LINCS

Linking Information for Nonfatal Crash Surveillance

A guide for integrating motor vehicle crash data to help keep Americans safe on the road

Centers for Disease Control and Prevention
National Center for Injury Prevention and Control

U.S. Department of Health and Human Services
Centers for Disease Control and Prevention (CDC)
National Center for Injury Prevention and Control (NCIPC)

Note for accessibility: For complete explanations of figures with titles, see Appendix R.

LINCS

LINCS

Linking Information for Nonfatal Crash Surveillance

A guide for integrating motor vehicle crash data to help keep Americans safe on the road

EXECUTIVE SUMMARY

The *Linking Information for Nonfatal Crash Surveillance* (LINCS) Guide is intended to help states start a data linkage program or expand their current program to help prevent motor vehicle crash-related injuries and deaths. The guide discusses the key components of successful linkage programs and details each step in the data linkage process.

Motor vehicle crashes (MVCs) are a leading cause of death for people aged 1-54 years in the United States (U.S.). More than 100 people die in MVCs each day and thousands more are injured. Understanding the risk factors and ways to address them can help prevent MVC-related injuries and deaths and reduce costs.

One method to better understand MVCs is to effectively use existing data sources, such as police, hospital, and emergency medical services (EMS) records. These data sources contain different information and the data sets are generally collected and stored separately. Therefore, linking the data sets together can create a more comprehensive understanding of MVCs by pulling all of the data together into one linked data set. A linked data set will include information about what happened before (e.g., impaired driving), during (e.g., seat belt was being used), and after a crash (e.g., medical outcomes and costs).

The Centers for Disease Control and Prevention's (CDC's) LINCS Guide focuses on establishing and improving linkage programs at the state level, with the following objectives:

- Understand how linked data can be used
- Document challenges and successes in implementing linkage programs
- Explore methods and tools available for data linkage
- Help states to start linking data or to expand and improve current linkage programs

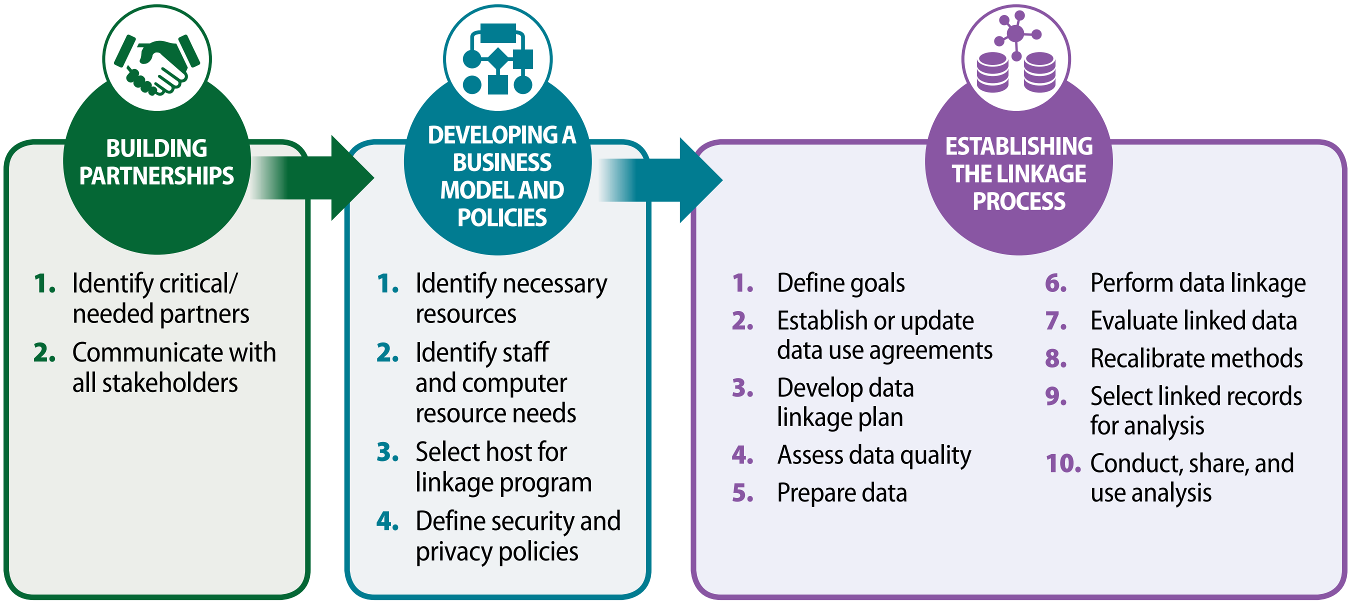
The guide was based on:

- Lessons learned from previous efforts
- Best practices of successful linkage programs
- Updated environmental scans for data linkage research, methods, and tools
- State data linkage pilot efforts

How Do States Start or Enhance a Data Linkage Program?

Establishing a linkage program consists of three major components: building partnerships, developing a business model and policies, and establishing the linkage process as shown in Figure 1.

Figure 1. Components of a Motor Vehicle Crash Data Linkage Program



BUILDING PARTNERSHIPS

1. Identify critical/needed partners
2. Communicate with all stakeholders

DEVELOPING A BUSINESS MODEL AND POLICIES

1. Identify necessary resources
2. Identify staff and computer resource needs
3. Select host for linkage program
4. Define security and privacy policies

ESTABLISHING THE LINKAGE PROCESS

1. Define goals
2. Establish or update data use agreements
3. Develop data linkage plan
4. Assess data quality
5. Prepare data
6. Perform data linkage
7. Evaluate linked data
8. Recalibrate methods
9. Select linked records for analysis
10. Conduct, share, and use analysis

Building Partnerships

Building partnerships and communicating with stakeholders are key to the success and sustainability of a linkage program. Partnerships and coalitions provide a way for organizations to:

- Expand the scope of injury prevention opportunities
- Build credibility
- Share resources
- Leverage knowledge and skills
- Disseminate findings and recommendations
- Raise MVC safety awareness with stakeholders

Developing a Business Model and Policies

There are four essential parts of business model development:

Establish funding. Consistent and sustained funding is essential for long-term success. Partnerships and coalitions can potentially provide channels for new funding and revenue generation.

Identify resources. Successful linkage programs depend on often hard-to-find, skilled staff and technology infrastructure.

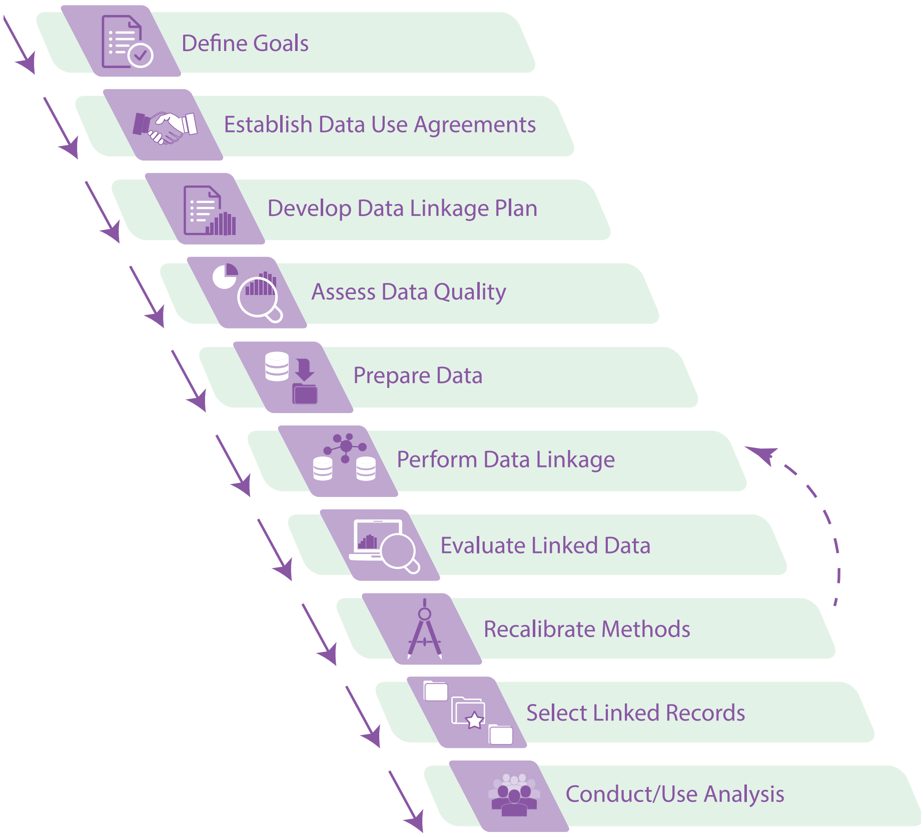
Explore business models. States use different business models to ensure successful and sustainable linkage programs. Deciding where to house the linkage program is an important decision (e.g., at a state agency or trusted third party, such as a university).

Define privacy and security policies. Determine the applicable state and federal privacy laws and consult with the appropriate state offices to leverage existing policies, procedures, and guidance for data protection.

Establishing the Linkage Process

This is the final component of a successful and sustainable program. Figure 2 provides a summary of the 10 steps for establishing the linkage process.

Figure 2. Process for Motor Vehicle Crash Data Linkage



The CDC's National Center for Injury Prevention and Control (NCIPC) enlisted the Centers for Medicare & Medicaid Services (CMS) Alliance to Modernize Healthcare (CAMH)—a federally funded research and development center operated by The MITRE Corporation—to create a guide to help states start or enhance data linkage programs. Linking MVC data sets creates a more comprehensive set of linked data for each MVC incident and for each individual involved in the MVC. Comprehensive MVC linked data can enable analysis of the relationships among contributing factors, interventions, outcomes, and impacts. For example, one advantage of linking police MVC records to hospital records is to assess the magnitude of nonfatal MVC injuries and associated healthcare costs.

CONTENTS

APPENDICES

LIST OF FIGURES

LIST OF TABLES

Page intentionally left blank.

MOTOR VEHICLE CRASHES AND LINCS

INTRODUCTION

Motor vehicle crashes (MVCs) are a leading cause of death for people aged 1-54 years in the United States (U.S.) [1]. In 2017, MVCs accounted for 37,133 deaths [2], and more than 3 million injuries [1]. Furthermore, MVCs are a leading cause of injury-related emergency department visits [3]; the fourth leading cause among all ages in 2017 [1]. Across the globe, the U.S. MVC death rate is twice the average of other high-income countries [4]. Although many studies have been conducted to identify risk factors that contribute to MVCs (e.g., speeding, alcohol use) and to determine the impact of interventions (e.g., seat belts, law enforcement), most of the research has been based on data related to fatal injuries.

Existing national surveillance programs (Appendix A) capture data on all MVC-related deaths, but the same level of coverage is not available for nonfatal MVC-related injuries. With thousands of people injured in MVCs each day, MVCs are a significant public health problem. Understanding MVC risk factors and ways to mitigate them can help prevent deaths and injuries and reduce associated costs. One method to better understand MVCs is to effectively use existing data sources, such as police, hospital, and emergency medical services (EMS) records, in combination for improved surveillance. These data sources contain different information (e.g., risk factors and outcomes) and the data sets are generally collected and stored separately. Therefore, linking the data sets together can create a more comprehensive understanding of MVCs by pulling all of the data together into one linked data set. A linked data set will include information about what happened before (e.g., risk factors such as impaired driving), during (e.g., protective factors such as seat belt use), and after a crash (e.g., medical outcomes and costs).

Linking these data sets helps states explore surveillance-related questions such as:

How many people are nonfatally injured in MVCs? What is the severity of MVC-related injuries?

Linking police and medical data will will ensure medically diagnosed injuries are attributed to crashes.

What risk factors are associated with the most severe/costly injuries?

Police crash reports include risk factor information (e.g. seat belt use); medical data capture medically determined injury severity, diagnoses, and the costs of treatment.

Among drivers who were nonfatally injured in a MVC, what proportion tested positive for alcohol, marijuana, opioids, or other drugs? Are there differences in risk factors or injury severity among those who test positive for substances versus those who test negative?

Police crash reports include information about substance use and testing and risk factors; toxicology data can give more extensive substance use information; medical data will give information on injury severity.

Another benefit of linkage programs is to facilitate evaluations. Listed below are some examples of questions (see more examples in Appendix B) that states can explore with linked data.

- Are MVC injury prevention efforts effective at reducing serious MVC injuries?
- How effective are interventions, such as improvements to graduated driver licensing (GDL) programs, for preventing and reducing MVC-related injuries among teens?
- Have increases in motorcycle helmet use resulted in fewer injuries or a reduction in injury severity?
- Has the accuracy of injury assessment captured in police reports improved over time?
- Does legalizing marijuana impact MVC injuries and deaths?

Linked data have been used to identify the risk for MVC-related injuries among specific populations, the economic impacts of MVCs on populations, and the impacts of preventive interventions on MVC occurrences and MVC-related injuries and deaths.

The utility of linked data for improving our understanding of MVCs is considerable, yet states face numerous challenges in starting or sustaining a data linkage program. Specific state challenges include a lack of funding, limited access to data sets, a lack of trained staff, and difficulties in developing a robust data linkage process. To help address these challenges, the Linking Information for Nonfatal Crash Surveillance (LINCS) Guide was developed in consultation with state and federal transportation safety and health agencies.

The guide was designed to help states establish or improve existing linkage programs. The long-term goal is to use linked data to inform strategies to reduce MVC-related injuries and deaths, which supports the National Road to Zero Initiative [5], Toward Zero Deaths, and Vision Zero Network.

The goals of this guide are as follows:

- Enable states that have never linked, or are no longer linking data sets, to begin a linkage program. This includes considering best approaches to partnership building and developing a business model.
- Enable states with existing linkage programs to expand and improve their data linkage process, such as adding new data sets to capture more information related to nonfatal MVCs.
- Enable states and partner organizations to improve their data quality and make use of linked data to support organizational goals.

These goals can be accomplished by reviewing the LINCS guide and applying the recommendations. The guide also provides in the various appendices background materials, technical information, and additional resources available to help states start or expand linkage programs.

Definitions

A list of acronyms and a complete glossary are provided at the end of this guide.

Why States Implement Motor Vehicle Crash Data Linkage Programs

The National Highway Traffic Safety Administration (NHTSA) and Federal Highway Administration (FHWA) work with states to reduce the number of MVCs on public roads and the severity of crash impacts through education, engineering, enforcement, and EMS. In keeping with that mission, the 2012 Moving Ahead for Progress in the 21st Century Act (MAP-21) and the continuation of the 2015 Fixing America's Surface Transportation Act (FAST Act) required the establishment of performance measures for states to assess and measure serious injuries due to MVCs [6, 7]. As of April 15, 2019, states will be required to adopt definitions, variables, and coding conventions outlined in the Model Minimum Uniform Crash Criteria (MMUCC) for reporting suspected serious injuries on police crash forms and in statewide MVC databases [7, 8]. Historically, sharing, comparing and linking data between localities, states and the federal government has been difficult because data used by separate agencies to describe the same crash characteristics have different definitions and variables [9]. Therefore, uniform definitions, variables, and coding conventions for serious injuries should make it easier for states to use nonfatal MVC outcome data.

MVC risk factors (e.g., speeding, alcohol use), protective factors (e.g., seat belt use), and outcomes (i.e., health or economic) are typically collected across four domains:

- Person
- Vehicle
- Crash
- Roadway

Factors and outcomes are captured in multiple data sets such as police crash reports and police citations (traffic tickets, etc.) and hospital discharge records. Another way to consider MVC data is by timeframe, as each MVC can be classified into three phases: pre-crash, crash, and post-crash [10, 11].

Figure 3 provides examples of relevant information associated with each crash phase, most of which might already be collected in many states. Linking data sets from across the four domains or three crash phases can provide more-robust MVC data for analysis, evaluation, surveillance purposes, and to focus prevention efforts.

Figure 3. Motor Vehicle Crash Phases and Examples of Associated Information

Pre-Crash

- **Driver characteristics**
- **Vehicle characteristics**
- Driver behaviors
- Driving laws
- Road design, including presence of embankments, guardrails, and median barriers

Crash

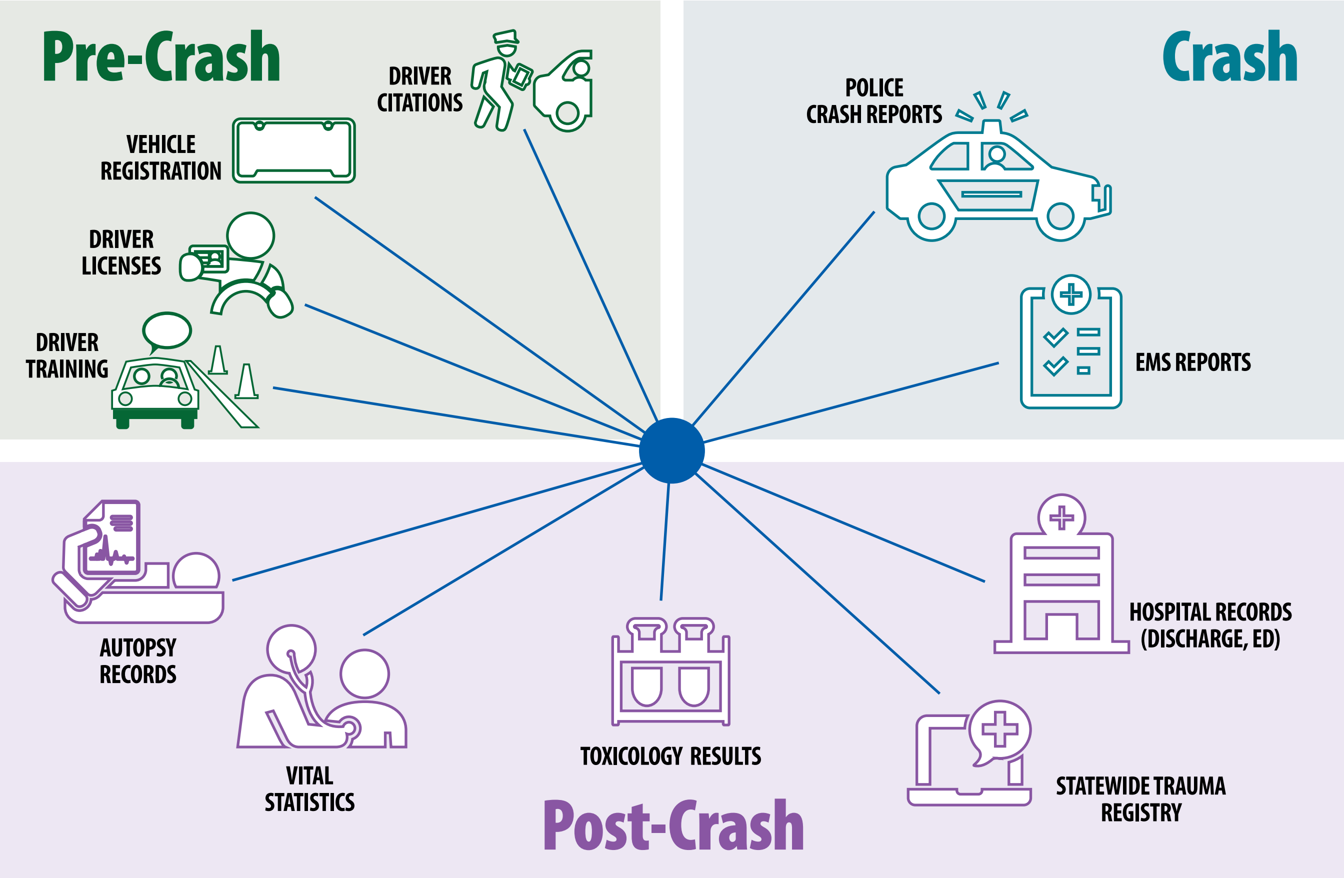
- **Driver characteristics**
- **Vehicle characteristics**
- Human factors, including restraint use, impaired status, and speed
- Road and traffic conditions, including other road users
- Number of vehicles, drivers, and passengers
- Vehicle trajectory
- Injury mechanism(s)

Post-Crash

- **Driver characteristics**
- **Vehicle characteristics**
- Emergency management assessments and interventions at crash scene
- Medical transport
- Injury treatment
- Outcomes of interest, such as health diagnoses and medical costs

Figure 4 shows the range of data sets that can be linked to provide a comprehensive view of each MVC. Historically, states with linkage programs have leveraged existing data sets from police records, hospital records, and EMS. When starting a linkage program, a reasonable place to begin is linking police crash reports and hospital or emergency department discharge records.

Figure 4. Existing Motor Vehicle Crash Data: Starting a Linkage Program



There are many benefits to implementing a linkage program:

- Linking MVC data sets creates a more comprehensive set of linked data for each MVC incident and for each individual involved in the MVC.
- Comprehensive MVC linked data can enable analysis of the relationships among contributing factors, interventions, outcomes, and impacts.

For example, one advantage of linking police MVC records to hospital records, particularly hospital emergency department and inpatient records, is the ability to assess the magnitude of nonfatal MVC injuries and associated healthcare costs. Historically, MVC injury data collected by the police have not reliably captured the injury severity or the actual injury outcomes [12–16]. Measuring the magnitude of a public health problem, such as MVC-related injuries, can help states prioritize prevention programs, strategies, and resources.

Appendix B provides an overview of the literature review conducted to inform the guide, and a detailed listing of MVC publications that have used linked data to demonstrate the utility of linked data.

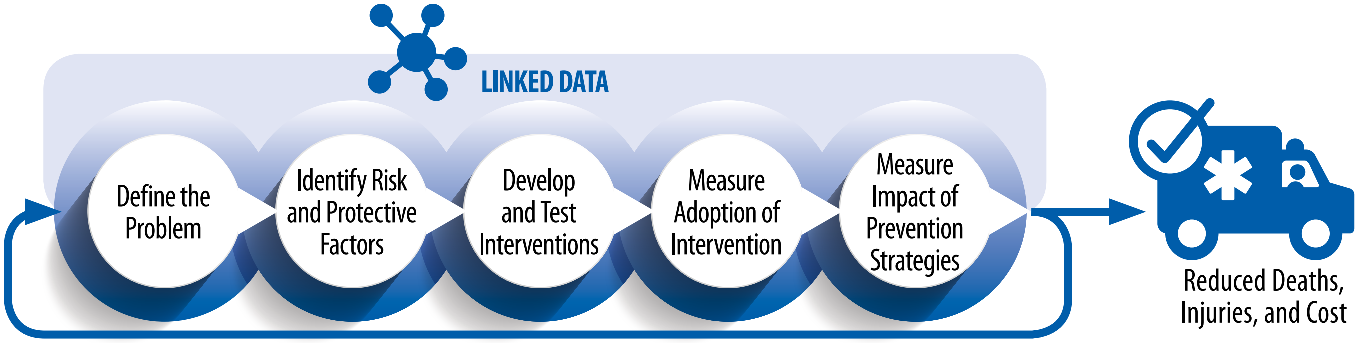
Linked data can be used for program monitoring and evaluation. For example, the Road to Zero Initiative, a collaboration within the Department of Transportation including NHTSA, FHWA, and the Federal Motor Carrier Safety Administration (FMCSA), and chaired by the National Safety Council (NSC), brings together a coalition of more than 900 members and other stakeholders, including the Centers for Disease Control and Prevention, to use a data-driven, interdisciplinary approach to end roadway deaths in the U.S. by 2050 [5, 17]. States can monitor progress toward the initiative's milestones and goals that can be enhanced by using linked data from public health and transportation safety stakeholders in a coordinated effort.

Linkage programs also have the potential for increased operational efficiency by improving data quality through common data standards and by reducing redundant data collection processes [18]. However, for most states, agencies, and organizations, realizing the full benefits of a shared linkage program to improve operational efficiency is a long-term goal.

Role of Linked Data in Injury Prevention and Control

Linked MVC data are critical, not only to inform and evaluate injury prevention strategies, but also to understand other negative impacts related to MVCs, such as injury, disability, and healthcare costs. The Injury Surveillance Training Manual [10] developed by CDC includes a high-level model of the public health approach to injury prevention and control, which is adapted in Figure 5. This iterative approach defines five steps necessary to prevent and reduce the severity of MVCs. In this section, the role of linked data in each of these steps is discussed.

Figure 5. Linking Motor Vehicle Crash Data in a Public Health Approach to Injury Prevention


LINKED DATA
Define the Problem
Identify Risk and Protective Factors
Develop and Test Interventions
Measure Adoption of Intervention
Measure Impact of Prevention Strategies
Reduced Deaths, Injuries, and Cost

Define the Problem. Identifying and analyzing the characteristics of MVCs provides insight into crashes, the magnitude of their impacts, and how these impacts are distributed across a population. Linking data across multiple sources enables a richer identification and characterization of MVCs because more information can be extracted about the incidents once they are linked. With more information on MVCs, such as road conditions, impairment status, and patient outcomes, stakeholders can prioritize among public health and transportation safety programs and resources to better target interventions for maximum impact. As an example, showing MVCs by location and time enables local law enforcement to better position patrols and road engineers to prioritize road segments for improvement to prevent crashes [19]. A future road safety application might include monitoring and evaluating the impact of self-driving cars (i.e., autonomous vehicles). By linking vehicle information with other data sources we can better understand how automation and other safety features are impacting safety.

Identify Risk and Protective Factors. There are many contributing factors to why crashes occur and how severely people are injured. Analysis of MVCs can help identify and measure risk factors (e.g., speeding, alcohol and drug use) and protective factors (e.g., use of car seats and booster seats for children), and their association with health and economic outcomes. Linking across multiple data sets enables analysis of a broader set of contributing factors present in MVCs, which in turn allows for better data-driven decision making. For example, state lawmakers might want to know whether age or driving at night (i.e., risk factors) increases the rate of MVCs and the costs associated with MVC injuries [20]. Using only police-reported MVC data, Massachusetts determined that drivers aged 16–17 years accounted for almost twice the proportion of crashes from 9 p.m. to midnight, compared with adult drivers [20]. By linking data from police-reported MVCs with hospital data, Massachusetts showed that, for MVCs that occurred at night (defined as 9 P.M. to 5:59 A.M.), inpatient hospitalization costs were 33% more ($20,000 versus $15,000) for drivers aged ≤17 years, compared with drivers aged 18–20 years [20]. These findings provided the evidence needed to restrict unsupervised night driving by newly licensed drivers and to expand the restricted night driving period to 9 P.M. to 5:59 A.M. [20].

Develop and Test Interventions. Once contributing risk and protective factors are better understood through analysis of linked data, interventions can be designed or implemented to prevent MVCs and to reduce injuries, injury severity, and resulting impacts. There are many types of evidence-based interventions available to states to prevent and mitigate MVCs. NHTSA publishes and updates Countermeasures That Work, a guide to assist states in selecting effective, evidence-based interventions to address traffic safety problem areas [21]. CDC offers an interactive calculator, Motor Vehicle Prioritizing Interventions and Cost Calculator for States (MV PICCS 3.0), to help states prioritize and select effective motor vehicle injury prevention interventions based on interventions for safer road users in NHTSA's Countermeasures That Work [21, 22]. An example of developing and testing interventions comes from South Carolina where the state used police-reported MVC data linked to hospital inpatient and emergency department data to evaluate GDL, a proven intervention to reduce crashes among teenage drivers [23]. South Carolina demonstrated that, despite reductions in teenage driver crashes associated with GDL, crashes were twice as likely to result in serious injury when teenage passengers were present than when the teenage driver was alone [23]. This finding provided the evidence for South Carolina to consider updating their GDL to restrict the number of passengers aged <21 years, a known effective intervention [23].

Measure Adoption of Interventions. The magnitude of an intervention's success is typically proportional to the rate of adoption. There might be many challenges to implementing an intervention widely, and analysis of linked data can help measure to what extent the intervention was adopted and identify potential obstacles to adoption. For example, universal motorcycle helmet laws are highly effective in protecting motorcycle drivers and passengers and reducing the risk of death and head injuries in a crash [24]. Michigan evaluated the impact of a partial repeal of the universal helmet law on serious injuries by linking police-reported crash data to a statewide trauma registry to compare head injuries before and after the repeal [25]. Helmet use decreased by 24–27% following the repeal [25]; the percentage of motorcycle drivers and passengers with head injuries increased by 14%, with significantly more injuries requiring neurosurgical intervention following repeal [25].

Measure Impact of Prevention Strategies. Linked data can be used to measure the impact of prevention interventions. State and federal agencies are increasingly interested in measuring health outcomes and associated costs of both MVCs and interventions. For example, using linked transportation and health data available in NHTSA's former Crash Outcome Data Evaluation System (CODES), CDC analyzed data from 11 states to examine how restraint use, among children aged 1–12 years who were involved in MVCs, was associated with injuries, medical outcomes, and hospital charges [26]. In another example, Washington state used linked police-reported crash data and roadway data with barrier presence information to investigate the motorcycle-to-barrier crash frequency on curved roadways [27]. The findings in Washington led to better identification and prioritization of sections of curved roadways for the placement of barriers as a crash countermeasure [27].

Linking Information for Nonfatal Crash Surveillance Guide: Purpose

The CDC's National Center for Injury Prevention and Control (NCIPC) enlisted the Centers for Medicare & Medicaid Services (CMS) Alliance to Modernize Healthcare (CAMH), a federally funded research and development center operated by The MITRE Corporation, to create a guide which builds on previous efforts and best practices for establishing and improving linkage programs at the state level. The audiences for the guide are linkage program managers, data analysts, and other technical support team members (e.g., statisticians, epidemiologists, software developers). The guide is built upon previous work (see Appendix C for more detail) and is supported by findings from newly conducted environmental scans, which addressed the following objectives:

- Develop an understanding of how linked data can be used (Appendix B)
- Document challenges and successes in implementing linkage programs (Appendix D)
- Explore alternative methods and tools available for data linkage (Appendix E and Appendix F)

To accurately represent the challenges faced by existing linkage programs, over a 6-month timeframe listening sessions were conducted with staff from seven states, as shown in Table 1. Appendix D provides additional information about the listening sessions approach as well as detailed findings from the sessions.

Table 1. List of States Participating in Listening Sessions

State	Entity Name
Georgia	Injury Prevention Program (IPP), Georgia Department of Public Health
Kentucky	Kentucky Injury Prevention and Research Center (KIPRC)
Maryland	National Study Center for Trauma & Emergency Medical Systems (NSC), University of Maryland, Baltimore
Massachusetts	UMassSafe Traffic Safety Research Program
New York	Bureau of Occupational Health and Injury Prevention, New York State Department of Health
Utah	Intermountain Injury Control Research Center, University of Utah School of Medicine
Virginia	Traffic Records Management, Reporting and Analysis Division of the Department of Motor Vehicles (DMV), Virginia Highway Safety Office (VAHSO)

Between December 2012 and May 2013, CDC and NHTSA conducted an evaluation of states that were known to have linkage programs [28]. Of the 25 states that responded to the evaluation, 14 states were part of NHTSA's CODES, which provided partial funding and technical assistance to linkage programs through cooperative agreements that ended in 2013, and 11 states had other data linkage systems unaffiliated with CODES [28]. The evaluation found that most of the responding states identified MVC injury as a high-priority health problem, and states consistently reported having police-reported MVC, hospital inpatient, and emergency department data available to link in their programs [28].

As of March 2017, an online review determined that 19 states had linkage programs, including

- Alaska
- California
- Georgia
- Iowa
- Kentucky
- Maine
- Maryland
- Massachusetts
- Michigan
- Minnesota
- Nebraska
- New York
- Ohio
- Oklahoma
- Tennessee
- Utah
- Virginia
- Washington
- Wisconsin

Based on the online review, all states, except Massachusetts and Virginia, linked police MVC data with hospital data for analysis, evaluation, or surveillance. Appendix G provides a list of states with linkage programs as of March 2017 and highlights which programs link MVC and hospital data.

Based on lessons learned from linkage programs, this guide describes both a step-by-step approach to start and maintain a new linkage program and how to expand and improve an existing linkage program. This guide was created to be relevant for all states; however, each state will need to make adaptions to accommodate its specific laws, regulations, priorities, and populations.

A wide audience should find this guide useful to perform any of the following actions:

- Consider or start a new linkage program
- Revive a defunct linkage program
- Expand or improve an existing linkage program
- Explore how to contribute data sets to a linkage program
- Evaluate the quality of data sets
- Conduct analysis, evaluations, monitoring, surveillance, and research with linked data
- Improve linked data products, such as presentations, reports, and publications

The guide is intended to be a detailed resource that provides the background and history of data linkage, a literature review, state success stories, and the inner workings of the data linkage methods—all of which can be referenced at different stages of starting or expanding a linkage program. The detail provided in the 17 appendices is intended for readers who want to understand the complexities of a given process or step. Multiple examples are included throughout the guide to demonstrate the range of programs and to showcase potential solutions to common data linkage issues. The examples presented are based on the literature review (Appendix B) and stakeholder listening sessions (Appendix D) conducted to inform the development of this guide.

LINCS

THE LINCS GUIDE

SECTION 1. ESTABLISHING A MOTOR VEHICLE CRASH DATA LINKAGE PROGRAM

The purpose of this section is to provide a high-level overview of the components necessary to start a linkage program at the state level. Successful linkage programs include functions that span partnerships, business models and polices, and the process and technology of data linkage, as shown in Figure 1. If you already have a successful linkage program set up, Section 4 describes potential enhancements to the data linkage process.

Building partnerships and establishing communication with both internal and external stakeholders are key to the success and sustainability of a linkage program. Section 2 focuses on how to build partnerships and coalitions and how to communicate effectively with stakeholders.

Consistent and sustained funding and resources are critical to starting and maintaining a successful linkage program. Section 3 discusses the development of a business model with examples of best practices and privacy and security policy considerations for data protection.

The final component of a successful and sustainable linkage program is the linkage process, which the guide presents in discrete steps, from defining the goals for the process to conducting analysis on linked data. The 10 steps for establishing a data linkage process are provided in Section 4.

SECTION 2. BUILDING PARTNERSHIPS

BUILDING PARTNERSHIPS

This section explains how to build effective partnerships and coalitions along with the considerations for long-term linkage program planning.

2.1 Build a Coalition

Prior to creating a coalition, it will help to answer the following questions:

- ▶ What are the current issues that need to be resolved?
- ▶ Who is affected by the current issues?
- ▶ Who will benefit if the issues are resolved?
- ▶ Who can help resolve the issues?

Coalitions are people and groups that come together for a common purpose. Creating coalitions focused on MVC safety and public health can expand data sharing opportunities and scope, establish greater credibility among stakeholders, and increase resources [29]. Although coalitions and partnerships might initially require effort, resources, and time to establish, thoughtful strategic planning can lead to broader positive change and wider impact. When building a coalition, start with existing partnerships and relationships, when possible, and expect that new relationships will take time, communication, and information sharing [30]. Staff continuity in collaborating agencies and organizations is beneficial to successful long term working relationships. Coalitions are more likely to be successful if the following six steps are incorporated early in the process:

1. Understand coalition needs. Prior to creating a coalition, consider conducting an assessment to determine the needs and to gain a common understanding of the issues or barriers related to data and information sharing. Identify the levels of need and incentives for stakeholders, and the barriers to participation. An assessment can also determine what partnerships already exist and identify the processes and procedures already in place to address data sharing.

2. Identify relevant stakeholders. A coalition, at minimum, will need to include those who want to access the data and the agencies or organizations that own the data. There are many potential stakeholders to consider when building a coalition. For example, public or state stakeholders include law enforcement, public health, public safety, transportation, judicial, vital statistics, and policy agencies within local and state governments. Examples of private stakeholders include academic institutions and universities, hospitals and healthcare organizations, insurance companies, safety advocacy groups, commercial associations, and potential donors. At the federal level, the Department of Transportation (DOT) and the Department of Health and Human Services (HHS) agencies have missions aligned with improving MVC safety and public health, and supporting linkage programs. Appendix H has a list of resources to help identify potential partners and stakeholders.

3. Identify a shared vision. To build a successful coalition, members should create a clear vision and goals that are shared by stakeholders and form the foundation for coalition activities. Successful linkage programs work collaboratively with stakeholders to identify core values that are important to the coalition's members [31]. An example of a successful coalition built on a shared vision of reducing MVCs, injuries, and deaths is Minnesota's Toward Zero Deaths (TZD), the state's integrated traffic safety program established in 2003 by the Minnesota Departments of Public Safety, Transportation, and Health [32]. Minnesota's TZD includes public and private partners across the advocacy, education, engineering, judicial, law enforcement, and health sectors; TZD developed a Roadmap of Partners for communities wanting to build partnerships, coalitions, and community support [33].

4. Establish an organizational structure. To maintain effective and ongoing planning, a governance structure should be established for the coalition. The organizational framework of a coalition depends on several factors, including leadership roles, membership considerations, coalition size, and goals. The lead agency or organization(s) should consistently engage members and use a shared decision-making approach to achieve success. Common governing and supporting entities might include a board of directors, steering committees, support committees, and task forces. Successful coalitions generally have active planning groups or subcommittees that carry out coalition activities. An example of existing organizational structures is the state Traffic Records Coordinating Committee (TRCC), which has members from state agencies across six core record systems: crash (e.g., law enforcement crash reports), vehicle (e.g., vehicle registration data), driver (e.g., driver license and history), roadway, citation/adjudication, and injury surveillance (e.g., data from EMS and hospital records). The state TRCC is responsible for coordinating state organizations involved in the administration, collection, and use of highway safety data and traffic records [34].

5. Keep coalition members informed. Coalitions are built on trust and information. Regular meetings and communications serve to build trust and sustain momentum for the coalition's goals and activities. Regularly scheduled meetings can increase member participation. Subcommittees might meet more frequently or communicate through conference calls or other channels.

6. Advocate for the coalition. As the newly formed coalition expands, there will be opportunities for advocacy work. Coalition members who are in a position to influence and advocate for legislative policies may play a key role in strengthening the coalition. Designated advocates may be able to use their skills to build external partnerships and to garner support for issues affecting the MVC data linkage programs. Effective planning and monitoring of legislative activity may be necessary to sustain support and to meet the desired outcomes of the coalition.

2.2 Communicate with Stakeholders

The formation and maintenance of any successful coalition requires effective and continuous communication both within and outside the coalition. The following approaches can facilitate communication:

Raise awareness to key stakeholder groups. Communication to external stakeholders can raise awareness about the coalition, bring attention to motor vehicle safety issues, and share successful strategies and stories. Communication channels might include traditional media, social media, conferences, publications, and other activities [35]. Communication materials should be client oriented and should focus on the coalition's goals and activities related to motor vehicle safety and linked data. Common communication tools are a well-designed website, Frequently Asked Questions (FAQs) sheets, fact sheets, and brochures.

Leverage expertise and technical assistance. States with experience and expertise in linking MVC data-sets should be leveraged by states that are starting new, or re establishing existing, linkage programs. States with functional linkage programs can assist states that do not have a program and can share best practices. As of 2017, 19 states performed data linkage of two or more MVC data sets for surveillance, evaluation, and research purposes; Appendix G shows which states linked police crash reports with hospital data. Technical assistance is available to states through participation in national professional societies and associations; Appendix H contains a list of available resources. Formal technical assistance is also available through the federal government; Appendix I contains information about technical assistance resources available from the DOT agencies.

Communicate findings and recommendations. Communicating the findings from linked MVC data can serve to meet the informational needs of end users, share the value of the linkage program, solidify existing relationships, and develop new relationships. In addition to describing the problems, methods, and results, the stakeholders might be interested in recommendations for additional evaluation and analysis topics, ideas for measuring and monitoring interventions, and potential programmatic and policy initiatives.

When presenting findings, presentations should be tailored to the audience by first understanding their priorities and what the audience most wants to know, technical experience, and preferred communication modality. Visualizations, such as tables, graphs, and maps, can be powerful ways to summarize findings, but it is important to provide narrative interpretations of summary statistics, especially in the context of relative magnitude, trends over time, crude versus adjusted rates, and any limitations. Appropriate communication might impact ongoing access to data sets and linkage program funding.

Below are some options for communicating linkage program findings.

Final or preliminary published reports. Findings can be incorporated into local, state, and federal agency reports, such as CDC and NHTSA reports [28]. For operational or policy purposes, stakeholders might want the findings reported before all required data are received. In these cases, the linkage program might consider issuing a preliminary report with clear language stating that the findings might differ when final data are received and when analysis is complete and has been reviewed.

In-person presentations. State legislatures might request in-person presentations of the findings in addition to written briefs or reports. Informal presentations might be more effective at routine meetings with data owners, coalition members, and stakeholders.

Open access via website. Even if findings are regularly incorporated into government reports, most linkage programs also have websites for dissemination purposes; Appendix G provides links to active linkage program websites for 19 states. Reports can be posted across the websites of linkage programs, local and state governments, and advocacy and funding organizations, for maximum dissemination and exposure. Some linkage program websites (e.g., Georgia) have interactive tools that enable the public to create custom reports or visualizations of the data per user-defined parameters.

Private access. Subscription services to data sets and informational products, such as reports, might be desirable for covering costs. For example, Utah's Department of Public Health shares data sets for a fee; the application requires details on how the data will be stored. However, state laws differ regarding whether fees can be charged for publicly collected data and information products generated by specific data sets.

Peer-review journals and scientific meetings. Dissemination of findings through peer-reviewed journal publications and scientific conferences or meetings can reach broad audiences across specialty domains, professional fields, and sectors. For example, the Traffic Records Forum is an annual meeting, sponsored by the Association of Transportation Safety Information Professionals, of data analysts, state and local law enforcement officials, engineers, EMS providers, judicial administrators, and highway safety professionals from across the U.S. and international communities. Specific sessions are focused on data integration to improve and expand the use of linked MVC data. Appendix H includes a partial listing of scientific conferences and association meetings that might be of interest to linkage programs.

By building partnerships and engaging stakeholders with an interest in motor vehicle safety, states can leverage support to start or expand a linkage program. Working with a coalition and community stakeholders can lead to additional opportunities and resources, and a wider awareness of motor vehicle safety issues.

SECTION 3. DEVELOPING A BUSINESS MODEL

DEVELOPING A BUSINESS MODEL AND POLICIES

The purpose of this section is to provide information about how to address the funding, resourcing, and policy aspects of establishing a linkage program.

3.1 Establish Funding

Sustainable funding is necessary for the long-term success of a linkage program. Many states with linkage programs have noted that the lack of consistent funding to cover operational costs has been a challenge (see Appendix D). Examples of ongoing operational costs are data fees, staff training, software licensing fees, and technical support fees. Potential funding sources are listed below.

Grants or Cooperative Agreements. One example of federal funding to states is the Core State Violence and Injury Prevention Program (Core SVIPP) provided by the CDC NCIPC. Core SVIPP funding helps 23 states implement, evaluate, and disseminate strategies that address high-priority injury and violence issues, including MVC-related injuries and deaths [36].

- If grants or cooperative agreements (federal or non-federal) are only a single year in duration, long term sustainability should be considered. States that have relied on grants to fund linkage programs have experienced staff turnover due to gaps or uncertainty in funding. In addition, there can be limitations on grant funding use.

Fees or subscription charges. Linkage programs can charge fees to fulfill data requests or to provide consulting services to assist users. State regulations frequently specify how data can be used, which can limit how a linkage program can generate revenue through fees or subscription. In the past, the UMassSafe Traffic Safety Research Program maintained a data repository for multiple government organizations and charged fees for services to link or analyze data.

Shared resources. Labor (e.g., faculty, students), facilities, and computing infrastructure can be shared across coalition partners and academic institutions.

Services diversification. Providing data services for non-MVC domains can open opportunities to other funding sources. Consistent funding helps maintain program continuity and skilled staff to operate linkage programs. External revenue can offset operational costs of the linkage program.

Some grant or cooperative agreement applications might require estimated project budgets to determine award amounts. It is important to leave adequate margins in these cost estimates, as the actual costs typically exceed the expected costs due to unforeseen circumstances, especially when a linkage program is just getting started. When developing proposed budgets to submit for funding opportunities, other costs to consider include:

- Fees paid to use specific data sets
- Staff time as in-kind services in return for sharing data
- Staff time and expenses involved with staff training
- Technical support fees to software vendors selected to process, link, and analyze the data sets

3.2 Identify Staff and Computer Resources Needed

Successful linkage programs depend on skilled staff and technology infrastructure. States indicated that finding data analysts with data linkage expertise was more challenging than finding staff with proficiency in statistical methods for the analysis of linked data [28]. Successful linkage programs depend on having staff with the appropriate skills, including "a high level of expertise with the linkage software packages, epidemiology, statistics, knowledge of traffic safety, data sets that were being linked, and presentation and marketing skills to ensure data are used" [28].

Table 2 provides a list of the business and staffing models used by selected state linkage programs (see Appendix D). Although models differ substantially among the states, often one or two staff take on many roles and responsibilities in linkage programs. In general, the following roles and services are involved in a linkage program:

- Program leadership to interact with stakeholders, procure funding, and manage projects
- Principal investigators to design evaluations, surveillance plans, and research studies based on the linked MVC data, and to write grants/funding proposals
- An information technology (IT) engineer to assist with procuring and maintaining the computing infrastructure and services, such as the linkage program website
- Computing infrastructure and services to perform data linkage and analysis, including hardware, software, storage, and network connectivity
- Data analysts, data scientists, epidemiologists, or statisticians to use data linkage tools, perform data linkage, and analyze linked MVC data
- Graphic-design and communications services to develop materials and products for dissemination and the linkage program website

Table 2. Examples of Motor Vehicle Crash Data Linkage Program Models

State Program	Funding Source	Staff Types	Total Staff
Injury Prevention Program of the Georgia Department of Public Health	State funding and annual Highway Safety grants (Technology TRCC grants)	▸ Director ▸ Program Manager ▸ Data Scientist ▸ Part-time Data Scientist	1.5 FTE[1]
Kentucky Injury Prevention and Research Center (KIPRC)	Two grants/cooperative agreements (DOT and CDC)	▸ Director ▸ Data Scientist ▸ Statistician	1.35 FTE
National Study Center for Trauma and Emergency Medical Systems (NSC) at the University of Maryland, Baltimore	Shared resources using various funding sources	▸ Database Coordinator ▸ Epidemiologist ▸ Statistician	1.5 FTE
UMassSafe: University of Massachusetts Traffic Safety Research Program	Shared resources using various funding sources	▸ Data Scientist ▸ Statistician ▸ Database Coordinator	1.5 FTE
Minnesota Department of Health—Injury and Violence Prevention	State funding and annual grants/ cooperative agreements (e.g., CDC and Highway Safety Office)	▸ Director ▸ Statistician/Data Scientist ▸ Epidemiologists	1.5 FTE
Bureau of Occupational Health and Injury Prevention, New York State Department of Health	State funding and annual grants (e.g., CDC and Highway Safety Office)	▸ Director ▸ Epidemiologist ▸ Statistician	2 FTE
Intermountain Injury Control Research Center at the University of Utah School of Medicine	Shared resources using various funding sources	▸ Statisticians	2 FTE
Traffic Records Management, Reporting and Analysis Division of the Department of Motor Vehicles, Virginia Highway Safety Office (VAHSO)	Shared resources using various funding sources	▸ Deputy Director ▸ Data Manager ▸ Data Analysts ▸ Developers ▸ Tester	2 FTE and 6 contractors

[1] Full-time equivalent (FTE)

One of the biggest challenges that states noted during the listening sessions was staff turnover. If a single individual who worked in isolation left a linkage program, key institutional knowledge was lost, which could negatively impact the program [28, 37]. Replacing people can be difficult, time-intensive, and costly [28]. In addition, based on information shared during the listening sessions, there are limited opportunities for staff to take formal training for data linkage. The federal government provides technical assistance and training opportunities for states; Appendix I provides a listing of DOT programs.

Minimize impacts of staff turnover by creating and maintaining documentation:

- Repeatable analytic and data linkage steps
- Operating procedures
- Lessons learned

When selecting software resources, it is important to consider the full spectrum of the data-related functions required for the linkage program. Multiple software packages might be needed to meet the needs of database management, duplicate removal, data linkage, statistical analyses of the linked data, and the presentation of findings. Before purchasing tools, it is important to have a clear plan of what methods will be required to perform data linkage (see Section 4). The linkage program should take advantage of any discounts on software for government, academic institutions, or nonprofits that apply, and should explore whether evaluation licenses are available so that states can test data linkage tools prior to purchasing them.

3.3 Decide Who Will Host Linkage Program

States should consider what aspects of data linkage and analysis they can realistically perform with available resources. Many states with linkage programs use a trusted third party for data repositories, data linkage, and analyses. Trusted third parties can be academic institutions or nonprofit or for-profit institutions. When considering a business model, there are two important functional roles—data stewardship and analysis—which might be performed by a single entity or by different entities.

Data steward: Receives data from the data owners and performs data linkage. Because state agencies are one of the main sources of MVC data, one approach is for a single state agency to function as the data steward. That state agency would receive data from other state agencies and perform the data linkage.

Analyst: Performs analyses using linked MVC data. If this function is performed by an entity different than the data steward, each agency that receives the linked data is responsible for analyses. Therefore, each agency must maintain proficient data analysts to conduct analyses and disseminate findings.

Table 3 presents the benefits and challenges of the two business models that are utilized by states for linkage programs.

Table 3. Benefits and Challenges of Business Models for Data Stewardship and Analysis

Host/Entity Performing the Data Steward and Analyst Role	Benefit(s)	Challenge(s)	Organization(s) Using this Business Model
State agency	▶ State DMVs, transportation, or public health have public trust to serve as the data steward.	▶ Greater focus on the priorities of the lead agency. ▶ Public might react negatively if the data steward is law enforcement. ▶ Might be limited in the scope of operations to the lead agency's primary mission.	▶ Injury Prevention Program of the Georgia Department of Public Health ▶ Minnesota Department of Health—Injury and Violence Prevention ▶ Bureau of Occupational Health and Injury Prevention, New York State Department of Health ▶ Traffic Records Management, Reporting and Analysis Division of the Department of Motor Vehicles, Virginia Highway Safety Office (VAHSO)
Trusted third party	▶ Operational freedom to pursue a range of activities using linked MVC data. ▶ Can foster trust between public-sector and private-sector data owners. ▶ Might be viewed as less biased or less likely to have a specific agenda. ▶ Can leverage existing infrastructure to handle sensitive data and to reduce operational costs. ▶ Can leverage an interdisciplinary approach through collaboration across academic departments. ▶ Faculty and student salaries might be partially covered by an academic institution.	▶ Using students to perform the data linkage or analysis might provide inexpensive, skilled analysts, but also contributes to high turnover when students graduate or move to other opportunities. It is recommended that students not be used to do the data linkage, so that staff can have long-term continuity to create relationships and understand the complexities of the data sets.	▶ Kentucky Injury Prevention Research Center (KIPRC) ▶ National Study Center for Trauma and Emergency Medical Response (NSC) at the University of Maryland, Baltimore ▶ UMassSafe: University of Massachusetts Traffic Safety Research Program ▶ Intermountain Injury Control Research Center at the University of Utah School of Medicine

3.4 Define Security and Privacy Policies

Privacy and security are two distinct disciplines that are mutually supportive; security mechanisms are used to protect privacy, and privacy requirements are used to identify appropriate security protections.

Security focuses on protecting information and information systems, such as ensuring the availability of systems, malicious code detection and prevention, configuration and patch management, intrusion detection and mitigation, and physical protection. Two parts of the MVC data linkage process have critical security considerations: storage of data sets, and linked data transfer methods between the data owner, data steward, and analyst. The linkage program should collaborate with the state's agency responsible for cybersecurity to ensure compliance with state policies, procedures, and guidance, as discussed in Appendix J. A linkage program should identify the best security approaches to use for protecting data appropriately. The three core principles behind security are identified below [38].

- **Confidentiality.** Preserving authorized restrictions on information access and disclosure, including the means for protecting personal privacy and proprietary information.
- **Integrity.** Guarding against improper modifying or destroying of information and ensuring information non-repudiation and authenticity.
- **Accessibility.** Ensuring timely and reliable access to, and use of, information.

Each organization that is responsible for a linkage program should consult legal counsel to determine which privacy laws apply, at both the state and federal levels. Listed below are some potential questions to consider in collaboration with the state agency or organizational entity responsible for privacy policies, procedures, and guidance.

The following questions should be considered when planning security activities related to the data set and linked data transfer method(s) and data storage:

- What data set and linked data transfer methods should be used to meet security polices?
- What security mechanisms are in place to protect the data sets and linked data while stored?
- What roles and individuals within the organization should have access to the data sets and linked data?
- How will access control be managed?

- Is the agency or organization considered a covered entity under the Health Insurance Portability and Accountability Act of 1996 (HIPAA)?
- What requirements does the agency or organization need to meet to maintain Personally Identifiable Information (PII) and Protected Health Information (PHI)?
- Are there state laws or regulations pertaining to specific data sets (e.g., court records, medical examiner data)?
- Is a Data Use Agreement (DUA) or Memorandum of Understanding (MOU) needed for each data set sharing arrangement?
- What is the minimum amount of PII and PHI that the linkage program should receive and store to effectively link data sets?
- How and when should consent be obtained and updated for the linkage program to use PII and PHI?
- How and when should public notifications on the use of PII and PHI be shared?
- Are there more privacy concerns with the linked data than the original data sets (i.e., are individuals more easily identified in the linked data)? If so, are additional privacy and security measures necessary? Are de-identification steps needed?

Examples of selected PII relevant to linkage programs:

- Name
- Address
- Date of birth
- Social security number
- Photographs of persons
- Driver's license number
- Vehicle license plate numbers

- What are appropriate retention periods for the PII and PHI data sets used for data linkage?
- What disposition and destruction processes and policies must be followed when the retention periods end?

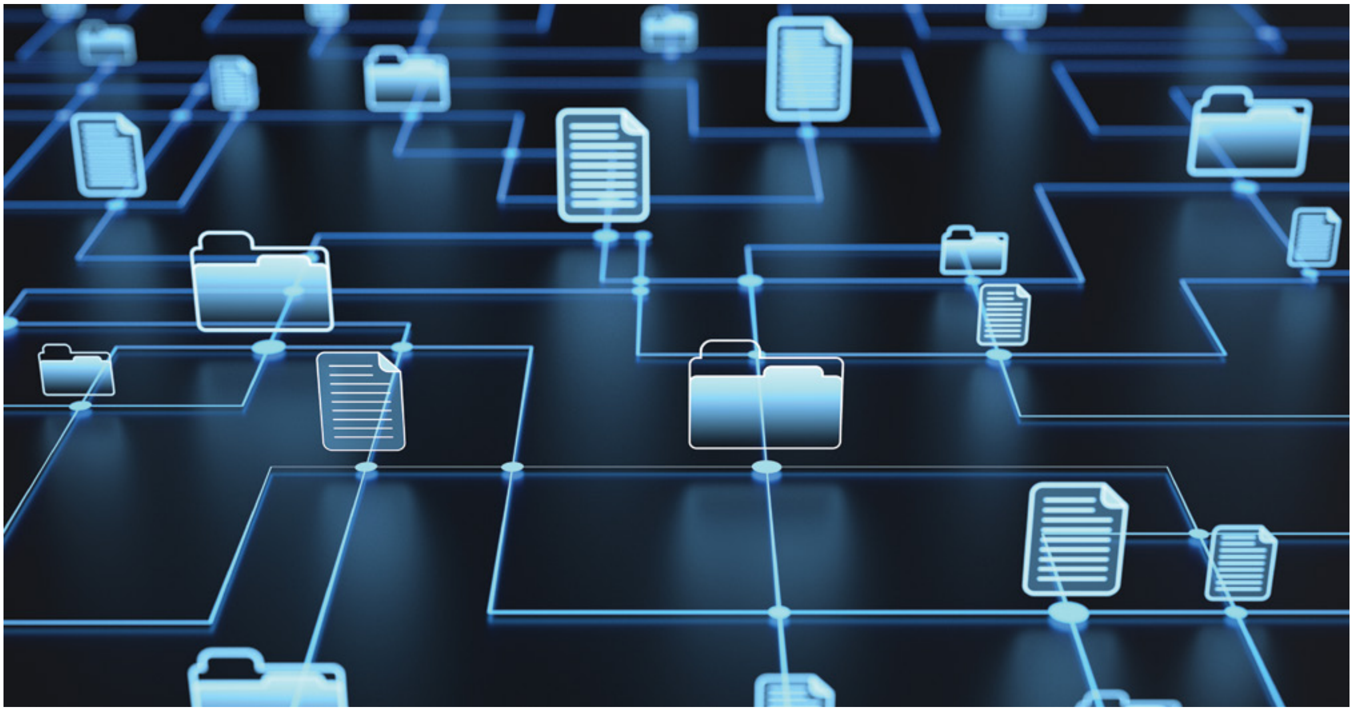
Collaboration with the state's Office of Privacy to leverage state policies, procedures, and guidance is prudent. Appendix K presents typical privacy-related activities for linkage programs.

Sustainable funding, skilled staff, and technology infrastructure are necessary for the long-term success of a linkage program. Although linkage programs are hosted by either a state agency or a trusted third party (e.g., a university), most are small with one or two "champions" responsible for the program. Linkage programs can leverage existing state or host-institution security and privacy policies, procedures, and guidance to ensure data protection.

SECTION 4. ESTABLISHING THE DATA LINKAGE PROCESS

ESTABLISHING THE LINKAGE PROCESS

Ten steps are identified for conducting the MVC data linkage process, as shown in Figure 2. If the linkage has been done before, then the process will start at the Prepare Data step. Each step is described in detail in this section.



4.1 Define Goals of the Linkage Program

An important first step in establishing a linkage program involves defining the goal(s) of the program. Analyses of linked MVC data can yield many practical applications, including those listed below:

- Understand the nature, number, severity, and outcomes of injuries and costs associated with MVCs, and their impacts in specific locations or statewide [18].
- Inform the operations of various stakeholders focused on motor vehicle safety [18].
- Improve data owners' data collection, management, processing, and quality assurance methods.
- Improve data sharing between agencies that need/use the same data sets [18].
- Help design and appropriately target interventions to prevent and mitigate deaths, injuries, and associated costs from MVCs.
- Inform state and federal motor vehicle safety priorities.
- Drive data-based and evidence-based decision making at state and federal levels for motor vehicle safety efforts.

Clearly articulating the data linkage goals is essential to the next steps of identifying data sets and tools for linkage. Table 4 shows the current status of data linkage and data sets used by states who participated in the stakeholder listening sessions from October 2016 to March 2017 (Appendix D).

Table 4. Examples of State Data Linkage Status and Data Sets among States that Participated in Listening Sessions

State Program	Data Linkage Status	Data Sets
Injury Prevention Program of the Georgia Department of Public Health	Link crash and health records	▶ Police crash reports ▶ Emergency department and hospital discharge records
Kentucky Injury Prevention and Research Center (KIPRC)	Link crash and health records for state	▶ Police crash reports ▶ Emergency department and hospital inpatient billing records ▶ Trauma registry data ▶ Highway information system data ▶ Death certificates ▶ Fatality Analysis Reporting System (FARS) data
National Study Center for Trauma and Emergency Medical Systems (NSC) at the University of Maryland, Baltimore	Link crash record, driver and vehicle characteristics, and health records for state	▶ Police crash reports ▶ EMS reports ▶ Hospital discharge, inpatient, ambulatory care reports ▶ Trauma registry reports ▶ Medical examiner reports ▶ Driver license data ▶ Driver citation data ▶ Motorcycle safety training data ▶ Vehicle registration data
UMassSafe: University of Massachusetts Traffic Safety Research Program	No current data linkage (only in past projects)	▶ No current data linkage
Minnesota Department of Health – Injury and Violence Prevention Section	Link crash and health records	▶ Police crash reports ▶ Hospital discharge (inpatient and emergency department) reports ▶ Trauma registry data ▶ Traumatic brain and spinal cord injury registry ▶ Death certificates ▶ EMS ambulance run reports
Bureau of Occupational Health and Injury Prevention, New York State Department of Health	Link crash, driver characteristics, and health records	▶ Police crash reports with damages exceeding $1,000, and injury ▶ Hospital, emergency departments records ▶ EMS records ▶ Trauma registry ▶ DMV accident information system data
Intermountain Injury Control Research Center at the University of Utah School of Medicine	Research on linking crash and health records	▶ Police crash report ▶ Hospital reports (i.e., emergency department, inpatient, and discharge)

Continued

State Program	Data Linkage Status	Data Sets
Traffic Records Management, Reporting and Analysis Division of the Department of Motor Vehicles, Virginia Highway Safety Office (VAHSO)	Traffic Records Electronic Data System (TREDS) interfaces with the following systems: ▶ Roadway system (DOT) through a web service ▶ Virginia Polytechnic Institute and State University ▶ NHTSA systems (i.e., Fatality Analysis Reporting System [FARS], Crash Report Sampling System [CRSS], and Crash Investigation Sampling System [CISS]) ▶ DMV Citizen Services System (CSS) (i.e., driver, conviction) ▶ Motorcycle student training module ▶ Driving under the influence (DUI) Checkpoint Strikeforce module ▶ Click It or Ticket (CIOT) module ▶ Efforts are underway to interface electronically with EMS, medical examiner, and vital records. Currently, all three agencies provide the DMV with data via a file transfer, which is manually linked and then uploaded to TREDS.	▶ Police crash records ▶ Motorcycle student training data ▶ Crash site roadway location data ▶ Driver history and driver conviction(s) records ▶ Driving under the influence or toxicology data (only from Office of the Chief Medical Examiner) ▶ Ignition interlock system data (Virginia Alcohol Safety Action Program) ▶ FARS data ▶ CRSS data ▶ CISS data ▶ EMS data ▶ Blood alcohol content data ▶ Vital statistics records ▶ State Police's SafetyNet system data ▶ DUI CheckPoint Strikeforce data ▶ CIOT data ▶ Medical review data ▶ Uninsured vehicle data

4.2 Establishing Data Use Agreements

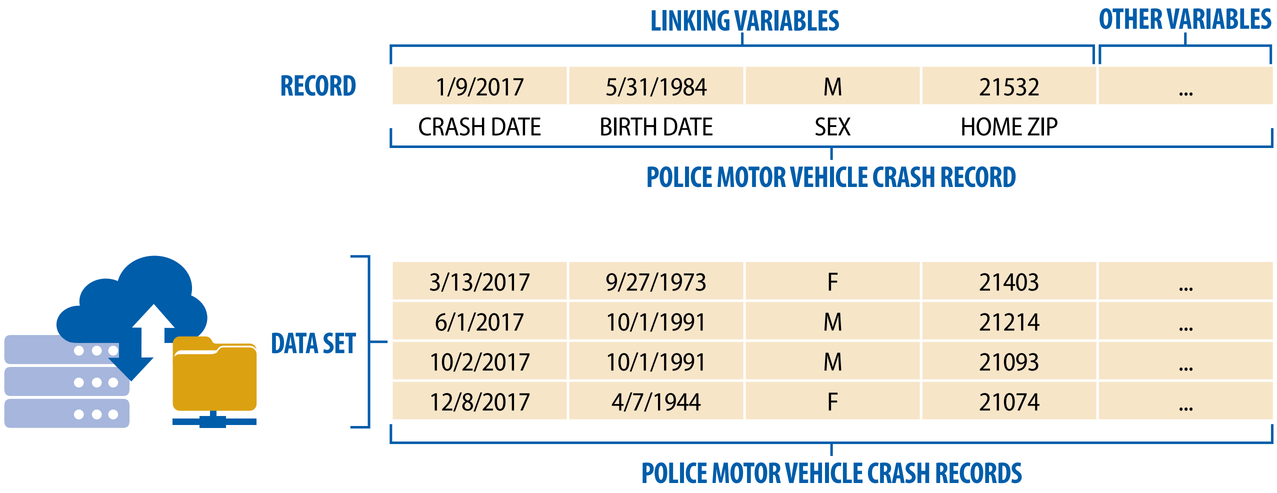
Linkage programs can take advantage of data that are regularly collected for operational purposes, such as police crash reports and hospital records, and data linkage can enhance the utility of the original data sets [39].

Data owners within the coalition should list the potential data sets that might have the information of interest. Data owners will understand the variables (see Figure 6) contained in their data sets. When the same variables that have the potential for identifying unique cases (such as name, date of birth, date of event, etc.) are present in two or more data sets (see Figure 7), there is potential that the data sets can be linked (see Figure 8). Variables must have comparable formats and be reliably populated in each data set. In some cases, a bridging data set serves to link two or more data sets that otherwise could not be reliably linked. Data owners and users can help identify bridging data sets for successful data linkage.

Successful Data Owner Engagement Strategies

- Try a bottom-up approach. Before reaching out to the head of the organization that owns the data set, talk to the person most directly responsible for maintaining the data set; they typically know more about the data/data set and could point out potential barriers and facilitators for linkage.
- Understand the data owner's needs. What do they want in exchange for providing the data set? Can this process provide analyses or insights on data set quality that will benefit the data owner?
- Address legal and security concerns early in the process.
- Agree to staffing and time commitments for both organizations. This should include time to learn about the nuances of the data sets and data linkage process, and to address any data set quality issues.

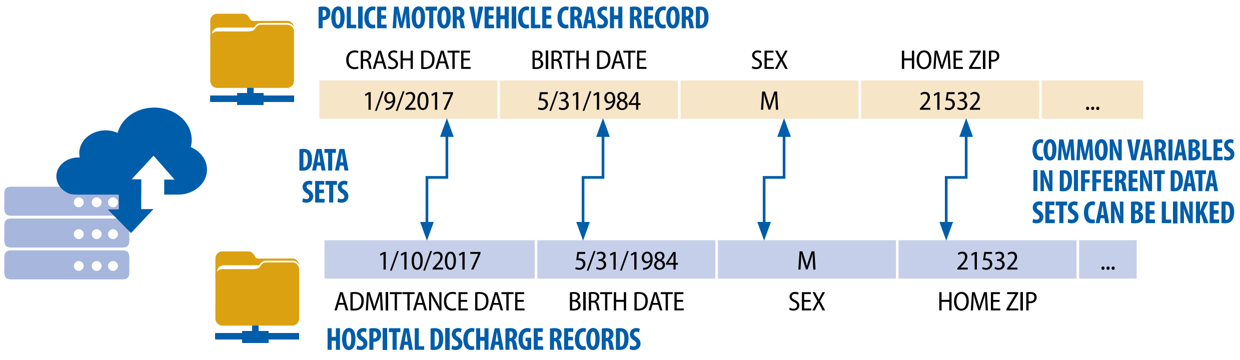
Figure 6. Linkage Variables and Records in a Data Set


LINKING VARIABLES
OTHER VARIABLES
RECORD
POLICE MOTOR VEHICLE CRASH RECORD
DATA SET
POLICE MOTOR VEHICLE CRASH RECORDS

CRASH DATE	BIRTH DATE	SEX	HOME ZIP	
1/9/2017	5/31/1984	M	21532	...

3/13/2017	9/27/1973	F	21403	...
6/1/2017	10/1/1991	M	21214	...
10/2/2017	10/1/1991	M	21093	...
12/8/2017	4/7/1944	F	21074	...

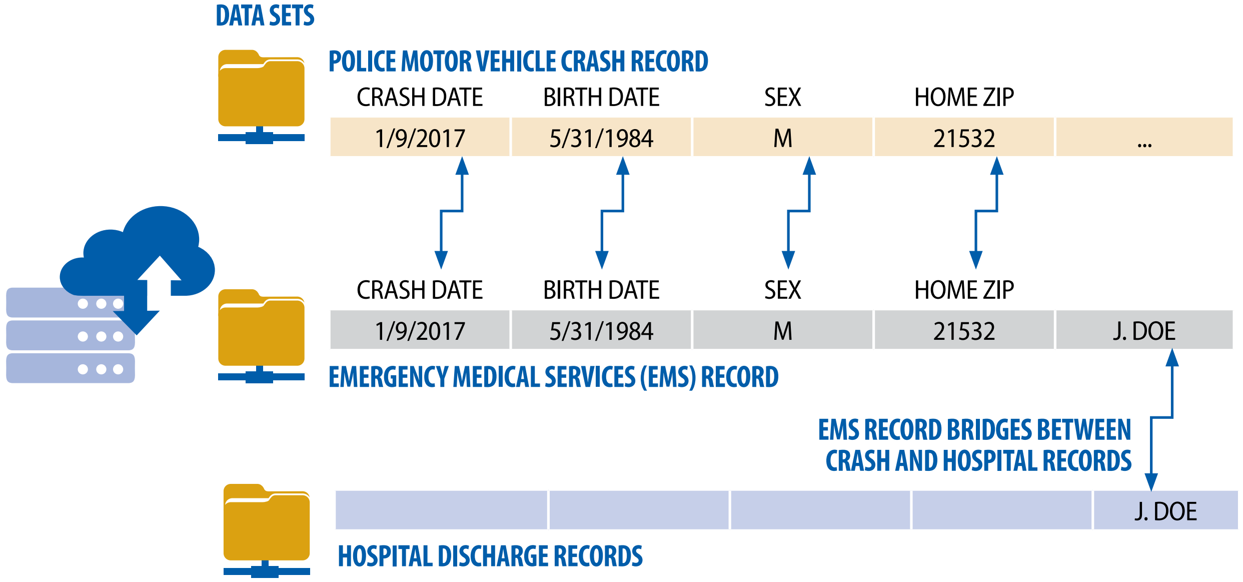
Figure 7. Linking Records Across Data Sets Using Common Variables


POLICE MOTOR VEHICLE CRASH RECORD
DATA SETS
COMMON VARIABLES IN DIFFERENT DATA SETS CAN BE LINKED
HOSPITAL DISCHARGE RECORDS

CRASH DATE	BIRTH DATE	SEX	HOME ZIP	
1/9/2017	5/31/1984	M	21532	...

ADMITTANCE DATE	BIRTH DATE	SEX	HOME ZIP	
1/10/2017	5/31/1984	M	21532	...

Figure 8. Bridging Data Set to Link Two Other Data Sets


DATA SETS
POLICE MOTOR VEHICLE CRASH RECORD
EMERGENCY MEDICAL SERVICES (EMS) RECORD
EMS RECORD BRIDGES BETWEEN CRASH AND HOSPITAL RECORDS
HOSPITAL DISCHARGE RECORDS

CRASH DATE	BIRTH DATE	SEX	HOME ZIP	
1/9/2017	5/31/1984	M	21532	...

CRASH DATE	BIRTH DATE	SEX	HOME ZIP	
1/9/2017	5/31/1984	M	21532	J. DOE

				J. DOE

4.2.1 Identify Data Set Options

The goal of the data set selection step is to determine if data set sharing is feasible and viable to support data linkage. This initial assessment of candidate data sets can usually be done without having to transfer any data sets. Data owners often have documentation about the data set (e.g., data dictionary), or technical staff can explain how the data sets are collected and used.

Table 5 provides a list of data sets that might be considered for data linkage, including benefits, limitations, and potential owners of each data set. For example, evaluating interventions could make use of all data sets identified. The availability of each state's data sets for data linkage, and the type of information within the data sets, might vary. Data set availability also changes over time as data collection and storage processes are digitized and as laws about data sharing change.

If data linkage is not currently conducted in your state, it would be worth reviewing and understanding the reasons why. Reviewing and updating policies, data use agreements, MOUs, etc., to meet current needs could help a state begin data linkage or enhance ongoing data linkage efforts.

Table 5. Potential Data Sets for Linkage Programs

Motor Vehicle Crashes (MVC)

Potential Data Sources	Benefits	Limitations	Potential Data Owners
Police crash reports	▸ Might describe how the MVC occurred based on eye-witness account(s), people involved in the crash, and other evidence at the scene. ▸ Might identify underlying risk and protective factors of the crash or injury. ▸ Might provide information on the vehicles and the people involved in a crash. ▸ Can be useful for linking the drivers' licensing information of multiple drivers involved in a crash, or the medical records of multiple people sustaining injuries, to a single crash incident.	▸ About half of the MVCs in the country are not reported to the police (these are typically minor property damage and/or non-injury crashes). ▸ Might be biased toward more-serious crashes.	▸ State or local police
Auto insurance carrier records	▸ Includes less-serious crashes and crashes that might not have been reported to the police.	▸ Might be more resource-intensive to obtain.	▸ Auto insurance organizations

Motor Vehicle Crash-Related Deaths

Potential Data Sources	Benefits	Limitations	Potential Data Owners
Police crash records	▸ Typically includes information about deaths that occurred at the scene of a crash.	▸ Might underestimate the true MVC-related death rate because some deaths occur later in time and away from the crash scene.	▸ State or local police
Medical examiner, coroner records or death certificates	▸ Captures the cause(s) of death. ▸ Can provide toxicology results.	▸ There is a data lag for vital statistics.	▸ Coroner or medical-examiner office ▸ State vital records
Health records (e.g., hospital or EMS)	▸ Captures MVC-related deaths that occurred at the site of the crash, during transport, or in the ED/hospital. ▸ Includes diagnoses, procedures, treatments, and length of stay related to ED/hospital encounter prior to death. ▸ Can provide information on related medical diagnoses, injury severity, and healthcare costs.	▸ Scope of information will depend on the source.	▸ Individual hospitals ▸ State EMS ▸ State health information exchange (HIE) ▸ State hospital association

Continued

Motor Vehicle Crash-Related Injuries

Potential Data Sources	Benefits	Limitations	Potential Data Owners
Police crash records	▸ Might provide injury severity level for each person involved in the crash. ▸ Can be useful for linking the drivers' licensing information of multiple drivers involved in a crash, or the medical records of multiple people sustaining injuries, to a single crash incident.	▸ About half of the MVCs in the country are not reported to the police [3]. ▸ Might be biased toward more-serious crashes.	▸ State or local police
EMS records	▸ Provides information about injury mechanisms (e.g., crush, burns, ejection from vehicle, drowning, inhalation injury). ▸ Helpful in linking crash data and other healthcare-related data because of the intersection between the crash location and the health record.	▸ Lacks comprehensive clinical assessment. ▸ Underestimates MVC injuries because not all MVC injuries are transported or treated by EMS.	▸ State EMS
Hospital emergency department records	▸ Provides definitive assessment of the scope and severity of MVC injuries. ▸ Might provide information on healthcare costs.	▸ Lacks information on the treatment of MVC injuries once the patient is discharged from the emergency department.	▸ Individual hospitals ▸ State hospital association ▸ State Health Information Exchange (HIE)
Hospital admission or discharge records	▸ Might identify MVC injuries after the initial assessment. ▸ Includes diagnoses, procedures, treatments, and length of stay related to hospital encounter to treat MVC injuries. ▸ Provides information on the patient's status (e.g., Did the condition worsen?). ▸ Might include the location to which the patient was discharged (e.g., long-term or post-acute care, home). ▸ Might provide information on healthcare costs.	▸ Might be formatted differently than the admission or discharge record. ▸ Underestimates MVC injuries because not all MVC injuries are seen in the emergency department or admitted to the hospital. ▸ Final functional status and the total cost of medical care related to treating the MVC injuries are not known at the time of record generation. ▸ Health diagnosis codes do not describe the degree of impairment associated with an MVC injury.	▸ Individual hospitals
Trauma registry records	▸ Includes detailed medical information about MVC injuries abstracted from hospital records. ▸ Might be more standardized (e.g., delimited files) than hospital billing records or medical records.	▸ Provides a select set of MVCs and does not provide a good baseline.	▸ Individual hospitals ▸ Hospital networks or systems

Continued

Potential Data Sources	Benefits	Limitations	Potential Data Owners
Health insurance records	▶ Contains records of all major medical care an individual receives while covered by insurance, regardless of the provider. ▶ Can provide a longitudinal view of the health treatments and costs associated with MVC injuries that result in permanent or short-term disabilities. ▶ Consistent coding across providers.	▶ Does not include self pay health encounters or uninsured individuals.	▶ Individual insurance providers ▶ State All-Payer Claims Database (APCD)
Disability records	▶ Assesses relationships between MVCs and short-term and long-term disabilities.	▶ Unknown/undetermined	▶ Unknown/undetermined

Driver Characteristics

Potential Data Sources	Benefits	Limitations	Potential Data Owners
Police crash records	▶ Might identify driver and occupant characteristics, such as age, sex, seating position, and restraint use.	▶ About half of the MVCs in the country are not reported to the police [3]. ▶ Might be biased toward more-serious crashes.	▶ State or local police
Police citations	▶ Provides information on high-risk behaviors leading to police intervention (e.g., history of reckless driving, impaired driving, history of non-use of seat belts, helmets, and car seats).	▶ Does not include information on high risk behaviors for which no citation exists. ▶ Missing information for individuals who have driving experience outside the state.	▶ State or local police
Toxicology reports	▶ Provides information on the presence and level of chemical substances (e.g., alcohol, drugs, prescription opioids).	▶ Tests, methods, results, and interpretations can vary substantially, depending on the entity conducting the test and the purpose of the test (e.g., as part of criminal proceedings).	▶ Criminal laboratories ▶ Individual hospitals ▶ State and local police ▶ State medical examiner or coroner
Autopsy records	▶ Provides information on the type and severity of injuries, toxicology results (e.g., alcohol, drugs, prescription opioids), and other medical diagnoses.	▶ Only includes MVC related fatalities.	▶ State medical examiner or coroner
State driver's licensing data	▶ Provides information about the driver's age, driver education, and years of driving experience.	▶ Missing information for individuals who have driving experience outside the state. ▶ Can be difficult to access and/or link to other data sets for legal and regulatory reasons.	▶ State motor vehicle administration
Citations, convictions, and legal penalties	▶ Provides information on individual's legal interventions.	▶ Might be legal limitations on linking with other data.	▶ Court system ▶ State police
State motor vehicle registration	▶ Provides the age and safety features of the individual's vehicle.	▶ Unknown/undetermined	▶ State motor vehicle administration

Continued

Environmental Characteristics

Potential Data Sources	Benefits	Limitations	Potential Data Owners
Police crash records	▸ Might provide characteristics such as time of day, light conditions, and roadway and traffic information.	▸ About half of the MVCs in the country are not reported to the police [3]. ▸ Might be biased toward more-serious crashes.	▸ State or local police
Roadway data	▸ Provides information on the design of the roadside environment, such as roadway inventories.	▸ Assigning exact locations to MVC data sets can be challenging.	▸ State DOT
Traffic data	▸ Provides information on the traffic patterns on that roadway. ▸ Provides contextual information on the traffic level at the time of the crash. ▸ Can enhance the quality of linkage by adding additional context to the linked data sets.	▸ Assigning exact locations to MVC data sets can be challenging.	▸ State DOT

Evaluate Interventions

Potential Data Sources	Benefits	Limitations	Potential Data Owners
Citations, convictions, and legal penalties	▸ Provides information on legal intervention efforts.	▸ Might be legal limitations on linking with other data.	▸ Court system ▸ State police
State motor vehicle registration	▸ Provides the age of the vehicle.	▸ Unknown/undetermined	▸ State motor vehicle administration

4.2.2 Understand Access, Storage, and Use Limitations

Understanding the access, storage, and use limitations of a candidate data set will assist in linkage planning. Answers to the following questions will inform an approach to accessing, storing, and using the data set:

- Where is the data set to be linked stored?
- Is the data set stored centrally or is it distributed?
- Is the data set hosted locally or on the cloud?
- Will the data set be hosted by the data owner(s) or data steward?
- What is the estimated storage capacity required to host the data set?
- Is the data set publicly available?
- Is there a fee to access the data set?
- What are the allowable uses of the data set?
- Is a data set sharing or use agreement (DUA) required? A sample DUA is provided in Appendix L.
- What security measures are required to access the data set (e.g., transfer and storage)?
- How frequently will the data set updates be shared?

4.2.3 Estimate Data Quality

Before a final decision to obtain a data set is made, an initial estimate of data quality should be considered. NHTSA's Model Performance Measures for State Traffic Records Systems identified six characteristics (timeliness, accuracy, completeness, uniformity, integration, accessibility) that can be used to describe and assess the quality of MVC-related data sets. Some of these characteristics are also useful for describing and assessing data set quality for data linkage purposes. During the initial review, data set characteristics that can be examined to assess data quality are data set accuracy, completeness, integration, and uniformity.

Accuracy. The accuracy of a data set refers to whether the stored values are correct. Data set accuracy problems might be caused by user errors during the data collection process (e.g., typos), which can be difficult to detect. A subset of data set accuracy is data corruption, which refers to errors that occur during the writing, reading, storage, or processing of the data set. An initial data set quality estimate of the potential for accuracy, including corruption, can be performed by interviewing staff and reviewing a sample of the data set or some aggregate measures of data set quality.

Completeness. Data set completeness refers to the amount of information that is missing from the data set. Completeness can address records missing from the data set and values missing from record fields. Issues with data set completeness might arise if data sharing constraints restrict the amount or types of data that can be obtained for data linkage purposes. Data set completeness issues can greatly increase the complexity of data linkage tasks.

Integration. Integration is the degree to which different data fields reflect the same concept within the same data set or across multiple data sets. A well-structured data set will not have overlap in the meaning of different data fields. If multiple fields have similar meanings, then it can lead to variability in the way that the fields are populated with records of a data set. There are times, however, when different data fields are needed to collect information about similar concepts, which might be confusing to whomever is performing the data collection (e.g., sex vs. gender, race vs. ethnicity). In these cases, the data owner should be able to provide documentation regarding good data hygiene practices (e.g., enumerated fields, data validation, data dictionary definitions of fields).

Uniformity. Data set uniformity refers to the consistency with which values are stored. For each data field that is potentially useful for data linkage, its uniformity can be estimated by understanding the data owner's record-keeping practices. Inconsistencies might arise due to different definitions of variables within and across organizations or to misalignment of data input design to data collection processes (e.g., not enough values for enumerated fields, poor labeling of data fields). Inconsistency within and across organizations in the way the values for an important data field are collected will lead to inaccurate analysis or data linkage errors if the data field serves as a linking variable [18]. For example, individual hospitals and even individual coders at the same hospital might vary in their medical coding practices. Adherence to data coding standards or use of a common data dictionary within and across organizations providing data sets to be linked can improve the accuracy of data pooled or linked across multiple data sets.

The data owner might be able to perform some rudimentary aggregate analyses of the data set (e.g., the percent of records that have values for each data field) or provide a sample of the data set for analysis. Table 6 lists questions that can be used when working with the data owner to identify any potential data set quality issues.

Table 6. Questions to Inform Considerations of Data Quality Issues

Categories of Data Quality to Consider	Sample Relevant Questions
Data collection process	▶ What sampling method was used? ▶ What is the data set collection process (e.g., recorded on paper and manually transcribed electronically, entirely electronic process)? ▶ Does the data set collection process incorporate any quality checks (e.g., dates of birth cannot occur in the future, certain fields are required)? ▶ Are values self-reported, expert-gathered, or automatically generated? ▶ What training is offered to data collection staff? ▶ What is the potential for duplicate records (e.g., the same person might have multiple crash records)?
Data input design	▶ Are data sets entered as free text or structured data (e.g., checkboxes, pick lists, enumerated fields)? ▶ What standard terminologies are used (e.g., medical diagnoses in hospital records are coded using International Classification of Diseases [ICD])? ▶ What default data field values are used? ▶ What quality or completeness controls or validations are used during data entry? ▶ What training is offered to data entry staff? ▶ What is the range of potential formats for each data field (e.g., numerical, free text, codes)? ▶ Do these data set entry fields align with the information collected? ▶ How well does the data set reflect the intended operational purposes of the data owner?
Data processing flow	▶ Is the data set abstracted from a primary source? ▶ How are data set quality issues addressed?

4.3 Develop Data Linkage Plan

A data linkage plan includes the selection of variable(s), data linkage method(s), data linkage tool(s), and an approach to organize and select the match results.

Choose variable(s). Variables are needed to effectively link data sets (see Glossary section). When deciding which variables to use, consideration should be given to available data linkage methods and tools and to data quality assessment (Section 4.4).

Select data linkage method(s). Methods include direct, deterministic, probabilistic, hybrid, or machine learning methods. There are two main data linkage methods used by existing data linkage tools: deterministic matching and probabilistic data linkage (Appendix E has more details). These methods differ in the criteria used for matching.

> Direct matching. Direct matching, the strictest form of deterministic matching, can be performed when there is a single unique identifier present (e.g., a social security number) that must be matched exactly for the data to be linked.

Deterministic matching. In deterministic or "rule-based" matching, the user creates or selects rules that identify record matches based on which variables in the records can be matched. For example, the rule could be: if age and sex and crash date match, then the records are considered a match and the data are linked. These rules might be based on the probability of matches in other sample data sets (e.g., common error lists), but they are not derived empirically from the frequencies of values observed in the data sets being matched.

Probabilistic matching. In probabilistic matching, pairs of records are assigned values between 0 and 1 which indicate the probability that the records match. The probabilities are computed based on the patterns of match variable values in the full data sets. For example, a system based on probabilistic matching would assign lower probability to records that match on a frequently occurring value for a variable (such as the last name "Smith" in a data set of U.S. origin) than to records that match on a value occurring infrequently in the data set. There are multiple algorithms that can be used for probabilistic matching including the seminal approach developed by Fellegi & Sunter [40].

Hybrid approach. Hybrid matching approaches can be developed that use different combinations of deterministic and probabilistic methods. For example, deterministic methods could be used to match records that could be confidently linked using rules, and a probabilistic approach could be used for the remaining records. Also, methods can be applied in sequence for different subsets of the data sets. For example, if there is a unique identifier in two data sets to be linked, but it is not always populated in every record, then the tool might first match the values of the unique identifier variable, as in direct linkage, followed by deterministic or probabilistic methods using different variables to link the subset of records that had missing, corrupted, or incomplete values for the unique identifier variable. Another record matching algorithm might compute match scores using several methods, with a final score determined by the values from all methods.

Machine learning approach. Recent machine learning matching, such as clustering and neural networks, has not been formally compared with deterministic and probabilistic approaches and has not yet been incorporated into available data linkage tools. Clustering techniques are motivated by the desire to increase the speed of performing data linkage and have shown promise in improving efficiency [41]. A data linkage model was trained by using neural networks and improved linkage accuracy over traditional methods, when applied to genealogical data [42]. Supervised machine learning approaches require a large amount of data to train the data linkage model and are not currently feasible for use in applications of data linkage.

Select data linkage tool(s). The data linkage tool and methodology might be tightly coupled with tools specializing in certain data linkage methods. For example, the software from the CODES program (i.e., LinkSolv) [43] specializes in probabilistic data linkage. Some data linkage tools allow the user to select among various data linkage methods to construct a customized approach. Many commercial data linkage tools use hybrid linkage approaches. Parameter configuration in the data linkage tool is informed by the decisions made concerning variables, methods, and tools. After completing a data linkage task and performing either a reassessment of the data quality or a validation of the match results, it might be necessary to revise various aspects of the data linkage plan, including applying different data linkage tools or using a different set of variables for matching records. Appendix F provides details on available tools and gives some information about how to select the appropriate tool.

Reduce computational complexity. As data sets grow, it becomes computationally intensive and impractical to compare all records to one another. Blocking and filtering are optional methods to reduce the number of records compared by selecting subsets of records that are more likely to match (Appendix M provides more detail).

Choose match results. Once data linkage is complete, a list of possible record matches and associated scores will be generated. For example, after probabilistic methods compute match probabilities, each record will be matched to many records with varying probabilities. One method to select the matched records that will be used for subsequent analyses is multiple imputation, in which multiple sets of matched records are analyzed and the results are combined using appropriate statistical analyses. Appendix N discusses this approach in greater detail. Another option is manual inspection of record pairs associated with lower probabilities to judge whether they should be included in the matched set for analyses. Manual review of matches can focus on potentially challenging matches present in the linked data, allowing the discovery of patterns of shared variables associated with matching (or non-matching) records. Modifying matching algorithms or parameters to account for these patterns might improve performance on difficult cases. Review of the match results permits the selection of true match record pairs. Section 4.9 and Appendix M discuss this in detail.

Different states have different policies on what PII (Section 3.4) and unique identifier variables can be obtained and used to perform data linkage. More informative variables, such as names and social security numbers, are increasingly protected by privacy legislation and often must be de-identified after linkage, when available. Table 7 shows the data linkage methods used by the states that participated in the listening sessions (Appendix D has more details). Reviewing the methods adopted by previous linkage programs can be useful when planning a new program or study.

Table 7. Data Linkage Method Used by State from Listening Sessions

State Linkage Program	Data Linkage Method	Data Linkage Tool
Injury Prevention Program of the Georgia Department of Public Health	Probabilistic linkage matching	CODES2000
Kentucky Injury Prevention and Research Center (KIPRC)	Probabilistic linkage matching and deterministic matching	LinkSolv, SAS
National Study Center for Trauma and Emergency Medical Systems (NSC) at the University of Maryland, Baltimore	Probabilistic linkage matching and hybrid deterministic-probabilistic data linkage	LinkSolv, SAS
UMassSafe: University of Massachusetts Traffic Safety Research Program	Hybrid deterministic-probabilistic data linkage	Custom-built software
Minnesota Department of Health – Injury and Violence Prevention Section	Probabilistic linkage matching and hierarchical/deterministic linkage between hospital/trauma/death data sets	LinkSolv, SAS
Bureau of Occupational Health and Injury Prevention, New York State Department of Health	Probabilistic linkage matching	LinkSolv
Intermountain Injury Control Research Center at the University of Utah School of Medicine	Probabilistic linkage matching	LinkSolv
Traffic Records Management, Reporting and Analysis Division of the Department of Motor Vehicles, Virginia Highway Safety Office (VAHSO)	Direct linkage matching	Custom-built software

4.4 Assess Data Quality

The purpose of a data set quality assessment is to determine the accuracy and "cleanliness" of the data set to be used in data linkage. The quality of any given variable within a data set must be sufficient across all data sets that will be linked by that variable to ensure valid results of the linking process.

The following questions should be considered when conducting a data set quality assessment:

- Which data set has the most complete or accurate variables needed for data linkage?
- Which data set has the required linkage variables that will be the easiest to isolate and standardize (when appropriate)?
- What is the feasibility of data linkage with other data sets that contain other variables of interest?
- What methods can be used to improve data set quality?
- Is the data set reliable—are the variables of interest reasonably complete and accurate, and does the data set meet the intended purpose?
- Is the data set valid—how well does the data set represent the real-world phenomena that it is designed to measure?

For example, the Model Minimum Uniform Crash Criteria (5th edition) and American National Standard Institute (ANSI) D.16 (8th edition) both use the updated KABCO injury scale, which is used by police to assess injured persons at crash scenes [9, 44]. The KABCO injury scale is categorical: "K" denotes fatal injury; "A" denotes suspected serious injury; "B" denotes suspected minor injury; "C" denotes possible injury; and "O" denotes no apparent injury [9, 44]. Prior to the 2017 update, "K" denoted fatal injury; "A" denoted incapacitating injury; "B" denoted non-incapacitating injury; "C" denoted possible injury; and "O" denoted no injury [45]. Generally, states collect data based on the updated KABCO scale as they make changes to their MVC data systems. However, the KABCO scale might not reflect the true injury severity as determined by medical professionals [12].

Variables of interest for the data linkage process should be assessed for the following data set quality issues:

Completeness vs. missing data. Data sets could be missing individual variable values for a variety of reasons, including poor data collection, corruption during data transfer, or low frequency of occurrence. If values within variables are frequently missing or erroneous, then the data linkage might fail to identify many true matches [3]. If missing or erroneous variable values are related to some mechanism, such as injury severity, then biases might be unintentionally introduced into the linked data. Appendix N describes approaches to assessing missing data. At a minimum, the percent of records that have values for each variable should be calculated. The following questions should be asked regarding missing and complete values:

- Will the populated linkage variable in the original data set provide an acceptable level of completeness in the final linked data?
- If the indicator (e.g., MVC fatalities per mile driven vs. MVC fatality per 1,000 people) is calculated based on information across several data fields, what is the completeness level of the calculated indicator?
- Are there biases or patterns in the missing values?

Duplicate records. When a data set has more than one record that represents a single entity or event, it is considered a duplicate record, unless multiple people are involved. Duplicate records might be intentional or unintentional; however, their existence might create multiple records that will be matched to one (i.e., many-to-one match) or more (i.e., many-to-many matches) records in another data set. The record-matching tool might be configured to handle these many-to-one or many-to-many matches.

Variations in values. When assessing the quality of data within each potentially useful variable, it is important to assess the variation in the values of the variables that will be used for linkage. Values such as names, addresses, and dates can be represented in multiple ways by using abbreviations and different orderings of the parts. Variables that are coded according to standard terminologies tend to have less variation, making data linkage easier and more robust. It is important to understand whether different data sets use different values in variables sharing the same concept. Through an understanding of the interaction between the data quality and the capabilities of the potential data linkage tool (e.g., specific algorithms used by the record-matching tools might be more susceptible to certain types of variation, leading to reduced match accuracy), it is possible to focus on normalizing variation. A variant taxonomy (the set of variations for a given variable) conceptually organizes and categorizes the types of variation within the values of a variable type, such as name, address, telephone number, or vehicle make or model. As described in Appendix O, variant taxonomy enables the measurement of the quantity and types of variation in the data sets. Variant taxonomies are also useful when building synthetic data sets as the ground truth, which are intended to mimic real-world data (evaluation using the ground truth allows for a calculation of performance metrics related to accuracy). See Appendix P for more on ground-truth data sets.

4.5 Prepare Data

There are many ways to improve data quality prior to data linkage. Even if the data owner provided metrics on each variable, it is critical to analyze the values of potentially useful variables. Best practice dictates that action should be taken to optimize the data quality of the lowest-quality data set first, and then to reassess the impact of the data quality improvement. If the data quality improved sufficiently to permit data linkage and indicator calculation, then data quality optimization should proceed with other data sets. If, however, the reassessment indicates that the data quality improvement was insufficient to permit data linkage or indicator calculation, then further improvements to data quality are necessary.

One such method, such as normalization, might be needed for the following purpose:

- To address differences in formatting that might prevent a computer algorithm from recognizing that values are the same.
- To standardize information through conversion to a finite set of coded values or a data content standard (see Appendix Q for examples of data content standards used in MVC related data sets).

EXAMPLES OF NORMALIZATION

Dates

Original Values	Normalized Values
January 24, 2017	01/24/2017
24-1-17	01/24/2017
1-24-17	01/24/2017

Social Security Numbers

Original Values	Normalized Values
123.45.6789	123-45-6789
123456789	123-45-6789

Avoid False Precision

Data Set 1: Time interval rounded to 15-minute intervals
Data Set 2: Time interval recorded at 1-second intervals

Data Set 1	Match/No match	Data Set 2
15 min.	Matches to	900 seconds
30 min.	Matches to	1,800 seconds
30 min.	No match to	1,905 seconds
60 min.	No match to	3,708 seconds

High missed match rate because of the precision measurements in Data Set 2. Rounding to the precision of the least precise measurement will avoid false precision.

For every new record, a new field is created that contains the normalized value of an original variable that had problematic values. For documentation purposes, it is better to preserve the original value of any variable that is normalized.

Caution should be used when normalizing data fields that contain identity information, as normalization might remove some useful information. For instance, it could be important that two data sets encode variables differently and normalizing the data sets would remove that helpful information and might reduce a data linkage method's effectiveness. For example, if in the crash data there is only first and last name, but in the medical data there is also the middle initial or full middle name, if you normalized the data to just first and last name, but then later wanted to link to a third data source such as toxicology, then if you didn't have the middle name/initial you might get a false positive match if someone else has the same first and last name in the data sources. If the data linkage tool has more-sophisticated data linkage methods, then it might be more effective to allow the data linkage tool to interpret the variables' variation.

Many data manipulation techniques are included as modules in common software packages. SAS, for instance, has modules that can convert values into the same case (e.g., all uppercase) and into the same format (e.g., 01SEP2013), and that can even alter the content (e.g., strip of all punctuation and digits). This works on a variety of element types:

- Dates of birth can be parsed into the month, day, and year of each birth.
- Addresses can be parsed into the street, city, state, and zip code.
- Names can be parsed into a variety of fields, including first name, middle name, and last name.

It is important to consider cultural and regional differences in the variable that is being normalized. Some languages, for example, do not always have analogues for concepts, such as the middle name, which can lead to data variation issues like those mentioned in the previous section.

Pre-processing procedures are designed to mitigate data variation and errors. After normalizing data, it might be useful to repeat data quality assessments, this time using the normalized data to determine if the updated data sets meet the necessary data quality standard. Additional data cleanup procedures can be used to improve data quality, including removing unrealistic values or "N/A," or removing invalid values (e.g., alphanumeric characters where only numeric characters are allowed). If the data quality of the normalized data improved sufficiently, then the data linkage process can proceed. It is possible, however, that extensive data errors can make the data sets unreliable for some information types; for example, if variables in MVC and hospital data sets are not compatible for linking the two files, then EMS records might be able to bridge the two data sets.

4.6 Perform Data Linkage

After the data quality has been determined to be adequate after pre-processing the data sets, the next step is to use one or more data linkage tools to link one or more data sets. Figure 9 demonstrates the basic components of a data linkage system, which includes the necessary data linkage tasks, the efficient sequence for those tasks, and a data linkage strategy to accomplish each task.

A data linkage task identifies the data sets and processes that the data linkage tool will perform to accomplish the deduplication or data linkage. Different data linkage tools might be used, depending on each tool's strengths and weaknesses, the complexity of the data sets involved in the data linkage task, and the available resources. To achieve the best results, different matching parameter configurations might be applied for different tasks (see Appendix F for tool parameters).

A Stepwise Approach for Data Linkage Tasks

In many states, data are available from all hospitals in the state, but the data have not been aggregated at the state level. Because patients receive multiple treatments from one hospital, and treatments from different hospitals, you must develop a state hospital file that can be linked to other data sets.

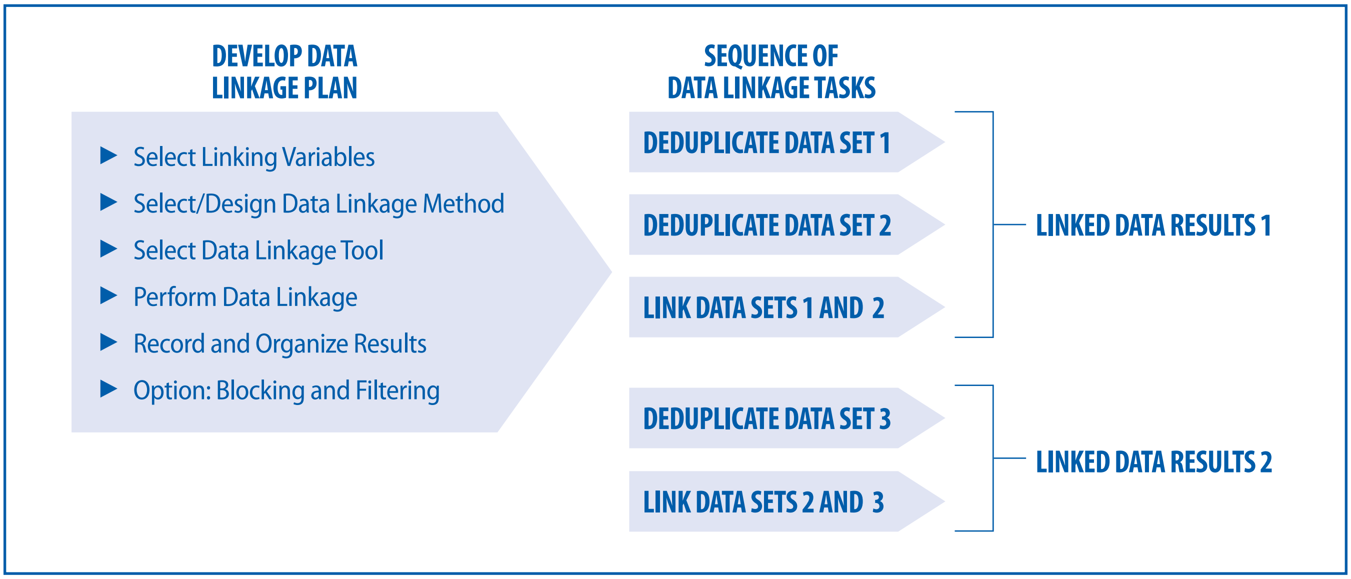
Your record matching tasks might include those listed below.

1. Deduplicate within each data set.
2. Consolidate deduplicated data sets to create one state hospital data set.
3. Link the new state hospital data set to the MVC data set.

Duplicate identification occurs during the data quality assessment and data processing steps. It might be necessary, however, to perform preliminary data linkage to estimate the potential number of duplicate records within a data set. This information will inform the decision to use data linkage to remove duplicates. Once potential duplicate records have been identified, they must be removed. Alternatively, with near-duplicate records, merging the records is more appropriate to preserve information when empty values in one record might have values in the near-duplicate record. After duplicates have been removed, it might be useful to reassess for data quality improvements.

If three or more data sets are being linked, then the sequence for linking each pair of data sets must be determined. One sequence might apply a "fail fast, fail often" approach by deduplicating and linking the lowest-quality data sets. This approach might identify obstacles that prompt the consideration of alternative data sets. Another sequence might take the approach of first deduplicating and linking the data set with many of the critical linkage variables. Yet, another sequence might first deduplicate and link data sets with similar content (e.g., data from multiple hospitals within one state) before linking with data sets from other content.

Figure 9. Sample Data Linkage System Linking Three Data Sets



4.7 Evaluate Linked Data

The evaluation of linked data quality mainly involves assessing the accuracy of individual match results from one or more data linkage tools and the consistency of match results over time. Accuracy relates to the tool correctly designating record pairs as matches or non-matches. Periodic evaluation determines whether data linkage strategies are not performing as expected and, therefore, should be modified. Linkage programs that monitor changes or trends over time by using data sets rely heavily on consistency in the data linkage process and its linked data to ensure the validity of comparisons over time. If the linked data quality is low, then it might be necessary to recalibrate (Section 4.8) by re-doing the pre-processing (Section 4.5) and data linkage steps (Section 4.6).

Several approaches may be used to assess the quality of the match results which are described in more detail in Appendix P:

- Manual inspection of a sample of matched records
- Manual inspection of the distribution of match scores
- External corroboration of match rates
- Comparison of multiple match results that used different data linkage strategies
- Evaluation with the ground truth (Section 4.9.1) to enable the calculation of performance metrics (Appendix P)

These approaches are not mutually exclusive. Tasks that should always be performed include the manual inspection of a sample of matched records and the distribution of match scores, when applicable. The calculation of match rates is another check that should always be performed. Only an evaluation using the ground truth, however, allows for a calculation of performance metrics related to accuracy. The evaluation approach that is selected must balance the accuracy of the data linkage with the available resources and time.

Regardless of which validation approach is selected, the degree of the effort spent on evaluation must be tailored to the amount of records to be linked, the data refresh rates (i.e., this is often annually for states), and the available computing resources. Listed below are several areas to focus on when assessing linked data quality.

Low accuracy of matches. Accuracy generally refers to the degree to which records that should be matched are matched (true positives), and records that should not be matched are not matched (true negatives). Inaccurate matching affects the data linkage results by undercounting events when records that should have been matched were not matched (false negatives) or by overcounting events, such as when duplicate records are retained, or records were matched that should not have been matched (false positives). Methods that employ approximate or probabilistic matching are especially susceptible to false matches (i.e., two or more records that should not have been matched were matched), but even exact matching can result in false matches. An example of a false match would be if the date of birth and the last name are used as linking variables for twins; this might lead to records being incorrectly matched to the wrong twin. A missed match occurs when a data linkage tool does not match two records that should have been matched.

Data sets change over time. If variable distributions change over time, then previously optimized data linkage parameters become outdated or no longer applicable, resulting in data drift. Other opportunities for data drift occur from changes in data characteristics as data sets are updated or as new types of variation are introduced (e.g., the racial and ethnic composition of the population changes over time, changing the types of names encountered and the variations in name types; population age distributions change as people live longer and have fewer children). Data drift might require additional data cleaning or normalization, as well as changes in the weights assigned to certain data fields. Some data linkage tools account for frequencies associated with specific data field values. These frequencies might change as the population changes; however, this drives the need for constant evaluation of the matching process.

Variables for analysis change over time. Changes in relationships between data fields might lead to linkage patterns that bias analyses. For example, motor vehicle injuries might be more likely to cause death in elderly victims, though this relationship will weaken if the elderly population grows healthier. Any data linkage system and analyses must account for these variable dependencies. Variable dependencies also pose a risk if analyses are based only on records with the highest match probabilities, as related variables tend to result in higher match scores. For instance, because more crashes involve younger people in urban areas, data linkage probabilities computed using age and county as linking variables will assign greater weight to records matching older people in rural counties because they are less common. Researchers selecting only the highest match scores will therefore analyze samples that over-represent incidents involving older people in rural areas [43]. This risk can be lowered by using multiple imputation to select the matched sets for analysis (see Appendix N).

Complexity reduction issues. Blocking and filtering (see Appendix M) are associated with risks because they remove records from consideration by selecting match candidates based on a single field or a few characters within a field. If, for example, a blocking method selects records that match the first letter of the surname, then records with typographical errors on the first letter of the surname will be removed from consideration. To mitigate this concern, multiple blocking strategies should be employed.

4.8 Recalibrate Methods

When match results are not satisfactory, every step of the data linkage strategy should be reviewed, and modifications should be explored. This might require that certain parts of the overall process need to be repeated. After the data linkage process has been recalibrated, new match results should be reassessed. The iterative process of validation and recalibration underscores the benefits in having a ground-truth evaluation data set, against which improvements can be measured and weaknesses can be identified. Periodic spot-checks of the data can be used to build that evaluation data set: Pairs that are spot-checked are flagged with the manually adjudicated match values, and those values can be reviewed with each iteration. Once the match results meet the expectations, the set of matched records can be selected. If, however, the match results do not align with the desired results after every option has been exhausted, it might be necessary to reconsider the data linkage goals.

Even if data linkage is only performed once, establishing a regular recalibration schedule is an important step because it is highly unlikely that the first attempt at linkage will capture the data well enough to satisfy every goal. With periodic data linkage, validation (to monitor the consistency of the match results) will identify if recalibration is necessary to accommodate data changes. If the evaluation has employed a ground-truth data set, then performance can be tested on the same data at each iteration, which is a strong motivation for investing in the production of a suitable evaluation data set. When the data linkage system uses a threshold to delineate the match status, and when match results have been stored, those results can be reviewed to determine if changing the threshold would result in more acceptable false-positive and recall levels.

Recalibration typically involves the iterative processing of one or more or the following steps, which were described previously:

- Assess data quality (Section 4.4): The conventions for encoding information should be reviewed to ensure that the field values are comparable across data sources.
- Prepare data (Section 4.5): More data cleaning or data standardization might be needed.
- Perform data linkage (Section 4.6): Data linkage recalibration tactics include adjustments in the threshold

used to identify acceptable matches, the tolerances for variable-level matches, or the field comparison weights; or switching data linkage methods or rules. If blocking is used, then it might be too restrictive; adding additional blocks might improve the percent of the correct matches returned by the algorithm recalibration tactics include adjustments in the threshold used to identify acceptable matches, the tolerances for variable-level matches, or the field comparison weights; or switching data linkage methods or rules. If blocking is used, then it might be too restrictive; adding additional blocks might improve the percent of the correct matches returned by the algorithm.

Data Linkage Recalibration Tips

- **Too many false positives:** Adjust threshold match scores higher.
- **Too many false negatives:** Adjust threshold match scores lower.
- **Weighted field has a high missing values rate:** Exclude it or assign a lower weight to it.
- **Broaden or narrow tolerance specifications** when matching variables with numerical values.

4.9 Select Linked Records

This section describes some decisions that might be required to create a subset of linked MVC data to be analyzed from the match results from each of the data linkage tasks. If an analysis uses only a single data set, then decisions about which records to include in the analysis must be made around data quality (e.g., excluding records with missing values). However, when an analysis uses linked data, the matching accuracy must also be considered (e.g., excluding records that have been incorrectly matched). The matching accuracy must be addressed first, before addressing the data quality issues, such as missing data, within the linked records. The next two subsections discuss selection of which matched record pairs (or "linked records") to consider true matches and false matches ("match status") and ways to address missing values in the data fields.

4.9.1 Determining Match Status

After a list of matched records is generated, it is important to select a subset of matched records that are likely to be true matches, as there will always be some uncertainty regarding whether the data linkage tool has made a true match. If a matching method is used that generates dichotomous match results (e.g., a record pair is either a match or not a match), then there is little choice but to select all the records matched by the data linkage tool, unless subsequent evaluation shows that the match status of some records should be changed. There are, however, unique challenges and options if the data linkage tool calculates continuous match scores for each record pair. In this case, there are different approaches to selecting a set of matched record pairs. A commonly used method is to select only those records with match scores above a specified value. Regardless of the record matching method used, for analyses of MVC data, the most important criterion for selecting sets of matched records is that the records are representative of the total population of linked records. The assumption that linked records are representative of the total population is essential for generalizing the results of studies based on the linked data. Section 4.7 provides an example of how selecting records with only high probability matches could result in data with disproportionate numbers of crashes involving older people in rural areas. Evaluation of matching results (see Appendix P) and multiple imputation of links (see Appendix N) reduce the risk of selecting unrepresentative sets of linked records.

4.9.2 Addressing Missing Data

After the data linkage process is completed and a set of matched records has been determined to match, the linked MVC data can be analyzed. Although records with missing values can be linked, missing data are problematic for analyses based on the linked records.

Complete Case Analysis. Traditionally, if data are missing from a field that is needed to calculate an indicator, then the record is excluded from the aggregate analysis. Complete case analysis excludes all cases that have one or more missing data points and is the default of many procedures in statistical software. Excluding records with missing values reduces the size of the data set and the statistical power; it also introduces a significant potential for bias. Two examples of this bias are (1) the date of birth might be missing more often for passengers than for drivers, and (2) dispositions for transferred or deceased patients are not available in hospital records [43]. When data are systematically missing, the exclusion of those records results in a sample that no longer reflects the population.

Imputation of Missing Data. Instead of excluding records due to missing data fields, imputation of missing data (not to be confused with imputation of match status, see Appendix N) can be used to populate missing values so that all records can be used in the analysis [43]. The purpose of missing data imputation is to maximize the predictive value of the data by enabling the inclusion of more records for analysis. Appendix N describes how to identify patterns and mechanisms of missing values. A review of the content will help the analyst determine if a method for missing data imputation is appropriate for the data set, and which method to use.

The methods to impute missing values in linked data can drastically change the results of the analysis of linked data. This is because the statistical inferences on which MVC data linkage rely require assumptions about the observed data that might be undermined by attempts to manage the missing data. Therefore, imputation of missing values should be used when the theory behind the methods and the design of the analyses are well understood. Imputing missing values might not be appropriate for all data linkage efforts. Imputing values is generally not appropriate for individual-level analyses but can be appropriate when doing aggregate analyses with multiple imputation.

4.10 Conduct/Use Analysis

Linkage programs often use existing data sets that are collected by states for operational purposes. MVC linked data are therefore used to support operational objectives. In addition, linked data can be used to support MVC analyses conducted by states or other stakeholders (i.e., researchers). The linkage program should develop objectives for data linkage and the variables that will be used to answer questions of interest. Objectives can include the questions that the MVC analysis is trying to answer. The more specificity about the design of the indicator (e.g., MVC fatalities per mile driven vs. MVC fatality per 1,000 people) will help ensure that data sources which have the data fields necessary to calculate that indicator can be obtained. The MVC analysis might require the use of many variables spanning the pre-crash, crash, and post crash phases, including characteristics of the crash, the short-term and long-term outcomes of MVCs, and the risk and protective factors for MVCs, injuries, and deaths.

It is important to develop a data linkage plan, even a preliminary one, to determine the level at which the data sets will need to be linked to support the analysis, or whether multiple data sets can be analyzed, but not linked. For example, if the MVC analysis requires an evaluation of each crash's impact, then it is not enough to link records for everyone; rather, the multiple individuals involved in the same crash must also be linked. If, instead, the objective of the MVC analysis is to understand how an individual's driving history is related to outcomes, then it is necessary to link multiple crash and/or citation incidents to one individual.

While data linkage will result in a more comprehensive understanding of MVCs and their outcomes, there are often limitations in using existing data sets for epidemiological studies. Listed below are some common limitations of epidemiological studies conducted using medical records and registries that apply to MVCs [46]:

- Missing necessary information (e.g., blood alcohol concentrations)
- Missing data quality information (e.g., methods to collect data are not recorded or described)

Because of these limitations, linked MVC data will not necessarily eliminate the need for large-scale, cohort studies that collect new data to answer specific questions or new and emerging issues that can eventually be captured in the MVC data sets. However, these studies are often expensive, time consuming, and have their own limitations.

Ideas for research questions can be identified from data users, state priorities, or existing MVC literature [47]. Table 8 lists possible contributing factors to the frequency and severity of MVC injuries, which could be integrated into the analysis of linked data [10].

Table 8. Factors Related to the Likelihood of Motor Vehicle Crash Injury [45]

Phases	Human Factors	Vehicle Factors	Physical and Social Environmental Factors
Pre-crash	Alcohol and/or drug impairment Fatigue Experience and judgment Driver vision Speed	Brakes, tires Center of gravity Jackknife tendency Ease of control Load weight Speed capability	Laws related to traffic safety Visibility of hazards Road curvature and gradient Surface coefficient of friction Divided highways, one-way streets, intersections, access control Signalization Speed limits
Crash	Seat belt use Age Sex	Speed at impact Vehicle size Vehicle safety features Hardness and sharpness of contact surfaces Load containment	Recovery areas Guardrails Characteristic of fixed objects Median barriers Roadside embankments
Post-crash	Age Physical condition Disabilities	Fuel system integrity	Emergency communication and transport systems Distance to, and quality of, medical services Rehabilitation programs

In summary, building a data linkage program from scratch can be a daunting, but rewarding task. This section describes a series of steps that can act as a roadmap during this process. The 10 steps that are shown in Figure 2 walk through the process from the very beginning, assuming no prior data linkage experience or program. It is also possible to skip ahead several steps if the linkage has been done previously (e.g., the Prepare Data step might be a good starting point). However, reviewing the initial steps might be worthwhile to see if new ideas emerge.

While performing the actual data linkage can be very technical, this section also covers less technical, but necessary, components of a successful linkage process. Defining clear and achievable goals, engaging data owners to build a trusting relationship, and planning the details of the data linkage process are important steps that pave the way toward a successful outcome. Understanding the research questions driving data linkage leads to a better understanding of the data required and provides a basis for conversations with data owners.

As covered in this section, remaining steps in the process go into detail about preparing the data, performing the linkages, ensuring that those linkages are accurate, and ultimately using the linked data to perform analysis. Understanding the data to be linked is very important prior to linking, as is recalibrating your process periodically after reviewing the linkage results. This portion of the process can be improved upon iteratively over time.

In conclusion, the utility of linked data for improving our understanding of MVCs is considerable, yet states face many challenges starting and at times sustaining data linkage programs. Overcoming these challenges enables states to efficiently use existing data in new ways. Linked data can be used to develop programmatic and policy recommendations by identifying the risk for MVC-related injuries among specific populations, the economic impacts of MVCs on populations, and the impacts of preventive interventions on MVC occurrences and MVC-related injuries and deaths. If linked data are used more for program monitoring and evaluation they can assist in strategic planning. States can monitor progress toward public health and transportation safety milestones and goals that can be enhanced by using linked data from stakeholders in a coordinated effort.

CONCLUSION

The utility of linked data for improving our understanding of MVCs is considerable, yet states face many challenges starting, and at times sustaining, data linkage programs. Overcoming these challenges enables states to efficiently use existing data in new ways. Linked data can be used to develop programmatic and policy recommendations by identifying the risk for MVC-related injuries among specific populations, the economic impacts of MVCs on populations, and the impacts of preventive interventions on MVC occurrences and MVC-related injuries and deaths. If linked data are used more for program monitoring and evaluation, they can assist in strategic planning. States can monitor progress toward public health and transportation safety milestones and goals that can be enhanced by using linked data from stakeholders in a coordinated effort.

APPENDICES

APPENDIX A. NATIONAL SYSTEMS FOR MOTOR VEHICLE CRASH DATA

Since the early 1970s, the National Highway Traffic Safety Administration (NHTSA) has collected crash data to support its mission to reduce motor vehicle crashes (MVCs), injuries, and deaths. Table 9 lists five national systems that collect MVC data to support data analysis; four are NHTSA systems and the fifth is a collaborative system that collects MVC-related injury data from hospital emergency departments.

Table 9. National Systems for Motor Vehicle Crash Data

Description	Federal Agency Name	National Program Name	Manually Link Data from Different Data Sets?
Census of all MVCs that occur on a public roadway and involve a fatality	National Highway Traffic Safety Administration (NHTSA)	Fatality Analysis Reporting System (FARS)	Yes
Sample of fatal, serious, and minor MVCs from police crash reports that involve at least one towed passenger vehicle	NHTSA	Crash Investigation Sampling System (CISS)	Yes
Sample of MVCs with fatalities, injuries, and property damage from police crash reports	NHTSA	Crash Report Sampling System (CRSS)	No
National database of emergency medical services patient care data resulting from an emergency 9-1-1 call for assistance	NHTSA	National Emergency Medical Services Information System (NEMSIS)	No
Sample of nonfatal MVCs that involve an emergency department visit, among injuries from other mechanisms. Data are weighted to provide national estimates	Centers for Disease Control and Prevention and United States Consumer Product Safety Commission	National Electronic Injury Surveillance System—All Injury Program (NEISS-AIP)	No

Fatality Analysis Reporting System (FARS)

The NHTSA's National Center for Statistics and Analysis (NCSA) operates the Fatality Analysis Reporting System (FARS), which contains data derived from a census of fatal traffic crashes within the 50 states, District of Columbia and Puerto Rico beginning in 1975 [48]. To be included in FARS, a crash must involve a motor vehicle traveling on a public traffic way and must involve the death of at least one person within 30 days of the crash (vehicle occupant or a non-motorist) [49]. Because FARS lacks data on crashes that do not involve a fatality, it is limited in its ability to support the analysis of nonfatal MVC injuries.

Each state employs FARS analysts to extract MVC data from source documents, code it, and enter it daily into NHSTA's central FARS database [48]. This ensures that all data sets from each state are standardized, complete, and consistent. The analysts obtain documents needed to complete the FARS forms including from police crash reports, state vehicle registration files, state driver licensing files, state highway department data, vital statistics, death certificates, coroner or medical examiner reports and emergency medical services reports [48]. FARS data are used extensively throughout NHTSA, state and local governments, research organizations, industry, Congress, the media, and the public. However, crash deaths are just the tip of the iceberg since more than three million people are nonfatally injured in MVCs each year in the United States.

Crash Investigation Sampling System (CISS)

The NCSA's Crash Investigation Sampling System (CISS) collects detailed crash data to help scientists and engineers analyze MVCs and injuries through a representative sample of minor, serious and fatal crashes [50]. CISS randomly selects thousands of cases from police-reported crash reports; to be eligible for the sample, a crash must involve at least one towed passenger vehicle [50]. Trained crash technicians obtain data from crash sites by documenting evidence from the scene such as skid marks, fluid spills, and struck objects [50]. They locate the vehicles involved, document the crash damage, and identify interior components that occupants made contact with during the crash [50]. On-site inspections are followed-up with confidential interviews of the crash victims and a review of medical records [50]. Collectively, information from these sources provides a detailed picture of a crash – from just before the crash through medical care received by the injured. CISS uses emerging technology and methods to acquire quality data; the de-identified data is available to other federal agencies, state and local governments, universities, research institutions, industry, and the general public [50].

Crash Report Sampling System (CRSS)

The NCSA's Crash Report Sampling System (CRSS) is used to estimate the overall crash picture, identify highway safety problem areas, measure trends, drive consumer information initiatives, and form the basis for cost and benefit analyses of highway safety initiatives and regulations [51]. Data are obtained from a sample selected from the estimated 5 to 6 million police-reported MVCs that occur annually [51]. To be eligible for the CRSS sample, a police crash report must be completed and reported to the state, must involve at least one motor vehicle traveling on a traffic way, and must result in property damage, injury, or death [51]. No other data are collected beyond what is contained in the selected police crash reports [51]. Approximately 120 linkage variables are coded into a common format electronic data file by trained personnel annually, which is made available to other federal agencies, state and local governments, universities, research institutions, industry, and the general public [51].

National Emergency Medical Services Information System (NEMSIS)

The National Emergency Medical Services Information System (NEMSIS) is operated by NHTSA's Office of EMS in collaboration with the University of Utah, which hosts the Technical Assistance Center [52]. NEMSIS is a national database that is used to store EMS data from states and territories. NEMSIS collects patient care information resulting from an emergency 9-1-1 call for assistance [52]. The mission of NEMSIS is to improve patient care through standardization, aggregation, and utilization of point-of-care EMS data at the local, state and national levels [52]. The 2017 NEMSIS Public-Release Research Dataset includes 7.9 million EMS activations submitted by over 4,000 EMS agencies serving 35 states and territories.

National Electronic Injury Surveillance System—All Injury Program (NEISS-AIP)

The United States Consumer Product Safety Commission's National Electronic Injury Surveillance System – All Injury Program (NEISS-AIP), is a collaborative effort with the Centers for Disease Control and Prevention (CDC), National Center for Injury Prevention and Control (NCIPC), since 2000 [53]. The NEISS-AIP estimates are based on the National Electronic Injury Surveillance System (NEISS) national probability sample of hospital emergency departments in states and territories [53]. NEISS-AIP collects data on nonfatal MVC injuries from a subsample of NEISS hospitals [53]. The CDC's Web-based Injury Statistics Query and Reporting System (WISQARS) online database uses data from NEISS-AIP to generate national estimates of nonfatal injuries, including MVC injuries [1].

Each of these national systems has strengths and limitations. Only FARS is representative at the state level because it is a census of all MVCs involving a fatality. FARS and CISS both have more comprehensive data about each MVC, including information from interviews and medical records, because each case includes an investigation by trained examiners, and analysts who code and input standardized data into the systems. Healthcare costs associated with MVCs are not included in any of the existing national systems.

APPENDIX B. LITERATURE REVIEW OF PUBLISHED MOTOR VEHICLE CRASH RESEARCH USING LINKED DATA

A literature review of motor vehicle crash (MVC) research publications by the states that participated in the National Highway Traffic Safety Administration's (NHTSA's) Crash Outcome Data Evaluation System (CODES) from 2009 to 2012 has previously been published and is available in Crash Outcome Data Evaluation System (CODES): Program Transition and Promising Practices [47]. For this guide, a literature review was conducted to build on the existing body of published research to identify MVC-related studies that used linked data. Two approaches were used including:

- A bibliometric analysis (which uses statistics to analyze trends in the literature) to provide a historical perspective of data linkage literature and to identify the major challenges and successes of data linkage research as of March 24, 2017.
- A PubMed search was performed to identify published literature from January 2013 to July 2017 to update the previous literature review. The updated literature identified in the review demonstrates the benefits and purposes of linking MVC data. A secondary benefit of the updated literature review was to identify methods and tools used for data linkage.

Articles identified through the PubMed search, prior CDC/NHTSA evaluations, and from key informants were also used to create these examples of previous questions that have been answered using linked MVC data with a short description included.

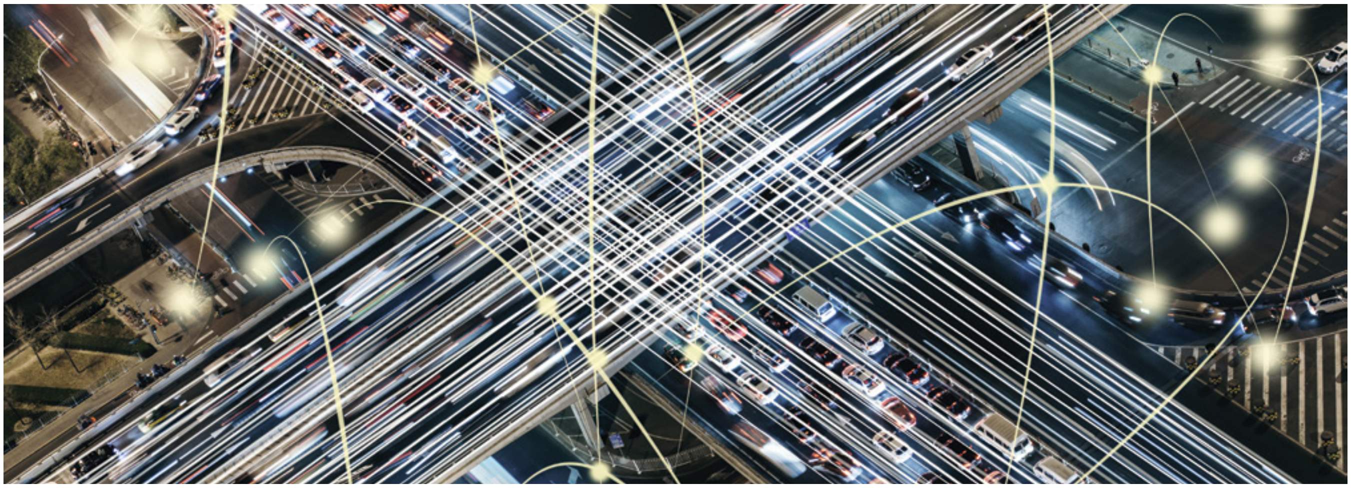
Examples of Questions Answered Using Linked Motor Vehicle Crash Data

Q: How are specific **medical conditions or characteristics associated with MVCs**? What are risk factors for MVCs among those with specific medical conditions?

A: Attention-deficit/hyperactivity disorder (ADHD). A study linking data from the New Jersey Department of Transportation crash database for police-reported crashes, with electronic health record (EHR) data from the Children's Hospital of Philadelphia outpatient clinics, showed that the adjusted risk for first crash among teen and young-adult licensed drivers with ADHD was 1.36 times higher than for those without ADHD, and did not vary by sex, licensing age, or over time. Through linking data, many limitations in previous studies were overcome [54].

Sedative hypnotic medications. An analysis of linked crash report, driver's license, and health plan data of about 600,000 patients served by a single integrated health delivery system in Washington state showed that new use of sedative hypnotic medications was associated with an increased MVC crash risk. The study results led to recommendations to educate prescribers and users of sedative hypnotic medications about the increased likelihood of MVCs [55].

Pregnancy. A North Carolina statewide analysis linked individual vital records with state crash records. The study showed a higher risk of crashes among pregnant women who were young (i.e., 18–24 years), black, moderately educated, unmarried, or used tobacco. The linked data in the study enabled the linkage of maternal characteristics with crash severity [56].



Q: How do **rates of MVCs or related outcomes** compare among different demographic groups?

A: Rates of MVCs. A statewide analysis of linked crash and licensing records in New Jersey showed that first-month MVC rates were higher among the youngest drivers licensed at 17 years, 0 months. Drivers who were licensed later experienced lower crash rates in the critical initial months of driving, but the benefit of later licensure plateaued once drivers had 6 months of driving experience [57]. A statewide analysis of linked crash and hospital records in Massachusetts studied whether age or driving at night increases young drivers' rate and severity of MVCs, so that lawmakers can design effective protective factors, such as graduated driver licensing policies [20, 28].

Aggressive driving. A statewide analysis of linked crash and hospital records in Ohio showed that teen drivers have higher rates of aggressive driving compared with other age groups and that teen drivers account for 30% of the total hospital charges for aggressive driving-related injuries [58].

Q: Among people involved in MVCs, which **populations have higher costs?**

A: Unbelted vehicle occupants. A statewide analysis of linked crash report and hospital discharge data from Nebraska showed that, compared with belted occupants, unbelted occupants had higher total medical charges primarily due to differences in the injury profiles between these two groups [59].

Inadequately restrained children. An 11-state analysis of linked crash reports, emergency department and hospital discharge reports showed that, compared with restrained children, inadequately restrained children are associated with increased hospital charges [26].

Q: Does being in a nonfatal MVC increase the risk of **other longer-term adverse health outcomes?**

A: Adverse pregnancy outcome. A statewide analysis of crash reports linked to vital records for North Carolina showed that, compared with pregnant women who were not involved in crashes, pregnant drivers involved in crashes had elevated rates of adverse pregnancy outcomes, and multiple crashes were associated with even higher rates of adverse pregnancy outcomes. Crashes were especially harmful if drivers were unbelted [60].

Q: What **interventions** can prevent MVC occurrences?

A: Road lighting. A statewide analysis of linked crash report and roadway inventory data in Washington state showed that one-sided lighting is associated with an increased occurrence of property damage, compared with both-sided lighting; 5–9 foot shoulder widths are associated with decreased occurrences of property damage and possible injury occurrences, compared with 2-foot right-shoulder widths; and five-lane and higher cross-sections are associated with an increased occurrence of property damage, compared with four-lane cross-sections [19].

Road curvature. A statewide analysis of linked crash report and roadway inventory data of Washington state showed that, compared with other roadway characteristics, the strongest predictor of motorcycle-to-barrier crash frequency was found to be the curve radius. Features that are likely to improve the effectiveness of a motorcycle-to-barrier crash countermeasure are roads with longer curves, roads with higher traffic volume, and roads that have no adjacent curved sections within 300 feet of either curve end [27].

Graduated driver licensing (GDL). A statewide analysis of linked crash report and hospital inpatient and emergency department records in North Carolina showed that, even after implementing GDL restrictions, teen crash rates are still substantially higher than adult rates. Further, the same risk factors for fatal crashes also exist for crashes with nonfatal injuries: speeding, having other teen passengers, and not using restraints [23]. A statewide analysis of linked crash report and automobile licensing data from New Jersey suggests that the use of decals to identify vehicles of young probationary drivers positively affects their safety and provides valuable information to U.S. and international policymakers who are considering adding decal laws [61].

Q: What **interventions** can prevent MVC deaths and injuries?

Motorcycle helmets. An analysis of linked crash report and a trauma registry data set showed that, among motorcyclists involved in MVCs treated at Level I trauma centers in Michigan, after the partial repeal of the Universal Motorcycle Helmet Law, although fatalities did not change overall, helmet use decreased in crashes and head injuries and neurosurgical intervention increased [25]. An 11-state analysis of linked crash report, emergency department and hospital discharge reports showed that, compared with motorcyclists involved in MVCs in partial helmet law states, motorcyclists involved in MVCs in states with universal helmet laws had lower rates of head, face, and brain injuries [62]. A statewide analysis of linked motorcycle crash data and emergency department and hospital records using probabilistic linking from Utah showed that a 42% reduction in median hospital charges was associated with motorcycle helmet usage [63]. A statewide analysis of linked police collision reports and hospital inpatient records in Kentucky showed that motorcycle helmet use was associated with a 69% reduction in skull fractures, 71% reduction in cerebral contusion, and 53% reduction in intracranial hemorrhage. This study found that current motorcycle helmets do not protect equally against all types of head injury [64].

Seat belts. A statewide analysis of linked crash report and hospital discharge data from Nebraska showed that, compared with unbelted vehicle occupants, seat belt use was associated with decreased rates of traumatic brain injury (10.4% no seat belt; 4.1% seat belt) and other head, face, and neck injuries (29.3% no seat belt; 16.6% seat belt), but increased rates of thoracic-to-coccyx spine injuries (17.9% no seat belt; 35.5% seat belt) [65].

Child restraints. An 11-state analysis of linked data from crash reports, emergency department reports, and hospital discharge reports showed that proper car seat, booster seat, and seat belt use among children in the back seat prevents injuries and deaths, as well as averts hospital charges [26].

Vehicle type. A statewide analysis of linked police reports and hospital discharge data in Minnesota determined that, in crashes between a car and truck, drivers of light trucks were less likely to be hospitalized relative to the drivers of cars. Passengers of light trucks also had a lower likelihood of hospitalization [66].

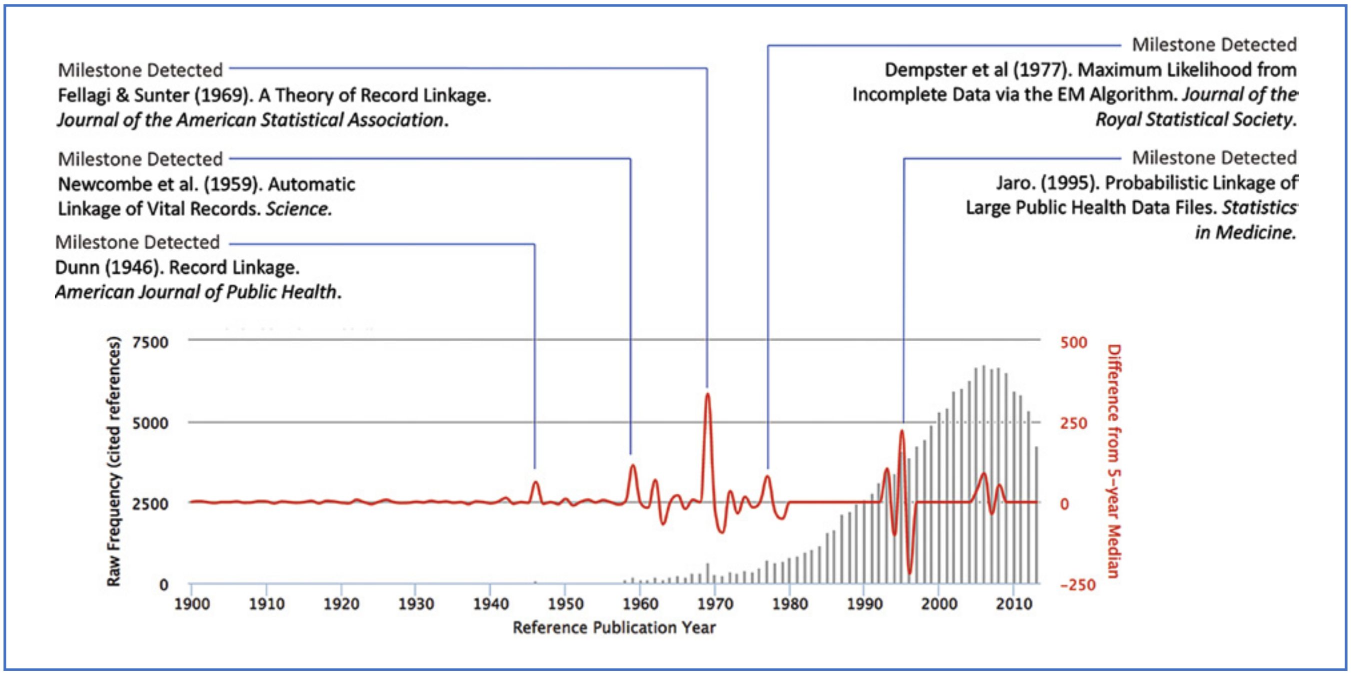
Bibliometric Analysis

A bibliometric analysis employed automated search and analytic methods over a large set of relevant scientific literature to create a broad overview of data linkage publications relevant to many domains, including, but not limited to, public health. The Reference Publication Year Spectroscopy (RPYS) bibliometric analysis method was used to automatically identify five milestone publications among the set of all publications related to data linkage. The RPYS algorithm defines milestone publications as those which were cited an unusually high number of times by other articles, given the year of publication. The RPYS algorithm has been successfully applied to numerous scientific fields [67, 68, 69], and has been shown to converge with independently-derived expert opinion on historical milestones for topics in biomedical research [70].

To generate a list of publications on data linkage for the RPYS algorithm to analyze, first, a search of the Web of Science (WoS) Core Collection scholarly database was performed. WoS contains metadata on more than 55 million scientific publications, proceedings, books, and editorials from the physical, life, and social sciences. To retrieve an appropriate set of publications related to data linkage, WoS was queried for scientific publications containing the term data linkage* (where * serves as a wildcard to represent any alphanumeric characters, as in data linkages), record linkage*, or record match*, within the title, abstract, or keywords to generate a list of scientific publications and their metadata, which is subsequently referred to as the "WoS data set."

Next, the RPYS algorithm was employed to mine the citations in the WoS data set to identify milestone publications in the record matching and data linkage. The RPYS algorithm bins cited references by their publication year and incorporates a normalization procedure to identify historical windows with unexpectedly high amounts of citation activity as compared with the 5 year median. The RPYS algorithm controls for the fact that, when a new idea is first published in the literature, the number of publications per year on that topic is very low, causing the number of articles that can be cited by a new publication to be lower. The results of the normalization procedure are then plotted against the reference's publication year, as demonstrated by the red line in Figure 10.

Figure 10. Historical Milestones in Data Linkage Literature



As of March 24, 2017, the search of WoS generated a WoS data set of 4,922 scientific publications related to data linkage and record matching. The 4,922 publications in the WoS data set contained over 90,000 cited references. The RPYS analysis of the over 90,000 cited references generated a list of 5 milestone publications, which are graphically depicted on a timeline in Figure 10. The red line in Figure 10 represents the results of the RPYS normalization procedure to detect milestones (right y-axis). When a peak is generated, the cited references underlying the peak were investigated. The absolute count of cited references binned by their publication year (x-axis) is shown in the gray bars (left y-axis). The absolute count of cited references in the most-recent years always drops from the peak because there is a time lag for recent publications to be included in a new publication.

Note that the year with the highest count of cited references occurs in 2006 and that cited reference counts decrease slightly from then onwards. This pattern of results is typical in RPYS analyses, as it takes several years for new scholarly work to become codified into the scientific community [70].

Based on the bibliographic analysis, the earliest milestone in the record matching and data linkage literature is Halbert Dunn's article "Data Linkage," which was published in *American Journal of Public Health* in 1946. Dunn, who served as Chief of the National Office of Vital Statistics for the U.S. Public Health Service, introduced the idea of a data linkage through a concept dubbed the "Book of Life." This notional Book of Life would record the key events of a person's life from birth to death. Dunn suggests that such an event log would provide large benefits both to individuals and to the public. Dunn notes that key events of a person's life are often recorded in disparate places, and, as such, there is a need to "link the various important records of a person's life"—thereby introducing the concept of data linkage across different data sets [71].

The next three most cited publications are focused on how to perform record matching to link data sets. The second most cited publication was the Newcombe et al. 1959 article in Science, which suggested a key role of computers in connecting disparate records on the same individual and in generating useful statistics, from a public health perspective, based on these linked records. These concepts were formalized by the milestone occurring in 1969, which introduced the mathematical models of computer-based data linkage derived by Fellegi and Sunter [72]. Subsequent milestones include the Dempster et al. publication that introduced a general algorithm for computing maximum likelihood in the face of incomplete data in 1977 [73]. The Dempster et al. algorithm was further advanced by the 1995 milestone associated with Matthew Jaro's methods for linking data sets under conditions of uncertainty [74]. These articles both represent a focus on probabilistic methods for data linkage to mitigate concerns associated with incomplete data or uncertainty.

PubMed Search and Analysis

A PubMed search was conducted to identify 61 publications using linked MVC data from January 1, 2013 to July 17, 2017, using the search terms ("link*") AND ("motor vehicle crash"). Of note, these search terms did not identify articles that focused primarily on outcomes other than MVCs, such as citations or licensing rates. Publications were excluded in the following order:

- If the MVC study focused on non-U.S. populations (Australia had the most publications compared with other countries) – 24 excluded leaving 37 articles.
- If the MVC study did not involve the analysis of linked data – 8 excluded so 29 articles.
- If the MVC study did not relate to MVCs– 1 excluded so 28 articles.
- If the MVC study used only the Fatality Analysis Reporting System (FARS) data set, and no data set including MVC injuries – 4 excluded so 24 articles.
- If the MVC study used only manual methods to link the data sets– 1 excluded so 23 articles.

Therefore, 38 publications were excluded. The MVC studies that used automated data linkage methods, even if the researchers did not actually perform the data linkage themselves as part of the study, were included. The 23 articles that met these criteria are shown in Table 10. Google Scholar was then used to identify the number of other publications that cited each of these 23 publications as of July 24, 2017.

The journals with the most publications were Traffic Injury Prevention and Accident Analysis & Prevention. Two notable studies involving 11 CODES states analyzed the states' combined populations [26, 62]. One study analyzed a nationally representative sample [75]. An analysis of the number of other publications that cited the 23 publications from the literature review revealed that 11 (48%) had two or fewer citations and five (21%) had no citations. Of the five articles with no citations, four (80%) were recently published in 2017 which might account for the lack of citations.

Table 10. Publications Using Linked Crash Data, January 2013 to July 2017

Publication	No. of Citing Articles	Population Studied	Study Objective	Data Linked	Findings
Bunn, T., et al. (2013). Concordance of motor vehicle crash, emergency department, and inpatient hospitalization data sets in the identification of drugs in injured drivers. *Traffic Injury Prevention*. 14(7), 680-689.	1	Drivers treated at Kentucky hospitals for MVC	Outcomes: hospitalization, costs, death Linkage methods	Crash; emergency department and inpatient hospital	Drug involvement in crashes recorded by law enforcement underreported the rate, compared with inpatient hospital drug testing. The most common drugs identified in the linked data were marijuana and cocaine. Opioids, amphetamines, and sedatives were less commonly detected.
Carter, P. M., et al. (2017). The Impact of Michigan's Partial Repeal of the Universal Motorcycle Helmet Law on Helmet Use, Fatalities, and Head Injuries. *American Journal of Public Health*. 107(1), 166-172.	1	Motorcyclists treated at Michigan Level I trauma centers for MVC	Effect of intervention on outcomes: helmet use, fatalities, head injuries, neurosurgeries	Crash; trauma registry	After the partial repeal of the Universal Motorcycle Helmet Law in Michigan, helmet use decreased in crashes (93.2% vs. 70.8%; P < .001), head injuries, neurosurgical interventions increased, and fatalities remained the same.
Conderino, S., et al. (2017). Linkage of traffic crash and hospitalization records with limited identifiers for enhanced public health surveillance. Accident Analysis & Prevention. 101, 117-123.	0	Drivers and occupants treated at New York City hospitals for MVC	Linkage methods	Crash; emergency department and inpatient hospital	From 2009–2013, there were 1,054,344 individuals involved in MVCs in New York City and 280,340 emergency department visits and hospitalizations from MVC-related injuries. There were 145,003 linked pairs, giving a linkage rate of 52% of the total MVC-related hospital records. The linkage had a sensitivity of 74% and a specificity of 93%. Linkage rates were comparable by age, sex, crash role, collision type, hospital county, injury location, hospital type, and hospital status, indicating no apparent biases in the match by these variables.
Conner, K. A. & Smith, G. A. (2014). The impact of aggressive driving-related injuries in Ohio, 2004–2009. *Journal of Safety Research*. 51, 23-31.	5	Aggressive drivers in Ohio involved in MVC	Risk factors for outcomes: costs, injury severity, death	Crash; hospital	Teen drivers have higher rates of aggressive driving and account for 30% of total hospital charges for aggressive driving-related injuries.
Conner, K. A. & Smith, G. A. (2017). An evaluation of the effect of Ohio's graduated driver licensing law on motor vehicle crashes and crash outcomes involving drivers 16 to 20 years of age. *Traffic Injury Prevention*. 18(4), 344-350.	0	Young drivers in Ohio and their occupants involved in MVC	Effect of intervention on outcomes: crash, injuries, death	Crash; emergency department and inpatient hospital; trauma registry	Compared with the pre-graduated driver licensing law period, the post-graduated driver licensing period was associated with lower crash, injury crash, and fatal crash involvement among drivers and occupants ages 16–17 years, but higher overall crash involvement for drivers and occupants ages 19 years, 20 years, and 18–20 years combined.

Continued

Publication	No. of Citing Articles	Population Studied	Study Objective	Data Linked	Findings
Curry, A. E., et al. (2017). Motor Vehicle Crash Risk Among Adolescents and Young Adults with Attention-Deficit/ Hyperactivity Disorder. *JAMA Pediatrics*. 171(8), 756-763.	0	Southern New Jersey young drivers treated at Children's Hospital of Philadelphia for MVC	Outcome: first crash	Hospital EHRs; driver's license; crash	The adjusted risk for first crash among teen and young-adult licensed drivers with attention-deficit/hyperactivity disorder (ADHD) treated at Children's Hospital of Philadelphia was 1.36 times higher than for those without ADHD, and did not vary by sex, licensing age, or over time.
Curry, A. E., et al. (2015). Young driver crash rates by licensing age, driving experience, and license phase. *Accident Analysis & Prevention*. 80, 243-250.	16	Young drivers in New Jersey involved in MVC	Risk factor on outcome: MVC	Crash; driver's license	First-month crash rates were higher among the youngest drivers licensed at 17 years, 0 months. Drivers who were licensed later experienced lower crash rates in the critical initial months of driving, but the benefit of later licensure plateaued once drivers had 6 months of driving experience. Further, at each age, those with more driving experience had lower crash rates; however, the benefit of increased experience was greatest for the proportion of teens licensed immediately after becoming eligible at 17. Finally, independent of age and experience, teen drivers' crash risk increased substantially at the point of transition to a full license, while drivers of a similar age who remained in the intermediate phase continued to experience a decline in crash rates.
Gabauer, D. J. & Li, X. (2015). Influence of horizontally curved roadway section characteristics on motorcycle-to-barrier crash frequency. *Accident Analysis & Prevention*. 77, 105-112.	4	Motorcyclists involved in crashes in Washington	Risk factors for outcome: motorcycle crash	Crash; roadway inventory	Compared with other roadway characteristics, the strongest predictor of motorcycle crash frequency was curve radius. Features that are likely to improve the effectiveness of a motorcycle-to-barrier crash countermeasure include longer curves, those with higher traffic volume, and those that have no adjacent curved sections within 300 feet of either curve end.
Han, G., Newmyer, A., & Qu, M. (2015). Seat belt use to save face: impact on drivers' body region and nature of injury in motor vehicle crashes. *Traffic Injury Prevention*. 16(6), 605-610.	2	Drivers hospitalized in Nebraska for MVC	Effect of intervention on outcomes: body region injured, nature of injury	Crash; hospital discharge	Seat belt use significantly reduced the proportions of traumatic brain injury (10.4% no seat belt; 4.1% seat belt) and other head, face, and neck injury (29.3% no seat belt; 16.6% seat belt) but increased the proportion of thoracic-to-coccyx spine injury (17.9% no seat belt; 35.5% seat belt). Although the proportion of thoracic-to-coccyx spine injury was increased in drivers with seat belt use, the severity of injury was decreased, such as fracture (22.0% no seat belt; 4.2% seat belt). Furthermore, the total medical charges decreased, due to the change of injury profiles in drivers with seat belt use, from a higher percentage of fractures (average cost per case: $26,352) to a higher percentage of sprains and/or strains (average cost per case: $1,897) with thoracic-to-coccyx spine injury.
Han, G., Newmyer, A., & Qu, M. (2017). Seatbelt use to save money: Impact on hospital costs of occupants who are involved in motor vehicle crashes. *International Emergency Nursing*. 31, 2-8.	3	Occupants hospitalized in Nebraska	Effect of intervention on outcome: costs	Crash; hospital discharge	Seat belt use is significantly associated with reduced hospital costs among injured MVC occupants, even when adjusting for relevant factors. Mean hospital costs were significantly lower among motor vehicle occupants using a lap-shoulder seat belt ($2,909), lap-only seat belt ($2,289), children's seat belt ($1,132), or booster seat belt ($1,473), when compared with those not using any type of seat belt ($7,099).

Continued

Publication	No. of Citing Articles	Population Studied	Study Objective	Data Linked	Findings
Hansen, R. N., et al. (2015). Sedative hypnotic medication use and the risk of motor vehicle crash. *American Journal of Public Health*. 105(8), e64-e69.	23	Washington members of Group Health Cooperative treated for MVC	Risk factors for outcome: MVC	Crash; health plan administrative, medical encounter and pharmacy records; driver's license	New use of sedative hypnotics is associated with increased MVC risk among members of the Group Health Cooperative integrated delivery system. Hazard ratios for three sedative drugs analyzed ranged from 1.27 to 1.91. The risk estimates are equivalent to blood alcohol concentration levels between 0.06% and 0.11%. Clinicians should advise patients of this driving risk and consider the length of treatment.
Karaca-Mandic, P. & Lee, J. (2014). Hospitalizations and fatalities in crashes with light trucks. *Traffic Injury Prevention*. 15(2), 165-171.	1	Drivers and passengers hospitalized in Minnesota for MVC	Risk factors for outcomes: hospitalization, costs, death	Crash; hospital discharge	Drivers and passengers of light trucks are less likely to be hospitalized or killed, and incurred lower hospitalization charges, than drivers and passengers of cars. Among hospitalized occupants, there were no differences in hospital charges between light truck drivers and car drivers, but hospital charges for hospitalized light truck passengers were 59% (95% CI, 40–87%) of the hospital charges of hospitalized car passengers.
Olsen, C. S., et al. (2016). Motorcycle helmet effectiveness in reducing head, face and brain injuries by state and helmet law. *Injury Epidemiology*. 3(1), 8.	6	Motorcyclists hospitalized in 11 CODES states: Connecticut, Georgia, Kentucky, Maryland, Minnesota, Missouri, Nebraska, New York, Ohio, South Carolina, and Utah	Effectiveness of intervention on outcomes: reported helmet use, head and face injuries, median hospital charges, payer type	Crash; emergency department and inpatient hospital discharge	MVC-related behavioral, medical, and economic outcomes were compared among 11 CODES states during the period of 2005–2008. Compared with the 5 states with universal laws requiring all motorcyclists to wear a helmet, the 6 states with partial laws requiring only a subset of motorcyclists to wear helmets have lower rates of reported helmet use (42% vs. 88%), 1.5 times higher rates of injuries to face, higher rates of severe traumatic brain injury (6% vs. 4%), and higher median hospital charges adjusted for inflation and differences in state income (emergency department $1,987 vs. $1,443, inpatient $31,506 vs. $25,949), and the payments were more likely to be borne by the driver or government vs. private health insurance.
Olsen, C. S., Thomas, A. M., & Cook, L. J. (2013). Hospital charges associated with motorcycle crash factors: a quantile regression analysis. *Injury Prevention*. 20(4), 276-280.	3	Motorcyclists treated in Utah for MVC	Risk factors for outcome: costs	Crash; emergency department and inpatient hospital	Motorcycle helmets were associated with reduced median hospital charges of $256 (42% reduction) and reduced 98th percentile of $32,390 (33% reduction). Helmets were associated with reductions in charges in all upper percentiles studied after adjusting for other factors.
Papa, L., et al. (2014). A method for linking motor vehicle victim and collision data collected by multiple county agencies. *Traffic Injury Prevention*. 15(1), 18-24.	0	Medical examiner cases due to MVC from one county in Florida	Risk factors for outcome: death <48 hours	Crash; EMS; trauma registry; medical examiner	Among MVC occupants who died, the most significant factors associated with early mortality (<48 hours) include hemothorax, liver injury, hypotension, cardiopulmonary resuscitation, involvement of drugs and/or alcohol, total fatalities, speed of vehicle, and number of lanes at the crash scene.

Continued

Publication	No. of Citing Articles	Population Studied	Study Objective	Data Linked	Findings
Ponicki, W. R., Gruenewald, P. J., & Remer, L. G. (2013). Spatial panel analyses of alcohol outlets and motor vehicle crashes in California: 1999–2008. *Accident Analysis & Prevention*. 55, 135-143.	25	California	Risk factors for outcome: MVC	Crash; alcohol license	Zip codes with higher restaurant outlet densities are positively related to total injury crash risks, and bar densities are positively related to the risk of crashes being alcohol-related.
Sauber-Schatz, E. K., Thomas, A. M., & Cook, L. J. (2015). Motor vehicle crashes, medical outcomes, and hospital charges among children aged 1–12 years—crash outcome data evaluation system, 11 states, 2005–2008. Centers for Disease Control and Prevention (CDC). *MMWR Surveillance Summaries*. 63(8), 1-32.	9	Hospitalized children in 11 CODES states: Connecticut, Georgia, Kentucky, Maryland, Minnesota, Missouri, Nebraska, New York, Ohio, South Carolina, and Utah	Risk factors for outcomes: child restraint use, costs, injury, injury type	Crash; hospital	Optimal restraint use in the back seat declined with child's age (1 year: 95.9%; 5 years: 95.4%; 7 years: 94.7%; 8 years: 77.4%, 10 years: 67.5%, 12 years: 54.7%). Unrestrained children were associated with unrestrained drivers (41.3% of children riding with unrestrained drivers were unrestrained vs. 2.2% of children riding with restrained drivers) and with impaired driving due to alcohol or drug use (16.4% of children riding with drivers suspected of alcohol or drug use were unrestrained vs. 2.9% of children riding with drivers not suspected of such use). Optimally restrained and sub optimally restrained children were less likely to sustain a traumatic brain injury than unrestrained children. Hospital charges for children who were optimally restrained were less than for those who were sub optimally restrained or unrestrained. Proper car seat, booster seat, and seat belt use among children prevents injuries and deaths, as well as averts hospital charges.
Singleton, M. D. (2017). Differential protective effects of motorcycle helmets against head injury. *Traffic Injury Prevention*. 18(4), 387-392.	0	Motorcyclists treated in Kentucky for MVC	Effect of intervention on outcomes: head injury, types of head injury	Crash; emergency department and inpatient hospital	Motorcycle helmets do not protect equally against all types of head injury. Motorcycle helmets were associated with a 69% reduction in skull fractures, 71% reduction in cerebral contusion, and 53% reduction in intracranial hemorrhage. Efforts to improve rotational acceleration management in helmets should be considered to reduce cerebral concussions.
Thomas, A. M., et al. (2014). The Utility of Imputed Matched Sets. *Methods of Information in Medicine*. 53(3), 186-194.	2	Treated in Utah hospitals for MVC	Linkage methods	Crash; hospital	When linking data sets for population-based analysis, the level of information available when selecting data linkage methods should be considered. High probability matched sets are suitable for high-to-moderate information settings, whereas imputed matched sets are preferable for low information settings.
Venkataraman, N., Ulfarsson, G. F., & Shankar, V. N. (2013). Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. *Accident Analysis & Prevention*. 59, 309-318.	65	MVC in Washington	Risk factors for outcomes: MVC, injury severity, property damage Statistical methods	Crash; roadway inventory	Among the roadway characteristics observed, one-sided lighting is associated with an increased occurrence of property damage, compared with both-sided lighting; 5–9 foot shoulder widths are associated with decreased occurrences of property damage and possible injury occurrences, compared with 2 foot right-shoulder widths; and five-lane and higher cross-sections are associated with an increased occurrence of property damage, compared with four-lane cross-sections. Significant improvement in fit when using random parameter negative binomial (RPNB) traffic crash frequency models, compared with fixed parameters.

Continued

Publication	No. of Citing Articles	Population Studied	Study Objective	Data Linked	Findings
Vladutiu, C. J., et al. (2013). Adverse pregnancy outcomes following motor vehicle crashes. *American Journal of Preventive Medicine*. 45(5), 629-636.	11	Pregnant drivers involved in MVC in North Carolina	Risk factors for outcome: adverse pregnancy outcome	Crash; vital records	Compared with pregnant women who were not involved in crashes, pregnant drivers involved in crashes had elevated rates of adverse pregnancy outcomes, and multiple crashes were associated with even higher rates of adverse pregnancy outcomes. Crashes were especially harmful if drivers were unbelted.
Vladutiu, C. J., et al. (2013). Pregnant driver-associated motor vehicle crashes in North Carolina, 2001–2008. *Accident Analysis & Prevention*. 55, 165-171.	6	Pregnant drivers involved in MVC in Washington	Risk factors for outcome: crash while pregnant	Crash; vital records	North Carolina has a relatively high pregnant driver crash rate (12.6 per 1,000 pregnant women) and a relatively high pregnant driver crash risk of 2.9%, compared with rates published in studies of Washington (2.8%) and Pennsylvania (1.1%). Pregnant women are at higher risk of being drivers in a crash if they are young, black, moderately educated, or unmarried, and are at lower risk of being drivers in a crash during the last few weeks of pregnancy.
Zhang, Y. & Lin, G. (2013). Disparity surveillance of nonfatal motor vehicle crash injuries. *Traffic Injury Prevention*. 14(7), 697-702.	1	Drivers in Nebraska involved in MVC	Risk factors for outcome: injury severity	Crash; driver's license	Black drivers had 31.6% and 87% more-severe and non-severe injuries, respectively, than white drivers in Nebraska. Rural residents were more likely than urban residents to have severe MVC injuries.

APPENDIX C. CRASH OUTCOME DATA EVALUATION SYSTEM (CODES)

For more than 20 years, linkage programs have existed in some states and longer in fields outside of the public health domain. There is a rich history of data linkage practice that includes tools, techniques, and publications about the process. Understanding the historical landscape and the current state of data linkage practice was integral to the development of the guide. Examples and insights from the National Highway Traffic Safety Administration's (NHTSA's) Crash Outcome Data Evaluation System (CODES)-related publications were incorporated throughout the guide. Below specific CODES-related publications are listed and mapped to the relevant sections of the guide for ease of reference.

Assessment of Characteristics of State Data Linkage Systems (2015). [28] A joint assessment between the Centers for Disease Control and Prevention and the National Highway Traffic Safety Administration to increase knowledge about state data linkage systems. Used to develop content for Section 2 and Section 3.

The Crash Outcome Data Evaluation System (CODES) and Applications to Improve Traffic Safety Decision-Making (2010). [29] Overview of the CODES program and data linkage methodology with details on participating states. Used to develop content for Section 2 and Section 3.

CODES: Program Transition and Promising Practices (2015). [47] Describes the CODES transition to state responsibility, and catalogs four years of state-linked data application. Used to develop content for Section 2 and Section 3.

CODES: An Examination of Methodologies and Multi-State Traffic Safety Applications (2015). [43] Shows technical work in CODES including topics related to probabilistic linkage and missing data imputation and demonstration projects pooling multiple states' CODES data for traffic-safety topics. Explains the development of a standard data model (General Use Model [GUM]) for mapping state-specific fields in motor vehicle crash (MVC) and hospital records to a standard format. Used to develop content for Section 4.

"EMS Data Linkage presented by Larry Cook in Virginia (2010)." [76] Utah CODES Project, University of Utah, Department of Pediatrics, presentation on applying record-matching methods to linking MVC data sets. Used to develop content for Section 4.

APPENDIX D. STAKEHOLDER LISTENING SESSIONS

Listening Session Methodology

Between December 2016 and May 2017, listening sessions were conducted with staff from five state linkage programs in person (Maryland, Georgia, Massachusetts, Virginia, and Kentucky) and two state linkage programs via conference call (Utah and New York). To prepare for these listening sessions, publicly available information about each state's linkage program, including their official website and their affiliated state and federal government stakeholders, was reviewed. Table 11 outlines the discussion topics and additional discussion probes from the stakeholder listening sessions.

Table 11. Stakeholder Listening Sessions Discussion Topics

Discussion Topic	Additional Discussion Points
State priorities	Does the program focus on motor vehicle crash (MVC) records and hospital records?
Main stakeholders	Who in state government is responsible for the data linkage, reporting and research? Do they have a board of directors? How is the linkage program funded?
Staff and training	How many staff are part of the program, what are their roles, and how are staff trained?
Data sets	What data sets are linked? What is their impact on linkage? What data are wanted but not linked?
Data linkage method(s)	What linkage methods have been tried, and what has been learned? What were the specific variables that were used to perform the linkage?
Validation of data linkage	How are matches reviewed? What are the review guidelines? What are the metrics and thresholds that are used to assess the quality of linkage?
Computing infrastructure	What software tools are used? What is the software system security?
Application of results	How are the data being used? Are the data being used in public health analysis, evaluation, surveillance or research?
Partnerships with states	Would it be useful to share data with a neighboring state? Has it been tried? Would a national program for MVC data linkage be useful?
Successes and best practices	What has worked? What has been learned?
Current challenges	What are the main challenges to increasing the scope of data linkage?

Listening Session Results

During the listening sessions, states expressed similar issues about implementing and maintaining a linkage program. States indicated that funding was an issue, either due to a lack of consistent funding or because applying for grants is a very time consuming process. Data linkage programs in Maryland, Massachusetts, Kentucky, and Utah are all located at universities (i.e., trusted third party model) and have access to staff and students with the expertise and skills that data linkage requires. Georgia and New York have linkage programs in the department of health, and Virginia's linkage program is in the Department of Motor Vehicles, Virginia Highway Safety Office. Details for the state programs listed are presented in Table 12.

Maryland has state-mandated, centralized data collection processes and the linkage program has been in existence for over 20 years with many of the same staff in program positions. For many states, data access is a challenge. Kentucky expressed difficulty in acquiring department of motor vehicle data. Utah shared that when the Health Insurance Portability and Accountability Act of 1996 (HIPAA) was enacted, agencies became more cautious with the types of data shared. Massachusetts does not currently link health data; Virginia does not currently have access to or link with emergency department or hospital data. Massachusetts cites a lack of clear statewide policies related to health data stewardship and usage. Virginia and Kentucky have laws that limit or restrict data access.

The listening sessions expanded to incorporate attendance at state Traffic Records Coordinating Committee (TRCC) meetings in Maryland and Georgia with staff from the respective linkage programs. The TRCC meetings led to requests to use state data and work collaboratively on an evaluation tool, which is described in Appendix P.

The states that took part in the listening sessions use different software in their linkage programs. Georgia uses the original CODES2000 software (Strategic Matching, Inc., Morrisonville, NY, USA). CODES2000 and LinkSolv software (Strategic Matching, Inc., Morrisonville, NY, USA) were the same software used when NHTSA funded CODES. Kentucky, Maryland, New York, and Utah use newer versions of the LinkSolv software. Both Kentucky and Maryland use SAS software (SAS Institute, Inc., Cary, NC, USA) and LinkSolv; Maryland also uses ArcGIS software (ESRI, Redlands, CA, USA). Finally, Massachusetts and Virginia use custom software. Table 13 indicates how each state uses each software package.

Table 12. Motor Vehicle Crash Data Linkage Program by State

GEORGIA

Georgia Department of Public Health (DPH), Atlanta, GA

- Point of Contact
  - Lisa Dawson: Director, Injury Prevention Program, Georgia DPH, 20 years, lddawson@dhr.state.ga.us
  - Denise Yeager: Program Manager, Injury Prevention Program, Georgia DPH, 20 years, dmyeager@dhr.state.ga.us
- Information current as of January 11, 2017
- URL: Georgia Department of Public Health

Area	Comments
Funding	▸ Annual grant from Highway Safety Office for traffic records program approved by executive Traffic Record Coordinating Committee (TRCC). ▸ Need to apply annually
Staffing	▸ Georgia has one full-time and one half-time position funded by Georgia TRCC grants for linking data, data requests, and providing data support for Strategic Highway Safety Plan (SHSP) task teams. ▸ The half-time position only includes Federal Insurance Contributions Act (FICA) tax. Turnover is high for this position. ▸ Ideally, the half-time person has epidemiology and coding experience.
Data and Linkage	▸ Georgia DPH links police crash reports with emergency department and hospital discharge records; emergency medical service (EMS) was previously linked but have not received statewide data. Trauma reports, Department of Driver Services (DDS) license and conviction records, and death records are also being considered. ▸ LongID (i.e., parts of first and last name, date of birth, and sex) is used for data linkage when possible, especially when name is not included. ▸ Georgia DPH prepares and cleans the linkage variables for data sets (e.g., name in the crash data), which can be time-consuming. ▸ CODES2000 used for data linking. ▸ GA DPH houses emergency department and hospital discharge data that it receives from Georgia Hospital Association (GHA); it also housed the EMS, Trauma, and vital records data.
State-Specific Information	▸ MVC and hospital data are provided as row-level data (one row of data per person), which has personally identifiable information (PII), and EMS trauma data are provided as aggregate data. ▸ Georgia's data linkage group meeting runs concurrently with the TRCC.
Partnerships and Organization	Georgia's standing data linkage group meets bimonthly. Stakeholders include: ▸ National: National Highway Transportation Safety Administration (NHTSA), Federal Highway Administration (FHWA). ▸ Georgia: DPH, Georgia Department of Transportation (GDOT), Georgia Administrative Office of the Courts (GA AOC), Department of Driver Services (DDS), Governor's Office of Highway Safety (GOHS). ▸ Local: Atlanta Regional Commission. ▸ Nonprofit: Emory University, SAFE KIDS, Brain and Spinal Cord Trust Fund Commission. ▸ Industry: Appriss and DEKRA. ▸ DPH includes representation from each data source (emergency department / hospital, EMS, trauma, and vital records) plus Injury Prevention Program.
Additional Comments	The Georgia DPH acted as the CODES representative for Georgia while the CODES program was active and continued to perform data linkage after the CODES program was fully funded by the Highway Safety Office.

KENTUCKY

Kentucky Injury Prevention and Research Center (KIPRC), Lexington, KY

- Point of Contact
  - Dr. Michael Singleton: University of Kentucky, Assistant Professor, Biostatistics, 20 years, msingle@uky.edu
- Information current as of March 10, 2017
- URL: Kentucky Injury Prevention and Research Center (KIPRC)

Area	Comments
Funding	▶ There are limited funding sources for core, ongoing record linkage activities. For example, 405c grants have a 1-year project period and can only be used for improvements to data quality, e.g. integration of previously unlinked data sets. ▶ Short-term grants make it challenging to train and retain core data linkage staff. This is exacerbated by the fact that record linkage is a skill set that is not typically taught in statistics programs. ▶ KIPRC is largely grant-funded, with a small amount of recurring state funding that covers essential core staff and activities (unrelated to record linkage). ▶ The Kentucky Office of Highway Safety has provided support for several data integration projects, through Section 405 grants.
Staffing	▶ Six faculty and 22 staff with the following roles: Administration Business Officer, Project or Program Manager, Principal Investigator, Epidemiologist, Statistician or Analyst, Case Investigator, Program Evaluator, Health Educator, and KDPH Liaison.
Data and Linkage	▶ Integrates traffic records data with emergency department and hospital billing records, trauma registry records, death records, highway information system, and FARS. ▶ KIPRC has not been able to obtain access to driver's licenses, vehicle registrations, or traffic citation/adjudication records for integration with police accident report records. ▶ KIPRC computing resources are secured behind the University firewall.
State-Specific Information	▶ Kentucky does not provide full PII in data sets, only the date of birth and zip code are used for most traffic record linkages. ▶ Kentucky has access to statewide EMS data, but collection was slow to develop, as they did not have a statute that required it until 2012. ▶ Kentucky state laws prevent charging fees for data analysis purposes.
Partnerships and Organization	▶ Being a trusted third party addresses the challenge of conflicting priorities and conflicts of interest. ▶ KIPRC provides study findings to several components of state government: KDPH, Transportation Cabinet, Governor's Executive Committee on Highway Safety, and Department of Aging. ▶ Other key partners for traffic records linkage include the Kentucky State Police, Kentucky Office of Health Policy, University of Kentucky Transportation Center, Kentucky Office of Highway Safety, Federal Highway Administration, Kentucky Trauma Advisory Committee and the Kentucky Board of Emergency Medical Services. ▶ Founder of Kentucky Safety and Prevention Alignment Network (KSPAN). ▶ KIPRC meets with state TRCC quarterly.
Additional Comments	Current staff ran the Kentucky CODES and have been doing data linkage for 20 years. KIPRC is a partnership between the KDPH and the University of Kentucky's College of Public Health.

MARYLAND

National Study Center for Trauma and Emergency Medical Systems (NSC) at the University of Maryland, Baltimore School of Medicine, Baltimore, MD

- Point of Contact
  - Cindy Birch: NSC, Program Manager, 15 years, cburch@som.umaryland.edu
  - Dr. Tim Kerns, NSC, Epidemiologist, 30 years
- Information current as of December 1, 2016
- URL: National Study Center for Trauma & Emergency Medical Systems (NSC)

Area	Comments
Funding	▶ Need to apply annually for funding opportunities. ▶ Diversifying funding sources and projects helped to ensure that the NSC staff could be sustained. ▶ Conflicting priorities between local and federal funding sources. ▶ NSC applies to Maryland TRCC for NHTSA 405c grants. ▶ Current grants or cooperative agreements from: Maryland Department of Health, Maryland Highway Safety Office, Centers for Disease Control and Prevention, National Highway Traffic Safety Administration. ▶ In addition to public sources, NSC has also received funding from private sources, including major corporations, and foundations.
Staffing	▶ Located at University of Maryland, Baltimore with onsite access to technical experts and university infrastructure, policies, and staff to ensure privacy and data security. ▶ NSC has epidemiologists, physicians, statisticians, and database coordinators on staff for their data linkage projects. ▶ Continuity in staffing at NSC and at state offices throughout the state of Maryland. This enables the development of a strong network of stable and long-standing relationships with other organizations, and a deep level of understanding of the data and its many applications. Most representatives participating in the state's Traffic Records Coordinating Committee. ▶ NSC staff performed data linkage on various health data sets prior to participating in CODES, which has evolved into a mature and robust infrastructure for linking and analyzing data that has been in continuous use for more than 20 years. ▶ With a larger project team, more members retain the data sharing information and institutional knowledge, thus stabilizing the program and mitigating the risk of staff turnover.
Data and Linkage	▶ Limited availability of unique identifiers for performing linking and analysis. ▶ Access to other data sets such as EMS records from Maryland's central repository and patient data from the Chesapeake Regional Information System for our Patients (CRISP) Health Information Exchange (HIE). ▶ CODES2000/LinkSolv used for probabilistic data linking. ▶ SAS is used for a hybrid deterministic-probabilistic data linking method that allows for the incorporation of weighting different fields in the data. ▶ ArcGIS is used to clean and validate location information contained in the data sets to be linked; the resulting data are de-identified prior to analysis or sharing.
State-Specific Information	▶ Maryland's data collection processes are highly centralized. For example, Maryland has access to data from all hospitals in the state. ▶ There has been unusual continuity in the staff in Maryland agencies. NSC understands the needs of their data owners. ▶ Maryland has mandated certain data sets to be stored in a central state repository. NSC hosts and works with Maryland state crash and medical data sets.
Partnerships and Organization	▶ Being a trusted third party addresses the challenge of conflicting priorities and conflicts of interest. ▶ NSC uses an Institutional Review Board (IRB), which outlines all security requirements related to accessing, handling, and sharing data sets. ▶ NSC meets with its data owners as a group annually. ▶ NSC meets with TRCC bimonthly.
Additional Comments	NSC has acted as the CODES representative for the state of Maryland since 1996 and collaborated with CDC and NHTSA to assess the CODES system in 2013. NSC's advice for implementing and achieving successful data linkage and subsequent analysis includes establishing data sharing and usage partnerships with large data owners, such as state agencies and hospital networks rather than state divisions and individual hospitals, to maximize the amount of data available and optimize time spent on creating such partnerships.

MASSACHUSETTS

UMassSafe Traffic Safety Research Program in the University of Massachusetts Transportation Center (UMTC), College of Engineering, Amherst, MA

- Point of Contact
    - Dr. Mike Knodler, Jr.: Director of University of Massachusetts Transportation Center (UMTC), mknodler@ecs.umass.edu
    - Robin Reissman: Deputy Director of UMTC; UMassSafe Founding Member
- Information current as of December 13, 2016
- URL: UMassSafe Traffic Safety Research Program (UMassSafe)

Area	Comments
Funding	▸ Lack of consistent funding, both types and availability vary from year-to-year. ▸ Acted as a data repository for multiple organizations and charged fees for service to link or analyze data in the past. ▸ Grants from Massachusetts TRCC for NHTSA 405c grants with the Center for Health Information Analysis (CHIA).
Staffing	▸ Located at University of Massachusetts, Amherst with access to on-site technical experts that help mitigate technical challenges. (e.g., computer scientist that created customized code). ▸ Twenty faculty and staff on the team, with 30 graduate students, most of whom are in transportation and transportation safety programs.
Data and Linkage	▸ Need a thorough understanding of the data sets. ▸ Taking steps to improve the data quality of reported data. ▸ Lack of access to health data. ▸ CODES2000 used for probabilistic linking in the past. ▸ Custom MATLAB and Structured Query Language code for a hybrid deterministic-probabilistic method aimed at incorporating fuzzy matching instead of multiple imputation of data fields.
State-Specific Information	▸ Lack of clear, statewide policies defining roles and responsibilities related to health data stewardship and usage.
Partnerships and Organization	▸ Being a trusted third party addresses the challenge of conflicting priorities and conflicts of interest. ▸ Has access to private insurance data (e.g., Arbella Insurance). ▸ Highlights the lack of a control entity (e.g., board of directors) that could establish priorities or champion causes. ▸ UMassSafe meets individually with their data owners, ranging from weekly to very seldom.
Additional Comments	UMassSafe was the CODES representative for the state of Massachusetts for 10 years. UMassSafe's primary focus is on transportation safety with other projects that only occasionally intersect with the health domain. UMassSafe has no experience with traffic-related health data sets. UMassSafe eventually transitioned the data set repository and associated responsibilities to a private contractor. UMassSafe would like to grow its current partner network to include health-related agencies and make the most of its university affiliation to expand its role as a trusted third party. UMassSafe's advice for implementing a nationwide framework for linking MVC data included making participation in the program mandatory for states, employing an advisory board of directors, and discontinuing the use of manual entry and paper forms in crash reporting.

MINNESOTA

Minnesota Department of Health, St. Paul, MN

- Point of Contact
  - Mark Kinde: Director, Injury and Violence Prevention Section, Health Promotion and Chronic Disease Division, Minnesota Department of Health, mark.kinde@state.mn.us
  - Anna Gaichas: Statistician / Data Scientist, Injury and Violence Prevention Section, Health Promotion and Chronic Disease Division, Minnesota Department of Health, anna.gaichas@state.mn.us
- Information current as of September 23, 2017.
- URL: Minnesota Department of Health, Injury and Violence Prevention

Area	Comments
Funding	▶ Traffic Records Coordinating Committee Section 405c Data Improvement Funding. ▶ CDC Core State Violence and Injury Prevention Program. ▶ MN State Traumatic Brain Injury Registry funding.
Staffing	▶ The statistician / data scientist is funded across multiple projects. ▶ The director, epidemiologist supervisor and other epidemiologists contribute time as available. ▶ New TRCC funding will help pay for much of the statistician's time, part of one epidemiologist and one entry-level research analyst.
Data and Linkage	▶ Minnesota has a new crash data system starting with 2016 data, which allows electronic reporting only. Human factors studies guided the design of the system to improve usability and data quality. ▶ The Minnesota Department of Health links police crash reports and hospital discharge records (inpatient and emergency department), trauma registry, traumatic brain and spinal cord injury registry and death certificates; emergency medical service (EMS) data will also be linked in the coming year. ▶ LinkSolv used for data linking (SAS used for hierarchical linkage combining hospital/trauma/death data and preparing crash and health data for linkage).
State-Specific Information	▶ Crash and some health data are provided as row-level data with personally identifiable information (PII); hospital discharge data do not contain PII.
Partnerships and Organization	▶ The MN CODES Board of Governors meets intermittently. Meetings occur regularly when a special TRCC project is in place. ▶ Member agencies: MN Departments of Health, Public Safety and Transportation, MN Hospital Association, MN EMS Regulatory Board.
Additional Comments	The Minnesota Department of Health acted as the CODES representative for Minnesota while the CODES program was active and continued to perform data linkage after the CODES program ended.

NEW YORK

New York State Department of Health

- Point of Contact
  - Michael Bauer: Section Chief, Epidemiology and Surveillance, michael.bauer@health.ny.gov
- Information current as of May 31, 2017
- URL: Bureau of Occupational Health and Injury Prevention, New York State Department of Health

Area	Comments
Funding	▸ Staff are funded by both state and grants to do work on injury, violence, and occupational health. ▸ Grant funding is a challenge. Currently have two grants: CDC injury program funds an epidemiologist (full-time) and Highway Safety Office funds two grants that included data linkage. Grant funding pays for a contract with LinkSolv for yearly updates.
Staffing	▸ Staff contingent on grant funding. ▸ Reduced funding causes staff turnover. ▸ One full-time epidemiologist (grant-funded). ▸ One full-time data scientist performs data linkage (grant-funded). ▸ New York looks for staff with the following skills: epidemiology, Master of Public Health (MPH), biostatistics, and statistics.
Data and Linkage	▸ New York data sets are maintained in the New York Department of Health (DOH). Access is only granted to those with data use agreements (DUAs). ▸ Has access to other data sets from: hospitals, emergency departments, EMS, DMV accident information system, police crash reports with over $1,000 in damages and injury. ▸ LinkSolv is used for probabilistic data linking. ▸ Waits for data to be confirmed final, and links only once per year. ▸ Linking EMS data has not been done since 2008 due to a major change from paper to electronic. The first complete year was 2015. Just bringing this into linked data.
State-Specific Information	▸ New York DOH provides base-level health summary statistics. ▸ The DOH does not share individual level health data; data must be de-identified before sharing. ▸ DOH only shares data with organizations with whom they have DUAs (e.g., EMS, DMV, SPARCS). ▸ DOH uses data for epidemiological studies.
Partnerships and Organization	▸ New York DOH meets with TRCC three times a year. ▸ CODES advisory board has not met in 2 years. ▸ New York DOH meets with other individual stakeholders frequently.
Additional Comments	New York receives data requests from academia, local county health departments, and traffic safety boards. When asked what type of national system would be beneficial, New York said that it depends on the activity. For example, trying to identify high-crash locations might need data that are delivered in a timelier fashion. The opioid crisis also requires more real-time data.

UTAH

Department of Pediatrics, University of Utah

- Point of Contact
  - Dr. Larry Cook: Department of Pediatrics for 20 years, Mathematics, PhD statistics, larry.cook@hsc.utah.edu
- Information current as of February 21, 2017
- URL: Intermountain Injury Control Research Center, University of Utah School of Medicine

Area	Comments
Funding	▶ Lack of consistent funding sources. ▶ Currently has funding from UDOT and UDPS.
Staffing	▶ Not discussed.
Data and Linkage	▶ Most difficult challenge of any data linkage study is acquiring the data. ▶ Utah's data linkage program links to numerous data sets (e.g., poison, burns, firearms). Data from crash to hospital (i.e., emergency department, inpatient, and discharge) are being used in the past. ▶ LinkSolv is used for probabilistic data linking.
State-Specific Information	▶ State mandated submission from hospitals and EMS agencies.
Partnerships and Organization	▶ After the state developed centralized state databases, the CODES board of directors decreased in size, and the meetings eventually stopped. ▶ Has had to rebuild stakeholder relationships.
Additional Comments	HIPAA has made agencies protective of their data.

VIRGINIA

Department of Motor Vehicles (DMV): Virginia Highway Safety Office (VAHSO), Richmond, VA

- Point of Contact
  - Angelisa Jennings: Deputy Director, DMV: VAHSO, angelisa.jennings@dmv.virginia.gov
  - Lam Phan: Data Manager, DMV: VAHSO, lam.phan@dmv.virginia.gov
- Information current as of January 25, 2017
- URL: Traffic Records Management, Reporting and Analysis Division of the Department of Motor Vehicles (DMV), Virginia Highway Safety Office (VAHSO)

Area	Comments
Funding	▶ Virginia has applied for and been awarded funds to develop, implement, maintain, and operate Traffic Records Electronic Data System (TREDS). ▶ Virginia uses both federal and state funding. ▶ Virginia shares IT resources with other partner agencies to implement data linkages.
Staffing	▶ Staffing has been relatively stable with no major turnover. ▶ Deputy Director, Data Manager and TREDS Operations Manager ▶ Data Analysts ▶ TREDS Operations Staff for Data Quality ▶ Fatality Analysis Reporting System (FARS) Analysts ▶ TREDS IT staff
Data and Linkage	▶ Data sources: police crash records, motorcycle student training, roadway location, driver history and convictions, driving under the influence (DUI) toxicology (only from the Office of the Chief Medical Examiner), Ignition Interlock System (Virginia Alcohol Safety Action Program), FARS, EMS, vital statistics, SafetyNet, DUI CheckPoint Strikeforce, and Click It or Ticket. ▶ TREDS transmits fatal crash record information to the FARS/EDT daily. ▶ TREDS transmits crash record information to National Highway Transportation Safety Administration's Crash Reporting Sampling System (CRSS) and Crash Investigation Sampling System (CISS). ▶ DMV: VAHSO partners with the Center for Geospatial Information Technology (CGIT) at Virginia Polytechnic Institute and State University (Virginia Tech) to locate all crashes on Virginia's roadways. TREDS interfaces with Virginia Tech's system to transmit bidirectional crash and location information. Through this process, Virginia Tech verifies and validates location data and transmits data to TREDS after completing the coding process. ▶ TREDS exports weekly commercial motor vehicle reportable crash data to Virginia State Police's SafetyNet system. ▶ TREDS exports weekly reportable medical-related crash data to Virginia DMV Medical Review. ▶ TREDS exports weekly reportable uninsured-related crash data to Virginia DMV the Uninsured Vehicle Group. ▶ Linkage is conducted on the name, date of birth, and crash date, crash document number, and driver license number. ▶ Custom-built software that performs direct data linkage.
State-Specific Information	▶ Virginia has strong partnerships with a multitude of agencies. ▶ DMV: VAHSO has access to specific sets of health data including EMS, vital records, and medical examiner records. ▶ Access and release of Virginia EMS, vital records, and medical examiner records are governed by memorandums of understanding. ▶ Legislation would be needed to allow DMV: VAHSO to have access to and store emergency department, hospital discharge and citation data.
Partnerships and Organization	▶ Strong TRCC and partnerships with various agencies including Virginia Departments of Motor Vehicles (i.e., crash, driver and vehicle), Transportation, State Police, Health (i.e., EMS, medical examiner, vital records) and Forensic Science; Virginia Tech; Supreme Court, Office of the Executive Secretary; Metropolitan Planning Commission; FARS; other highway safety advocates.
Additional Comments	None

Table 13. Software Used by State Motor Vehicle Crash Data Linkage Programs

State Program	Software Program Used for Linkage	Links Hospital Data
Injury Prevention Program (IPP) of the Georgia Department of Public Health	▸ CODES2000 used for data linkage.	Yes
Kentucky Injury Prevention and Research Center (KIPRC) at the University of Kentucky	▸ LinkSolv used for probabilistic data linkage. ▸ SAS used for data preprocessing and deterministic matching.	Yes
National Study Center for Trauma & Emergency Medical Systems (NSC) at the University of Maryland, Baltimore	▸ CODES2000/LinkSolv used for probabilistic data linkage. ▸ SAS used for a hybrid deterministic-probabilistic data linkage. ▸ ArcGIS used to clean and validate location information contained in the data sets to be linked. ▸ Resulting data are de-identified prior to analysis or sharing.	Yes
UMassSafe: University of Massachusetts Traffic Safety Research Program	▸ Custom MATLAB and Structured Query Language code for hybrid deterministic-probabilistic data linkage.	No
Bureau of Occupational Health and Injury Prevention, New York State Department of Health	▸ LinkSolv used for data linkage.	Yes
Intermountain Injury Control Research Center at the University of Utah School of Medicine	▸ LinkSolv used for data linkage.	Yes
Traffic Records Management, Reporting and Analysis Division of the Department of Motor Vehicles (DMV), Virginia Highway Safety Office (VAHSO)	▸ Custom-built software that performs direct data linkage.	No

APPENDIX E. SELECT DATA LINKAGE METHOD(S)

Automated data linkage is performed by comparing data sets using data linkage tools to determine whether data from multiple sources correspond to a single entity or different entities. Data linkage tools often perform multiple comparisons when matching records. Some comparisons determine the degree to which two values for the same variables match (i.e., variable match), while others combine variable-level comparisons to determine the degree to which two records match (i.e., record match) [77].

If only one variable is being compared when matching records, then a record pair for which that variable matches is considered by the tool to be a matched record pair. If multiple variables are being compared to determine a record match, then the tool must first perform all the relevant variable-level comparisons before deciding whether a potential record pair should match. The output of this comparison at the variable level can be a dichotomous variable-level match status (match vs. non-match) or a continuous variable-level match score.

Variable Matching

Comparison at the variable level can employ either exact matching or approximate matching methods. Exact matching is a deterministic method that requires that the values within a variable across two data sets are string matches. Exact matching generates dichotomous variable-level match results. If data sets contain a unique identifier, such as social security number, then it is possible to perform direct linkage based on exact matching of the identifier; however, even social security numbers might be missing or affected by typographical errors. The data quality assessment of the variable containing the unique identifier in all data sets will aid in deciding how reliable the unique identifier is for data linkage.

Approximate matching is an alternative to exact matching, in which values within a given variable can be matched when spelling variants, abbreviations, typographical errors, or other types of variation are present. Approximate matching can employ deterministic or probabilistic methods.

Deterministic variable matching. A common deterministic method that is based on probabilities calculated in data samples other than the MVC data set. Deterministic variable matching can use the following features:

Common-error lists. Lists of values judged to be equivalent due to common errors. The population that is used to derive the list must be representative of the population in the MVC data set.

Methods for approximate or fuzzy matching. Algorithms for computing the similarity between two values. Some data linkage software packages include multiple algorithms for determining approximate matches between names, numbers, and dates.

Rounding thresholds. Allowing rounding to help avoid false precision. Some data linkage software packages allow values to be rounded to a certain degree of information (e.g., time might be reported in 15-minute intervals if one source uses that as the minimum increment).

Similarity measures. Algorithms that are used to calculate a similarity measure comparing two values within a variable. The *Jaro* string comparator is commonly used in data linkage software; it computes the number of common characters in two strings, the lengths of both strings, and the number of transpositions to compute a similarity measure between 0 and 1. There are also many specialized similarity measures for different variables (e.g., names, numbers, dates, times, distances). However, similarity measures can lead to errors in matching variables. For example, 'Tammy' and 'Tomm'" might have a higher similarity measure than 'Thomas' and 'Tommy,' but 'Thomas' and 'Tommy' are more likely to be an actual match.

TIGER algorithm. Developed by the United States Postal Service, this algorithm uses the variable of address to match an address that contains typographical errors and variations in the street suffix (e.g., 'Lane,' 'Ln,' 'La') to a standard address within their database.

Probabilistic variable matching. Calculates a variable-level match score, or probability, for each match being compared. This variable-level match score reflects the probability that the variables are real matches by taking into consideration the reliability of a match on a rare value, compared with a match on a very common value, in the population. For example, matching the last name of 'Smith' has a lower variable-level match score than matching the last name of 'Jacobsen' because 'Smith' is the more common last name.

Data Linkage Methods

Exact matching with a single variable (e.g., unique identifier) to generate dichotomous match results for two data sets results in a variable-level match score equal to the record-level match results; this is commonly referred to as direct linkage. When multiple variables are compared, the data linkage method must provide a way to combine the results of matching the different variables for each record. Usually, better results can be achieved if some variables are assigned more weight than others when computing a final score. These weights can be determined by the user, independently of the distribution of values in the samples (deterministically); empirically by the tool, based on the distribution of values within each variable (probabilistically); or by a hybrid approach.

Deterministic data linkage. Deterministic data linkage (or "rule-based") algorithms specify a step-by-step procedure for weighting and combining variable-level comparisons into a record-level match score or a dichotomous match result. Deterministic data linkage methods that generate a dichotomous match result often combine the variable-level match scores into a single match score and specify a threshold match score as the criteria for a record match. Alternatively, the record match might be determined in terms of conjunctions and disjunctions of specified fields.

Probabilistic data linkage. Probabilistic data linkage algorithms construct a model that weights each variable based on the probability that matched records will match on that variable and the probability that non-matched records will match on that variable by chance. The matching algorithm calculates a variable weight for each of the variables, when multiple variables are being compared, to determine whether two records should be matched. The relative weighting of each variable will be used to generate a record-level match score by combining the variable-level match scores for a candidate record pair. The algorithm considers the distribution of data values for these variables in the data sets being linked to calculate the relative weighting of the variables with respect to each other. Thus, a variable is weighted higher when fewer records have the same value for that variable compared with another variable. For example, the variable weighting, from high to low, of the following is: date of birth, city, county, state. If the total score exceeds the tool's threshold, then the records are linked. Most tools will allow the user to adjust the threshold parameter. The match score that is generated from probabilistic data linkage is often referred to as the match probability. There are multiple algorithms that can be used for probabilistic matching. A seminal approach to probabilistic matching was developed by Fellegi & Sunter [72].

Hybrid data linkage. Deterministic and probabilistic methods can be combined in hybrid systems that leverage the advantages of both methods. The output of such hybrid methods is a score rather than probabilities. For example, a common approach is using deterministic matching to link records that can be matched with confidence using unique or good identifiers such as social security number or name plus date of birth. Then probabilistic matching is used for records with missing, variable, or unreliable values.

Pros and Cons of Data Linkage Methods

Selecting the best data linkage methods is dependent upon the characteristics and data quality issues of the data sets being linked.

Direct Linkage. The simplest approach to link two different data sets is the strictest form of deterministic matching, called direct linkage. Direct linkage can be performed when there is a *single* unique identifier, present in two or more data sets, that can be used (e.g., social security number) as the sole variable for matching records, such that exact matches on the values for that unique identifier can be made. The direct linkage approach critically depends on the accuracy of the data values for the single identifier used for linkage. The presence of a unique-identifier variable common to two or more data sets is not, by itself, enough to guarantee good results. Direct matching requires equivalent collection and coding of the unique-identifier variable across the data sets to be linked. If standards are not implemented during data collection and entry, then it might be necessary to develop data normalization processes to eliminate discrepancies between varying formats of the unique identifier to support system interoperability. Direct linkage is often used in interfaces between two data sets that allow them to seamlessly interact in near-real time. For example, "a police officer scans a driver license and is immediately provided with information regarding the driver and vehicle. Different hospitals under the same ownership may have their databases interfaced using a medical record number" [43].

Direct linkage is only appropriate when unique identifiers are reliably populated in all data sets to be linked. Another limitation of direct linkage methods is that, for many data linkage applications, including MVC data linkage, the data sets to be integrated do not have reliable unique-identifier variables in common. Direct matching interfaces between "MVC and healthcare data rarely exist due to the complex nature of data ownership and how the data are compiled. Often, the most informative attribute available is the social security number. This variable is typically not present in at least one of the sources, making direct linkage much more difficult to perform well" [43].

Deterministic Linkage. An advantage of deterministic matching systems is that they are typically easy to understand, and there are relatively obvious cause-and-effect relationships between the weights assigned to variables chosen for matching and the resulting linkages and scores produced by the systems. Thus, it is often possible to expose configuration parameters to users, thus providing a fair amount of control over the matching process. Deterministic matching is also less computationally intensive than probabilistic matching and is often used when data are being updated regularly, because, unlike probabilistic matching, deterministic rules do not require recalculating probabilities as the data change over time, producing results that are more stable when faced with a changing data environment.

One disadvantage of deterministic matching is that users set match weights, and the process can be somewhat arbitrary and potentially biased. This can also lead to inconsistency in applying the same algorithm (i.e., configured by different user) to the same data sets with different match weights leading to different match results. These problems do not occur with probabilistic matching systems, in which weights are determined empirically from the data sets. A second disadvantage of deterministic matching comes from assigning weights to evidence at the variable level, independent of the distribution of the values of those variables in the overall data sets. This results in giving equal weight to matches on variable values that appear frequently, and those that appear infrequently.

Probabilistic Linkage. Probabilistic matching approaches can be applied to data sets without requiring the user to employ much knowledge about the data sets to configure the matching algorithm. In contrast to deterministic approaches, probabilistic approaches automatically compute and optimize variable matching weights based on the actual data set on which they operate. Furthermore, the record- level match scores produced by probabilistic-matching approaches can be interpreted in terms of match probabilities, and methods such as multiple imputation can be used to select representative sets of linked records for analysis (see Appendix N).

Conversely, some of these advantages give rise to qualities of probabilistic matching that might be regarded as disadvantages. It can be said that the probabilistic-matching system appears to the user as a black box, without obvious cause-and-effect relationships between the system configuration and the system outputs. This can make it difficult to tune the system or to remedy problems of known false positives or false negatives in any straightforward way. Also, given the mathematical basis of probabilistic-matching systems, it is possible to configure (or tune) a system to perform extremely well on a data set with a certain composition. This can lead to overfitting of that data, resulting in a system that performs poorly when faced with phenomena not present in the original data sets. Such systems might be brittle as the data sets changes over time or if the characteristics of incoming data are unpredictable or do not mirror the characteristics of the data sets on which the system was trained.

APPENDIX F. SELECT DATA LINKAGE TOOLS[2]

The literature review for data linkage methods and tools secondarily identified many applications of data linkage tools for public health research that focused on using data linkage to link data sets, rather than identifying/merging duplicate records within a single data set. For example, data linkage tools can be used to create linked data sets for health surveillance systems based on disease registries [78]; for health indicators, such as maternal and child health outcomes [79]; or for studies of cohorts for environmental epidemiological research [80].

Data linkage is also widely used outside of the public health arena. Examples of applications of data linkage in other domains are listed below. The data linkage methods used in these domains are the same as those used in public health, but the data sources and motivations for matching records differ.

Healthcare. Data linkage can be used to reduce health information technology (IT)-related medical errors by matching the medical records of a single patient across different medical systems and identifying duplicate records for the same patient within a medical record database [81].

National defense. Data linkage can be used when screening airline passengers to compare a passenger's information to a list of known bad actors and to inform whether that individual should be allowed to board an aircraft [82].

Crime prevention. Data linkage can be used to help detect insurance fraud or identify criminal activities. For example, data linkage could be used to consolidate information from multiple casinos, and the linked data could be analyzed to find patterns related to money laundering [82].

Business. Data linkage can be used in customer record management by collating information from multiple sources to better understand a customer's shopping and buying habits [83]. Businesses can use this information to customize marketing campaigns and to tailor promotions to specific customers to optimize advertising spending.

To link data sets, a software tool that can read input data sets, perform data linkage, and output results is necessary. This appendix outlines a set of steps for selecting a tool, as there are many existing tools that can perform the data linkage, but operational constraints on resources might make some tools a better fit than others.

Many organizations have existing processes for selecting software that can guide the selection of a tool. If your organization does not have a process, however, following the guidelines in the remainder of this appendix can assist with choosing an effective tool.

Understand Data Linkage Tool Requirements

In Section 4.1, the first step in building a linkage program is defining the goals. Each organization performing data linkage has a unique set of requirements that vary based on several factors including size of the data sets, characteristics of the linkage variables, the types of data sets available for linkage, and computing resources. With this list of requirements in mind, an organization can identify and select from a set of potential software tools.

The program should build a list of requirements for each of the following categories to help in selecting a tool. Understanding functionality of the tool, data management capabilities, available budgets, licensing restrictions (e.g., can you use open source?), and computing hardware available will lead to a better selection of tool amongst the many options.

Functions: Tool functionality includes data linkage method and data pre-processing capabilities. What functionality does your program require?

Data management: The degree the software can handle data of all types. *Does your program have special data management needs a tool must support?*

Costs: Any costs related to the software and tool, including subscriptions and technical support. *What is your program budget?*

Licensing: Terms of use for software distribution (e.g., open-source or limited distribution) and installations (e.g., number of installations). Does your program have any restrictions on licensing, especially open source?

Hardware requirements: Software or operating-system requirements for use. *What computing resources does your program have or will it have?*

[2] Results from data linkage assessment of selected software packages were based on specific criteria. This was not an exhaustive assessment of all available software features, and is intended only to provide states with information about the findings, and does not represent an endorsement of any software product by CDC or MITRE.

Explore Data Linkage Tools

Once basic requirements are understood, the program can review existing tools. There will be commercial off-the-shelf (COTS), government off-the-shelf (GOTS), and open-source data linkage tools available. Ideas for discovering these tools include:

- Literature review of MVC data linkage publications—many have a list of software included.
- Publicly available material, such as marketing material provided for different data linkage tools.
- Consult with subject matter experts (SMEs) in data linkage and those with expertise in using data linkage for public health research.
- Industry reports (e.g., Forrester) on data linkage, data integration, identity or entity resolution, and related technologies.

A list of data linkage tools compiled using this approach is in Table 14. This table can be used as a starting point for your own tool selection process.

Table 14. Survey of Data Linkage Tools

Tool	Functions	Data Content	Costs	Licensing	Requirements	Notes
CODES2000	DL	N/A	N/A	Can install or use on single "primary" computer. Additional installations allowed under some circumstances.	Access and Windows	Uses Microsoft Access; software used in Crash Outcome Data Evaluation System (CODES) program.
DataMatch	M	N/A	$	No licensing information specified.	N/A	Claims to be as fast and accurate as more expensive tools (SAS and IBM); offers consulting and case studies.
Febrl	DL	N/A	Free	Mozilla Public License 1.1 (MPL 1.1) allows non-exclusive rights to use, reproduce, and modify.	Python	Febrl (Freely Extensible Biomedical Record Linkage) is an open-source GitHub project; has a graphical interface.
FRIL	DL	N/A	Free	No licensing information specified.	Java 6	FRIL (Fine-Grained Records Integration and Linkage Tool) maintained by Emory University, open-source tool that enables fast and easy data linkage—last update in 2011.
IBM InfoSphere	M	N/A	$	License Agreements.	SaaS	Product suite that includes temporal customer management; IBM QualityStage is the tool in InfoSphere.
Link Plus	DL	N/A	Free	No licensing information specified.	Windows	Link Plus is a probabilistic data linkage program developed at the Centers for Disease Control and Prevention's (CDC's) Division of Cancer Prevention and Control in support of CDC's National Program of Cancer Registries (NPCR).
LinkageWiz	DL	N/A	$	Each license covers installation on a single PC only.	Windows	Probabilistic data linkage tool; also handles data pre-processing, address standardization, and list management.
LinkSolv	DL	N/A	$	Can install or use on single "primary" computer. Additional installations allowed under some circumstances.	Access and Windows	Sister to CODES2000 software; same company, but with no requirement to have a National Highway Traffic Safety Administration (NHTSA) grant.
LiveRamp	M	Customer marketing data	$	No licensing information specified.	Access and Windows	Focus is on customer management for marketing purposes; suite of tools for pulling customers from multiple data channels.

Continued

Tool	Functions	Data Content	Costs	Licensing	Requirements	Notes
NextGen	M	Healthcare data (medical records and claims)	Free version and $	No licensing information specified for Mirth Match Enterprise Master Patient Index.	Java with dependencies	NextGen offers a large suite of tools for medical health record interoperability management; it is used by Maryland's Chesapeake Regional Information System for our Patients (CRISP) Health Information Exchange (HIE). Support is only offered for the paid (commercial) product.
NetOwl EntityMatcher	M	N/A	$	There are multiple different licensing models that are available with the NetOwl products. Please contact NetOwl to help determine the most appropriate model for your particular intended application(s).	Java with dependencies	Distinct from NetOwl's entity extraction tool Extractor; performs identity resolution based on any combination of available record variables; handles structured and unstructured data; winner of the MITRE Multicultural Name Matching Challenge.
OpenEMPI	DL	Patient registration records	Free	Open source solutions can use AGPLv3 license. Contact OpenEMPI to acquire commercial license for commercial solutions.	Java jar file	Deterministic and probabilistic; has dependencies on Postgres, Tomcat, and OrientDB for Windows.
OpenMRS	M	Medical records	Free	Mozilla Public License 2.0 with Healthcare Disclaimer (MPL 2.0 HD).	Mostly Java	Electronic medical record system that includes a patient matching module.
OYSTER	DL	N/A	Free	GNU General Public License version 2.0 (GPLv2) allows copying and distribution with correct copyright and warranty disclaimer.	Windows	OYSTER (Open sYSTem Entity Resolution) supports probabilistic direct matching, transitive linking, and asserted linking.
SAP	M	N/A	$	Access and use governed by software license agreement. If no license agreement accompanies software, it may be used for personal, informational, noncommercial purposes only; software may not be modified or altered in any way and may not be redistributed.	N/A	Full suite of analytic tools, with entity matching API service.
SAS	M	N/A	$	Terms and Conditions.	SAS license	SAS is a large analytical software suite that includes many statistical functions, many types of data cleaning and pre-processing functions, and customizable SAS code scripts that can fully automate these steps; performs both deterministic and probabilistic linkage.
SERF	DL	N/A	Free	BSD license allows distribution with correct copyright.	Java jar file	SERF (Stanford Entity Resolution Framework) employs the SWOOSH algorithm.
Syncsort	M	N/A	$	Services are for personal and non-commercial use only, unless otherwise specified. Terms of Use.	N/A	Formerly Trillium software, Large enterprise customer management system includes entity matching.

Continued

Tool	Functions	Data Content	Costs	Licensing	Requirements	Notes
STATA	M	N/A	$	Government purchasing prices available.	Windows, Mac, Linux	Data analysis and statistical software; similar features as SAS.
The Link King	DL	N/A	Free	No licensing information specified.	SAS base license v9.0	Deterministic and probabilistic; interface for manual review of matches; generates random samples for validation.
WinPure	M	N/A	$	No licensing information specified.	Windows	Data pre-processing and matching software suite; cost per license ranges (Clean & Match Lite).

Functions: DL = performs data linkage only, M = performs multiple functions, including data linkage
Cost: $ = cost to purchase software, a technical service, or a subscription
Licensing: includes any available information on the licensing model (purchase, software, technical service, and/or subscription)
Requirements: software or operation system required to use data linkage tool
Notes: additional information

These 38 data linkage tools identified in the survey were organized into two categories: data linkage specific tools and multi-purpose tools, as shown in Table 15.

Table 15. Data Linkage Specific Tools

Tool Category	Tool Name
Data Linkage Specific Tools	360º Science
	Black Oak Analytics
	ChoiceMaker
	CODES2000
	DataMatch
	Febrl
	FRIL
	LinkageWiz
	Link Plus
	LinkSolv
	OYSTER
	RLT-S
	SERF
	TAILOR: A Data Linkage ToolBox
	The Link King

Tool Category	Tool Name
Multipurpose Tools	Accumatch Property Tax Intelligence (mostly about taxes and escrow)
	AGIL ONE
	DataTools (Australian addresses)
	GBG Datacare
	HOPEWISER (UK/Australia product)
	IBM InfoSphere
	Intech Solutions (Australia/New Zealand product)
	Intellexer
	LiveRamp
	Manthan (business software, for sales)
	Mastersoft Harmony
	Matrix
	Mirth Match
	NetOwl EntityMatcher
	OpenEMPI
	OpenMRS
	Oracle and Datalogix
	Pacific East (mostly addresses, phone numbers)
	Pitney Bowes (mostly for addresses)
	R
	SAP
	SAS
	STATA
	Strategic Matching
	Syncsort

Select Data Linkage Tool

This section reviews the process for selecting a data linkage tool. After the initial survey of tools, there might be a long list of possible tools that needs to be shortened to something more manageable. Using the requirements generated initially, a quick review of the longer tool list can reveal many that would not work for linkage programs. For example, if there are cost considerations the most expensive tools would likely not be good options. After pruning the longer list, your program will end up with a shorter list that can be reviewed in more careful detail. Below is a description of each tool for such a shortened list.

Twelve (12) tools were identified as potentially useful for a linkage program because these tools are agnostic to data content and have sufficient documentation. These 12 tools include:

- **Free tools:** FRIL, The Link King, Link Plus, Febrl, and OYSTER.
- **Commercial tools:** LinkSolv, LinkageWiz, NetOwl EntityMatcher, SAS, SAP, IBM InfoSphere, and WinPure.

THESE 12 TOOLS ARE DESCRIBED IN DETAIL BELOW.

LinkSolv is the continuation of the original CODES2000 software (no longer available). LinkSolv is built by the same company, but, unlike CODES2000, there is no requirement for users to have an NHTSA grant. Like CODES2000, LinkSolv requires Microsoft Access and Windows. The software allows some data standardization of field formats. There is a cost for purchasing the LinkSolv software, however, compared with a data analytics suite, this would be a low cost.

LinkageWiz is commercial software that includes functionality for data pre-processing, matching, and deduplication. The cost is relatively low, with a current price that starts at $199, and there is a free trial available via the internet. The software is capable of different types of data pre-processing and standardization, deduplication, geocoding of addresses, and some simple data transformations like merge and concatenate. LinkageWiz uses probabilistic matching and advertises that it can handle up to 4-5 million records. LinkageWiz requires Windows. It does not have dependencies on other software, and thus runs standalone, with a professional-quality user interface.

Link Plus is a probabilistic data linkage program developed at CDC's Division of Cancer Prevention and Control in support of CDC's National Program of Cancer Registries (NCPR). Link Plus requires Windows. This tool can be used with any data in a standard format for import, even though it was originally designed for cancer data. Link Plus is freely available from the CDC.gov website in a stable-release (2.0) or beta (3.0) version. The probabilistic linkage is based on the framework by Fellegi & Sunter [72]. It is possible to do simple blocking and certain types of approximate matching of variables, including phonetic matching. Phonetic matching assigns high match scores to words that sound alike but are not spelled the same.

SAS is a large analytical software suite that includes many statistical functions and packages. A SAS license is required and comes with a higher cost. SAS can work on multiple operating systems, including Windows and Linux. SAS is capable of many types of data-cleaning and pre-processing steps that are applied before the data linkage process. Additionally, the user can write SAS code scripts that can fully automate these steps. It is possible to perform deterministic and probabilistic matching in SAS. There are several published papers available online that have example SAS code that show how SAS performs data linkage. Because of the versatility of the software, it can initially be more complex than other tools, resulting in a steep learning curve for users who are unfamiliar. SAS offers training and support that would be expected with an enterprise software solution.

SAP is similar to SAS. SAP offers a large analytic software suite that comes with a higher price and a lot of functionality. SAP can work on Windows or Mac operating systems. There are many built-in functions for data cleaning and pre-processing, as well as approximate matching capabilities. Within SAP, there is a tool called Data Cleaning Advisor that allows for both data cleaning and data linkage, as well as an entity-matching API service. There are advanced graphical interfaces for the SAP tools as well. Like SAS, SAP offers the training and support that would be expected from a widely used commercial product.

IBM InfoSphere: IBM QualityStage is an "enterprise-level solution" with a large amount of functionality and support. The cost is on the high end. IBM QualityStage is the data linkage tool in IBM InfoSphere suite. IBM QualityStage offers what IBM terms "entity analytics." The suite of tools can run on IBM public or private cloud and use other IBM InfoSphere components, such as big data and analytic tools. The QualityStage component of InfoSphere includes a probabilistic data linkage system.

WinPure is a data pre-processing and data linkage software suite that requires Windows. WinPure offers several products with varying degrees of functionality. Some versions are free, while others have licensing fees; however, overall, the costs could be expected to be lower than enterprise tools like SAS, SAP, and IBM. The software offers data pre-processing tools and data matching. WinPure claims to have fuzzy-matching capabilities to aid with finding duplicates. Additionally, there are built-in reporting and charting capabilities. The free and trial versions are available to download via the internet.

FRIL is an open-source tool that enables fast and easy data linkage by Emory University. The tool was last updated in

2011. FRIL is free to download and requires Java 6 (the link for download is available on the website). FRIL has limited data-cleaning and pre-processing abilities. When importing data files, the deduplication and filtering of records are possible. Also, the data fields can be merged or cast into a particular format. The tool includes algorithms for determining the edit distance between names, numbers, and dates to measure the similarity of two values within these variables. In addition, to an exact string match and a simple string comparator value, FRIL includes Soundex, Jaro-Winkler distance, or Q-Gram distance, all of which are options for determining if two names are an approximate match. The user can edit the variable weight, which determines the likelihood that pairs are judged matches. Users can adjust the value of the cutoff threshold based on the results that they receive, allowing for more-precise or more-sensitive results.

The Link King is available for free download on their website and requires SAS (base license v9.0). The Link King can perform deterministic and probabilistic matching. The tool includes pre-composed collections that offer a one-step solution to decide point values for deterministic matching, as well as different levels of intensity of blocking for probabilistic matching. This tool also includes nickname or common-error lists by default and allows for the addition and subtraction of items from the lists, so that a user can update the provided list to include variation from their specific data sources. There does not appear to be much pre-processing or data cleaning available during data import.

Febrl is an open-source GitHub project that has a graphical interface and requires Python. This software was developed in Australia and is freely downloadable. It uses the Mozilla Public License 1.1 (MPL 1.1). It is possible to do data cleaning and pre-processing by using Febrl, as well as data linkage with fuzzy matches.

OYSTER is a free-entity resolution system that supports probabilistic direct matching, transitive linking, and asserted linking. OYSTER uses the GNU General Public License version 2.0 (GPLv2) and requires Java. The functionality that OYSTER supports includes probabilistic and direct matching. Additionally, this tool can perform blocking and has a built-in entity management system that forces entities in the system to maintain uniqueness (with identifiers that cannot be matched). It is also possible to fix false-positive and false-negative linkages.

NetOwl EntityMatcher is an "intelligent identity resolution software" that performs matches "based on any combination of available record variables by utilizing its unique proprietary search and indexing engine that allows combination of evidence from multiple matching variables in a highly robust, scalable, and intuitive fashion. "This tool uses NetOwl's NameMatcher to enable multi-cultural name matching across different languages to perform scalable high-accuracy identity resolution. In addition to names, it can use variables, such as the date of birth, place of birth, and nationality, as well as social-network information, such as employer, spouse, and associate, to match an individual's identity across records. In addition to matching an individual's identity, it can also match places, and addresses. NetOwl EntityMatcher can be used in the domains of electronic medical record; targeting fraud, waste, and abuse (e.g., Medicare or Medicaid, tax, insurance); homeland security (e.g., terrorist watch lists, visa screening); law enforcement; regulatory compliance; anti-money laundering; customer relationship management; and master data management.

COMPARE A SUBSET OF TOOLS

States have significantly different resources, management structures, and data sets. The variety of circumstances poses a challenge in performing an evaluation of tools and methods. While it is possible to consult with other data linkage programs and to review software evaluations others have performed, it might be necessary to perform an in-house analysis of the short list of tools to make a final selection.

After careful examination of candidate data sources and the development of a plan, schedule, and strategy, the software required to implement a data linkage system must be selected. The system should have the following basic functionality to accomplish the tasks required for data linkage:

Database software with appropriate access controls. Database software often includes some capabilities for data characterization, cleaning, normalization, and indexing, though these might not be sufficient for data linkage.

Capabilities for data cleaning and normalization. The software should handle a broad range of data types (e.g., names, addresses, dates) and many types of variation or errors. It should be easy to implement cleaning and normalization routines that are not already available in the package. Tools to facilitate a comprehensive characterization of the variation in the data are desirable.

Capabilities for indexing, blocking, and filtering. (See Appendix M) Capabilities for indexing are usually included in database software, but some data linkage software includes built-in functions for indices commonly used in blocking methods, such as Soundex.

Data linkage and deduplication capabilities. Capabilities for each step of the data linkage process are needed:

- Tools to select fields for comparison, comparison functions, and parameters, should be easy to use. Resources, such as approximate string comparison functions, lists of nicknames, or frequency look-up tables, are desirable.
- Deterministic data linkage rules, and the assignment of

weights to results of the comparison functions, should be easy to specify and should be flexible.

- Tools for analyzing and visualizing the models and probability distributions obtained by probabilistic data linkage and record match scores obtained by deterministic data linkage are useful.
- Built-in classifier functions that automatically combine the field comparison scores (and weights) in various ways can be useful for experimenting with different options.

Capabilities for producing and maintaining ground-truth evaluation data (i.e., data used to test if the expected outcome/linkages result). These capabilities are not typically included in data linkage software, but they are needed, and some functionality for reviewing and validating matches should be included in the data linkage software.

Capabilities for data analysis. Although not strictly part of the data linkage process, data analysis is an essential step.

Other important factors in choosing data linkage software are listed below.

- What software is available for the data linkage method(s) in the strategy?
- For each software package considered, the following factors should be assessed:
    - Is the software customized to specific data content (e.g., medical records, customer management), or can it handle a wide array of data?
    - Does the software only perform data linkage, or does it provide multiple functions, such as data management, statistical analysis, and visualizations of findings?
    - Does the software require additional software packages to run?
    - Are the instructions well-documented?
    - What degree of technical support will be provided by the software company?
    - Which parts of the software are under user control?
    - To what extent does the software allow the user to tune the tool?
    - How long will it take to perform the data linkage described in the data linkage plan, and does this meet the schedule demands?
    - What are the program factors?
    - What are the skills and experience of the staff performing data linkage?
    - What other software tools does the program already have access to?
    - What is the budget for all the software needs of the program?

CRITERIA FOR COMPARISON

Many criteria for comparing tools are possible. Below is a list of criteria specific to data linkage tools that can be used as a starting point when selecting the program tool. Note: not all criteria might be needed in the analysis or tool scoring step.

Characteristics for Motor Vehicle Crash Data Linkage System Assessment

CRITERIA FOR SELECTING AND COMPARING TOOLS

APPLICABILITY TO STATE DATA LINKAGE SYSTEMS

- Versatility of data linkage capabilities (customizable parameters; capabilities for inexperienced and experienced users; deterministic and probabilistic methods)
- Multiple functional modules other than data linkage (data pre-processing; statistical analysis; visualizations)

EASE OF USE

- Learning curve (quality of user manual and documentation; technical support; training courses available; familiar terminology to typical user)
- General usability considerations
- New tools not used by states performing data linkage

PRICE

- License
- Other computing requirements (e.g., operating system)

SOFTWARE MATURITY

- Technical support provided by company
- Features of software
- History of innovation and updates

LONG-TERM POTENTIAL

- Sustainability of vendor
- Enterprise tools that could be provisioned at the national level

INTEROPERABILITY

- Data sharing and analysis
- Interoperability with other tools for pre-processing and analysis

USABILITY OF DATA LINKAGE TOOLS WITH RESPECT TO THE POSSIBLE USER BASE

- Learnability of the software by a user
- Operability of software by multiple users with varying degrees of expertise
- Documentation

EFFICIENCY

- Pre-processing
- Total end-to-end processing time
- Resource utilization

When selecting software, it is important to consider the full spectrum of data-related functions required for the linkage program so that software costs can be within budget. In some cases, it might be more economical to invest in a single software package that performs multiple functions than to purchase multiple tools for analyzing the quality of the data, normalizing the data, merging duplicates, linking data sets, managing data sets, performing statistical analysis, generating visualizations of findings, and so on. However, a tool that only performs data linkage might have more-sophisticated data linkage abilities, greater customization of linkage parameters, and other features not included in the multi-function systems.

Also, if a tool requires other software that the linkage program does not already have licenses for and experience with, then these costs must also be considered. Generally speaking, the greater the skill level and experience of the available staff, the greater the potential for limiting the software cost by selecting free or less-expensive tools. Of course, experienced staff can also take advantage of the additional features and time-saving functionality in a more-expensive full-featured software.

Savings in the direct software cost might be offset by indirect costs, such as reduced accuracy of data linkage or increased staff time required to learn a tool with little technical-support service and to perform more-manual processes for tools with less functionality. The free and open-source software options are less likely to have detailed instructions, whereas the more-expensive commercial software is more likely to provide detailed instructions, tutorials, and on-demand technical support. Also, the lack of data linkage software features, such as lists of nicknames or common types of data linkage errors, might require that the data analyst perform more review of match results, normalization of data, or reconfiguration of the data linkage parameters. For maximal flexibility, it is important for the data linkage tool to be able to handle a wide array of data content. A data linkage tool that allows the user to specify which variables to compare for matching is more useful than a tool that is limited to specific variables (e.g., names, social security numbers), especially if the variables for which the tool is customized are not present in the data sets of interest to the linkage program. Also, if a tool that is customized for using specific variables is selected, it might not meet the linkage program's needs in the future if different data sets are added. For instance, linkage programs that need to link medical records across multiple providers might benefit from the use of a general medical records software. For example, if a study wanted to estimate the costs of healthcare to treat injuries from a single MVC incident over time, then it would be necessary to link a patient's medical record across hospitals, post-acute care facilities, and outpatient providers. It is likely that the information across these providers would reside in different data sets, with no unique identifier across all the data sets.

FINAL SELECTION

After scoring the short list of tools, a program is ready to make a final selection. The scores give a method for ranking the data linkage software in a way that applies to a set of unique requirements. Should questions arise later about the use of this tool, it is possible to return to this scoring analysis and perform it again using a newly understood set of requirements.

One method for this would be creating a score for each tool in the different important categories for your program. A score is given for every category for each tool, and a final score summed up to indicate which tool is best for your program. An example of this method is shown in Table 16 for many data linkage tools (note: it is only an example and does not represent actual scoring).

Table 16. An Example Scoring Matrix

TOOL A

Criteria (Applied to Evaluation)	Weight	Score	Weighted Score
Price	80	5	400
Applicability	90	5	450
Ease of use	80	3	240
Software maturity	50	2	100
Long term potential	20	2	40
		Total Score	1230

TOOL B

Criteria (Applied to Evaluation)	Weight	Score	Weighted Score
Price	80	4	320
Applicability	90	5	450
Ease of use	80	4	320
Software maturity	50	4	200
Long term potential	20	4	80
		Total Score	1370

TOOL C

Criteria (Applied to Evaluation)	Weight	Score	Weighted Score
Price	80	5	400
Applicability	90	5	450
Ease of use	80	3	240
Software maturity	50	5	250
Long term potential	20	4	80
		Total Score	1420

Continued

Table 16, continued

TOOL D

Criteria (Applied to Evaluation)	Weight	Score	Weighted Score
Price	80	3	240
Applicability	90	5	450
Ease of use	80	5	400
Software maturity	50	5	250
Long term potential	20	4	80
		Total Score	1420

Configure Data Linkage Parameters

Parameters are the options that data linkage tools provide to match records within a single data set or multiple data sets. The actual matched pairs of records generated by the software will vary depending on the user-specified values of those parameters. Depending on the data linkage method (deterministic versus probabilistic) and the tool selected, parameters for the data linkage tool might be different. However, the principles involved in configuring the parameters are generalizable across tools and data linkage methods.

Determining the optimal value for the parameters is achieved through an empirical and iterative process. The user sets the parameters initially, and then tweaks the parameters between runs of the data linkage tool to affect the resulting matched pairs. If a ground-truth data set is available, then matching quality measures can be computed for each configuration of parameters. The following parameters are typically considered, for both deterministic and probabilistic matching methods, when assessing a data linkage software:

Variable matching weight (or "variable weight"). The relative importance of each variable for matching. Variable weights can be used in deterministic linkage, depending on the tool. Some software packages that perform deterministic linkage, such as The Link King, include pre-composed collections with a one-step solution to decide point values for variable weights. Weights are computed by probabilistic data linkage tools, based on user-provided probabilities, but tools like FRIL allow the user to edit the variable weight. The following weight parameters might be included:

- Agreement weight. Some tools' default value might be set to 1.
- Disagreement weight. Some tools' default value might be set to 0.
- Default weight. Instructs the tool how to match two records with the default value in a given variable.
- Empty value weight. Instructs the tool how to match two records with an empty value in a given variable. If a data set is sparsely populated, then matching empty values will result in false-positive matches. Some tools treat empty values the same way, while other tools, such as FRIL, allow the user to customize the matching weight for empty values in a given variable.

Nickname or common-error lists. Lists of values that are judged to be equivalent. For example, The Link King includes these lists by default and allows for the addition and subtraction of items from the lists, so that users can update the provided lists to include variation from their specific data sources.

Methods for approximate or "fuzzy" matching. Algorithms for computing the similarity between two values. Some software packages, such as FRIL, include multiple algorithms for determining approximate matches between names, numbers, and dates. In addition to an exact string match, FRIL includes Soundex, Jaro-Winkler distance, and Q-Gram distance.

Blocking protocols. Procedures to reduce the number of records considered for matching by using variable criteria to select smaller subsets of records or "blocks" for matching (see Appendix M for more information). Some software packages, such as The Link King, include different levels of blocking.

Deterministic matching rules might have parameter thresholds that depend on the variables and field comparisons used for optimal results matching. Another deterministic approach is summing the match scores from all the selected field comparisons, and setting a threshold match score, so that records with match scores above the threshold can be classified as matches, and records with scores below the threshold can be classified as non-matches.

Probabilistic matching might also require thresholds, or cut-points, to discriminate matches from non-matches. Traditionally, two points are used: the first point separates the most-probable matches from the matches needing review, and the second point delineates the matches needing review from the most-probable non-matches. Some software packages, such as FRIL, allow the analyst to adjust the value of the cutoff threshold based on their results for more-precise or more-sensitive results. Multiple imputation of match status (see Appendix N) avoids the need for thresholds and manual review to select linked records for analysis. This method and the statistical packages required to analyze multiple imputations require parameter configurations that depend on the data and the researcher's goals.

APPENDIX G. STATE MOTOR VEHICLE CRASH DATA LINKAGE PROGRAMS

Historically, 31 states were known to have motor vehicle crash (MVC) data linkage programs. As of 2017, 19 states had active data linkage programs based on an online search. Table 17 shows the states with historical linkage programs and states with active linkage programs in 2017, including links to their websites and denoting if they link MVC records to hospital data. States considering starting a linkage program should leverage the experience and expertise in linking data sets available in neighboring states.

Table 17. Online Evidence of State Motor Vehicle Crash Data Linkage Programs, 2017

State	Motor Vehicle Crash Data Linkage Program Name and Website	Link to Hospital Data?
Alabama	No online evidence of linkage program	n/a
Alaska	Alaska Crash Outcomes Pilot Project (ACOPP)	Yes
California	California Department of Public Health, EpiCenter California Injury Data Online	Yes
Colorado	No online evidence of linkage program	n/a
Connecticut	No online evidence of linkage program	n/a
Delaware	No online evidence of linkage program	n/a
Florida	No online evidence of linkage program	n/a
Georgia	Georgia Crash Outcome Data Evaluation System (CODES)	Yes
Illinois	No online evidence of linkage program	n/a
Indiana	No online evidence of linkage program	n/a
Iowa	Iowa Department of Public Health Crash Outcome Data Evaluation System (CODES)	Yes
Kentucky	Traffic Injury Prevention and Research Program (TIPRP)	Yes
Maine	Division of Public Health Systems: Data, Research and Vital Statistics	Yes
Maryland	Maryland Crash Outcomes Data Evaluation System (CODES) National Study Center for Trauma & Emergency Medical Systems (NSC) at the University of Maryland, Baltimore	Yes
Massachusetts	UMassSafe Traffic Safety Research Program	No
Michigan	University of Michigan Transportation Research Institute (UMTRI)	Yes
Minnesota	Minnesota Crash Outcome Data Evaluation System (CODES) Minnesota Department of Public Health	Yes
Missouri	No online evidence of linkage program	n/a
Nebraska	Nebraska DPHHS Crash Outcome Data Evaluation System (CODES) Injury Surveillance	Yes
New York	New York Crash Outcome Data Evaluation System (CODES) New York State Department of Health Injury and Violence in New York State	Yes
North Carolina	No online evidence of linkage program	n/a
Nevada	No online evidence of linkage program	n/a
Ohio	Nationwide Children's Ohio Crash Outcome Data Evaluation System (CODES) Program	Yes
Oklahoma	Oklahoma State Department of Health (OSDH) Traffic Data Linkage Project	Yes

Continued

State	Motor Vehicle Crash Data Linkage Program Name and Website	Link to Hospital Data?
Rhode Island	No online evidence of linkage program	n/a
South Carolina	No online evidence of linkage program	n/a
Tennessee	Tennessee Crash Outcome Data Evaluation System (CODES)	Yes
Utah	Intermountain Injury Control & Research Center (IICRC)	Yes
Virginia	Traffic Records Management, Reporting and Analysis Division of the Department of Motor Vehicles (DMV), Virginia Highway Safety Office (VAHSO)	No
Washington	Washington State Department of Health Injury & Violence Prevention Centers for Disease Control and Prevention (CDC) State Violence and Injury Prevention Program (Core SVIPP)	Yes
Wisconsin	Wisconsin Crash Outcome Data Evaluation System (CODES)	Yes

APPENDIX H. MOTOR VEHICLE CRASH DATA LINKAGE PROGRAM RESOURCES

Numerous resources are available to states that are considering a motor vehicle crash (MVC) data linkage program and to states that want to expand or enhance existing linkage programs. Table 18 captures a range of resources that might be useful to state linkage programs. Potential government partners at the local, state, and federal levels, as well as national and international stakeholders, are listed. Table 18 includes organizations focused on motor vehicle safety awareness, professional societies and associations providing technical assistance, and a partial listing of scientific conferences and association meetings that might be of interest to linkage programs.

Table 18. Resources for Motor Vehicle Crash Data Linkage Programs

LOCAL

Entity	Description
County and city governments	Local government is usually composed of counties, towns, cities, etc. The website allows users to find contact information for local governments by state.
County and municipal courts	There are different types of courts at the state, county, and municipal levels. These can include small claims courts, traffic courts, juvenile courts, and family courts.
Hospital networks and hospital systems	A hospital network is a group of hospitals that work together to coordinate and deliver a broad spectrum of services to their community. A hospital system or healthcare system is two or more hospitals owned, sponsored, or contract managed by a central organization. The website can be searched by city.

STATE

Entity	Description
Department of Motor Vehicles (DMV)	State agency that provides motor vehicle services to the public. The website allows users to find state DMVs via a drop-down menu.
Department of Transportation (DOT)	State transportation agency. The website hosts an alphabetical listing of the DOT for all 50 states.
Emergency Medical Service (EMS) agencies	EMS provides out-of-hospital acute medical care and transport to a hospital for persons with illness or injury. The website is an alphabetical listing of general phone numbers and website addresses of EMS agencies by state.
Health agencies	An agency or department of state government focused on public health. The website has an alphabetical list of state health agencies.
Highway patrol and law enforcement organizations	State highway patrol and law enforcement agencies. The website lists police and law enforcement organizations for each state.
Legislatures	State lawmaking assemblies, which are made up of elected legislators who make laws for the state. The website lists the legislatures for all states.
State Highway Safety Office (SHSO)	SHSOs are usually located in state DOT or public safety departments. SHSO directors are part of the Governors Highway Safety Association (GHSA).
State Traffic Records Coordinating Committee (TRCC)	TRCC has members from state agencies across all six core systems: crash, vehicle, roadway, citation and adjudication, and injury surveillance. The state TRCC has responsibility for coordinating state organizations involved in the administration, collection and use of highway safety data and traffic records.
Vital records offices	Legal authority for the registration of births, deaths, marriages, and divorces resides with the states. The website provides information for each state.

Continued

FEDERAL GOVERNMENT

Entity	Description
Centers for Disease Control and Prevention (CDC)	CDC works 24/7 to protect America from health, safety and security threats, both foreign and in the U.S.
CDC National Center for Injury Prevention and Control (NCIPC)	NCIPC is responsible for preventing injuries and violence through science and action; the Transportation Safety Team maintains an extensive list of resources on the website.
CDC National Institute of Occupational Safety and Health (NIOSH)	NIOSH is responsible for promoting productive workplaces through safety and health research, and the Center for Motor Vehicle Safety provides research-based guidance to prevent crashes among workers.
Department of Transportation (DOT)	DOT is responsible for developing and maintaining the nation's transportation systems and infrastructure.
DOT Office of the Secretary of Transportation (OST)	OST is an office within the DOT that oversees the formulation of national transportation policy and promotes intermodal transportation.
DOT Traffic Records Coordinating Committee (TRCC)	DOT TRCC is a multi-modal group with members from NHTSA, FHWA, FMCSA, and the OST, that works to improve the collection, management, and analysis of traffic safety data at the state and federal levels.
Federal Highway Administration (FHWA)	FHWA is an agency within the DOT that supports state and local governments in the design, construction, and maintenance of the nation's highway system.
Federal Highway Administration (FHWA) Office of Safety	The FHWA Office of Safety offers technical assistance on policy, program, and technical issues to state and local roadway agencies to assess, develop, implement, or evaluate effective strategies and programs that reduce roadway fatalities and serious injuries on public roads.
Federal Motor Carrier Safety Administration (FMCSA)	FMCSA is an agency within the DOT that is responsible for regulating and providing safety oversight of commercial motor vehicles to reduce crashes, injuries, and fatalities involving large trucks and buses.
Indian Health Service (IHS)	IHS is an agency within the Department of Health and Human Services that is responsible for providing health services to American Indians and Alaska Natives. The Injury Prevention Program website has a list of injury prevention contacts by area.
National Highway Traffic Safety Administration (NHTSA)	NHTSA is a DOT agency with the mission to save lives, prevent injuries and reduce economic costs due to road traffic crashes, through education, research, safety standards and enforcement activity. NHTSA provides grants to state governments to conduct effective highway safety programs.
NHTSA Office of EMS	NHTSA Office of EMS advances a national vision for EMS through projects and research, fosters collaboration among federal agencies involved in EMS planning, measures the health of EMS systems, and delivers the data to help advance systems.
National Information Exchange Model (NIEM)	The NIEM Surface Transportation domain is under the stewardship of the DOT and supports information sharing and promotes interoperability among transportation regulators, operators, and stakeholders including law enforcement, courts, health, and emergency management partners.

NATIONAL AND INTERNATIONAL

Entity	Description
Academic institutions	Academic institutions can partner with states as a trusted third party to perform data linkage. The Department of Education lists accredited academic institutions at this website.
Administrative Data Research Network (ADRN)	ADRN works to enable evidence-based policy development and research and to share knowledge, methods and insight for public benefit across the United Kingdom. ADRN offers training podcasts on data linkage. No prior knowledge is necessary for the introductory course, which is a mixture of lecture and hands-on exercises.
American Automobile Association (AAA)	AAA is a federation of North American motor clubs and an advocate for motor vehicle safety. The AAA Foundation for Traffic Safety is a not-for-profit, publicly supported charitable research and education organization dedicated to saving lives by preventing traffic crashes and reducing injuries when crashes occur.
American Hospital Association (AHA)	AHA is a national organization that provides representation and advocacy for hospitals, healthcare networks and individual members.

Continued

Entity	Description
American Public Health Association (APHA)	APHA is an organization that provides representation and advocacy for public health professionals.
Association of Transportation Safety Information Professionals (ATSIP)	ATSIP is a professional association devoted to furthering the development and sharing of traffic records system procedures, tools, and professionalism.
Governors Highway Safety Association (GHSA)	GHSA is an association that represents state and territorial highway safety offices, which implement federal grant programs to address behavioral highway safety issues. GHSA also sponsors an annual conference and research program.
International Population Data Linkage Network (IPDLN)	IPDLN is a network of global data linkage centers, whose purpose is to facilitate communication across groups performing data linkage.
Life Savers Conference	The largest gathering of highway safety professionals committed to sharing best practices, research, and policy initiatives that are proven to work.
National Center for State Courts (NCSC)	NCSC provides information on the administrative office of the courts, the court of last resort, any intermediate appellate courts, and each trial court level. The website provides a list of judicial branch links for each state.
National Emergency Medical Service Management Association (NEMSMA)	NEMSMA is an association that provides leadership, education and advocacy for EMS professionals.
National Governors Association (NGA)	NGA is a bipartisan organization of the nation's governors. The website has a listing of current governors for each state.
National Safety Council (NSC)	NSC, in partnership with the NHTSA, FHWA and FMCSA, coordinates the Road to Zero Initiative, an effort to bring together multiple stakeholders to use a data-driven, interdisciplinary approach to eliminating MVC-related deaths and serious injuries.
Safe Kids	Safe Kids works to keep children safe from preventable injury through research, public policy, and education and awareness programs. The organization works with an extensive network of more than 400 coalitions in the United States and with partners in more than 30 countries.
Toward Zero Deaths (TZD)	Led by the TZD Steering Committee, the National Strategy on Highway Safety provides a platform for state agencies, private industry, national organizations and others to develop safety plans that prioritize traffic safety culture and promote the national vision of a highway system free of deaths.
Traffic Records Forum	Annual meeting of data analysts, state and local law enforcement officials, engineers, motor vehicle officials, emergency medical services providers, judicial administrators, and highway safety professionals from across the United States and international communities sponsored by the ATSIP. Specific sessions are focused on data integration to improve and expand states' use of linked motor vehicle crash data.
Transportation Research Board (TRB)	TRB is a program unit of the National Academy of Sciences, Engineering and Medicine that provides innovative, research-based solutions to improve transportation.
Vision Zero Network (VZN)	VZN is a collaborative campaign of local leaders in health, traffic engineering, police enforcement, policy and safety advocacy to build momentum for and advance the strategy of eliminating all traffic deaths and severe injuries while increasing safe, healthy, equitable mobility for all.

APPENDIX I. DEPARTMENT OF TRANSPORTATION TRAFFIC RECORDS COORDINATING COMMITTEE TECHNICAL ASSISTANCE RESOURCES

The mission of the Department of Transportation Traffic Records Coordinating Committee (DOT TRCC) is to maximize the overall quality of safety data and analysis based on state traffic records data across crash, vehicle, driver, roadway, citation and adjudication, and injury surveillance systems [84]. DOT TRCC offers a variety of technical assistance and training programs that can help states build the needed traffic safety data collection, management, and analysis capacity as listed below:

National Highway Traffic Safety Administration: Crash Data Improvement Program (CDIP). "NHTSA's CDIP technical assistance program examines the quality of a state's crash data and provides the state with specific recommendations to improve the quality, management and use of that data to support safety decisions. This program is free to the states and made available on a first-come, first-served basis given available funds. States that wish to request a CDIP at no expense should complete the "Application for Traffic Records Programs" [85]. The CDIP Guidelines, updated in 2017, can be downloaded here."

Federal Highway Administration: Data and Analysis Technical Assistance Program. "The Data and Analysis Technical Assistance Program is available through the Roadway Safety Professional Capacity Building Peer-to-Peer Technical Assistance web application. Any public transportation agency can request technical assistance. The program can address any roadway safety data and analysis challenge in a variety of formats" [86].

Federal Highway Administration: Roadway Data Improvement Program (RDIP). "RDIP is a technical assistance program to help transportation agencies improve the quality of their roadway data to better support safety improvement initiatives. RDIP focuses on the process and practices used by the agency for collecting, managing, and utilizing their roadway data to support safety investment decision making" [85].

National Highway Traffic Safety Administration: GO Teams. "NHTSA's Traffic Records GO Team program helps states improve their traffic records systems by deploying tailored technical assistance and training based on states' actual needs. The program provides support from subject matter experts who work with the state's traffic records professionals to improve a specific aspect of their traffic records data collection, management, or analysis capabilities. Successful GO Team applications request specific technical assistance focusing on a discrete traffic records challenge or technical training need as identified by state traffic records program managers. States are encouraged to submit GO Team requests that address a specific traffic records improvement need, either highlighted during a state's traffic records assessment or identified by the state's Traffic Records Coordinating Committee and Highway Safety Office" [86].

National Highway Traffic Safety Administration: Traffic Records Executive Training. NHTSA worked with the Governor's Highway Safety Association (GHSA) to develop a "course that provides new Governor's Representatives for Highway Safety (GRs) and Highway Safety Coordinators with an understanding of the critical role traffic records data plays in a State Highway Safety Office's planning and evaluation efforts" [87]. The course is suitable for new state managers and executives responsible for traffic safety data.

APPENDIX J. SECURITY PROGRAM ACTIVITIES

A security plan is a formal document that provides an overview of the security requirements for an information system or an information security program. A security plan also describes the controls that are in place or planned for meeting those requirements. No organization is immune to a security compromise or vulnerability that could result in confidential data, such as personally identifiable information (PII), being exposed to unauthorized people. The release of confidential information can have substantial adverse effects on an organization's reputation as well as legal standing and profitability. Therefore, each organization managing a linkage program should develop a security plan for the information systems used in the program so that the security risks are addressed appropriately.

Many information security activities are common, but state agencies and organizations should consult with the state or organizational entity responsible for information security (e.g., Chief Information Security Officer) to leverage existing policies, procedures, and guidance. Established security and control standards should be used when possible [38, 88]. The states that participated in the listening sessions have not found the software system security requirements hard to meet. One state was allowed to link data sets on work-provided laptops and another state stored data on dedicated computers not on the network, in a locked room in a building with security. Typical information security activities include:

Typical information security activities include:

- Security program management with assessment or audit and with authorization or accountability (e.g., approval of the systems' security controls).
- Physical and environmental protection.
- System and services acquisition, system maintenance, configuration management.
- System and information integrity.
- Incident response, risk assessment, contingency planning, including preparing a separate incident response plan that identifies how to report and respond to security incidents.
- System and data protection, including the use of encryption and information protection during transmission and storage.
- Access control (i.e., identification and authentication of users), including identifying methods to restrict individual access to sensitive information based on business need.
- Security awareness and training for users.
- Personnel security.

APPENDIX K. PRIVACY PROGRAM ACTIVITIES

States that establish and maintain a linkage program should implement a documented, comprehensive privacy program that describes their privacy risk management activities to prevent unlawful disclosure and use of an individual's information. State agencies and organizations that collect, use, share and store data on individuals, including personally identifying information (PII) or protected health information (PHI), are subject to federal and state laws, directives and policies such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Most state agencies and organizations will have existing policies, procedures and guidance in place to protect the privacy of PII and PHI. Linkage programs with question or concerns should consult with their state, agency or organization's Office of Privacy to leverage expertise and resources.

States that participated in the listening sessions mentioned HIPAA concerns with the release of the whole hospital data set for linkage purposes and the release of blood alcohol content (BAC) data from hospitals for specific patients. In these cases, identifiable data were not shared for research. Hospitals (or hospital associations) might use HIPAA as a reason not to provide that information. It is important to periodically meet with stakeholders to re-review policies, restrictions, and data use agreements since data sensitivities might change over time. What was not allowed previously might be allowed in the future.

Privacy programs typically include the following activities:

Leadership and Organization

- Develop and implement a Privacy Program Plan that provides an overview and description of the privacy program.
  - Program strategic goals and objectives
  - Program structure
  - Program management controls
  - Program controls for meeting applicable privacy requirements and managing privacy risks
  - Roles of privacy officials and staff
  - Dedicated resources
- Review and revise the plan annually and update as needed.
- Coordinate with stakeholders to document the data owners and points of contact for each data set received, including those which contain PII.

Privacy Risk Management

- Maintain and enforce privacy policy, procedures, and standards.
  - Use a risk-based approach to develop internal standards of privacy requirements in coordination with stakeholders.
  - Provide resources (e.g., manuals, guides, handbooks) to support consistent implementation of privacy policies, procedures, and standards.
  - Reference existing privacy standards and best practices [38].
  - Establish and enforce a policy on workforce privacy rules of behavior, addressing required PII handling procedures, and information system usage; as well as the consequences for violating those rules.
  - Ensure that contractors meet appropriate privacy requirements by including privacy language in their contracts.
  - Review and revise privacy policy, procedures, and standards documentation periodically to ensure that they remain current.
- Minimize PII collection and use.
  - Describe the purpose(s) for which PII is collected, used, maintained, and shared, in publicly-posted privacy notices, and require approval by a privacy official when changes to the identified purpose occur.
  - Identify the minimum PII elements that are relevant and necessary to accomplish the legally authorized purpose of collection.
  - Limit the collection and retention of PII to the minimum elements identified for the purposes described in the notice, and for which the individual has provided consent where appropriate.
  - Conduct PII inventories to identify what PII is collected and where it resides.
  - Use PII internally, only for the authorized purpose(s) identified in privacy laws, in other applicable authorities, and/or in public notices.
- Limit external sharing of PII.
  - Share PII externally only for the authorized purposes identified in applicable laws, directives, and guidance and/or as described in its notice(s), or for a purpose that is compatible with those purposes.
  - Evaluate any proposed new instances of sharing PII, including ad hoc requests, with third parties to assess whether the sharing is authorized and whether additional or new public notice is required.
  - Where appropriate, enter into agreements with third parties that specifically describe the PII covered, and enumerate the purposes for which the PII might be used as well as the allowable disclosures (if any) and the retention conditions.
  - Monitor, audit, and train staff on the authorized sharing of PII with third parties and on the consequences of unauthorized use or sharing of PII.

  - Implement a process to evaluate ad hoc requests for sharing PII. Establish a mechanism to track access to protected information, including health information within the purview of the organizations and as required by law.
  - Develop reporting procedures to allow qualified individuals to review access information.
- Conduct privacy risk and impact assessments.
  - Execute initial and periodic information privacy risk assessments and conduct related ongoing compliance monitoring activities, in coordination with other compliance and operational assessment functions.
  - Document and implement a privacy risk management process that assesses the privacy risk to individuals that results from the collection, sharing, storing, transmitting, use, and disposal of PII.
  - Conduct privacy control assessments. Strategically integrate privacy documentation, including risk assessments, into privacy risk management processes.
  - Integrate privacy risk management into the enterprise risk management function.
  - Implement a process to manage privacy risks and issues, prioritize resolution, and track through to closure.
- Meet PII retention schedules.
  - Retain each collection of PII to fulfill the purpose(s) identified in the notice, in accordance with the applicable retention schedule, and as required by law.
  - Dispose of, destroy, erase, and/or anonymize the PII, regardless of the method of storage, in accordance with an approved record-retention schedule and in a manner that prevents loss, theft, misuse, or unauthorized access.
  - Use the organization's pre-approved techniques or methods to ensure secure deletion or destruction of PII (including originals, copies, and archived records).
- Meet PII quality standards.
  - Confirm, to the greatest extent practicable upon collection or creation of PII, the accuracy, relevance, timeliness, and completeness of that information.
  - Check for, and correct as necessary, any inaccurate or outdated PII used by programs or systems, in accordance with organizational standards.
  - Document processes to ensure the integrity of PII through existing security controls.
- Monitor government privacy changes.
  - Monitor privacy laws and policy for changes that affect the privacy program.
  - Review regulations and reports from other entities to identify trends and best practices that might benefit the organization.
- Provide continuous monitoring of privacy.
  - Establish and maintain a privacy continuous monitoring strategy and program to ensure that PII is appropriately protected against changing threats.
- Manage contractor and third-party privacy risk.
  - Support the identification and management of privacy risks through the acquisition life cycle.
  - Establish privacy roles, responsibilities, and access requirements for contractors and service providers.
  - Include privacy requirements in contracts and other acquisition-related documents.
  - Assess contractors for compliance with privacy requirements.
  - Reinforce vendor and contractor responsibilities through contract awards, contract clauses, and contract performance assessments.
  - Modify contracts as needed to address new requirements.

Engineering and Information Security

- Support privacy engineering.
  - Support the identification and management of privacy risks through the system and program development and management life cycle.
  - Incorporate information privacy principles into the information life cycle.
  - Include privacy sections in relevant system development documents (e.g., requirements documents, interface control documents.)
  - Align information privacy principles and activities with cybersecurity processes and activities.
  - Maintain business processes that prevent the operation of information systems that have not met applicable privacy requirements.

Incident Response

- Develop incident management procedures.
  - Develop and implement a Privacy Incident Response Plan that enables the organization to respond promptly to privacy incidents, including data breaches.
- Regularly test incident response procedures.
- Provide incident response training and appropriate reporting venues.
  - Train the workforce on privacy incident notification and reporting procedures as documented in policy and in the Privacy Incident Response Plan.
  - Provide multiple channels for reporting privacy incidents.
- Provide effective response and integration.
  - Provide an organized and effective response

to privacy incidents in accordance with the organizational Privacy Incident Response Plan.

- Integrate incident management processes into existing business processes within the organization, where practicable.
- Ensure that appropriate stakeholders are informed when incidents occur, including oversight organizations and IRBs.

- Maintain an incident response team.
  - Maintain a group comprised of senior leadership with decision making authority from relevant offices within the organization to respond to high-impact privacy incidents, including data breaches.

Individual Participation, Transparency, and Redress

- Provide public notice.
  - Make public privacy documentation that is required by law to be publicly accessible.
  - Conduct periodic reviews of privacy documentation to ensure that it remains current, and revise documentation as needed.
  - Provide privacy notice to individuals when their PII is collected. Notices should include statements of what PII is collected, why it is being collected, how it will be used, with whom it will be shared, how long it will be retained, how it will be protected, and whom to contact for further information, including access to an individual's PII and the ability to correct it.
  - Communicate to individuals and stakeholders any changes to the identified purposes for which PII is collected and used.
  - Ensure that privacy practices are publicly available through organizational websites or otherwise.
- Provide opportunities for consent.
  - Provide means, where feasible and appropriate, for individuals to authorize the collection, use, maintenance, and sharing of PII prior to its collection.
- Provide individual access.
  - Where appropriate, provide individuals with the ability to have access to their PII that is maintained in systems.
  - Post access procedures in public notices.
- Provide individual amendment process.
  - Provide a process for individuals to have inaccurate PII maintained by the organization corrected or amended, as appropriate.
  - Establish a process for disseminating corrections or amendments of the PII to other authorized users of the PII within a reasonable timeframe, such as external information-sharing partners. Where feasible and appropriate, notify affected individuals that their information has been corrected or amended.
  - Work cooperatively with other organizations overseeing patient rights to inspect, amend, and restrict access to Protected Health Information and other personal information.
- Provide redress management.
  - Establish and administer a process for receiving, documenting, tracking, investigating, and acting on all complaints concerning the organization's privacy policies and procedures in coordination and collaboration with other similar functions and, when necessary, legal counsel.

Privacy Training and Awareness

- Provide privacy training.
  - Develop, implement, and update a comprehensive training and awareness strategy aimed at ensuring that personnel understand privacy responsibilities and procedures.
  - Oversee, direct, deliver, and ensure the delivery of privacy training and orientation to all employees, volunteers, medical and professional staff, contractors, alliances, business associates, and other appropriate third parties, including initial orientation and ongoing education and awareness campaigns.
  - Administer basic privacy training at least annually.
  - Administer targeted, role-based privacy training, for personnel having responsibility for PII or for activities that involve PII, at least annually.
  - Ensure that personnel certify (manually or electronically) acceptance of responsibilities for privacy requirements at least annually.

Accountability

- Provide reporting on the status of the privacy program.
  - Develop, disseminate, and update timely and accurate reports to oversight bodies, as appropriate, to demonstrate accountability with specific statutory privacy program mandates, and to senior management and other personnel with responsibility for monitoring privacy program progress and compliance.
- Conduct oversight and monitoring planning.
  - Implement procedures to coordinate across the organization to address privacy requirements at all levels of the organization and in all physical locations.
- Monitor and audit privacy program.
  - Monitor and audit privacy controls and internal privacy policy periodically to ensure effective implementation.

APPENDIX L. SAMPLE DATA USE AGREEMENT

Data Use Agreements (DUA) are contracts which define the terms and conditions of non-public data that are subject to restricted use. A DUA is normally used to assist agencies/organizations/data owners/data users wishing to share data to better understand important information regarding the data being exchanged, such as privacy rights that are associated with transfers of confidential or protected data, obligations to safeguard the data, limitations on use of the data, and any liabilities related to the use of the data.

Appendix L provides a sample Data Use Agreement from Maryland (used with permission). This sample demonstrates what can be included in a Data Use Agreement if your organization does not already have a DUA or wants to consider updating their DUA.

DATA USE AGREEMENT FOR THE STATEWIDE INPATIENT AND OUTPATIENT CONFIDENTIAL-LEVEL DATA SETS

This data use agreement pertains to requests for FY or CY (s) ____ of the Confidential-level statewide Hospital Discharge Data Sets (Inpatient) and Hospital Outpatient Data Sets (Outpatient) collected by the Health Services Cost Review Commission ("HSCRC") under COMAR 10.37.06 and COMAR 10.37.04. These data are considered protected health information (PHI). The undersigned gives the following assurances with respect to the HSCRC data sets ("Data"):

____ (the "Organization") considers the security and confidentiality of PHI as a matter of high priority. Any and all members of the Organization having access to patient medical files and information contained in the Data will be held responsible for safeguarding and maintaining strict confidentiality. In order to be granted access to PHI, you must agree unconditionally to the following standards:

1. I attest that I have received training in the protection of sensitive and private information;
2. I will not attempt to use or permit others to use the data set to learn the identity of any person included in the data set;
3. I will require others in the Organization, as well as any subcontractor to the Organization who uses the Data, to sign an agreement assuring full compliance with this data use agreement. The Organization will keep these signed agreements and make them available to the HSCRC during normal business hours and upon receipt of prior written notice;
4. A data security plan shall be maintained by any subcontractor employed by the Organization which adequately addresses the requirements contained herein;
5. I will not release or permit others to release any information that identifies persons, directly or indirectly;
6. I will not release or publicize or permit others to release or publicize statistics where the number of observations in any given cell of tabulated data is less than or equal to ten (10);
7. I will not release or permit others to release the Data or any part of it to any person who is not a member of or to any entity, without the prior written approval of the HSCRC;
8. I will ensure any that any subcontractors accessing the Data will use the Data only for the purposes identified in the HSCRC Data Request Form and will destroy the data once the project is complete per #20 of this DUA;
9. I will not attempt to link or permit others to attempt to link the hospital stay records of the persons in the data set with personally identifiable records from any source;
10. I will not disclose confidential records identified by the CRISP algorithm as being related to Substance Abuse treatment or disorders in accordance with 42 CFR Part 2, unless the data has been de-identified and the purpose is for research only;

Continued

11. I will only use the Data for the purposes identified in the HSCRC Data Request Form and will acknowledge in all reports based on these Data, by direct cite where space and/or publication guidelines permit, or by inclusion in a list of data contributors available upon request that the source is the Health Services Cost Review Commission;
12. I will not use the Data for purposes of penetration or vulnerability studies to test whether patients in the dataset can be identified using variables contained in the Data;
13. The HSCRC staff or agent thereof reserves the right to inspect the offices of the data user, during normal business hours and upon prior written notice, to ensure compliance with this Data Use Agreement;
14. I will ensure that that the transmission of PHI is in full compliance with the Privacy Act, Freedom of Information Act, HIPAA, and all other State and federal laws and regulations, as well as all Medicare regulations, directives, instructions, and manuals;
15. I will submit an approval letter from an Institutional Review Board;
16. I will give HSCRC written notice immediately or as soon as reasonably practicable upon having reason to know that a breach, as defined below has occurred;

Any unauthorized use of the Data by ______________________________________ shall constitute a breach of this Agreement. Any breach of security or unauthorized disclosure of the Data by ________________________ _____________ shall constitute a breach of this Agreement. Any violation of State or federal law with respect to disclosure of the Data by _____________________________________, including but not limited to, the HIPAA, shall constitute a breach of this Agreement. Notwithstanding the breaches specifically enumerated above, any other failure by _____________________________ or business associates, including its contractors, subcontractors or providers to comply with the terms and obligations of this Agreement shall constitute a breach of this Agreement. Any Breach of the Data by a third-party will promptly (i) be the subject of contractual termination or other action, as determined by _______________________ and (ii) will be reported to the HSCRC within two (2) business days of the day ________________________________becomes aware of the third-party violation.

17. Any alleged failure of ____to act upon a notice of a breach of this Agreement does not constitute a waiver of such breach, nor does it constitute a waiver of any subsequent breach(es);
18. In the event that the HSCRC reasonably believes that the confidentiality of the Data has been breached, the HSCRC may: investigate the matter, including an on-site inspection for which _________________________ ___________ shall provide access; and require ______________________________________ to develop a plan of correction to ameliorate or minimize the damage caused by the breach of confidentiality and to prevent future breaches of data confidentiality. In the event of a breach of this Agreement, HSCRC may seek all other appropriate remedies for breach of contract, including termination of this Agreement, disqualification of ___________________from receiving PHI from HSCRC in the future, and referral of any inappropriate use or disclosure to the Maryland Office of the Attorney General, Consumer Protection Division;
19. At its sole cost and expense, the Organization shall indemnify and hold the HSCRC, its employees and agents harmless from and against any and all claims, demands, actions, suits, damages, liabilities, losses, settlements, judgments, costs and expenses (including but not limited to attorneys' fees and costs), whether or not involving a third-party claim, which arise out of or relate to the Organization's, or any of its subcontractors' use or disclosure of Data that is the subject of this Agreement. The Organization shall not enter into any settlement involving third-party claims that contain an admission of or stipulation to guilt, fault, liability or wrongdoing by the HSCRC or that adversely affects the HSCRC's rights or interests, without the HSCRC's prior written consent.
20. I will provide a Certification of Data Destruction to the HSCRC once the source data are destroyed and the project is completed;
21. I will retain these data files until (date 5 year maximum)____.
22. This Agreement will remain in effect for the duration of the State Fiscal Year ______, which may be terminated by the HSCRC at any time, and for any reason.

Continued

If this project is not completed within a one year timeframe, resubmission and approval by the HSCRC will be required.

HIPAA

The HIPAA Privacy Rule sets national standards for patient rights with respect to health information. The rule protects individually identifiable health information by establishing conditions for its use and disclosure by covered entities. Further information on the HIPAA Privacy Rule can be found at:
http://www.hhs.gov/ocr/hipaa/ or http://privacyruleandresearch.nih.gov/

CONFIDENTIALITY OF SUBSTANCE USE DISORDER PATIENT RECORDS (42 CFR PART 2)

Federal Statute 42 CFR Part 2 impose restrictions upon the disclosure and use of substance use disorder patient records which are maintained in connection with the performance of any part 2 program. The regulations in this part prohibit the disclosure and use of patient records unless certain circumstances exist.

Further information on the 42 CFR Part 2 Federal Statute can be found at:
https://www.samhsa.gov/health-information-technology/laws-regulations-guidelines

HEALTH SERVICES COST REVIEW COMMISSION
DATA USE AGREEMENT FOR THE STATEWIDE INPATIENT, AND OUTPATIENT CONFIDENTIAL-LEVEL DATA SETS

My signature indicates agreement to comply with the above-stated requirements. I understand that failure to comply with the provisions specified herein may result in civil and/or criminal penalties in accordance with state law and policy.

Signed: ______________________________ Date: ______________

Print or Type Name: ______________________________

Phone: ______________________________

Title: ______________________________

Organization: ______________________________

Address: ______________________________

City: ______________________________

State: ______________________________

Zip Code: ______________________________

E-mail Address: ______________________________

DATA MANAGEMENT PLAN GUIDELINES

The data management plan guidelines contain four steps. You are being asked to describe the actions that you will take to address these protections specific to this confidential request.

The safeguards you describe should be reasonable and appropriate based on the organizational environment in which your research is conducted.

Please note that your explanation should fully describe the protections you have in place. We expect that some of your safeguard descriptions in response to a step may overlap with another step. HSCRC is not requesting documentation that supports your descriptions at this time, only a copy of your organization's Management Plan. Researchers should maintain documentation supporting this data management plan should HSCRC request a remote review or on-site visit.

The safeguards you describe in this plan are specific to this research request.

1. PHYSICAL POSSESSION AND STORAGE OF HSCRC DATA FILES

- Who will have the main responsibility for organizing, storing, and archiving the data? Please provide name(s) and job title(s).
- Describe how your organization maintains a current inventory of HSCRC data files being accessed (identify how the agency tracks users and the data being accessed per project).
- Describe how your organization binds all members (i.e., organizations, individual staff) of research teams to specific privacy and security rules in using HSCRC data files.
- Provide details about who and how your organization will notify HSCRC of any project staffing changes.
- Describe your organization's training programs that are used to educate staff on how to protect HSCRC data files.
- Explain the infrastructure (facilities, hardware, software, other) that will access the HSCRC.
- Describe the policies and procedures regarding access to HSCRC data files.
- Explain your organization's system or process to track the status and roles of the research team.
- Describe your organization's physical and technical safeguards used to protect HSCRC data files (explain the safeguards used to protect user ids/passwords, ensure users comply with HSCRC rules of operation, only download statistical results, etc.).

2. DATA SHARING, ELECTRONIC TRANSMISSION, DISTRIBUTION

- Describe your organization's policies and procedures regarding the sharing, transmission, and distribution of HSCRC data files.
- If your organization employs a data tracking system, please describe.
- Describe the policies and procedures your organization has developed for the physical removal, transport and transmission of HSCRC data files.
- Explain how your organization will tailor and restrict data access privileges based on an individual's role on the research team (HSCRC users shall include language to ensure they only request access to the minimum amount of data necessary for completion of their project. Additionally, if a user has access for multiple projects, language shall be included to specify that the user will only access the data files specific to each DUA).
- Explain the use of technical safeguards for data access (which may include password protocols, log-on/log-off protocols, session time out protocols, and encryption for data in motion and data at rest).
- Are additional organizations involved in analyzing the data files provided by the HCRC? If so, please indicate how these organizations' analysts will access the data files:
    - » VPN connection
    - » Will travel to physical location of data files at requesting organization
    - » Request that a copy of the data files be housed at second location
    - » Other: Click here to enter text.

Continued

- If an additional copy of the data will be housed in a separate location, please describe how the data will be transferred to this location. (Also, please ensure you have included information on this organization's database management under the appropriate subsections of the database management plan.)

3. DATA REPORTING AND PUBLICATION

- Who will have the main responsibility for notifying HSCRC of any suspected incidents wherein the security and privacy of the HSCRC data may have been compromised?
- Please describe and identify your organization's policies and procedures for responding to potential breaches in the security and privacy of the HSCRC data.
- Explain how your organization's data management plans are reviewed and approved.
- Explain whether and how your organization's data management plans are subjected to periodic updates during the DUA period.
- Please attest to the HSCRC cell suppression policy of not publishing or presenting tables with cell sizes less than 10.

4. COMPLETION OF RESEARCH TASKS AND DATA DESTRUCTION

- Describe your organization's process to notify HSCRC when the project is complete and access is no longer needed.
- Describe your organization's policies and procedures for notifying HSCRC if a current HSCRC user is no longer working on the project (particularly if a project involves multiple users).
- Describe policies and procedures your organization uses to inform HSCRC of access changes when staff member's participation in the research project is terminated, voluntarily or involuntarily.
- Describe your organization's policies and procedures to ensure original data files are not used following the completion of the project.

APPLICATION FOR ACCESS TO THE HSCRC CONFIDENTIAL INPATIENT AND OUTPATIENT DATA FILES

IT SHOULD BE NOTED THAT RELEASE OF CONFIDENTIAL DATA IS EXTREMELY RARE.
MOST DATA REQUEST ARE ACCOMMODATED THROUGH THE USE OF THE HSCRC PUBLIC USE DATA FILES.

All requests for confidential data are reviewed by the Health Services Cost Review Commission Confidential Data Review Board. The Board makes recommendations to the Commission at its monthly public meeting. The review process may take up to **90 days** from submission of the complete letter of request and supporting materials to the Commission for consideration at the public meetings. At the Commission's sole discretion, the investigator may be invited, at the investigator's expense, to appear before the Committee or the Commission to discuss their application. The Commission makes the final decisions on the release of the confidential data sets.

The role of the Board is to review applications and make recommendations to the Commission for final consideration. The following conditions apply to users of confidential data:

1. compliance with Health General Article Section 4-101 et. Seq.;
2. compliance with Health General Article Section 19-207, COMAR 10.37.04, COMAR 10.37.06 and COMAR 10.37.07;
3. the data shall only be used for the purposes specified by the Commission;
4. the results of data analysis and reports must be submitted to the Commission prior to the public release;
5. other restrictions may apply as deemed appropriate

All approved applicants will be required to file annual progress reports to the Commission, any changes in goals or design of project, any changes in data handling procedures, work progress and any unanticipated events related to the confidentiality of the data.

To access the HSCRC Confidential Data File, **a formal letter (on company letterhead)** of request must be submitted. The letter must contain in detail the information identified below. The HSCRC reserves the right to require additional information to determine whether access should be granted to the requesting organization.

Send completed letter of application to:

Health Services Cost Review Commission
Attn: Oscar Ibarra
Chief, Program Administration and Information Management
4160 Patterson Avenue
Baltimore, Maryland 21215
Ph: (410) 764-2566
Fax: (410) 358-6217
Email: oscar.ibarra@maryland.gov

1. Specify the data file(s) (i.e. Inpatient, Outpatient Data Files) and the period requested. For a complete description of the data file maintained by the HSCRC, please contact

   The St. Paul Group
   PO Box 0628
   Fulton, Maryland 20759-0628
   Ph: (410) 760-3447
   Fax: (410) 768-6519
   http://www.thestpaulgroup.com/

Continued

2. Identify the organization of individual requesting data access. Include the following information:

   Name and Title of Representative
   Name of the Organization
   Mailing Address
   Telephone and Fax Numbers
   E-mail address

3. Specify in detail the purpose for which the data are requested. (If a research project is proposed, please provide 5 copies of the research proposal including study design.)

4. State in detail the applicant's qualifications to perform proposed studies, analyses. Specify experience using sensitive medical information, HIPAA training, qualification of investigators, and funding source(s)

5. Identify the public benefit of the proposed research analysis. Please be **specific** as this is a crucial component of the Commission's review for access to the confidential data.

6. Identify the risks to individuals, the public, or other entities, such as specific institutions for the proposed research or analysis.

7. Identify the estimated time frame for completion of the project, and a timeline. If the project takes longer than the estimated time frame, you will need to resubmit your application and receive approval to continue to use the data.

8. List and describe proprietary interests in this research, if applicable.

9. Describe why the HSCRC Public Use Data Files are insufficient for your data needs.

10. Please list the **specific** confidential data elements required and justify why each is required (see data dictionary link http://hscrc.maryland.gov/Pages/hsp_info1.aspx).

11. Decisions about release are made for each confidential data element request, rather than for the data set as a whole.

12. Provide a detailed description of your data security and confidentiality plan as it pertains to the use and storage of the data requested (HIPAA implementation and security system, confidentiality regulations, encryption). If a computer vendor is used, please specify. Who will have the main responsibility for notifying HSCRC of any suspected incidents wherein the security and privacy of the HSCRC data may have been compromised? Who will be notifying HSCRC of any project staffing changes, tracking the status and roles of the research team?

13. Provide verification that your entity/business unit functions is covered by HIPAA Regulations, and it complies with HIPAA Privacy. Please explain in detail.

14. 42 CFR Part 2 (Part 2) cases will not be include in the approved request.

15. Read and sign the Data Use Agreement.

APPENDIX M. REDUCE COMPUTATIONAL REQUIREMENTS

Researchers are not consistent in using the terms blocking, filtering, and indexing. For this document, blocking refers to the practice of selecting a subset of records for data linkage (or other types of processing) by using keys. Filtering is another method of defining subsets of data that defines exclusionary criteria to discard dissimilar records, in contrast to the blocking key, which groups similar records by inclusion criteria [89]. Indexing organizes the records in a database, such as by blocking keys, so that subsets of records with similar properties can be quickly accessed.

As data sets grow, it becomes computationally intensive and impractical to compare all records to one another. Blocking and filtering are optional methods to reduce the number of records to compare by selecting subsets of records that are more likely to match. Matching every record in a data set against every record in another (or the same) data set can require billions of data linkage operations, each of which might require multiple variable comparisons. For example, identifying potential duplicates among a set of 10,000 records would require 49,995,000 comparisons. Linking 10,000 emergency medical services (EMS) records to a set of 100,000 police crash records would require 1,000,000,000 record comparisons to identify the linked records in those two data sets. Consequently, blocking and/or filtering performed before data linkage conserves computational resources and reduces the computing time.

The processing power and time required for data linkage depends on the number of record comparisons. The processing time can be reduced by dividing the data into multiple subsets of records that are more likely to match. Both deterministic and probabilistic methods can be used with blocking and filtering techniques. Essentially, strategies can be employed to divide a complex data linkage task into series of smaller data linkage tasks involving fewer record pair comparisons. Depending on the strategy selected, some records are excluded from being compared by the data linkage tool.

The process of designing criteria for creating subsets in the data are like designing a binary deterministic rule to identify whether records match in variable-specific criteria and combining the variable-specific criteria from multiple variables and the output of applying the criteria leads to a dichotomous result (either a record is in the subset or it is not).

Select Subsets of Records

Many strategies are available for reducing the number of records compared in the data linkage process [90]. Blocking and filtering are two such strategies and can involve one or more variables to create multiple subsets for matching.

Selecting a subset of records: Blocking. A *blocking* key is created based on the information in record variables. The blocking key can contain information based on a single variable or multiple variables. The subset of records with the same blocking key is called a *block*. Only those records within the same block are compared during the linkage process, so only records with the same or similar variable values can be linked. Blocking keys can be based on information from one or more variables. Multiple blocking keys can be used to represent different pieces of information about a record, meaning each record can be associated with multiple possible blocking keys.

The concept of a *blocking key* is very similar to a binary deterministic rule used by a data linkage tool for comparing values within a variable to determine a variable-level match. For data linkage, an index of which records correspond to which blocks is created for both databases, and data linkage operations are executed only on record pairs from corresponding blocks in both databases. Consequently, the number of data linkage operations executed is significantly reduced, which greatly decreases the processing power and time required to perform data linkage.

Blocking keys can be created using different techniques.

Block values of a variable. A blocking key is based on the values of one or more variables in a record. A blocking key can be based on the exact value of the variable or part of the information captured in that value. For example, using the month value of a birth date as a blocking key would create 12 different subsets of records from the database (one block for each of 12 months). In another example, assume a set of 100,000 crash records and a blocking key consisting of the first letter of the surname variable, or 26 blocks. Suppose that all the crash records in the set of 100,000 crash records contain surname values, and that 1 out of 26 (1 divided by 26) crash records have surnames beginning with the letter A. The block of records with surname values that begin with the letter A would contain 3,846 records (100,000 divided by 26). Using similar assumptions, the corresponding block in the ambulance database would have about 385 records (10,000 divided by 26). This blocking method reduces the number of record pairs selected for data linkage from 1 billion (100,000 multiplied by 10,000) pairs to 1,480,710 (3,846 multiplied by 385) pairs. After data linkage operations for the block of records with surname values that begin with the letter A is completed, the next block (i.e., letter B) is compared until all 26 blocks have been processed by the data linkage tool.

Transform values in a variable. Encoding is the process of transforming the values in a variable. The transformed value can be used to select records that are similar. The encoded values are also used in subsequent data linkage process. Frequently, surnames are encoded or transformed using methods that reduce the number of characters. For example, one of the oldest techniques that is still used is the Soundex method (Russell and Odell, 1918), which represents all names in a code consisting of the first letter in the name followed by three distinct numbers. Blocking on Soundex codes provides about 3,900 possible blocks of records for matching, each of which will have far fewer records than the 26 blocks provided by the first letter of the name alone (26 possible letters).

Other methods of encoding names based on pronunciation are listed below.

- *Phonex* (Lait and Randell, 1993)
- New York State Identification and Intelligence System (NYSIIS) method (Taft, 1970)
- *Oxford Name Compression Algorithm (ONCA)*, which combines the Soundex and NYSIIS methods (Gill, 2001)
- *Double-Metaphone* algorithm, which incorporates pronunciations for European and Asian names (Philips, 2000)

Numerical values can be blocked by defining intervals (e.g., decades for age values; the initial two or final two digits for zip codes). Dates can be blocked by combining the month and year into a month-year code (e.g., *mar2003*).

Selecting a subset of records: Filtering. Filtering criteria can be used independently to generate a subset of records for matching or can be applied after blocking. For example, Vastsalan et al. uses filtering to exclude record pairs from match comparisons when the number of letters in the surname of one record exceeds the number of letters in the surname of the other record by more than a specified threshold [89]. A record pair with surnames *Smith* and *Macmillan* would be excluded because *Macmillan* is four characters longer than *Smith*. Murray describes filtering as discarding "any pairs not excluded by initial indexing steps, but which are still unlikely to be a match" (p. 6); he provides examples that require comparing variable values. Record pairs selected by blocking on month-year values could be filtered by computing an approximate matching score for the surnames and excluding records with scores below a specified threshold [91]. This additional filtering step can reduce the number of record pairs processed by matching operations, but it requires additional computation for the approximate matching operations.

Risks of Blocking and Filtering

Because not all the records are compared, blocking can potentially lead to unidentified matches among records that were not compared. Because blocking and filtering methods will affect the quality of the data linkage, it is important to devise a blocking strategy that will balance the number of comparisons performed with the number of true matches that might be excluded from the linked data. An ideal strategy (1) minimizes the number of record pairs processed by data linkage operations and (2) includes all true matches in the record pairs selected for processing. However, as the number of record pairs decreases, the likelihood that true matches will be excluded increases. This tradeoff is reflected in identifying blocking keys that select higher numbers of small record sets, rather than blocking keys that select fewer numbers of large record sets.

Another important factor to consider when devising a strategy to reduce the number of record pairs selected for matching is variable data quality. Ideally, variables used for blocking will have no errors, and every record will have a value for the variable. Missing values do not provide useful blocks and raise issues, as discussed in Appendix N.

Variable value frequencies and the size of the data set are important to consider when devising a blocking strategy because these factors drive the number of record pairs compared for data linkage. Christen provides the example of blocks generated by common surname values, such as Smith and Miller [90]. If one percent of records contains the surname Smith, then a database of 1 million records will contain 10,000 records with that surname. If the database is linked to another database with the same properties, then blocking on Smith will generate 100 million record pairs, which might still be an impracticably large number of records to compare in a data linkage task.

The design of the blocking or filtering criteria that are used to generate the data subset might affect the performance of the data linkage tool and potentially create false positive matches. Therefore, it is important to understand which variable or data linkage method is used by the tool.

Blocking on one variable. An important limitation of blocking is that the data linkage tool will not have the opportunity to consider any true matches that do not agree on the blocking key. There are many data entry errors (e.g., spelling, typographical) that can lead to values in a variable not agreeing on the blocking key between two records that should be identified as a match. For example, if blocking on the first letter of the surname, surname misspellings, such as Calder versus Kalder, could cause a record match to be missed because the records would fall into different blocks, rather than the same block. A data linkage strategy using the same data linkage tool that does not employ blocking could potentially identify these records as matches. To address the problem of excluding records from comparison that differ on the blocking key, most data linkage analysts will employ a blocking strategy that selects records based on different variables, rather than just a single variable. For example, the record pairs selected for matching might consist of records with the same surname Soundex key value, month-year birth date value, and initial-two-digit zip-code value. Murray explains that this approach "is also computationally efficient since it can be implemented by merging the results of multiple blocking queries. For these reasons, it is widely used in practice" [91].

Encoding with soundex. Also, if encoding (e.g., Soundex) is used for blocking on the surname, then blocking on the code that includes Smith will generate even more record pairs because the blocking will include Smithfield, Smathers, and other surnames with the same Soundex code. In these cases, it is useful to incorporate multiple variables in the blocking key. For example, the blocking key might consist of the surname Soundex code followed by the initial two digits of the zip code. Alternatively, an additional filtering step might be used to reduce the number of record pairs.

APPENDIX N. MULTIPLE IMPUTATION AND MISSING DATA

Missing data is a problem for research because it reduces the confidence that results obtained in analyses of data will generalize beyond the sample used in the study. Analyses conducted without the missing data are not based on a representative sample of the population and might be biased, especially if there are systematic underlying reasons for the missing values of variables used in the study. This appendix addresses the significant data quality issue of missing data and imputation methods used to overcome the problem.

The method of multiple imputation is also used in studies based on linked records to select the sets of matched records that will be used for analysis. In this approach, the actual match status of record pairs is treated as missing data, and match status is imputed based on the probabilities assigned by probabilistic matching. Multiple imputation of match status is also described in this appendix.

Assessing Missing Data

Cook et al. recommends that the mechanisms giving rise to missing data should be characterized into one of three probability categories, based on how the missing values are distributed across the sample, and gives examples from motor vehicle crash (MVC) data [43]:

Missing Completely at Random (MCAR). The probability that a data point is missing is independent of all other observed and unobserved characteristics of the study sample. In other words, subjects with missing data are a random sample of the study population. For example, in an emergency-medical-services (EMS) data set sorted in a random order, transport time was deleted for the top x number of patients.

Missing at Random (MAR). The probability that a value is missing depends on the observed values in the sample but is independent of any unobserved or missing values. In other words, the observed data contains information that explains the mechanism of missing data up to an element of randomness. For example, one hospital failed to report charges.

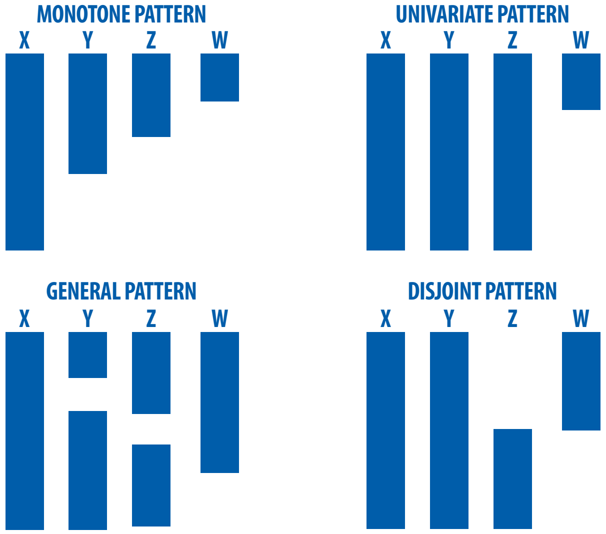
Missing Not at Random (MNAR). The probability that a value is missing depends on unobserved variables or the missing value itself. Consequently, it is impossible to estimate missing values by using other values that are present in the data set. There can be a procedural reason why some values are routinely missing, or a difference in the consistency of data collection methods among the different organizations or individuals collecting data. For example, transport time in an EMS data set might be missing more often for patients transported by a certain agency and the data set does not include information on which agency transported the patient.

Because there are no statistical analyses that can be used to distinguish MCAR, MAR, and MNAR, the data collection process must be investigated to determine the likely mechanism of missing data. If values are missing, the data owner should be contacted to discuss and understand possible causes for the missing data and to assess the missing data probability category by asking the following questions:

- Does an empty data field imply a null value (null is a placeholder in a database field indicating that the value is unknown)?
- Is there another way of indicating null values?
- What are the default data field values, if any? Actual missing values might be masked by default data field values.
- Are certain data fields reliably populated?
- Are there reasons that might cause the missing values to be distributed across the records in a specific pattern?

The discussion with the data owner might improve their data collection processes going forward or provide additional data to substitute the missing data. In addition, the next step is to conduct an aggregate analysis to understand the pattern of missing data (i.e., how the missing values are distributed across the records) to determine the most appropriate way to handle the missing data. Figure 11, which is from Cook et al., shows a graphical representation of four missing data patterns, where the columns represent data fields, the vertical axis represents observations, and a gray pattern represents the presence of observed data values within each data field [43].

Figure 11. Patterns of Missing Data



Imputation of Missing Data

There are many methods of imputing missing data, but all methods rely on imputing the values based on the estimated values determined by the distribution of the values observed in the data, which are randomly assigned to the missing data fields [43]. The number of imputed data sets generated through this process can be one or multiple.

Single imputation of missing data. Single-imputation methods replace missing values with estimated values to generate a single imputed data set, which is then analyzed. For example, a simple tactic is replacing all missing values with the mean for that value, which is computed from the other records in the data set. Gelman and Hill's (2007) critique of this method illustrates the kinds of consequences that researchers must consider when selecting imputation methods:

> Unfortunately, this strategy can severely distort the distribution for this variable, leading to complications with summary measures including, notably, underestimates of the standard deviation. Moreover, mean imputation distorts relationships between variables by "pulling" estimates of the correlation toward zero [92].

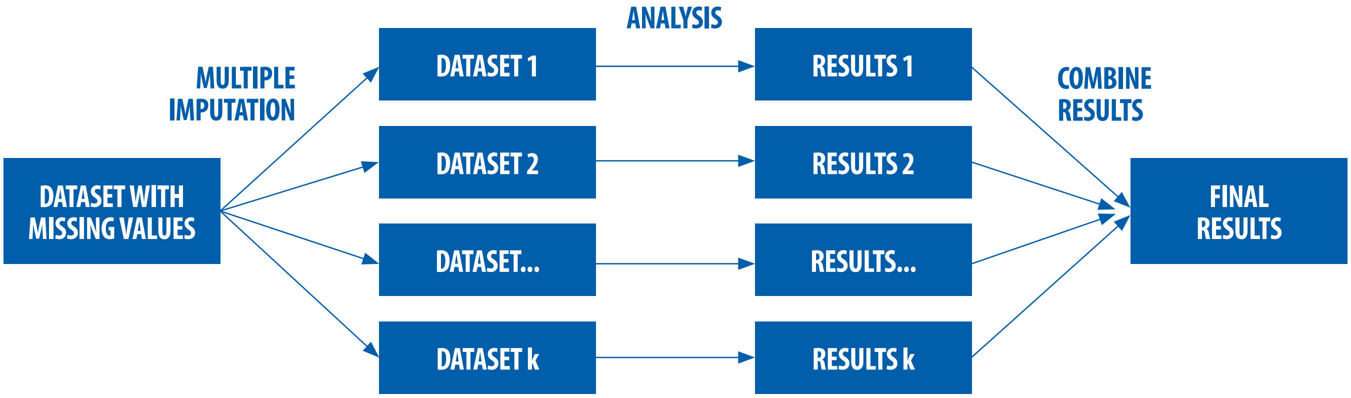
More sophisticated methods predict missing values based on regression models that consider the patterns of values among multiple variables in the data. For example, missing values for income might be predicted from a model that includes values for gender, age, and education level. This kind of modelling becomes complex when records have missing values for more than one variable. Alternatively, "hot-deck" [92] or cell [43] imputation matches records with missing values to records with similar values for the non-missing variables, and replaces the missing values with values from the matched records. Similar records might be matched from the same data set or from an external data set.

Cook et al. (2015, p. 43) describe the problems associated with single imputation: "Single imputation methods generally do not account for the uncertainty inherent in the imputed values, might result in underestimates of variances and inflated type I error, and might introduce additional bias into the data set" [43].

Multiple imputation of missing data. Cook et al. recommend using multiple-imputation methods, in which missing values are imputed multiple times to create multiple data sets for analysis [43]. According to Cook et al., multiple imputation can overcome some of the disadvantages of complete-case analysis and single imputation methods in many cases because multiple imputed data sets allow all cases to be included in analyses and accounts for the uncertainty inherent in the imputed value. In cases of missing data due to Missing Completely at Random (MCAR) or Missing at Random (MAR) mechanisms, multiple imputation leads to unbiased results and at least as much, if not more, statistical power than the exclusion of cases with missing data [43]. Like other imputation methods, multiple imputation methods cannot completely overcome bias introduced by Missing Not at Random (MNAR) mechanisms of missing data (e.g., bias due to unobserved variables or by the variable whose value is missing), but they have been shown to be less biased than other methods in these cases [93].

Data sets for multiple imputation are usually produced using models in which a random component is introduced when generating the missing values to account for the uncertainty of the prediction. Another common alternative is cell imputation, in which the matched record from which missing values are copied is randomly selected from a set of similar records. The data sets must be analyzed using special analytic methods described by Rubin, in which each imputed data set is individually analyzed, and then the results are combined, as illustrated in Figure 12 (Figure 1.4.2 from Cook et al.) [43, 94].

Figure 12. Overview of Analyzing Data by Using Multiple Imputation Methods [43, 94]



Multiple Imputation of Match Status

Like multiple imputation of missing data, multiple imputation of match status requires construction of multiple data sets for analysis. Each data set contains a different set of linked records, which are analyzed as if the records were true matches. Consequently, the match status of linked record pairs in the data sets is imputed, rather than based solely on the match probability. Instead, the linked data sets are selected based on match probabilities, and multiple sets are used for analysis. The results of analyses are then combined as illustrated in Figure 12 above using methods that consider the variability between imputations for final statistical parameter estimates [6, 95].

The imputed data sets are selected randomly for analysis from among all records with match probabilities that exceed a very low threshold, such as 0.01. Selection of record pairs for inclusion in the data sets is weighted by the match probabilities of the record pairs so that "pairs that have a probability of 0.90 of being correct are selected in about 90 percent of all samples… and pairs with a probability of 0.01 are selected in only about 1 out of 100 samples" [6, p. 10]. An additional Markov chain Monte Carlo (MCMC) step between selection of linked data sets is recommended to avoid certain dependencies and ensure that the sets of matched pairs are not related [6]. Cook et al. report that standard CODES practice is to take at least five samples. They also describe some alternative methods for constructing a single imputed data set.

Using multiple imputed match sets has several advantages. A significant advantage is that human review of the matches obtained from probabilistic matching is not needed. Because match status is imputed, it is not necessary to adjudicate records in the mid-range of probabilities between high probabilities which permit high confidence that the matches are correct and low probabilities which provide high confidence that the records do not match. Another benefit is that data sets for analysis can be larger than they would be if limited to records with high probabilities. Imputation is a useful tool when the information in the data is sparse, especially when very large numbers of records must be matched with relatively few common attributes. There is also evidence that multiply imputed match sets can avoid biases that might be introduced by selecting only records with high match probabilities, preserving the distributional features in the data more accurately [6].

A disadvantage of the method is the added effort and statistical sophistication required of analysts. The process requires specialized tools and skill with "additional SAS procedures (PROC MI and PROC MIANALYZE) and plug-ins (IVEware) in order to use the linked results" [6, p. 10]. Also, if there is plenty of information available to link with, the utility of imputation is negligible and can negatively affect the precision of the results. Finally, multiple imputation of match status is not appropriate for analyses that require tracking individuals: the advantages of the method apply only to population studies.

APPENDIX O. ASSESSING DATA QUALITY: VARIATION

To ensure effective data linkage, it is important to understand the types of variation in values that have the same meaning for each variable. A *variant taxonomy* is the set of variations for a given variable.

Table 19 describes the varying types of element variations and structural variations that can occur within a variable. Variation can occur at the following two different organizational levels within a variable, as depicted by the grey bars in Table 19 Table [96]:

- Data element variation: Variation that is limited to a single segment of a value of a variable type. The definition of a segment definition varies depending on the variable type.
- Structural variation: Variation that involves more than one single segment of a value and the interactions between those segments.
- Other variation: Variation that is not at the element or structural level.

Examples of each type of variation are provided in Table 19 when available. In Table 19, "N/A" in the variable columns signifies that the variation is not applicable to the variable.

Table 19. Types of Variations for Three Common Variables

Data Element Variations

Variation	Variation Subcategory	Variation Description	Name or Common Variables	Date-of-Birth or Common Variables	Passport Number or Common Variables
Data error	Optical Character Recognition (OCR)	Errors are caused by scanning paper documents.	Francis Huger becomes Francis linger		0909035509 becomes 0909035590
Data error	Truncation	Errors are missing characters at the end of a name, date, or passport number, and are usually the result of entering too many characters into a variable that is too short.	Carstanjen becomes Carstanj	04/08/1983 becomes 04/08/19	0909035509 becomes 09090
Data error	Typographical errors or transpositions, additions, deletions, and substitutions	Errors are from typing mistakes.	Rahman becomes Rahamn	07/06/1956 becomes 07/06/1965	0909035509 becomes 0909035590
Particles A name component that provides linguistic information, rather than name or identification information (e.g., de, al, and von generally indicate "from" or "of," while bin indicates "son of").	Particle segmentation	A segment is the building block of a name, such that at least one segment is used to construct a name. Segments can include particles. Concatenation with other name parts, which is often a transliteration issue. There are restrictions on permissible characters. There are spelling variations. There is case-based segmenting.	Abd al Rahman becomes Abdal Rahman Smith-Jones becomes SmithJones De Los Angeles becomes Delosangeles McHenry becomes Mc Henry	N/A	N/A

Continued

Variation	Variation Subcategory	Variation Description	Name or Common Variables	Date-of-Birth or Common Variables	Passport Number or Common Variables
Particles	Particle inclusion or omission	Name pairs that differ in the particles included (e.g., in Arabic (al or bin) and in Hispanic names (De)).	Saddam bin Husayn bin al-Majid becomes Saddam Husayn al-Majid al Tikriti becomes Tikriti Maria Rodriguez De Gonzalez becomes Maria Rodriguez Gonzalez	N/A	N/A
Short forms	Abbreviations	Common, shortened versions of certain words	Muhammad becomes Mhd James becomes Jas Maria del Carmen becomes Ma del Ca	March 1998 becomes Mar 1998	N/A
Short forms	Initials	The first letters of name parts.	Charles becomes C Mohamed Bin Ahmed Hosein becomes Mohamed B Ahmed Hosein	N/A	N/A
Short forms	Month numbers	Using month numbers versus spelling out the month.	N/A	March 1998 becomes 03/1998	N/A
Short forms	Dropping leading zeros or leading year digits	Leading zeroes might be present or absent in the date of birth.	N/A	03/1/1998 becomes 3/1/1998 3/1/1998 becomes 3/1/98	N/A

Continued

Variation	Variation Subcategory	Variation Description	Name or Common Variables	Date-of-Birth or Common Variables	Passport Number or Common Variables
Spelling	Alternative spellings	Common variations in the spelling of written names within a single language and script.	Catherine becomes Catharine or Katherine or Kathryn	N/A	N/A
Spelling	Transliteration	Alternate ways of Romanizing names from other scripts.	Husayn becomes Hussein or Hossein Wasim becomes Ouassim	N/A	N/A
Spelling	Non-word characters	Non-alphabetic characters appearing in a name, which could represent non-Roman characters, code, or vernacular.	‘Abd or Abd Ke!!y or Kelly	N/A	N/A
Nicknames and diminutives	N/A	Alternative names that are commonly used to refer to the same person.	Robert becomes Bob or Bobby Concepcion becomes Concha or Conchita	N/A	N/A
Translation variations	N/A	Equivalent names in different languages.	John becomes Jean or Juan Yahya and Joseph becomes Giuseppe or Yusif	N/A	N/A
Case variations	N/A	Equivalent names that differ according to the case used when writing the name.	Jurgen or JURGEN	N/A	N/A
Presence or absence of titles, affixes, qualifiers (TAQs)	N/A	Used to identify added or deleted TAQs, as well as spelling and translation variants of TAQs.	Mr. Schmidt or Herr Schmidt, and Frank Jones or Frank Jones Sr	N/A	N/A
Date range	N/A	N/A	N/A	N/A	N/A

Continued

Structural Variations

Variation	Variation Subcategory	Variation Description	Name or Common Variables	Date-of-Birth or Common Variables	Passport Number or Common Variables
Permutations	N/A	Reordering of name elements.	Clara Lucia Bolivar becomes Lucia Clara Bolivar	N/A	N/A
Element Segmentation	N/A	When two name elements are concatenated together.	Nur Mohammed Amin becomes Nur Mohammedamin	N/A	N/A
Field variations	N/A	A field is a section of a name that indicates some type of meaning, such as given name, family name, or patronymic. The order of name fields varies across cultures, such as given name followed by last name in English, but family name followed by given name in Korean. Name fields contain at least one segment, except in cases of missing data (see deletion below). When the full name has been divided into given name and surname in different places.	Maria GONZALES LOPEZ becomes Maria Gonzales LOPEZ Kim JONG JR becomes Jong Jr KIM	N/A	N/A
Deletion or Addition	N/A	The removal or addition of optional name elements like middle names, matronymics, or passport country of issuance. This variation can also be the result of incomplete data.	John C Smith becomes John Smith Maria Gonzales Lopez becomes Maria Gonzales	N/A	DEU F784756045 becomes F784756045
Placeholders	N/A	Missing name parts or country of issuance that have been marked in certain ways.	FNU, LNU, and XXX	N/A	DEU F784756045 becomes UNK F784756045

Continued

Other Variations

Variation	Variation Subcategory	Variation Description	Name or Common Variables	Date-of-Birth or Common Variables	Passport Number or Common Variables
Alias or AKAs (“also known as”)	N/A	When two names refer to the same individual but lack a linguistic connection.	Bruce Wayne becomes Batman	N/A	N/A
Date Range	N/A	Dates considered variants due to their proximity in a given span of time.	N/A	5 Aug 1972 = 7 July 1974 within a span of <3 years	N/A
Competing standards	N/A	Equivalent acronyms or abbreviations for a country that differ across standards or languages.	N/A	N/A	Germany = GER = DEU = DE
Other variations	N/A	When the variations are structural in nature, but not one of the categories listed above.	N/A	N/A	N/A
Undetermined	N/A	When the variations are structural in nature but cannot be precisely determined.	N/A	N/A	N/A

APPENDIX P. EVALUATING DATA LINKAGE PROCESSES

Evaluating data linkage processes is essential. Linkage programs can help all stages of the public health process, including designing new public health interventions, by producing high-quality linked data. Evaluation of data linkage processes allows decision makers to perform the following actions:

- **Determine** the suitability of data linkage processes for stakeholder needs.
- **Understand** the strengths and weaknesses of data linkage processes.
- **Monitor** the performance of the linked data over time.
- **Test and measure** improvements to existing data linkage processes.
- **Evaluate** interventions.

To assist states in developing an evaluation process, a general evaluation approach has been outlined. The purpose of this appendix is to describe the general evaluation approach and to provide an example of using the approach. The example is based on the tool evaluation (i.e., LinkageWiz, R, SAS, WinPure) performed by the authors of this guide with the Georgia Department of Public Health (GA DPH) and the University of Maryland, Baltimore (UMB).

The general evaluation approach can be adapted to account for other complexities that states, or other stakeholders might experience with their data agreements, data sets, methods, tools, and other areas of differentiation. The time and resources available are important factors to consider when scoping the evaluation; there might be characteristics that are important to the linkage program or its stakeholders that cannot be included in the formal evaluation due to limited resources.

Create an Evaluation Plan

The general evaluation approach was adapted from two theoretical frameworks: Expert Advisory Group on Language Engineering Standards (EAGLES) 7-step recipe [97] and Framework for the Evaluation of Machine Translation in ISLE (FEMTI) [98].

The seven steps below were based on the general approach and tailored for linkage programs. For ease of reference, the term "evaluation plan" is used in this guide to collectively refer to the first through sixth planning steps of the EAGLES recipe, as the seventh step is to execute the evaluation plan and interpret the results.

For the tool evaluation performed with GA DPH and UMB, the steps the General Approach for Evaluating Data Linkage Methods were performed and are highlighted in these boxes throughout this appendix.

General Approach For Evaluating Data Linkage Methods

1. Define evaluation objectives.

2. Define tasks that will be performed with the data sets.

3. Select high-level quality characteristics (e.g., accuracy and usability).

4. Define requirements (i.e., what is needed for the data linkage process?).

5. Define metrics to measure how well the requirements were met.

6. Plan the evaluation in detail.

7. Execute the plan.

Step 1: Define Evaluation Objectives. Objectives drive decisions during the evaluation process. There could be multiple goals for performing an evaluation, including identifying the cause of incorrect or missing links in data linkage results, determining whether a data linkage process is feasible to implement, or understanding how well a data linkage tool or method performs. A formal evaluation can allow decision makers to perform the following specific actions:

- Estimate the risk of using a data linkage process.
- Determine the suitability of a data linkage process for stakeholder needs.
- Compare data linkage tools.
- Understand the strengths and weaknesses of data linkage methods.
- Devise and test processes to leverage the strengths and/or mitigate the weaknesses of data linkage methods.
- Monitor the performance of data linkage results over time.
- Test and measure improvements to existing data linkage processes.

For the tool evaluation performed with GA DPH and UMB, the primary goal was to compare the capabilities of data linkage tools.

Step 2: Define Tasks. Tasks include all steps, as illustrated in Figure 2, and helps identify which aspect of the process will be considered during the evaluation. States have significantly different resources, management structures, data sets, and data linkage processes. The variety of circumstances pose a challenge in performing an evaluation of tools and methods. The tasks should be adapted to address the unique circumstances.

For the tool evaluation performed with GA DPH and UMB, each state selected their own data sets. Data quality issues were resolved prior to the evaluation. The evaluation focused primarily on the data linkage steps.

Step 3: Select High-Level Quality Characteristics. It is important to define the high-level characteristics of the data linkage process that is being assessed in the evaluation, so the results can be interpreted and used.

For the tool evaluation performed with GA DPH and UMB, the primary high-level characteristics assessed included: interoperability, accuracy, efficiency, and usability.

See page 84 for a list of possible characteristics that can be evaluated. FEMTI outlines multiple high-level characteristics [99, 100] that have been modified for data linkage evaluation in this list. Characteristics include the tool's applicability to state linkage programs, ease of use, price, software maturity, long-term potential, interoperability, usability, efficiency, and accuracy of the linked data.

Table 20. Important High-Level Characteristics That Can Be Assessed During an Evaluation

Characteristic	Details
Applicability to state linkage programs	▶ Versatility of data linkage capabilities (e.g., customizable parameters; capabilities for inexperienced and experienced users; deterministic and probabilistic methods) ▶ Multiple functional modules other than data linkage (e.g., data pre-processing; statistical analysis; visualizations)
Ease of use	▶ Learning curve (e.g., quality of user manual and documentation; technical support; training courses available; familiar terminology to typical user) ▶ General usability considerations ▶ Is tool used by states performing data linkage?
Price	▶ License ▶ Other computing requirements (e.g., operating system)
Software maturity	▶ Technical support provided by company ▶ Features of software ▶ History of innovation and updates
Long-term potential	▶ Sustainability of vendor ▶ Enterprise tools that could be used at the national level
Interoperability	▶ Data sharing and analysis ▶ Ingested by other tools for pre-processing and analysis
Usability	▶ Learnability of the software by a user ▶ Operability of software by multiple users with varying degrees of expertise ▶ Documentation
Efficiency	▶ Pre-processing ▶ Total end-to-end processing time ▶ Resource utilization
Accuracy of the linked data	N/A

Step 4: Requirements for Data Linkage. Requirements address the different features that a data linkage process should provide and are based on the evaluation characteristics selected in step 3. Requirements should be defined so that it is possible to measure whether, or how well, a data linkage process meets that requirement. Listed below is a general set of requirements for data linkage.

- Tool must be usable by staff with a base set of computer skills.
- Tool must accept data sets provided by data owners.
- Tool must perform data linkage.
- Tool must link data sets in a reasonable amount of time.
- Tool must create a data file that can be used for statistical analysis by other software.

For the tool evaluation performed with GA DPH and UMB, the data linkage tools selected included LinkageWiz, LinkSolv, R, SAS, and WinPure

Step 5: Define Metrics. Evaluation metrics are used to assess how well the data linkage processes and results meet the requirements defined in step 4. Step 5 consists of defining appropriate metrics and determining how those metrics will be assessed. The use of evaluation metrics makes it possible to compare whether one data linkage tool meets the requirements better or worse than another tool.

For the tool evaluation performed with GA DPH and UMB, CSV and Microsoft Excel formats were tested for interoperability.

Interoperability metrics addresses whether a tool can import data in a specific format and export results to a specific format. Common import and output formats include comma-separated values (CSV), JavaScript Object Notation (JSON), Microsoft Access, Microsoft Excel, text, and extensive markup language (XML). If a data linkage tool does not support interoperability, it might be necessary to convert data into a different format to be imported into the data linkage tool or convert output data into another format for further analysis and integration with other tools. Data linkage tools that handle more import and output formats meet system requirements better than tools that handle fewer formats.

Still other evaluation metrics require analysis of the output of the data linkage tool. Accuracy metrics indicate how well a tool links data records across multiple data sets that represent the same entities. Different accuracy metrics can provide insight into different aspects of data linkage performance. Accuracy metrics are obtained by reviewing the matches and non-matches produced by a data linkage tool.

To calculate accuracy metrics the linked data must be compared with ground-truth (or answer-key) data. The term ground-truth data implies that the classification of data into matching and non-matching records is known with certainty. In practice, however, it is nearly impossible to know with certainty whether two records should match, as this requires information that might not be present in the data sets being linked. Alternatively, simulated data sets can be used where the ground truth is known. Table 21 demonstrates how comparisons of the data linkage results to the ground truth are categorized as true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs).

Table 21. Measures of Performance for Data Linkage Methods

For accessibility, see full explanation of table shown below in Appendix R.

		Data Linkage Tool Results	
		Matched Records	Non-Matched Records
Known Match Status (Ground-Truth or Answer Key Data)	Match	True Positive (TP) "True Matches"	False Negative (FN) "Missed Matches"
	Non-Match	False Positive (FP) "False Matches"	True Negative (TN) "True Non Matches"

Performance measure values can be used to calculate the following accuracy metrics:

Positive predictive value (PPV). PPV, also known as precision, is defined as the percent of correct matches to the total matches (TP divided by the sum of TP and FP). PPV reflects the probability that a record pair determined to be a match is a true match. Higher PPV scores indicate that the data linkage method matches are accurate and include fewer false matches.

Sensitivity. Sensitivity, also known as recall, is defined as the percent of correct matches to the total true matches (TP divided by the sum of TP and FN). Sensitivity reflects the method's success in identifying all true matches. A low sensitivity score can be caused by a high frequency of missing values in the variable fields.

Specificity. Specificity, also known as true negative rate, relates to the tool's ability to identify true non-matches. Specificity is calculated as TN divided by the sum of TN and FP.

Negative predictive value (NPV). NPV reflects the probability that a record pair determined to be not a match is a true non-match (TN divided by the sum of TN and FN). Higher NPV scores indicate that method matches include fewer missed non-matches.

F1 Score. It might also be useful to calculate the F1 score, which is the harmonic mean of precision and recall. It provides a single metric for data linkage accuracy.

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

For the tool evaluation performed with GA DPH and UMB, the following accuracy metrics were selected to address how well and how thoroughly the data linkage tools link data:

1. Positive predictive value or precision
2. Sensitivity or recall
3. Specificity or true negative rate

Other metrics require use of the data linkage tool and observation of the tool during operation. To measure efficiency metrics, different types of information are collected each time that a data linkage tool is used, such as:

- Amount of time to perform any necessary data preprocessing.
- Data linkage tool used.
- Data linkage tool parameter configuration used. Deciding which methods to run in data linkage tools is discussed in Appendix E. Different parameter settings (e.g., data linkage thresholds 100% and 70%) can be used to assess which parameters will lengthen the data linkage run time. A linkage threshold of 100% requires perfect precision as records are only linked if linkage variables match exactly.

A linkage threshold of 70% loosens the restrictions.

- Number of data sets linked; more data sets require more run time.
- Number of records in each data set; more records require more run time.
- Amount of time to complete the entire data linkage process, including any preprocessing. This can provide insight into which parts of the process should be allotted the most time.
- Maximum Random-Access Memory (RAM) used during data linkage. This can be used to determine whether existing computing infrastructure is sufficient.

For the tool evaluation performed with GA DPH and UMB, the data linkage tools selected included LinkageWiz, LinkSolv, R, SAS, and WinPure. An attempt was made to evaluate each tool with linkage thresholds of 100% and 70%. Crash and hospital discharge records were linked (max 250,000 records). The amount of time to run the data linkage tool for each linkage threshold was recorded.

These efficiency metrics can be used to estimate whether data linkage can be successfully completed given existing staff availability and computing resources.

Usability metrics can provide information on how easy or difficult it is to use different data linkage tools. Methods for obtaining usability metrics can range from assessing user feedback to conducting controlled laboratory tests. The System Usability Scale (SUS) is an industry standard for understanding individual perception of usability. The Department of Health and Human Services (HHS) collaborates with other government, private companies, and universities to publish information about user experience [101, 102]. This site recommends SUS as one of the methods of performing usability evaluations, and notes that it works especially well in cases where there are only a few respondents. SUS is technology independent and can be used outside the bounds of a usability test. The System Usability Scale metric is a numeric value that can provide information on a single data linkage tool or can be used to compare the usability of multiple data linkage tools. If a tool has poor usability metrics, this could indicate that users will not be inclined to use that tool.

The following questions alternate between positive and negative interpretations of usability. This affects the scoring calculations. A Likert scale from 1 (strongly disagree) to 5 (strongly agree) is used to answer each of the questions [103]:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought that the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found that the various functions in this system were well integrated.
6. I thought that there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system to be very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

For the tool evaluation performed with GA DPH and UMB, the usability metrics included using System Usability Scale metric with the following steps, which were completed by each tool user:

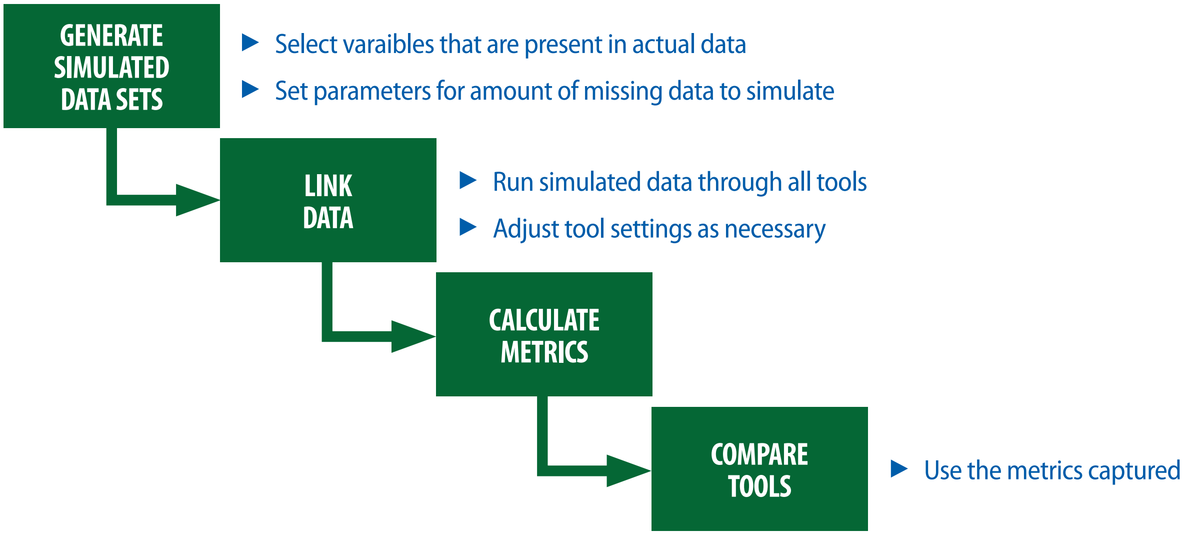
1. Carried out data linkage.
2. For each data linkage tool, reflected on personal experience relating to that survey item.
3. Documented responses for each tool.

When scoring the range of scores will be between 0 and 100, rather than providing a numerical score tool evaluation assigning scores of A (90-100), B (80-89), C (70-79), etc. is preferred due to the subjectiveness of the evaluation questions. For numerical scoring, sum odd and even adjustments. For the odd questions subtract 1 from the users' Likert scale responses. For even questions subtract the users' Likert scale responses from 5. Sum the odd and even adjustments then multiply the sum by 2.5. This will provide a numerical score between 0 and 100. Again, when interpreting the scores, it is best to define ranges of scores (e.g., A, B, C).

Step 6: Plan the Evaluation in Detail. In step 6, the detailed processes are determined to carry out the evaluation and obtain the metrics identified in step 5. It is necessary to consider the specific data sets and resources available when selecting an evaluation approach.

Even states that have largely identified data (e.g., GA DPH and UMB) confirm that there is not enough information in the crash and hospital records to assess if a match is correct when a human or clerical review is done. The only exception are clerical reviews that are funded to explore original data sources with access to other information if needed (e.g., hospital discharge record plus other medical records if variables are not conclusive). This is not feasible for routine data linkage. The states confirmed that the probabilistic features of the data linkage tools perform better than a human or clerical review of the crash and hospital match results. Since it is largely not possible to obtain ground-truth data for linkage programs, simulated data must be used to evaluate data linkage processes, which is shown in detail in Figure 13.

Figure 13. Evaluation Plan


GENERATE SIMULATED DATA SETS
Select varaibles that are present in actual data
Set parameters for amount of missing data to simulate
LINK DATA
Run simulated data through all tools
Adjust tool settings as necessary
CALCULATE METRICS
COMPARE TOOLS
Use the metrics captured

For the tool evaluation performed with GA DPH and UMB, synthetic data from LinkSolv was used, direct matching and 70% linkage threshold matching was run, and manual review of match pairs was done.

Simulated data, or data sets with known outcomes, should be as representative as possible of the real data being linked. Simulated data can be generated using look-up tables and rule sets to select values (and variants of values) for the fields in each record. Computer applications for producing simulated data are available and are included in some data linkage tools (Appendix F). The advantages of simulated data include:

- Results from linking specific variables is controlled through specific types of errors and the variation generated in the simulated data sets, which can indicate where adjustments to the tool procedures and parameters are needed to improve the data linkage quality.
- Types of errors and variants can be updated as new issues are identified in the real data sets.
- Size of the evaluation data sets can be controlled to test the scalability of the data linkage tools.

The creation of simulated data also has disadvantages. Understanding and operating the simulated data generation tools is challenging, and there are many parameters to adjust. The variants generated by the tools are limited to those that have been identified and included in the rules and look up tables; generating a data set that is truly representative of the records in data sets is not guaranteed. For example, does the distribution of ages within the population or the types of surnames commonly found in a particular geographic region reflect the real data sets? If the simulated and real data sets are very different, the metrics obtained using simulated data might not reflect data linkage accuracy on real data. Simulated data sets are designed to provide the ground truth or an answer key.

It is helpful to inspect the output of the data linkage tools when interpreting the metrics. Looking at the matched pairs can help with understanding the effect different parameter settings have and provide insight into configuration changes that might improve linkage. For example, inspection of the linked data might show that most of incorrectly linked records (i.e., false positives) contain dates of birth that differ by 10 or more days. In this case, it would be useful to update the tool parameters to only link records with dates of birth within 10 days of each other and re-run the data linkage tool.

There are many methods of reviewing match results, including:

Manual review of matches.
Manual review verifies that the methods are performing as desired. One approach is to manually inspect record pairs with match values that are close to the match thresholds or select pairs because of the presence of problematic values. For example, twins or nicknames might cause problems for a method and looking carefully at how these records are matched can reveal how well the data linkage tool performed. Another case for manual review occurs when a record appears

in multiple matched pairs. These multiple matched pairs should be inspected for issues, such as twins, dependents, and high-frequency names.

Using a threshold match score.

Using a threshold to identify true matches is a method that requires a pre-determined threshold that marks some records as true matches and others as false. There might be little or no evidence for where to set the threshold, making this approach less reliable and not recommended. Thresholds can be determined by algorithms within the data linkage tool (which might be configurable by the user) or can be selected by the user after the tool generates match scores. Thresholds might also be selected from visual examination of a match-score histogram; in this case, a threshold is selected based on the users' assessment of a threshold in keeping with the match-score distribution.

Manual inspection of the link scores' distribution.

If the linkage method generates numerical scores, rather than yes-or-no results, then the linkage quality can be examined by graphing the scores. Similar distributions—with peaks around the higher match weights (the linkages) and a larger distribution of records around a peak at very low match rates (the non linkages)—should result from any linkage method that uses continuous scores [43].

External corroboration of the linkage rates.

Linkage rates are the proportion of records, or subsets of records, that were linked by the data linkage tool. External sources of linkage rates can be used for comparison. External sources to consider include studies of similar populations, independent estimates from reporting agencies, or results of previous efforts to link similar types of records that report their linkage rates. Several data linkage evaluations that were performed between 2013 and 2017 reported linkage rates and other details of their data linkage methods [25, 58, 60, 61, 104, 105]. Keep in mind, however, that, because of differences in states' data collection methods, achieving similar linkage rates might be difficult, especially if their data sets are more standardized, are more centralized, or use consistent unique identifiers.

No-match test.

Another alternative is to match data sets that should have no matches (known true non matches) to assess the data linkages tool's rate of false matched records. For example, in an interview, Dr. Larry Cook from the University of Utah described an approach of using a data linkage tool to match real data from two different years that would have an expected match rate of zero. Any records that were matched using this approach would identify false matches and potential pitfalls of the matching methods.

Conclusion

When multiple types of metrics are used during an evaluation, it is important to interpret those metrics in combination. For example, one tool might have low usability metrics and high accuracy metrics, while another tool has high usability metrics and mediocre accuracy metrics. If data linkage accuracy is the highest priority, then it would probably be advisable to use the tool with better accuracy metrics. But if staff turnover is high, it might be a higher priority to incorporate the more user-friendly tool into a data linkage task and suffer the decreased accuracy. If budget constraints are an issue, then it might be worthwhile to examine which of the tools is less expensive and balance the priority of other types of metrics such as accuracy and usability. Interpreting multiple types of metrics at once is particularly important when comparing metrics for different tools that are being evaluated. In such cases, it is necessary to know which types of metrics weigh more heavily when interpreting evaluation results.

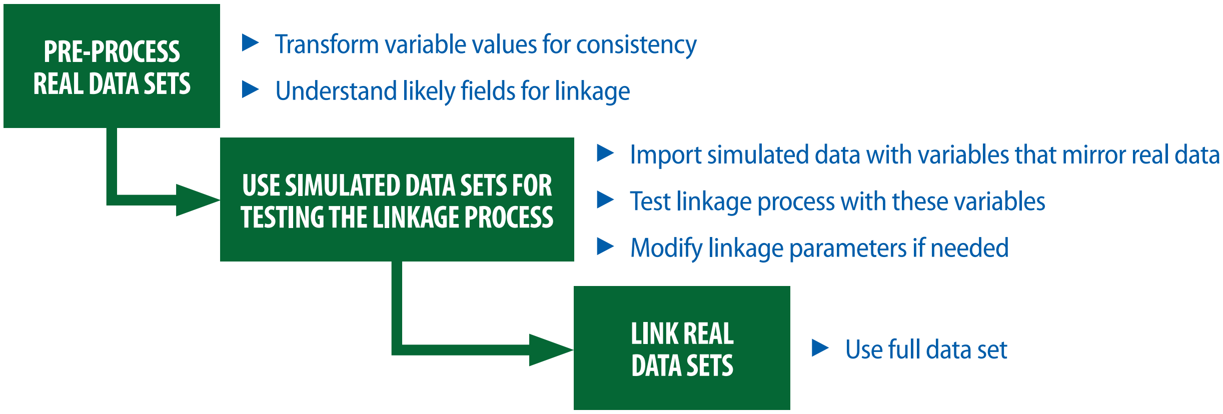
Evaluation Results

The results for the tool evaluation done with GGA DPH and UMB are described in this section. Table 22 indicates the results obtained for each data source (simulated and real data sets), method (70% and 100% linkage threshold; 100% linkage threshold requires perfect precision as records are only linked if linkage variables match exactly; 70% linkage threshold is less percise), tool (LinkageWiz, WinPure, SAS, and R), and state (GA DPH and UMB). Figure 14 shows the workflow that the states used. Pre-processing of real data sets occurred according to normal state practices. Simulated data were obtained from LinkSolv using Maryland-based assumptions. Each state linked both simulated and real data sets using their normal variables (see Table 22). Once lessons were learned with simulated data, the real data were linked.

Table 22. Results Obtained Using Simulated and Real Data Sets for Tool Evaluation, by State

Data Set	LinkageWiz	WinPure	SAS	R
Simulated Data Sets 70% Linkage Threshold	GA DPH, UMB	GA DPH, UMB	N/A	N/A
Simulated Data Sets 100% Linkage Threshold	GA DPH, UMB	GA DPH, UMB	UMB	GA DPH
Real Data Sets 70% Linkage Threshold	UMB	UMB	N/A	N/A
Real Data Sets 100% Linkage Threshold	UMB	UMB	UMB	N/A

Figure 14. State Use of Simulated Data Sets for Evaluation Prior to Linking Real Data Sets


PRE-PROCESS REAL DATA SETS
▶ Transform variable values for consistency
▶ Understand likely fields for linkage
USE SIMULATED DATA SETS FOR TESTING THE LINKAGE PROCESS
▶ Import simulated data with variables that mirror real data
▶ Test linkage process with these variables
▶ Modify linkage parameters if needed
LINK REAL DATA SETS
▶ Use full data set

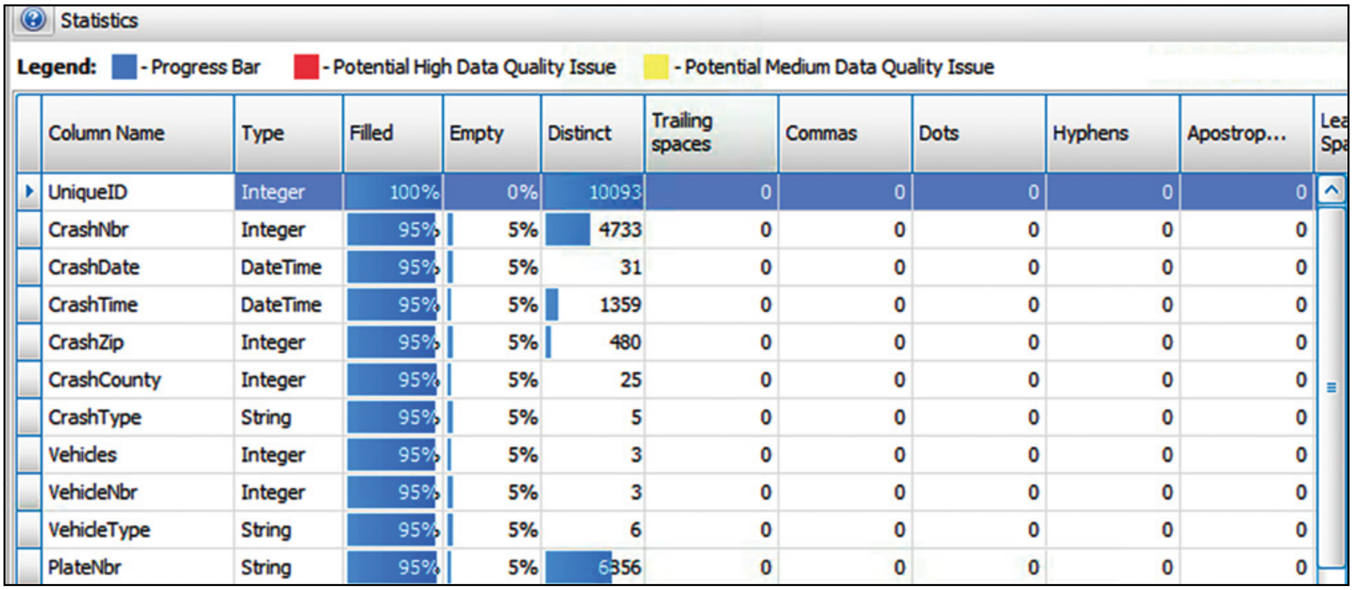
The following metrics were used for this tool evaluation:

1. Interoperability
2. Accuracy
3. Efficiency
4. Usability

1. Interoperability metrics. Both LinkageWiz and WinPure imported and exported data in CSV and Microsoft Excel formats, and did not require converting data formats.

2. Accuracy metrics. Simulated data (Figure 15) can be used as part of the data linkage evaluation if the simulated data are comparable in quality and uses similar variables to the real data (Figure 16). In the evaluation with UMB, simulated data were used to become familiar with the tools being tested (WinPure, LinkageWiz, and SAS). In the image below, the WinPure summary statistics of simulated and real crash data sets are shown. While the simulated data does not perfectly reflect actual crash data it was judged to be of high enough quality to use in the evaluation. For example, the simulated data all had the same percentage of filled fields (95%), but did not have the random dots, hyphens, and apostrophes seen in the actual data.

Figure 15. WinPure-Based Statistical Analysis of Simulated Crash Data Set (from LinkSolv)

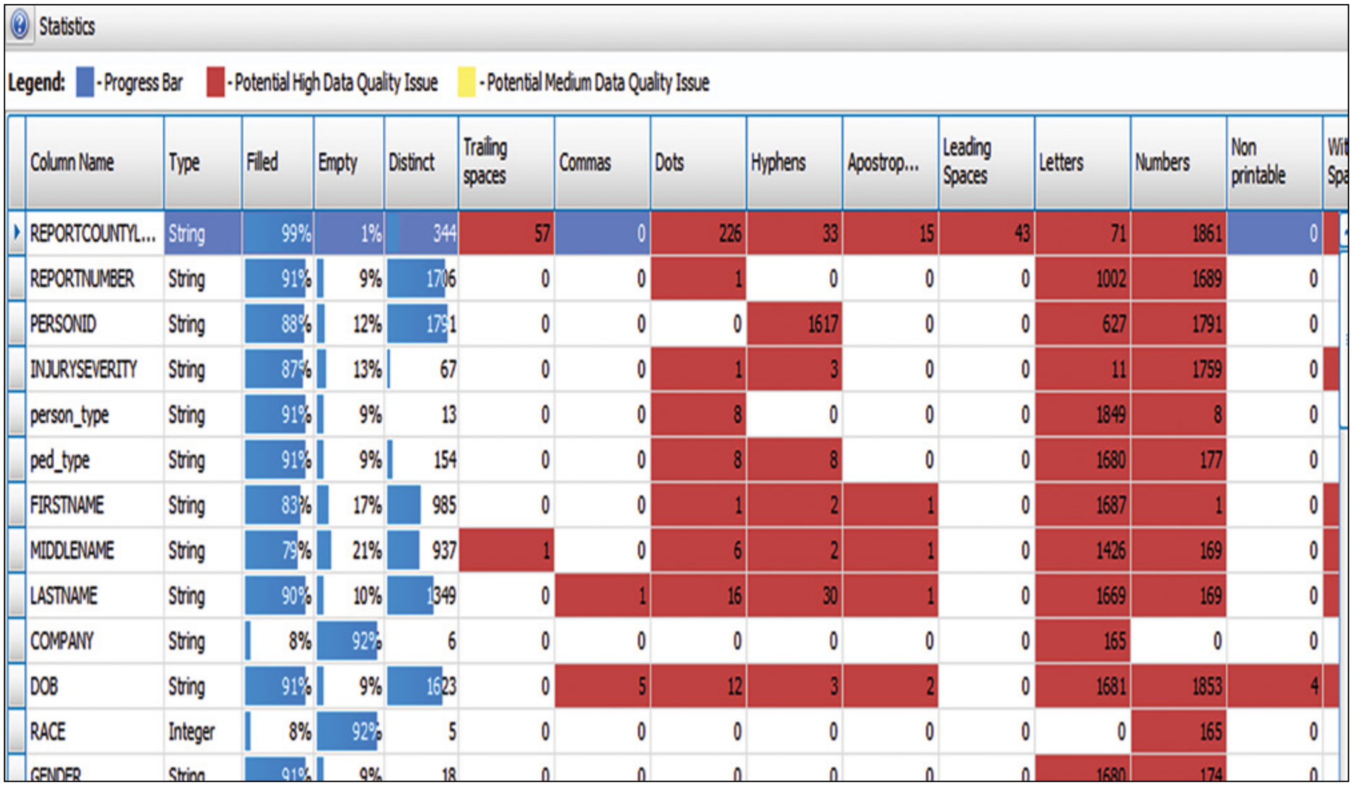


Statistics

Legend: 🟦 - Progress Bar 🟥 - Potential High Data Quality Issue 🟨 - Potential Medium Data Quality Issue

Column Name	Type	Filled	Empty	Distinct	Trailing spaces	Commas	Dots	Hyphens	Apostrop...	Lea Spa
UniqueID	Integer	100%	0%	10093	0	0	0	0	0	
CrashNbr	Integer	95%	5%	4733	0	0	0	0	0	
CrashDate	DateTime	95%	5%	31	0	0	0	0	0	
CrashTime	DateTime	95%	5%	1359	0	0	0	0	0	
CrashZip	Integer	95%	5%	480	0	0	0	0	0	
CrashCounty	Integer	95%	5%	25	0	0	0	0	0	
CrashType	String	95%	5%	5	0	0	0	0	0	
Vehicles	Integer	95%	5%	3	0	0	0	0	0	
VehicleNbr	Integer	95%	5%	3	0	0	0	0	0	
VehicleType	String	95%	5%	6	0	0	0	0	0	
PlateNbr	String	95%	5%	6356	0	0	0	0	0	

Figure 16. WinPure-Based Statistical Analysis of Real Crash Data Set for University of Maryland, Baltimore



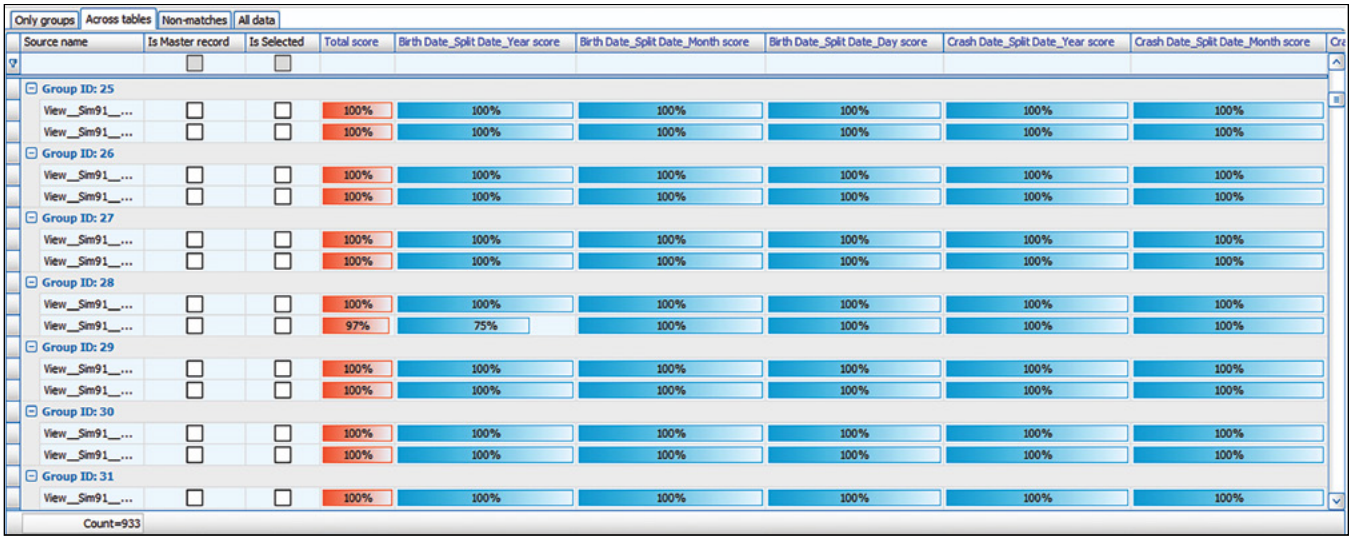
Statistics

Legend: 🟦 - Progress Bar 🟥 - Potential High Data Quality Issue 🟨 - Potential Medium Data Quality Issue

Column Name	Type	Filled	Empty	Distinct	Trailing spaces	Commas	Dots	Hyphens	Apostrop...	Leading Spaces	Letters	Numbers	Non printable	Wi Spa
REPORTCOUNTYL...	String	99%	1%	344	57	0	226	33	15	43	71	1861	0	
REPORTNUMBER	String	91%	9%	1706	0	0	1	0	0	0	1002	1689	0	
PERSONID	String	88%	12%	1791	0	0	0	1617	0	0	627	1791	0	
INJURYSEVERITY	String	87%	13%	67	0	0	1	3	0	0	11	1759	0	
person_type	String	91%	9%	13	0	0	8	0	0	0	1849	8	0	
ped_type	String	91%	9%	154	0	0	8	8	0	0	1680	177	0	
FIRSTNAME	String	83%	17%	985	0	0	1	2	1	0	1687	1	0	
MIDDLENAME	String	79%	21%	937	1	0	6	2	1	0	1426	169	0	
LASTNAME	String	90%	10%	1349	0	1	16	30	1	0	1669	169	0	
COMPANY	String	8%	92%	6	0	0	0	0	0	0	165	0	0	
DOB	String	91%	9%	1623	0	5	12	3	2	0	1681	1853	4	
RACE	Integer	8%	92%	5	0	0	0	0	0	0	0	165	0	
GENDER	String	91%	9%	18	0	0	0	0	0	0	1680	174	0	

Figure 17 shows sample data linkage results from WinPure, specifically the linkage results with respect to the date of birth parameter settings. The dates of birth values can have different years but still be considered a high-confidence match given the parameter settings for this example. The birth date year was not the same for one record but was within a defined 70% threshold and returned as a match.

Figure 17. Sample Data Linkage Results from WinPure



Only groups | Across tables | Non-matches | All data

Source name	Is Master record	Is Selected	Total score	Birth Date_Split Date_Year score	Birth Date_Split Date_Month score	Birth Date_Split Date_Day score	Crash Date_Split Date_Year score	Crash Date_Split Date_Month score	Cra
	□	□							
Group ID: 25									
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
Group ID: 26									
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
Group ID: 27									
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
Group ID: 28									
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
View_Sim91_...	□	□	97%	75%	100%	100%	100%	100%	
Group ID: 29									
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
Group ID: 30									
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
Group ID: 31									
View_Sim91_...	□	□	100%	100%	100%	100%	100%	100%	
Count=933									

Table 23 shows the precision (positive predictive value), sensitivity (recall), and specificity accuracy metrics for each tool used for the evaluation. These metrics show how accurately a tool can match record pairs. As expected, the results show a tradeoff between precision and sensitivity. This can be thought of as a tradeoff between returning a wide net of values at the sake of including a few wrong ones or restricting your net to only include pairs you are very sure about. The higher the score for each metric the better the matching algorithm performs for that case.

Data linked using a 100% matching threshold has perfect precision as records are only linked if linkage variables match exactly. As the linkage threshold is loosened (70%), the precision decreases. Sensitivity simultaneously increases because more true links are identified (more false links are identified as well)[3].

3. Efficiency metrics. Efficiency metrics were captured in the tool evaluation for LinkageWiz, WinPure, SAS, and R. The efficiency metrics captured when linking simulated data are shown in Table 23. Efficiency metrics can be challenging to measure because of variability in data size and the amount of computing power available. Using simulated data (~15,000 records), all the tools could complete the linkages in a matter of seconds to a few minutes if tuned. During evaluation, use of multiple blocking variables with LinkageWiz increased run times to approximately 10 minutes. A blocking variable is used to optimize comparison of records for efficient run times but using several can make the run time longer. This explains the longer run time results for GA DPH with LinkageWiz. Data preparation also took more time (on the order of 10 minutes) even though this was relatively clean simulated data. This was generally longer than run times for linkages, although, with larger data that might change.

The lessons learned using simulated data were applied to UMB's real crash and hospital data sets. Similar tool configurations were used for the simulated and real data sets, and metrics from linking UMB's and GA DPH real data (records) are shown in Table 24. Matched records obtained for exact matching (i.e., 100% linkage threshold) and using a lower, more permissive, linkage threshold (i.e.,70% fuzzy or probabilistic linkage) were reasonable with more matches being found with WinPure at 70% linkage threshold than LinkageWiz. No ground truth existed for UMB's real data, so it was not possible to calculate accuracy metrics. Time and resource constraints during the evaluation prevented some tools and tool configurations from completing data linkage on real data. It was not possible to optimize the R data linkage tool to complete data linkage given the available time and computing resources but could be done with enough time and resources in other scenarios. Similarly, it was not possible to complete fuzzy linkage using SAS given available time and computing resources, but this could be done given enough time and resources in other scenarios.

Table 23. Accuracy and Efficiency Metrics for Tool Evaluation

Tool Name	State	Linkage Variables	Linkage Thresholds	Blocking Variables	Parameter Settings	Linkage Run Time	Accuracy (Precision, Sensitivity, Specificity)
LinkageWiz	MD	Date of Birth, Gender, Date of Crash—Admittance Date, Home Zip	70, 100	DOB	"Minimum score" for merging groups and reporting true matches set at 25 for 100% and 17.5 for 70%	11 seconds 15 seconds	0.727, 0.564, 0.982 (70) 1.00, 0.564, 1.00 (100)
WinPure	MD	Date of Birth, Gender, Date of Crash—Admittance Date, Home Zip	70, 100	None	"Exact match" selected for all variables at 100. Date variables split to month, day, and year to allow for 70% fuzzy matching	2 seconds 2 seconds	0.973, 0.570, 0.999 (70) 1.00, 0.564, 1.00 (100)
LinkageWiz	GA	Last Letter in Last Name (LL), First Initial of First Name (FF), Sex, Crash Date—Admittance Date, Age, Day of Birthday, Date of Birth Month	70,100	LL, Crash Date—Admittance Date, Age	"Minimum score" for merging groups and reporting true matches set at 40 for 100% and 28 at 70%	9:02 minutes 8:27 minutes (fewer blocking variables reduces run time)	0.270, 0.792, 0.821 (70) 0.964, 0.549, 0.998 (100)
WinPure	GA	LL, FF, Sex, Crash Date—Admittance Date, Age, Day of Birthday, Date of Birth Month	70, 100	None	"Exact match" selected for all variables at 100 Date variables split to month, day, and year to allow for 70% fuzzy matching	< 10 seconds	0.973, 0.570, 0.999 (70) 1.00, 0.564, 1.00 (100)
SAS	MD	Date of Birth, Gender, Date of Crash—Admittance Date, Home Zip	100	None	SAS code needed for matching	5 seconds	0.738, 0.565, 0.982 (100)
R	GA	LL, FF, Sex, Crash Date—Admittance Date, Age, Day of Birthday, Date of Birth Month	100	All Variables	R code needed for matching	< 1 minute	1.00, 0.473, 1.00 (100)

Table 24. Measuring the Performance of Data Linkage Methods

Tool Name	Linkage Variables	Linkage Threshold	Blocking Variables	Parameter Settings	Matches
LinkageWiz UMB	Date of Birth, Gender, Date of Crash—Admittance Date, Home Zip	100	Date of Birth	"Minimum score" for merging groups and reporting true matches set at 25	873
LinkageWiz UMB	Date of Birth, Gender, Date of Crash—Admittance Date, Home Zip	70	Date of Birth	"Minimum score" for merging groups and reporting true matches set at 17.5	1099
LinkageWiz GA DPH	Last Letter in Last Name (LL), First Initial of First Name (FF), Sex, Crash Date—Admittance Date (split year, month, day), Age, Day of Birthday, Date of Birth Month	100	None	Not available	16,451
LinkageWiz GA DPH	LL, FF, Sex, Crash Date—Admittance Date (split year, month, day), Age, Day of Birthday, Date of Birth Month	70	None	Not available	23,257
WinPure UMB	Date of Birth, Gender, Date of Crash—Admittance Date, Home Zip	100	None	"Exact match" selected for all variables	873
WinPure UMB	Date of Birth, Gender, Date of Crash—Admittance Date, Home Zip	70	None	Date variables split to month, day, and year to allow for fuzzy matching as "datetime" variable cannot be fuzzy matched	1302
WinPure GA DPH	LL, FF, Sex, Crash Date—Admittance Date (split year, month, day), Age, Day of Birthday, Date of Birth Month	100	None	Not available	156
SAS UMB	Date of Birth, Gender, Date of Crash—Admittance Date, Home Zip	100	None	SAS code to perform the linkage	874

4. Usability metrics. SUS usability metrics from the tool evaluation are shown in Table 25. For this evaluation, WinPure usability is A+, SAS is B+, and LinkageWiz and R are B.

Table 25. Usability Results

Tool Name	SUS Score	Number of Tool Users
WinPure	A+	4
SAS	B+	2
LinkageWiz	B	4
R	B	1

Tool evaluation results. Results from this evaluation are based on specific criteria and are not an exhaustive trial of all available software features. Additional configurations and coding might allow some tools to improve performance. The results are intended to provide participating organizations information and not an endorsement of any tool.

Full results from the tool evaluation are given in Table 26. All tools performed well on direct matching with simulated data sets. LinkageWiz and WinPure were tested both for direct and partial matching. LinkageWiz configurability was well received by users. WinPure scored higher on usability.

This evaluation shows that multiple options exist for data linkage tools and that the details of a data linkage plan should be considered when selection a data linkage tool. For example, LinkageWiz could be useful if users have an existing understanding of data linkage, are knowledgeable about the data being linked, and want to control variable weights used when linking data. When users are already familiar with R and have time to write custom code partial linkage, then using R for data linkage makes sense, particularly if funding is a concern. If an organization already has experience with SAS and has licenses available, SAS is a good choice. Custom code is also needed for partial linkage with SAS. WinPure would be appropriate in data linkage scenarios that require few customizations to linkage weights or variables. Issues with the purchase and installation of both LinkageWiz and WinPure occurred due to the companies being non-U.S. companies and the use of older software for LinkageWiz.

Table 26. Tool Evaluation Results

Tool Name	When to Use	Advantages	Disadvantages
WinPure	▸ When performing one-to-one linkages, without many complications. ▸ When there are relatively clean starting data sets.	▸ User friendly. ▸ Easy to use out of the box. ▸ Video tutorial available. ▸ Many data set cleaning options.	▸ Hard to know what is happening behind the scenes. ▸ Harder to fine tune or use advanced options. ▸ Type conversion difficult. ▸ Partial linking difficult to understand. ▸ Might be hard to purchase due to foreign software and pricing.
LinkageWiz	▸ When total control of variable weights is preferred when matching. ▸ When data sets and linking process are well understood.	▸ Able to manipulate weights for fields. ▸ Can view other variables not in the linkage model in final output pairs. ▸ Similar to LinkSolv. ▸ Can run report to see details of the candidate pairs both above and below the threshold weights. ▸ Easy to adjust parameters.	▸ Random crashes during testing. ▸ Older software components might make harder to get IT approval. ▸ Need to have solid understanding of linking process. ▸ Only 2 custom fields; fields identified by user. ▸ Might be hard to purchase due to foreign software and pricing.
SAS	▸ When experienced with SAS and have licenses. ▸ When data sets are clean. ▸ When using direct matching. ▸ When prepared to work on the probabilistic coding.	▸ It is very easy to do linking without weights. ▸ Can handle a lot of data with no restrictions on variables and rows. ▸ Easy to manipulate the variables. ▸ Well supported.	▸ Could not find a way to weight different variables; another SAS package might be needed. ▸ How to do partial matching is also not clear. ▸ Annual fees for a license.
R	▸ When experienced with R. ▸ When cost is an issue. ▸ When using direct matching. ▸ When prepared to work on the probabilistic coding.	▸ Open source (free). ▸ Online support and community. ▸ Ability to write custom code.	▸ Needs a lot of memory; larger data sets ran into performance issues. ▸ Need to spend time learning R and installing libraries.

APPENDIX Q. EXAMPLES OF MVC DATA CONTENT STANDARDS

To help states that are considering a motor vehicle crash (MVC) data linkage program or expanding an existing linkage program, Table 27 captures examples of commonly used MVC related data content standards and includes links to more information about each standard. Using a content standard ensures consistency in data entry, which improves data quality. Data content standards usually accompany a data structure standard or metadata scheme and are often defined in the form of a data dictionary.

Table 27. Examples of Motor Vehicle Crash Data Content Standards

Traffic Data [95]

Responsible Agency	Data Standard	Description	Website
National Highway Traffic Safety Administration (NHTSA)	Model Minimum Uniform Crash Criteria (MMUCC)	MMUCC is a minimum, standardized data set for describing motor vehicle crashes (MVCs) and the vehicles, persons and environment involved. The Guideline is designed to generate the information necessary to improve highway safety within each state and nationally.	https://www.nhtsa.gov/mmucc-1 http://safety.fhwa.dot.gov/legislationandpolicy/fast/docs/ssds_guidance.pdf
Federal Highway Association (FHWA)	Model Inventory of Roadway Elements (MIRE)	The MIRE is a recommended listing of roadway inventory and traffic elements critical to safety management. The MIRE is intended as a guideline to help transportation agencies improve their roadway and traffic data inventories. The MIRE provides a basis for a standard of what can be considered a good/robust data inventory and helps agencies move toward the use of performance measures. More detailed roadway data are also needed by state and local departments of transportation as they implement their strategic highway safety plans and make safety assessments of various roadway treatments.	https://safety.fhwa.dot.gov/rsdp/mire.aspx
National Highway Traffic Safety Administration (NHTSA)	Model Performance Measures for State Traffic Records Systems	NHTSA compiled a collection of 61 performance measures to help states better quantify improvements to their traffic records systems. These performance measures were crafted with substantial input from a group of 35 experts with experience in at least one of the six core state traffic records systems. The measures are intended for use by federal, state, and local governments to monitor the development and implementation of traffic record data systems, strategic plans, and data improvement grant processes.	https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811441
Federal Highway Administration (FHWA)	Performance Measures for Roadway Inventory Data	Performance measures are tools that can help measure data quality and that can be used by states to establish goals for data quality improvement. The Model Performance Measures for State Traffic Records Systems—subsequently referred to as "the NHTSA Report"—is a publication that was released by NHTSA in February 2011. In this report, NHTSA defined measures for six performance variables that could apply to each of the six-core state traffic safety records systems. This report builds upon the roadway data performance measures presented in the NHTSA Report, providing a detailed review of each and suggesting modifications of, and possible additions to, that original list.	https://safety.fhwa.dot.gov/rsdp/downloads/performancemeasures.pdf

Continued

Medical Data

Responsible Agency	Medical Data Standard	Description	Website
American College of Surgeons (ACS)	The National Trauma Data Standard (NTDS)	The NTDS is a data dictionary that defines concepts that are used in the National Trauma Data Bank (NTDB), a registry of national trauma data. Standard definitions include: demographics, injury information, severity, pre-hospital, emergency department, and hospital procedure details, diagnosis, outcomes, financials, and measures for process of care.	https://www.facs.org/quality-programs/trauma/tqp/center-programs/ntdb/ntds/data-dictionary

APPENDIX R. EXPLANATION OF FIGURES FOR ACCESSIBILITY

Figure 1. Components of a Motor Vehicle Crash Data Linkage Program (page 4)

This graphic shows the three major components for establishing a linkage program: building partnerships, developing a business model and policies, establishing the linkage process.

Figure 2. Process for Motor Vehicle Crash Data Linkage (page 5)

This graphic provides a summary of the 10 steps for establishing the linkage process: define goals, establish data use agreements, develop data linkage plan, assess data quality, prepare data, perform data linkage, evaluate linked data, recalibrate methods, select linked records and conduct/use analysis. If the linkage has been done before, then the process will start at the Prepare Data step.

Figure 3. Motor Vehicle Crash Phases and Examples of Associated Information (page 14)

This graphic illustrates motor vehicle crash phases and examples of associated information. The graphic highlights how driver characteristics and vehicle characteristics are a common denominator for all 3 phases. For example:

For the Pre-Crash Phase, some of the information could be:

- **Driver characteristics**
- **Vehicle characteristics**
- Driver behaviors
- Driving laws
- Road design, including presence of embankments, guardrails, and median barriers.

For the Crash Phase, it can includev:

- Driver characteristics
- Vehicle characteristics
- Human factors, including restraint use, impaired status, and speed
- Road and traffic conditions including other road users
- Number of vehicles, drivers, and passengers
- Vehicle trajectory
- Injury mechanism(s)

For Post-Crash, some of the examples associated information could be:

- **Driver characteristics**
- **Vehicle characteristics**
- Emergency management assessments and interventions at crash scene
- Medical transport
- Injury treatment
- Outcomes of interest (e.g., health diagnoses, medical costs)

Figure 4. Existing Motor Vehicle Crash Data: Starting a Linkage Program (page 15)

This graphic shows the range of data sets that can be linked to provide a comprehensive view of each motor vehicle crash. Below are examples of data for each phase:

Pre-crash:

- Driver citations
- Vehicle registration
- Driver licenses
- Driver training

Crash:

- Police crash reports
- EMS reports

Post-Crash:

- Autopsy records
- Vital statistics
- Toxicology results
- Statewide trauma registry
- Hospital records (discharge, ED)

Figure 5. Linked Motor Vehicle Crash Data in a Public Health Approach to Injury Prevention (page 16)

This graphic shows the five steps necessary to prevent and reduce the severity of motor vehicle crashes. Linked data are part of each one of these steps:

- Define the problem
- Identify risk and protective factors
- Develop and test interventions
- Measure adoption of intervention
- Measure impact of prevention strategies

Figure 6. Linkage Variables and Records in a Data Set (page 31)

This graphic illustrates how variables and records can be linked in one data set. For example: Police motor vehicle crash records in one data set, with linking variables such as crash date, birth date, sex and home ZIP code.

Figure 7. Linking Records Across Data Sets Using Common Variables (page 31)

This graphic illustrates and provides examples of how to link records across data sets using common variables. For example, the data set from the Police Motor Vehicle Crash Record can be linked with the Hospital Discharge Records data set. Common variables such as crash date, birth date, sex and home ZIP code could be used to link these two different data sets if/when the common variables are the same across the data sets.

Figure 8. Bridging Data Set to Link Two Other Data Sets (page 31)

This graphic provides an example of how bridging data sets serves to link two or more data sets that otherwise could not be reliably linked. For example, the emergency medical services (EMS) record bridges between the police motor vehicle crash and hospital discharge records.

Figure 9. Sample Data Linkage System Linking Three Data Sets (page 44)

This graphic shows a sample of a data linkage system linking three data sets. It starts with the development of a data linkage plan, which includes:

- Selecting linking variables
- Selecting/designing data linkage method
- Selecting data linkage tool
- Performing data linkage
- Recording and organizing results
- Blocking and filtering (optional)

The sequence for linking each pair of data sets must be determined for example, first deduplicate and link data sets with similar content (e.g., data from multiple hospitals within one state) before linking with data sets from other content.

Figure 10. Historical Milestones in Data Linkage Literature (page 56)

This graphic shows the historical milestones in data linkage literature. The main milestones were detected between 1946–1995.

5. Dunn (1946). Record Linkage. *American Journal of Public Health*.
6. Newcombe et al. (1959). Automatic Linkage of Vital Records. *Science*.
7. Fellagie & Sunter (1969). A Theory of Record Linkage. *Journal of the American Statistical Association*.
8. Dempster et. Al. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*.
9. Jaro. (1995). Probabilistic Linkage of Large Public Health Data Files. *Statistics in Medicine*.

Figure 11. Patterns of Missing Data (page 108)

This figure shows a graphical representation of four missing data patterns, where the columns represent data fields, the vertical axis represents observations, and a gray pattern represents the presence of observed data values within each data field.

The following patterns can be discerned:

- **Monotone pattern.** Observations and data fields can be arranged so that there is sequential censoring by variable. For example, in Figure 13, a value within data field X is observed when a value in data field Y is observed, a value in data field Y is observed when a value in data field in data field Z is observed, and a value in data field Z is observed when a value in data field W is observed. There are no cases where data field W is observed, and data field Z is not. This pattern might be expected in a longitudinal study where all future observations of a case are missing after that case is censored.
- **Univariate missing data pattern.** This is a special case of monotone missing, where only one data field is missing values.
- **General pattern.** Missing data cannot be arranged into a monotone pattern. In Figure 13, there are cases with missing values for data field W and observed values for data field Z, and vice versa.
- **Disjoint pattern.** In a disjoint pattern, there are variables that are never observed at the same time. In Figure 13, values in data field Z are never observed with values in data field W.

Figure 12. Overview of Analyzing Data by Using Multiple Imputation Methods (page 109)

This graphic shows an overview of analyzing data by using multiple imputation methods. Datasets can be analyzed and all results combined to obtain the final results.

Table 21. Measures of Performance for Data Linkage Methods (page 119)

Table 21 shows how comparisons of the data linkage results to the ground truth are categorized as true positives (true matches), true negatives (true non matches), false positives (false matches) and false negatives (missed matches).

F1 Score. Formula (page 119)

This graphic presents the formula for calculating the F1 Score, which is the harmonic mean of precision and recall. It provides a single metric for data linkage accuracy. F1 equals precision multiplied by recall as the numerator; precision added to recall for the denominator. The numberator is divided by the denominator and the resulting number is multipled by 2 to calculate F1.

Figure 13. Evaluation Plan (page 121)

This graphic shows the steps of an evaluation plan. The process starts with generating simulated data sets, followed by linking data, then calculating metrics, and ends with the comparison of tools.

Figure 14. State Use of Simulated Data Sets for Evaluation Prior to Linking Real Data Sets (page 123)

This graphic shows steps for evaluation before linking real data. The first step is to pre-process real data sets, then use simulated data sets for testing the linkage process, and finally linking real data sets.

Figure 15. WinPure-Based Statistical Analysis of Simulated Crash Data Set (from LinkSolv) (page 124)

This image shows WinPure-Based Statistical Analysis of Simulated Crash Data Set (from LinkSolv). While the simulated data do not perfectly reflect actual crash data it was judged to be of high enough quality to use in the evaluation. For example, the simulated data all had the same percentage of filled fields (95%), but did not have the random dots, hyphens, and apostrophes seen in the actual data.

Figure 16. WinPure-Based Statistical Analysis of Real Crash Data Set for UMB (page 124)

This image shows WinPure-Based Statistical Analysis of Real Crash Data Set for UMB. In the evaluation with UMB, simulated data were used to become familiar with the tools being tested (WinPure, LinkageWiz, and SAS).

Figure 17. Sample Data Linkage Results from WinPure (page 125)

The image shows sample data linkage results from WinPure, specifically the linkage results with respect to the date of birth parameter settings. The dates of birth values can have different years but still be considered a high-confidence match given the parameter settings for this example. The birth date year was not the same for one record but was within a defined 70% threshold and returned as a match.

ACKNOWLEDGMENTS

CDC

- Erin Sauber-Schatz, ige7@cdc.gov
- Erica Spies
- Lisa Caucci
- Kate Shaw
- Ann Dellinger
- Yamile Underwood; Synergy America, Inc.

MITRE

- Jennifer Mathieu, jmathieu@mitre.org
- Mike Fulk
- Liz Richerson
- Keith Miller
- Ann M. Buff
- Suzette Stoutenburg
- Lisa Tutterow
- Mark Thomas

States

- Georgia, Georgia Department of Public Health
  Lisa Dawson and Denise Yeager
- Kentucky, University of Kentucky
  Mike Singleton
- Maryland, University of Maryland, Baltimore
  Cindy Burch and Tim Kerns
- Massachusetts, University of Massachusetts
  Michael Knodler and Robin Riessman
- Minnesota, Minnesota Department of Public Heath
  Mark Kinde and Anna Gaichas
- New York, New York State Department of Health
  Michael Bauer
- Utah, University of Utah
  Larry Cook
- Virginia, Virginia Department of Motor Vehicles
  Angelisa Jennings and Lam Phan

DOT

- National Highway Traffic Safety Administration (NHTSA)
  John Kindelberger, Sarah Weissman, and Tom Bragan

ACRONYMS

ACRONYM	DEFINITION
AAA	American Automobile Association
ACOPP	Alaska Crash Outcomes Pilot Project
ACS	American College of Surgeons
ADHD	Attention-Deficit/Hyperactivity Disorder
AHA	American Hospital Association
APHA	American Public Health Association
API	Application Programming Interface
ATSIP	Association of Transportation Safety Information Professionals
BAC	Blood Alcohol Concentration
CAMH	CMS Alliance to Modernize Healthcare
CDC	Centers for Disease Control and Prevention
CDIP	Crash Data Improvement Program
CDS	Crashworthiness Data System
CHIA	Center for Health Information Analysis
CISS	Crash Investigation Sampling System
CMS	Centers for Medicare & Medicaid Services
CODES	Crash Outcome Data Evaluation System
COTS	Commercial-Off-The-Shelf
CRISP	Chesapeake Regional Information System for our Patients
CRSS	Crash Reporting Sampling System
DDS	Department of Driver Safety
DMV	Department of Motor Vehicles
DMV: VAHSO	Department of Motor Vehicles: Virginia Highway Safety Office
DOH	Department of Health
DOR	Department of Revenue
DOT	Department of Transportation
DOT|TRCC	Department of Transportation Traffic Records Coordinating Committee
DPH	Department of Public Health
DUA	Data Use Agreement
EAGLES	Expert Advisory Group on Language Engineering Standards
EMS	Emergency Medical Services
FAQs	Frequently Asked Questions
FARS	Fatality Analysis Reporting System
FEMTI	Framework for the Evaluation of Machine Translation in ISLE
FHWA	Federal Highway Administration
FICA	Federal Insurance Contributions Act

ACRONYM	DEFINITION
FMCSA	Federal Motor Carrier Safety Administration
FN	False Negative
FP	False Positive
FTE	Full-Time Equivalent
GA AOC	Georgia Administrative Office of the Courts
GDOT	Georgia Department of Transportation
GES	General Estimates System
GHSA	Governor's Highway Safety Association
GOHS	Governor's Office of Highway Safety
GOTS	Government Off-The-Shelf
GR	Governor's Representative
GUM	General Use Model
HHS	Department of Health and Human Services
HIE	State Health Information Exchange
HIPAA	Health Insurance Portability and Accountability Act
ICD	International Classification of Diseases
IEC	International Electrotechnical Commission
IICRC	Intermountain Injury Control & Research Center
IML	Identity Match Laboratory
IPP	Injury Prevention Program
IRB	Institutional Review Board
ISO	International Organization for Standardization
IT	Information Technology
KDPH	Kentucky Department for Public Health
KIPRC	Kentucky Injury Prevention and Research Center
KSPAN	Kentucky Safety and Prevention Alignment Network
LINCS	Linking Information for Nonfatal Crash Surveillance
MADD	Mothers Against Drunk Driving
MAP	Mean Average Precision
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MIRE	Model Inventory of Roadway Elements
MMUCC	Model Minimum Uniform Crash Criteria
MNAR	Missing Not at Random
MOU	Memorandum of Understanding
MVA	Motor Vehicle Administration
MVC	Motor Vehicle Crash
N/A	Not Applicable

ACRONYM	DEFINITION
NASS	National Automotive Sampling System
NCIPC	National Center for Injury Prevention and Control
NSC	National Safety Council
NSC	National Study Center for Trauma & Emergency Medical Systems
NCSA	National Center for Statistics and Analysis
NCSC	National Center for State Courts
NEISS	National Electronic Injury Surveillance System
NEISS-AIP	National Electronic Injury Surveillance System – All Injury Program
NHTSA	National Highway Traffic Safety Administration
NIST	National Institute for Standards and Technology
NPCR	National Program of Cancer Registries
NPV	Negative Predictive Value
NSC	National Study Center for Trauma & Emergency Medical Systems
NTDB	National Trauma Data Bank
NTDS	National Trauma Data Standard
NYSIIS	New York State Identification and Intelligence System
OCR	Optical Character Reader/Recognition
ONCA	Oxford Name Compression Algorithm
OSDH	Oklahoma State Department of Health
OST	Office of the Secretary of Transportation
PAR	Police Accident Report
PHI	Protected Health Information
PII	Personally Identifiable Information
PPV	Positive Predictive Value or Perfect Precision
RAM	Random Access Memory

ACRONYM	DEFINITION
RDIP	Roadway Data Improvement Program
RPM	Regional Program Manager
RPYS	Reference Publication Year Spectroscopy (algorithm)
SHA	State Highway Administration
SHSO	State Highway Safety Office
SME	Subject Matter Expert
SP	Special Publication
SVIPP	State Violence and Injury Prevention Program
TAQs	Titles, Affixes, Qualifiers
TIPRP	Traffic Injury Prevention and Research Program
TN	True Negative
TP	True Positive
TRCC	Traffic Records Coordinating Committee
TREC	Text REtrieval Conference
TREDS	Traffic Records Electronic Data System
U.S.	United States
U.S.C.	United States Code
UMTC	University of Massachusetts Transportation Center
UMTRI	University of Michigan Transportation Research Institute
VA	Department of Veterans Affairs
VAHSO	Virginia Highway Safety Office
VDOT	Virginia Department of Transportation
VHA	Veterans Health Administration
VIN	Vehicle Identification Number
WISQARS	Web-based Injury Statistics Query and Reporting System
WoS	Web of Science

GLOSSARY

Blocking Protocols: Researchers are not consistent in using the terms blocking, filtering, and indexing. For this document, blocking refers to the practice of selecting a subset of records for data linkage (or other types of processing) by using keys.

Data Element: Parsing data values into their segments (e.g., "January 1, 2018" is the value in the "date" data field that can segmented into day, month, and year data elements). Creating more data fields to accommodate data elements can assist in probabilistic data linkage by allowing matches to occur on inexact values (e.g., within 2 days of a particular date).

Data Field: A column in a data set that contains values (e.g., "SSN" is the data field that contains the social security number for each record in a data set).

Data Linkage: The process of establishing a match between different records in one or more data sets, asserting that the records refer to the same person. When a single data set is used, data linkage can assist with removing duplicate records.

Data Linkage Method: A high level categorization of the way that a data linkage algorithm determines whether two values within a variable match, or whether a pair of records refers to a single person (e.g., deterministic, probabilistic, clustering, neural networks).

Data Linkage Software: The tool used to perform record matching (e.g., CODES2000 or LinkSolv, SAS, LinkageWiz).

Data Linkage System: The data linkage tool(s) and parameters required to accomplish one or more data linkage tasks required to create linked MVC data.

Data Linkage Tool: A software application that automatically compares all records using selected variables in all the data sets to identify potential matches for duplicate removal or record matching purposes.

Data Set: A set of records from one data source (e.g., MVC records).

Data Value: The information in a data field (e.g., "123-45-6789" is the value in the "SSN" data field).

Duplicate Records: Two or more records that exist for one person in a specific MVC. Duplicate records might exist intentionally, unintentionally, or both.

Entity: The people, animals, organizations, and vehicles involved in MVCs (e.g., person entity types include the driver, passenger, and pedestrian; animal entity types include pets and wild animals; organization entity types include local, state, national, university, and nonprofit; and vehicle entity type include car and bicycle).

Event: The MVC and subsequent related healthcare treatment (e.g., MVC event type, healthcare encounter event type).

False Negative: Incorrectly identifying a match as a non-match.

False Positive: Incorrectly identifying a non-match as a match.

Filtering: Researchers are not consistent in using the terms blocking, filtering, and indexing. Filtering is a method of defining subsets of data that defines exclusionary criteria to discard dissimilar records, in contrast to the blocking key, which groups similar records by inclusion criteria.

Indexing: Researchers are not consistent in using the terms blocking, filtering, and indexing. Indexing organizes the records in a database, such as by blocking keys, so that subsets of records with similar properties can be quickly accessed.

Linkage Variable: Specific data fields in data sets that are used to perform data linkage (e.g., date of crash or admission to hospital, date of birth, gender, and home zip code are used by Maryland).

Linked Data: The set of data produced by data linkage in which records matched across data sets refer to the same person or event.

Match Results: The output of using a parametrized data linkage tool. Can be simply a list of matched records (dichotomous match results) or a list of match record records with a continuous match score that reflects the degree of the records matched.

Match Status: Whether the matched record is considered to be a true match or a false match.

Matched Records: Linked data to a specific person that has been determined by a data linkage tool. This term can refer to duplicate records within a data set or "linked records" across two or more data sets. Matched records can represent either true matches or false matches.

Multiple Imputation: A statistical method for analyzing data sets in which some values are missing. Multiple imputation is used to replace missing values with statistically motivated values, and it is also used to assign the status of "true match" to pairs of records for record linkage.

Non-Matched Records: No linkage has been determined by a data linkage tool. Typically, not a specific output of a tool, unless the tool uses a threshold of match scores. Non-matched records can represent either true non-matches or false non-matches.

Record: A set of related data fields (e.g., an individual's name, date of birth, and address; the date, time, and location of a MVC) describing a person, event, transaction, or other item, typically thought of as a row in a data set.

Parameters: The assorted thresholds and values assigned by the user before running the data linkage software (e.g., cutoff threshold, match accuracy rules, minimum edit distances).

True Negative: Correctly identifying a non-match.

True Positive: Correctly identifying a match.

Variable: A data field in a data sets.

REFERENCES

1. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control WISQARS (Web-based Injury Statistics Query and Reporting System). (2018, February 5). Retrieved from www.cdc.gov/injury/wisqars.
2. U.S. Department of Transportation, National Highway Traffic Safety Administration. (2018–2017, October). *2017 Fatal Motor Vehicle Crashes, 2016 fatal motor vehicle crashes: Overview* (DOT HS 812 603). Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812603.
3. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. (2016, June). *Health, United States, 2015: With special feature on racial and ethnic disparities*. Retrieved from www.cdc.gov/nchs/data/hus/hus15.pdf#019.
4. Sauber-Schatz EK, Ederer DJ, Dellinger AM, Baldwin GT. Vital Signs: Motor Vehicle Injury Prevention — United States and 19 Comparison Countries. MMWR Morb Mortal Wkly Rep 2016;65. DOI: http://dx.doi.org/10.15585/mmwr.mm6526e1
5. Road to zero presents plan to eliminate roadway deaths: Getting to zero isn't impossible, it just hasn't been done yet. (n.d.). Retrieved from http://www.nsc.org/learn/NSC-Initiatives/Pages/The-Road-to-Zero.aspx.
6. Fixing America's surface transportation act, H. R. 22, 114 Congress, 1st Sess. (2015). Retrieved from https://www.gpo.gov/fdsys/pkg/BILLS-114hr22enr/pdf/BILLS-114hr22enr.pdf.
7. Moving ahead for progress in the 21st century act, H. R. 4348, 112 Congress, 2nd Sess. (2012). Retrieved from https://www.gpo.gov/fdsys/pkg/BILLS-112hr4348enr/pdf/BILLS-112hr4348enr.pdf.
8. National performance management measures, 23 C.F.R. § 490 (2016). Retrieved from https://www.gpo.gov/fdsys/pkg/CFR-2016-title23-vol1/pdf/CFR-2016-title23-vol1-part470-subpart490.pdf.
9. National Highway Traffic Safety Administration. (2017). *MMUCC guideline: Model minimum uniform crash criteria* (5th ed.). Washington, DC: Department of Transportation. Retrieved from https://www.ghsa.org/sites/default/files/publications/files/MMUCC_5thEd_web.pdf.
10. Espitia-Hardeman, V., & Paulozzi, L. (2005). *Injury surveillance training manual: participant guide*. Atlanta, GA: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. Retrieved from https://stacks.cdc.gov/view/cdc/11390.
11. Baker S., O'Neill B., Ginsburg M., & Li G. (1992). *The injury fact book* (2nd ed.). New York, NY: Oxford University Press.
12. Farmer, C. M. (2003). Reliability of police-reported information for determining crash and injury severity. *Traffic Injury Prevention*, 4(1), 38-44.
13. Burch, C., Cook, L., & Dischinger, P. (2014). A comparison of KABCO and AIS injury severity metrics using CODES linked data. *Traffic Injury Prevention*, 15(6), 627-630.
14. Compton, C.P. (2005). Injury severity codes: a comparison of police injury codes and medical outcomes as determined by NASS CDS investigators. *Journal of Safety Research*, 36(5), 483–484.
15. Miller, T. R., Calhoun, C., Kraugh, W. B., & Zegeer, C. (1991). The cost of highway crashes (FHWA-RD-91-055). Washington, DC: Federal Highway Association.
16. Popkin, C. L., Campbell, B. J., Hansen, A. R., & Stewart, R. R. (1991). *Analysis of the accuracy of the existing KABCO injury scale*. Chapel Hill, NC; University of North Carolina Highway Safety Research Center.
17. U.S. Department of Transportation. Federal Highway Administration. Safety culture and the zero deaths vision. Accessed March 29, 2019. Retrieved from https://safety.fhwa.dot.gov/tzd/.
18. Flannagan, C. A., Mann, N. C., & Rupp, J. D. (2015). *Development of a comprehensive approach for serious traffic crash injury measurement and reporting systems* (NCHRP 17 57). Washington, DC: The National Academies of Sciences, Engineering, and Medicine.
19. Venkataraman, N., Ulfarsson, G. F., & Shankar, V. N. (2013). Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type, *Accident Analysis & Prevention*, 59, 309-318.

20. UMassSafe: The University of Massachusetts Traffic Safety Research Program. (2006, June 27). *Teen driver safety and the junior operator law in Massachusetts* [PowerPoint slides]. Accessed March 29, 2019. Retrieved from www.umasstransportationcenter.org/images/umtc/NewsStories/UMassSafe/Teen-Driver-Policy.pdf.

21. Richard, C. M., Magee, K., Bacon-Abdelmoteleb, P., & Brown, J. L. (2018). *Countermeasures that work: a highway safety countermeasure guide for state highway safety offices* (9th ed., DOT HS 812 478). Washington, DC: National Highway Traffic Safety Administration. https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/812478_countermeasures-that-work-a-highway-safety-countermeasures-guide-9thedition-2017v2_0.pdf

22. Motor vehicle prioritizing interventions and cost calculator for states (MV PICCS). (2018, January 6). Retrieved from https://www.cdc.gov/motorvehiclesafety/calculator/index.html.

23. Shults, R. A., et al. (2017). Characteristics of single vehicle crashes with a teen driver in South Carolina, 2005-2008. *Accident Analysis & Prevention*. https://doi.org/10.1016/j.aap.2017.08.002.

24. Morris C. C. (2006). Generalized linear regression analysis of association of universal helmet laws with motorcyclist fatality rates. *Accident Analysis & Prevention*, 38(1), 142-147.

25. Carter, P. M., et al. (2017). The impact of Michigan's partial repeal of the universal motorcycle helmet law on helmet use, fatalities, and head injuries. *American Journal of Public Health*, 107(1), 166-172.

26. Sauber-Schatz, E. K., Thomas, A. M., & Cook, L. J. (2015). Motor vehicle crashes, medical outcomes, and hospital charges among children aged 1-12 years – crash outcome data evaluation system, 11 states, 2005–2008. *MMWR Surveillance Summaries*, 64(8), 1-32.

27. Gabauer, D. J., & Li, X. (2015). Influence of horizontally curved roadway section characteristics on motorcycle-to-barrier crash frequency, *Accident Analysis & Prevention*, 77, 105-112.

28. Milani, J., et al. (2015, August). *Assessment of characteristics of state data linkage systems* (DOT HS 812 180). Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812180.

29. U.S. Department of Transportation, National Highway Traffic Safety Administration. (2010, April). *The crash outcome data evaluation system (CODES) and applications to improve traffic safety decision-making* (DOT HS 811 181). Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811181

30. State and Territorial Injury Prevention Directors Association. (2007). *The national violent death reporting system: lessons learned from 17 states, 2002–2006*. Atlanta, GA: State and Territorial Injury Prevention Directors Association.

31. Raynor, J. (2011). *What makes an effective coalition? Evidence-based indicators of success*. New York, NY: TCC Group.

32. Center for Transportation Studies. (2013). *Minnesota TZD: 10 years of progress*. Minneapolis, MN: Center for Transportation Studies. Retrieved from http://www.minnesotatzd.org/whatistzd/mntzd/mission/documents/decade_report_tzd.pdf.

33. Minnesota Toward Zero Deaths. MN TZD partners. Accesses March 29, 2019. Retrieved from http://www.minnesotatzd.org/whatistzd/mntzd/partners/.

34. U.S. Department of Transportation, National Highway Traffic Safety Administration. (2011). *Model performance measures for state traffic records systems*. (DOT HS 811 441). Retrieved from https://safety.fhwa.dot.gov/rsdp/downloads/trcc_noteworthy.pdf.

35. Health Policy Project. (2014). *Capacity development resource guide: Networking and coalition building*. Washington, DC: Futures Group, Health Policy Project.

36. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. Accessed March 29, 2019. *Core State Violence and Injury Prevention Program (Core SVIPP)*. Retrieved from https://www.cdc.gov/injury/stateprograms/index.html.

37. Scopatz, R., & Goughnour E. (2016). *Maryland's data linkage and analysis to support decision making: roadway safety data and analysis case study* (FHWA-SA-16-049, p. 7). Washington, DC: Federal Highway Administration Office of Safety.

38. National Institute of Standards and Technology. (2013). *Security and privacy controls for federal information systems and organizations* (NIST Special Publication [SP] 800-53, Revision 4). Gaithersburg, MD: National Institute of Standards and Technology.

39. Harron, K. (2017). Introduction to data linkage. *Administrative Data Research Centre for England*. Podcast retrieved from https://www.adruk.org/.

40. Dusetzina, S. B., et al. An overview of record linkage methods. In *Linking data for health services research: a framework and instructional guide*. Retrieved from www.ncbi.nlm.nih.gov/books/NBK253313/.

41. Mamun, A., et al. (2014). Efficient sequential and parallel algorithms for record linkage. *Journal of the American Medical Informatics Association*, 21(2), 252-262.

42. Wilson, D. R. (2011). Proceedings of IJCNN 2011: *Beyond probabilistic record linkage: using neural networks and complex features to improve genealogical record linkage*. New York, NY: Institute of Electrical and Electronics Engineers.

43. Cook, L. J., et al. (2015, July). *Crash outcome data evaluation system (CODES): An examination of methodologies and multi-state traffic safety applications* (DOT HS 812 179). Retrieved from http://www-nrd.nhtsa.dot.gov/Pubs/812179.pdf.

44. Association of Transportation Safety Information Professionals. (2017). *Manual on classification of motor vehicle traffic crashes* (ANSI D16.1-2017, Eighth Edition). Mechanicsville, VA: Association of Transportation Safety Information Professionals.

45. National Safety Council. (1990). *Manual on classification of motor vehicle traffic accidents* (5th ed., ANSI D-16.1-1989). Itasca, IL: National Safety Council.

46. Thygesen, L. C., & Kjær Ersbøll, A. (2014). When the entire population is the sample: strengths and limitations in register-based epidemiology. *European Journal of Epidemiology*, 29(8), 551-558.

47. Kindelberger, J., & Milani, J. A. (2015, July). *Crash outcome data evaluation system (CODES): program transition and promising practices* (DOT HS 812 178). Retrieved from http://www-nrd.nhtsa.dot.gov/Pubs/812178.pdf.

48. U.S. Department of Transportation, National Highway Traffic Safety Administration. How to access FARS data. Accessed March 29, 2019. Retrieved from https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars.

49. U.S. Department of Transportation, National Highway Traffic Safety Administration. (2014, April). *Fatality analysis reporting system* (DOT HS 811 992). Retrieved from crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811992.

50. National Highway Traffic Safety Administration. (2014). *Crash investigation sampling system (CISS): Motor vehicle crash data collection* (DOT HS 812095). Washington, DC: National Highway Traffic Safety Administration. Retrieved from https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13148b-ciss_brochure_012218_v3_tag.pdf.

51. U.S. Department of Transportation, National Highway Traffic Safety Administration. (2014, December). *Crash report sampling system (CRSS): motor vehicle crash data collection* (DOT HS 812 096). Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812096.

52. NEMSIS—National EMS Information System. Accessed March 29, 2019. Retrieved from https://nemsis.org/.

53. Data Sources for WISQARS Nonfatal. Accessed March 29, 2019 from: https://www.cdc.gov/ncipc/wisqars/nonfatal/datasources.htm.

54. Curry, A. E., et al. (2017). Motor vehicle crash risk among adolescents and young adults with attention-deficit/hyperactivity disorder. *JAMA Pediatrics*, 171(8), 756-763.

55. Hansen, R. N., et al. (2015). Sedative hypnotic medication use and the risk of motor vehicle crash. *American Journal of Public Health*, 105(8), e64-e69.

56. Vladutiu, C. J., et al. (2013). Pregnant driver-associated motor vehicle crashes in North Carolina, 2001–2008. *Accident Analysis & Prevention*, 55, 165-171.

57. Curry, A. E., et al. (2015). Young driver crash rates by licensing age, driving experience, and license phase. *Accident Analysis & Prevention*, 80, 243-250.

58. Conner, K. A., & Smith, G.A. (2014). The impact of aggressive driving-related injuries in Ohio, 2004–2009. *Journal of Safety Research*, 51, 23-31.

59. Han, G. M., Newmyer, A., & Qu, M. (2017). Seatbelt use to save money: impact on hospital costs of occupants who are involved in motor vehicle crashes. *International Emergency Nursing*, 31, 2-8.

60. Vladutiu, C. J., et al. (2013). Adverse pregnancy outcomes following motor vehicle crashes. *American Journal of Preventive Medicine*, 45(5), 629-636.

61. Curry, A. E., et al. (2013). Graduated driver licensing decal law: effect on young probationary drivers. *American Journal of Preventive Medicine*, 44(1), 1-7.

62. Olsen, C. S., et al. (2016). Motorcycle helmet effectiveness in reducing head, face and brain injuries by state and helmet law. *Injury Epidemiology*, 3(1), 8.

63. Olsen, C. S., Thomas, A. M., & Cook, L. J. (2014). Hospital charges associated with motorcycle crash factors: a quantile regression analysis. *Injury Prevention*, 20(4), 276 280.

64. Singleton, M. D. (2017). Differential protective effects of motorcycle helmets against head injury. *Traffic Injury Prevention*, 18(4), 387-392.

65. Han, G. M., Newmyer, A., & Qu, M. (2015). Seat belt use to save face: impact on drivers' body region and nature of injury in motor vehicle crashes. T*raffic Injury Prevention*, 16(6), 605-610.

66. Karaca-Mandic, P., & Jinhyung, L. (2014). Hospitalizations and fatalities in crashes with light trucks. *Traffic Injury Prevention*, 15(2), 165-171.

67. Comins, J. A., & Hussey, T. W. (2015). Compressing multiple scales of impact detection by reference publication year spectroscopy. *Journal of Informetrics*, 9(3), 449-454.

68. Leydesdorff, L., et al. (2014). Referenced publication years spectroscopy applied to iMetrics: Scientometrics, Journal of Informetrics, and a relevant subset of JASIST. *Journal of Informetrics*, 8(1), 162-174.

69. Marx, W., et al. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology*, 65(4), 751-764.

70. Comins, J. A., & Leydesdorff, L. (2017). Citation algorithms for identifying research milestones driving biomedical innovation. *Scientometrics*, 110(3), 1495-1504.

71. Dunn, H. L. (1946). Record linkage. *American Journal of Public Health and the Nations Health*, 36(12), 1412-1416.

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

73. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J*ournal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1-38.

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

75. Metzger, K. B., et al. (2015). Association between NCAP ratings and real-world rear seat occupant risk of injury. *Traffic Injury Prevention*, 16(Suppl 2), S146-S152.

76. Cook, L. (2010). *EMS data linkage*. Presentation no longer available online.

77. Jurczyk, P., et al. (2008). FRIL: A tool for comparative record linkage. *AMIA Annual Symposium Proceedings Archive*, 2008, 440-444.

78. Hiatt, R. A., et al. (2015). Leveraging state cancer registries to measure and improve the quality of cancer care: a potential strategy for California and beyond. *Journal of the National Cancer Institute*, 107(5), djv047.

79. Zhang, Y., et al. (2012). Probabilistic linkage of assisted reproductive technology information with vital records, Massachusetts 1997–2000. *Maternal and Child Health Journal*, 16(8), 1703-1708.

80. Telfar Barnard, L. F., et al. (2015). Novel use of three administrative data sets to establish a cohort for environmental health research. *BMC Public Health*, 15(1), 246.

81. Grannis, S. J., Banger, A. K., & Harris, D. (2009). *Privacy and security solutions for interoperable health information exchange*. Washington, DC: Health Policy Institute & O'Neill Institute for National and Global Health Law.

82. Westphal, C. (2008). *Data mining for intelligence, fraud, & criminal detection: advanced analytics & information sharing technologies*. Boca Raton, FL: CRC Press.

83. Experian Information Solutions, Inc. (2014). *Finding insight through data collection and linkage: develop a better understanding of the consumer through consolidated and accurate data*. Den Haag, Netherlands: Experian.

84. U.S. Department of Transportation. Traffic records coordinating committee. Accessed March 29, 2019. Retrieved from https://www.transportation.gov/trcc.

85. U.S. Department of Transportation, Federal Highway Administration. Technical Assistance. Accessed on March 29, 2019. Retrieved from https://rspcb.safety.fhwa.dot.gov/technical.aspx.

86. U.S. Department of Transportation. Traffic Records: Technical Assistance and Training. Accessed March 29, 2019. Retrieved from www.transportation.gov/government/traffic-records/technical-assistance-and-training.

87. Governors Highway Safety Association. Traffic records training for state highway safety office leadership. Accessed March 29, 2019. Retrieved from www.ghsa.org/recordstraining.

88. International Organization for Standardization & International Electrotechnical Commission. (2005). *Information technology—Security techniques—Code of practice for information security management* (ISO/IEC 27002).

89. Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-preserving record linkage for big data: current approaches and research challenges. In A. Y. Zomaya & S. Sakr (Eds.), *Handbook of Big Data Technologies*. New York, NY: Springer.

90. Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. New York, NY: Springer Science + Business Media.

91. Murray, J. S. (2016). Probabilistic record linkage and deduplication after indexing, blocking, and filtering. *Journal of Privacy and Confidentiality*, 7(1), 3-24.

92. Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models* (vol. 1, pp. 532-533). New York, NY: Cambridge University Press.

93. Haukoos, J. S., & Newgard, C. D. (2007). Advanced statistics: Missing data in clinical research—Part 1: An introduction and conceptual framework. *Academic Emergency Medicine*, 14(7), 662-668.

94. Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons.

95. Data systems guidance. (2017, August 22). Retrieved from www.transportation.gov/government/traffic-records/data-systems-guidance.

96. Arehart, M. D., & Miller, K. J. (2008). Proceedings from LREC 2008: *A ground truth data set for matching culturally diverse romanized rerson names*. Retrieved from https://pdfs.semanticscholar.org/9833/35e7f33c8e75e8b822e69d8f1698f85153fb.pdf.

97. EAGLES Evaluation Working Group. (1999). *The EAGLES 7-step recipe*. Switzerland: Université de Genève. Retrieved from www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html.

98. Hovy, E., et al. (2003). *FEMTI: A framework for the evaluation of machine translation in ISLE*. Switzerland: Université de Genève. Retrieved from https://www.isi.edu/natural-language/mteval/.

99. Hovy, E., et al. (2003). *FEMTI: A framework for the evaluation of machine translation in ISLE*. Los Angeles, CA: University of Southern California, Information Sciences Institute. Retrieved from https://www.isi.edu/natural-language/mteval/.

100. Miller, K. J. (2008). Proceedings from LREC 2008: FEIRI: *Extending ISLE's FEMTI for the evaluation of a specialized application in information retrieval*. V. Arranz, K. Choukri, B. Maegaard, & G. Thurmair (Eds.). Retrieved from http://www.pdmpassist.org/pdf/LREC_Evaluation_Workshop_FIERI.pdf.

101. System usability scale (SUS). (n.d.). Retrieved from https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html.

102. Sauro, J. (2011). *Measuring usability with the system usability scale (SUS)*. Retrieved from https://measuringu.com/sus/.

103. Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 55. Retrieved from https://en.m.wikipedia.org/wiki/Rensis_Likert.

104. Bunn, T., et al. (2013). Concordance of motor vehicle crash, emergency department, and inpatient hospitalization data sets in the identification of drugs in injured drivers. *Traffic Injury Prevention*, 14(7), 680-689.

105. Conderino, S., et al. (2017). Linkage of traffic crash and hospitalization records with limited identifiers for enhanced public health surveillance. *Accident Analysis & Prevention*, 101, 117-123.

For more information please contact
Centers for Disease Control and Prevention
1600 Clifton Road NE, Atlanta, GA 33029-4027
Telephone: 1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348
E-mail: cdcinfo@cdc.gov
Web: cdc.gov/cdc-info

CS 302338-A