevents the injured person from / walking, driving, or continuing normal activities at the time of the / collision."
West Virginia	The WV Uniform Traffic Crash Report Student Manual defines an “Incapacitating Injury” as “Injury severe enough to require individual to be immediately transported from the scene. Injuries include bleeding wounds, distorted members, etc.”
Wisconsin	We use the KABCO injury scale. An 'A' (incapacitating) injury is considered a serious injury. It is defined on the crash report as "Any injury other than a fatal injury which prevents the injured person from walking, driving, or from performing other activities which he/she performed before the accident."
Wyoming	Incapacitating Injury - Any injury, other than a fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities the person was capable of performing before the injury occurred. Often defined as "needing help from the scene." / Includes: severe lacerations, broken or distorted limbs, skull or chest injuries, abdominal injuries, unconsciousness when taken from the accident scene.

12 Appendix B Identifiers for Linkage

Survey State	Identifiers-EMS Patient Care Data	Identifiers-ED Data	Identifiers-Hospital Discharge Data	Identifiers-Trauma Registry Data	Identifiers-Vital Records Data	Identifiers-Roadway Inventory Data
Alabama	Time of crash, crash location.		Patient Identifier Code.	Patient Identifier Code		Milepoint
Alaska	This has yet to be built but we are planning to use the APSIN ID as a unique identifier across databases.		This is currently being done using names and dates on both the crash data and the hospital discharge data.	This has yet to be built but we are planning to use the APSIN ID as a unique identifier across databases		The crash and roadway inventory data are in the same database and are directly linked
Arizona						
Arkansas			Name and other fields.			County, Route, Section, Log Mile
California		probabilistically linked-no common identifying field. Use demographic fields common to both, date and location of collision to increase match success.	probabilistically linked-no common identifying field. Use demographic fields common to both, date and location of collision to increase match success.		probabilistically linked-no common identifying field. Use demographic fields common to both, date and location of collision to increase match success.	

Survey State	Identifiers-EMS Patient Care Data	Identifiers-ED Data	Identifiers-Hospital Discharge Data	Identifiers-Trauma Registry Data	Identifiers-Vital Records Data	Identifiers-Roadway Inventory Data
Colorado					For Fatal Records (Crashes) only!	Direct interface with all Roadway data for exact location (coding) of both, On- and Off-System Crashes statewide!
Delaware	Patient's last name, first name, gender, age, crash date and time, ALS/BLS agency location (latitude and longitude).		Patient's last name, first name gender, age, position (driver, passenger, etc) crash date, hospital location (lat and long).			road name, road number, milepoint, latitude and longitude
Florida	Sex, Date of Birth, Home zip code, Incident Date.					GPS location, intersection information
Hawaii						milepost, intersections, GPS
Idaho	Name, Date Of Birth, Gender, Date of Incident, Hospital Transported to.					

Survey State	Identifiers-EMS Patient Care Data	Identifiers-ED Data	Identifiers-Hospital Discharge Data	Identifiers-Trauma Registry Data	Identifiers-Vital Records Data	Identifiers-Roadway Inventory Data
Illinois		age, gender, date of birth, date of birth, county, city, date of crash, date of admission.	age, gender, date of birth, date of birth, county, city, date of crash, date of admission.	age, gender, date of birth, date of birth, county, city, date of crash, date of admission		Location Codes
Iowa	A variety per CODES.			A variety per CODES.		Direct key field match.
Kansas	Indirect link with FARS database.					On Road/At Road information
Kentucky		Admit date; occupant date of birth, gender, and resident zip code; distance from crash location to hospital location; vehicle type; person type; crash type.	Admit date; occupant date of birth, gender, and resident zip code; distance from crash location to hospital location; vehicle type; person type; crash type.			I don't know
Louisiana	Not currently being performed, but we plan to do so and are looking into.	Not currently being performed, but we plan to do so and are looking into.	Not currently being performed, but we plan to do so and are looking into.	Not currently being performed, but we plan to do so and are looking into.	Not currently being performed, but we plan to do so and are looking into.	Lat/Long, street names, primary roadway, secondary roadway, distance, direction, milepoint
Maine			indirect.			Unknown

Survey State	Identifiers-EMS Patient Care Data	Identifiers-ED Data	Identifiers-Hospital Discharge Data	Identifiers-Trauma Registry Data	Identifiers-Vital Records Data	Identifiers-Roadway Inventory Data
Maryland	Time of day, day of week, county of crash, age, gender, mechanism of injury, person type (driver, passenger, pedestrian).	Time of day, day of week, county of crash, age/dob, gender, mechanism of injury, person type (driver, passenger, pedestrian).	Time of day, day of week, county of crash, age/dob, gender, mechanism of injury, person type (driver, passenger, pedestrian).	Time of day, day of week, county of crash, age/dob, gender, mechanism of injury, person type (driver, passenger, pedestrian)		Integrated through use of ArcGIS, mapping of crashes and roadways in the state
Massachusetts	Vehicular Injury Indicators, Area of the Vehicle impacted by the collision, Seat Row Location of Patient in Vehicle, Position of Patient in the Seat of the Vehicle, Use of Occupant Safety Equipment, Airbag Deployment, Barriers to	Encrypted SSN, Gender, DOB, Ecode, Org ID, Date, Time, Diagnosis, Status, Patient City, Patient Zip.	Encrypted SSN, Gender, DOB, Ecode, Org ID, Date, Time, Diagnosis, Status, Patient City, Patient Zip.	Protective Devices/Child Specific Restraint, Airbag Deployment, Alcohol Use/Drug Use Indicators, Diagnosis Code, Transport Mode, Injury Incident Date/Injury Incident Time, SiteOrgID, ED/Hospital Admission Time, Patient Street Address, Patient City, Patien	Encrypted SSN, Address, City, State, Date of Death, DOB, Cause of injury, Location of injury, Diagnosis	Street Name, City, Mile point, Route, Bridges, Speed Limit, Highway District, RPA (Regional Planning agency)

Survey State	Identifiers-EMS Patient Care Data	Identifiers-ED Data	Identifiers-Hospital Discharge Data	Identifiers-Trauma Registry Data	Identifiers-Vital Records Data	Identifiers-Roadway Inventory Data
	Patient Care, Alcohol/Dru.					
Michigan	Not yet linked.	Not yet linked.	Not yet linked.	Not yet linked		Not yet linked
Minnesota	Last name, first name, gender, DOB, crash date, location.	Hospital zip, home zip, injured (y/n), fatal (y/n), gender, DOB, admit hour, crash date, position, age, vehicle type.	Hospital zip, home zip, injured (y/n), fatal (y/n), gender, DOB, admit hour, crash date, position, age, vehicle type.	Last name, first name, hospital zip, home zip, injured (y/n), fatal (y/n), gender, DOB, admit hour, crash date, position, age, vehicle type	Last name, first name, hospital zip, home zip, injured (y/n), fatal (y/n), gender, DOB, admit hour, crash date, position, age, vehicle type	Route and reference point, roadway system type, county/city identifier
Mississippi	EMS agency and hospital in crash data base, as well as person name.	Person name		Person name	Person name	Linked with Crash by GPS Location
Missouri	Crash number, crash date, county, name, DOB.		Crash number, crash date, county, name, DOB.		Crash number, crash date, county, name, DOB	Crash number, location information, county, crash date
Montana						
Nebraska	Patient first name, last name, gender,	Patient date of birth, age, gender, occurrence date,	Patient date of birth, age, gender, occurrence date,		Patient first name, last name, gender, date of	

Survey State	Identifiers-EMS Patient Care Data	Identifiers-ED Data	Identifiers-Hospital Discharge Data	Identifiers-Trauma Registry Data	Identifiers-Vital Records Data	Identifiers-Roadway Inventory Data
	Date of birth, and occurrence date.	And occurrence county.	And occurrence county.		Birth, date of death, and occurrence county	
Nevada				Patient name, birthdate, incident date		Route ID and/or Route Full Name
New Mexico						
New York	Age, agency county, birth day, birth month, birth year, first two characters of first name, first two characters of last name, last two characters of last name, last four of social security, hour of call, injury flag, run date, sex, call location, call ty.	Facility county, first two characters of first name, first two characters of last name, last two characters of last name, last four of social security, state, zip code, role, age, date, date of birth day, birth month, sex, externalExternal-cause code.	Facility county, first two characters of first name, first two characters of last name, last two characters of last name, last four of social security, state, zip code, role, age, date, date of birth day, birth month, sex, externalExternal-cause code.	Unknown. We are in the process of obtaining trauma registry data. It is anticipated that variables such as: admission date, admission time, month, day and year of birth, age, sex, zip code, county, state, injury date, county of injury, place of occurrence	We do not currently have approval for using vital records in the CODES project.	We do not currently have access to Roadway Inventory data for the CODES project.

Survey State	Identifiers-EMS Patient Care Data	Identifiers-ED Data	Identifiers-Hospital Discharge Data	Identifiers-Trauma Registry Data	Identifiers-Vital Records Data	Identifiers-Roadway Inventory Data
North Carolina	date, county, time of day, sex, ethnicity, date of birth, person type, and name (if driver).	date, county, time of day, sex, ethnicity, date of birth, person type, and name (if driver).	date, county, time of day, sex, ethnicity, date of birth, person type, and name (if driver).	date, county, time of day, sex, ethnicity, date of birth, person type, and name (if driver)	name, date of birth, sex, county of residence	date, county, road on, road from, and road to
North Dakota						Roadway Location
Ohio	This data is provided by the Division of EMS, Ohio Department of Public Safety.	This data is provided by the Division of EMS, Ohio Department of Public Safety.	This data is provided by the Ohio Hospital Association.	This data is provided by the Division of EMS, Ohio Department of Public Safety	These records are provided by the Ohio Department of Health	This information is supplied by the Ohio Department of Transportation
Oklahoma	First and last name; DOB; age; sex; last four digits of SSN; incident date; soundex.		First and last name; DOB; age; sex; last four digits of SSN; incident date; soundex.		First and last name; DOB; age; sex; last four digits of SSN; incident date; soundex.	Year; County; Control Section; Subsection; Mile point.
Oregon	The following are fields that were tested during probabilistic linking of the EMS and Crash systems: Call #,					road number and milepoint

Survey State	Identifiers-EMS Patient Care Data	Identifiers-ED Data	Identifiers-Hospital Discharge Data	Identifiers-Trauma Registry Data	Identifiers-Vital Records Data	Identifiers-Roadway Inventory Data
	Incident #, Responding unit #, Unit dispatched (date/time), En route, Arrive at scene, Arrive at patient, Leave scene, Arrive destination, Dispatch.					
Pennsylvania						Currently we can only link state roads. This is done through County, Route, Segment, and Offset. The crash record and roadway record both contain this information.
Puerto Rico						
Rhode Island						We are in the planning stages of implementing a statewide LRS that will link crash data to roadway features/characteristics/etc via a geographic component.
Texas	Unknown	Unknown	Unknown	Unknown	Name	Latitude/Longitude Coordinates of the crash.

Survey State	Identifiers-EMS Patient Care Data	Identifiers-ED Data	Identifiers-Hospital Discharge Data	Identifiers-Trauma Registry Data	Identifiers-Vital Records Data	Identifiers-Roadway Inventory Data
Vermont	Likely names, DOB, EMS run information.					Location (town, route, route location), other?
Virginia	No direct linkage used in Virginia.		No direct linkage used in Virginia.	No direct linkage used in Virginia	No direct linkage used in Virginia	Document number per crash
Washington	Indirect linkage via Trauma Registry.	Indirect linkage via Trauma Registry.	first, middle, and last name, date of birth, date of incident, zip code, and gender	first, middle, and last name, date of birth, date of incident, zip code, and gender	first, middle, and last name, date of birth, date of incident, zip code, and gender	For state routes only: use WA unique SR ID information (LRS)
West Virginia						Location
Wisconsin						
Wyoming						Route and Milepost, Lat/Long