SO YOU WANT TO LINK YOUR STATE DATA

July 1996

NATIONAL ASSOCIATION OF
GOVERNORS' HIGHWAY SAFETY REPRESENTATIVES
750 First Street, NE Suite 720
Washington, D.C. 20002

National Highway Traffic Safety Administration
400 Seventh Street SW, Room 6125
Washington, D.C. 20590
TEL.: 202-366-5351
FAX: 202-366-7078

SO YOU WANT TO LINK YOUR STATE DATA

INTRODUCTION

Welcome to the excitement of data linkage and the opportunity to do more with less using routinely collected state data. Your state data are a valuable source of information to enhance decision making for highway safety and injury control activities. These data are usually collected to meet the specific needs of the collection agency. The collectors may be nonmedical personnel focusing on person, vehicle, or environmental specific data related to the cause of the injury at the time of onset. Or the data collectors may be medical personnel focusing on a particular phase of the response and treatment for a patient--at the scene and enroute, at the emergency department, as an inpatient, during rehabilitation, at the time of death. Each of these data sets alone lacks the comprehensive information required to support highway safety and injury control activities.

Population-based, computerized statewide data describing the outcome for all crash victims are generated from the linkage of state data. The linkage process itself has the added benefit of identifying problems related to data quality. It highlights where records are missing and data are incomplete. It promotes collaboration. Linkage is not a one time event. It should be repeated annually to monitor the scope of highway safety problems, target countermeasures, recommend prevention strategies, evaluate the cost effectiveness of these strategies, and support a multi-disciplinary and multi-organizational approach to the solution of highway safety and injury control problems.

Background: Health care reform demands that we not only decrease the volume of injuries, but also demonstrate a concomitant reduction in health care costs. Thus, it is important to know what makes a difference. It is important to evaluate the benefits of specific countermeasures after consideration of their utilization/implementation at the time of the event for both the injured and the noninjured.

Effective decision making depends on a systematic approach to data collection and analysis. This new approach to highway safety and injury control evaluation is represented below as a three-dimensional model (Exhibit 1) indicating the interaction of government, business, and health care with information about the environment, cause of injury, and its consequences. The interaction makes it possible to identify problems and implement effective programs at the national, state, and local levels. These programs are designed to prevent injuries, respond with appropriate acute care to those injuries that do occur, and return patients to work in a timely manner after effective rehabilitation. Feedback is required at all levels to target those resources which will have the most impact on reducing health care costs and improving medical outcome.

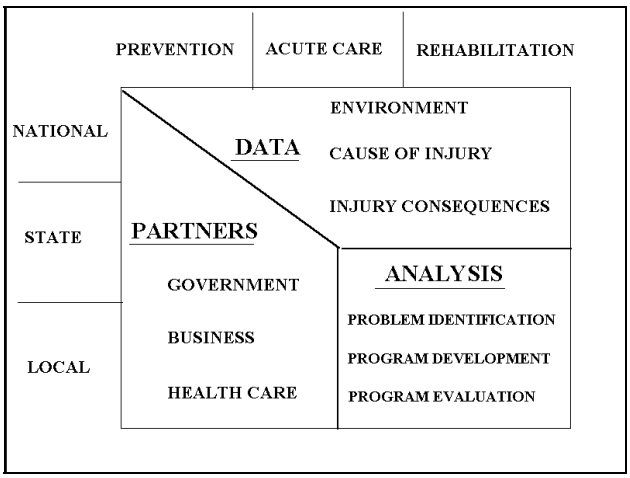

Exhibit 1: Injury Control Data Model

Linked state data are important to the highway safety and medical communities because they:

- Identify the financial consequences of injuries caused by crashes
- Indicate severity and medical outcomes for injuries caused by crashes
- Generate state specific data to support policy decisions and legislation
- Justify priorities based on reducing mortality, morbidity, severity, and costs
- Indicate EMS system performance
- Promote collaboration between highway safety and the health community
- Support community-based highway safety programs
- Support state-specific vehicle or crash-related analyses
- Support state-specific occupant-related analyses
- Support the safety management system
- Expand the usefulness of each data file being linked
- Improve the completeness and quality of state data

Format of Instruction Manual: This instruction manual focuses on how to obtain and prepare your data for linkage. It describes the linkage process but does not include instructions for implementing the linkage software. Finally it provides examples of how to use the linked data analytically. The instructions are presented as a series of **steps to be followed sequentially** for best results. The 22 steps are grouped into five sections. Section one describes the data, equipment, and personnel resources needed for linkage and the obstacles which may occur. Sections two and three discuss file preparation and selecting the cases for linkage. The actual linkage begins with section four which describes the concepts and phases of the linkage process. Finally, section five describes the output, the validation process, and several applications for the linked data.

SUMMARY OF THE STEPS

I. GETTING STARTED

Step 1	Identifying the data resources
Step 2	Accessing the data
Step 3	Identifying and resolving the obstacles
Step 4	Convening an advisory committee
Step 5	Obtaining other resources

II. PREPARING THE DATA

Step 6	Editing the data files
Step 7	Standardizing the file structures
Step 8	Standardizing the data elements
Step 9	Coding nonuniform data
Step 10	Creating new variables to support the analyses

III. CASE SELECTION

Step 11	Ancillary linkages
Step 12	Choosing the records for linkage

IV. PERFORMING THE DATA LINKAGE

Step 13	About probabilistic linkage
Step 14	Blocking the data files
Step 15	Assigning the weights
Step 16	Linking the files
Step 17	Defining match, nonmatch, clerical review
Step 18	Resolving problems

V. ANALYZING THE DATA

Step 19	Reviewing the linkage results
Step 20	Validating the linkage results
Step 21	Applying the linked data
Step 22	Documenting the linkage process

SECTION 1: GETTING STARTED

STEP 1: Identifying the Data Resources

State Data Sources and Their Characteristics: Exhibit 2 indicates the major sources of state highway safety and injury data useful for linkage. These data sources are arrayed in a flow chart to indicate the flow of events from the scene through the health care system to final disposition. The traffic records include the nonmedical data sources such as the driver license, vehicle registration, conviction, and roadway data files. Linkage of these data to the linked crash and injury data makes it possible to generate medical and financial outcome information for specific characteristics of the crash and its components.

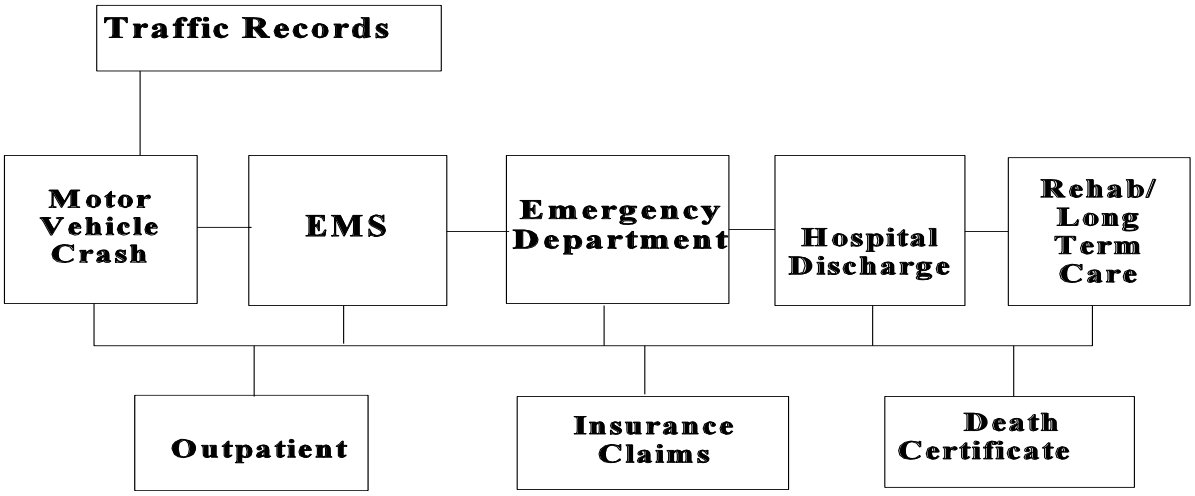

Exhibit 2: Crash and Injury Data Sources

In Exhibit 3, each source is described according to its characteristics. Population-based state data include everyone who is involved statewide compared to nonpopulation-based data which include everyone who is involved at a facility, a group of facilities, or as part of a group, such as an insurance group. The content of each state data file may focus on one component of the event (for example, person). Or the file (for example, claims) may include information describing several or all phases of the event.

Exhibit 3. Characteristics of the state data useful for highway safety and injury control

Data Source	Data Collector	Statewide	Population Based	Edited	Record Unit	Indicates CrasNh
Nonmedical Data Sources						
Crash **Report**	Department of Transportation/ Public Safety / Motor Vehicles	✔	✔		Crash	✔
Vehicle Registration	Department of Motor Vehicles	✔	✔		Vehicle	
Driver Licensing	Department of Motor Vehicles	✔	✔		Driver	
Census	Department of Health	✔	✔		Person	
Roadway/Infrastructure	Department of Transportation	✔	✔		Road	✔
Medical Data Sources						
EMS	Depts of Health or Public Safety		✔		Event	✔
Emergency outpatient	Hospital/Claims				Event	
Hospital discharge	Dept. of Health	✔	✔	✔	Event	
Registries: Trauma, Head & Spinal Cord, Poison	Hospital or Dept. of Health			✔	Person	✔
Death Certificates	Dept of Vital Statistics	✔	✔	✔	Person	✔
Insurance Claims Data						
Medicaid, Medicare	Dept of Health	✔		✔	Claim	
Private Health Insurance	Health Insurance Co				Claim	
Worker's Compensation	Dept. of Labor	✔		✔	Claim	
Private Vehicle Insurance	Vehicle Insurance Co				Claim	✔
National Auto Insurance Files	Association of Insurance Co				Claim	✔

- **Nonmedical Data Sources**

POLICE CRASH REPORT

The police crash record documents the characteristics of the vehicle, crash, and occupant at the time of a specific crash. This information includes crash time and location, type of crash, contributing factors, type of roadway, driver identifiers and actions, occupants and their injury status, description of the vehicles involved, safety belt/helmet/air bag use, and sometimes whether the occupant was transported by ambulance. Police may over report safety belt utilization for some occupants, particularly the less seriously injured in states with mandatory safety belt legislation. Not all states computerize information from the crash diagram. Thus it may be difficult to determine the direction of the impact or which vehicle was involved, for example in a roll over. Police crash data may include documentation of the time of onset for the crash which can be used to indicate the time of onset for the injury. When crash data document both the uninjured and injured occupants, police data become a potential source of information describing the success stories (such as those occupants who are not injured or who suffer minor injuries because they were wearing safety belts).

Crash records are more likely to link to an injury record as injury severity increases. However, injury severity is not a factor in the linkage of crash records to claims records. Police document injury severity using a functional measure of severity consisting of five levels including killed (K), severe or incapacitating injury (A), nonincapacitating injury (B), possible injury (C), and not injured (0). Because police evaluation of severity is based on level of functioning, injuries which are minor in terms of survivability may be included with the severe injuries and vice versa. Often, just the transport of a crash victim for treatment is enough for the police officer to code "incapacitating injury." In contrast, some types of head injuries are not evident at the scene but may become life threatening within hours of the crash. Police do not have the time or training to collect detailed medical information at the scene or to obtain other medical data generated either en route or at the hospital. The police severity score, KABC0, is useful for predicting linkage to an injury record, but is associated with survivability only for those who are killed at the scene. Implementation of the KABCO scale may vary among states. However, linkage makes it possible to standardize, for inter state comparisons, the severity levels by redefining them as died, inpatient, transported and/or ED, slightly injured, no injury.

All states have reporting thresholds so that not all motor vehicle crashes are reported or are reportable to the police. Persons in crashes involving no injury and a single vehicle with little damage may feel no obligation to notify authorities, particularly if the consequence might be higher insurance rates. In some instances, the crash may be reported but not computerized. The minimum reporting threshold excludes some or all of those crashes causing only minor property damage and no injuries.

The police crash data file includes a large volume of records which are usually stored on a mainframe computer. To facilitate access, police data may be split into smaller sub-files - occupant, vehicle, or crash. For linkage, the split crash-specific records must be combined and then converted to occupant specific records. The relinkage process may uncover a crash in the crash file which is not documented in the vehicle or person files or vice versa. In some cases, one of the records may be filled with default values making relinkage difficult. These records should be identified and then reviewed for accuracy. Some states also store, in separate files, the vehicle identification numbers and the points of impact for vehicles involved in crashes, and the types of fixed objects struck during crashes. All of these data files may be linked using an identification number unique to the crash.

DRIVER LICENSE FILE

Driver licensing data are driver-specific and include the driver license number, date of birth, social security number (SSN) and sometimes the driver's history of convictions. When driver information from the crash data are combined with medical cost and conviction information, this information is useful to assess the societal costs caused by repeat offenders. Linkage of the crash and driver licensing data files provides access to the SSN to facilitate linkage to insurance claims data, such as Medicaid.

VEHICLE REGISTRATION DATA

Vehicle registration data describe detailed characteristics of the vehicle being registered. This information includes vehicle identifiers including identification number (VIN). The VIN can be decoded to obtain information about the type of restraint system, vehicle weight and other vehicle characteristics useful for evaluating the consequences of particular types of crashes. When the VIN is also collected on the crash report, the crash and vehicle registration files can be linked directly. Linked crash, vehicle registration, census, and injury data generate information that relate specific types and characteristics of the vehicle to urban and rural crash patterns and their specific medical and financial consequences.

ROADWAY/INFRASTRUCTURE FILES

Roadway/infrastructure data are not crash specific. Instead they describe bridges, pavements, roadside inventories, etc. that describe the type of road where the crash occurred. These data generate information about the roadway function class for analytical purposes and can link directly to the crash record via a geocode (i.e. node number, latitude/longitude, etc.). These data, when linked to the crash and medical cost data, are useful to support cost-effective decisions on maintaining and upgrading streets and highways and for supporting implementation of the Safety Management System.

CENSUS

Census data are not crash specific but provide information about the geographic location where the crash occurred. These data generate population estimates for geographic areas, usually towns and counties. They can be linked to square mile estimates to standardize crash locations in terms of population density (population per square mile), such as metro, urban, suburban, rural or wilderness, for intra or inter-state comparisons.

- **Medical Data Sources**

EMERGENCY MEDICAL SERVICES (EMS)

The EMS record includes information about victims who are treated and transported to a hospital by the emergency medical services system (EMS). EMS records are the first medical records completed for persons injured in motor vehicle crashes who are transported by EMS and the first to indicate severity in physiological terms related to survival. They are the only source of routinely collected medical information indicating the patient's status and treatment provided at the scene and enroute to the hospital.

Severity is described in physiological terms, related to survivability, based on the patient's first set of vital signs at the scene, and also the eye opening, motor, and verbal responses to stimuli. The EMS record includes corroborating information about the utilization of occupant protection devices and the presence of alcohol/drugs.

A separate report is completed to record the status, treatment, and disposition of the victim by each EMS service which responds (first responder, basic life support, advanced life support, air transport). Thus the EMS data file may include many records for a patient for a single emergency event. None of the EMS records include information about crash victims not transported by EMS.

EMERGENCY DEPARTMENT

The victim's arrival at the emergency department is first recorded in the emergency department log and then subsequently in the patient medical record completed by the triage nurse, the attending physician and nurse, and the medical and mental health consultants who provide treatment. Billing data, including patient identifiers, are collected and usually computerized more frequently than the patient care data. When an emergency department patient is admitted as an inpatient, the billing record for this patient will be deleted from the emergency department data file and added to the hospital data file so that only one bill is generated for the patient. This factor must be compensated for before generating patient flow information from the computerized emergency department billing data. In addition, the emergency department data file may include more than one record for the same person because of readmissions for the same problem. Like the EMS report, severity is recorded in

physiological terms based on the patient's vital signs and the Glasgow Coma Score, plus detailed diagnostic test data.

The emergency department is the source of information about the treatment and disposition of crash victims who are not transported by EMS but who obtain outpatient medical treatment at a hospital. It also provides information about the additional treatment and disposition for those crash victims who were transported by EMS.

HOSPITAL DISCHARGE/REHABILITATION

Once admitted as an inpatient for acute care, a medical record is completed during the length of stay and abstracted into a discharge record for every patient. Patients who are discharged and then readmitted to receive rehabilitation services in the same acute care hospital are also included in the inpatient discharge data system. Thus, because of the readmissions, it is possible to have more than one record for the patient in the hospital discharge data file. Rehabilitation speciality hospitals required to submit discharge data to the state are also included in the hospital discharge data file.

Hospital data have been standardized for reporting to the Health Care Financing Administration for payment under Medicare/Medicaid. The data include patient, hospital and provider identifiers, procedures and diagnoses (International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) code), disposition, etc. The diagnosis codes recorded, assigned at the time a patient is discharged from a hospital, can be used to generate an injury severity score (ISS)[1] to standardize severity according to the types of injuries. The ISS is an anatomic measure based on body region injured as defined by a narrative description of the injury or by the ICD-9-CM code.

Hospital data provide a source of routinely collected financial information describing total charges and, in some states, hospital-based physician charges. Charges for other professional services are not included but are estimated to represent an amount equal to 25 percent of inpatient charges[2]. The charges reported reflect the price charged and do not represent the actual cost of providing care to that specific patient or the revenue received by the hospital.

---

[1] Copes, WS, Champion, HR, Sacco, WJ, et al: *The Injury Severity Score Revisited.* J Trauma 1988; 28: 69-77.

[2] Rice, D.P., MacKenzie, E.J., and Associates: *Cost of Injury in the United States: A Report to Congress 1989.* San Francisco, CA: Institute for Health & Aging (University of California), and Injury Prevention Center (The Johns Hopkins University), 1989.

Hospital discharge data do not computerize information about the utilization of occupant protection devices. Alcohol related information, although available in some instances, may be restricted from public access.

Because inpatient data are collected by licensed/certified trained medical records technicians and serve as the basis for payment, overall data quality is usually higher than other injury data. However, quality may vary for specific data elements, such as the E-codes or EMS run report numbers, not routinely used for billing purposes.

Access to hospital discharge data is usually restricted by legislated requirements. Data requests must be in writing. A fee may be required. And a data release must be signed indicating that the data will not be used except as stated. There may be other requirements specific to use of the data for linkage.

LONG-TERM HEALTH CARE (NURSING HOME) INFORMATION

More-seriously-injured crash victims may require long-term medical care in a nursing home where data are collected to meet the needs of the facility and for payment by Medicaid and Medicare. These data document the functional status of the patient receiving long-term care. They are rarely computerized statewide and must be accessed directly from the long-term care facility. Severity information is generated from data describing the patient's level of impairment and vital signs. Computerization of this information varies by facility.

OUTPATIENT DATA

Primary care data systems, such as those implemented by a health maintenance organization, include data collected when outpatient care is provided. These data are usually collected and maintained by the organization and not merged with other primary care data statewide. Outpatient data are characterized by more than one record per individual, and dates of service which may occur anytime during the week after the crash.

OTHER INJURY DATA SYSTEMS:

Medical status, treatment, and disposition information for injured victims of crashes may be obtained from other injury data generated by hospitals, health maintenance organizations, and government agencies. These data systems include trauma registries, Fatal Accident Reporting System (FARS), etc.

Trauma registry data are usually generated by designated trauma centers and, thus, are considered a subset of the EMS and hospital data for those patients with the most serious injuries. Coding practices may vary between the registries and the hospital discharge data. For example, the hospital may code a patient as a 23 hour observation compared to the registry which might code the same patient as a 1 day length of stay. Thus the two files may not include the same patients even though the patients have similar levels of severity.

FARS data are generated by states under contract to NHTSA from police and EMS data and include all victims of crashes who die within 30 days of the crash or who suffer nonfatal injuries during a fatal crash.

DEATH CERTIFICATE

The death certificate data describe the medical causes, time, location, and mechanisms of injury for all injury deaths, including those caused by motor vehicle crashes. They do not include standardized diagnosis codes describing the medical condition such as are recorded on the inpatient hospital data file; but they do use standardized codes to document the causes of death. The death certificate also records the time and location for the onset of an injury which can be used to corroborate information on the crash report. Unfortunately this latter information sometimes is not computerized. These files record all deaths, regardless of the residence of the victim, occurring within the state and all deaths of residents who die out of state. Death certificate data are computerized statewide according to standards that are uniform nationally.

- **Insurance Claims Data**

Limited medical information is generated as part of the claims process for health and vehicle insurance. Medical treatment and payment data describing injured crash victims over 65 years of age or disabled may be obtained from Medicare, for victims who are financially needy from Medicaid, for victims of occupational injuries from Worker's Compensation, and for victims whose care is paid by specific insurance groups such as Blue Cross/Blue Shield, Allstate, Aetna, State Farm, etc. The advantage of claims data is that they may include both outpatient (emergency department) and inpatient medical and financial information, and they are carefully edited to facilitate prompt payment. The disadvantage is that the data reflect information necessary to process an insurance claim and usually do not provide the detailed medical information, including injury severity, required to evaluate patient outcome. In addition, claims data files are usually very large since they include multiple claims records per event and multiple events per person. Records must be identified that relate to the specific event being studied.

Use of insurance claims data for linkage to crash data is complicated by the fact that more than one insurance company may be involved and not all pay at the same rate. In most cases, the no-fault insurance carrier or the automobile insurance company is liable for the health care charges. However, some victims file the claim with their health insurer to avoid having to pay higher automobile insurance rates. When the victim is also eligible for Medicare, the claim to Medicare will be filed last since Medicare pays at a lower rate. On the other hand for Medicaid eligible recipients, the claim may be filed first since Medicaid is often willing to take responsibility for recouping the expenses from the insurance payer who is liable for the costs. It is not surprising that the lag between billing and processing causes a delay in the availability of data for linkage and that the various co-payment arrangements complicate

the process of documenting the actual payers for analytical purposes. However, linked claims data are useful to audit and cross-check cases across different databases, and thus are significant to insurance companies and health providers interested in controlling costs.

National insurance data facilitate linkage. The American Insurance Services Group (AISG) describes its Insurance Index System as a national clearing house for bodily injury claims. It is administered by the AISG and is considered the leading industry-sponsored provider of loss data. This extensive system of claims records was initiated in the 1920's by the Association of Casualty Insurance Companies as a research tool to defend supporting insurance carriers against fraudulent bodily injury claims. The system is currently supported by 1,450 property/casualty insurance companies, 1,500 self-insurers, and 120 claims administrators that represent over 93 percent of the industry in premium volume. The Index System serves all of North America and the American possessions. These national data can be split into statewide files to support the linkage of state data.

STEP 2: Accessing the Data

Data useful for linkage are controlled by different owners. The owners, although usually public entities, may also include private entities such as hospitals. Access to each data file is governed by specific data release policies which have been legislated, mandated through regulation, or controlled by organizational policies.

The state public safety office usually owns the police crash data. Statewide EMS and hospital discharge data are usually owned by the state department of health, though sometimes the hospital discharge data are also available through a private, nonprofit data organization. Trauma registries are usually facility-owned, though some states require that trauma center registry data be merged at the state level. Exhibit 4 lists for each data source, the state organization which governs use of the data file. (Some of this information is also listed in Exhibit 3). Before obtaining access to a data file, it may be necessary to submit a written data request, pay a fee, and sign a data release form that restricts use of the data to the stated purpose.

Exhibit 4: State Organization in Charge by Data Source

Data Source	State Organization in Charge
Crash	DOT/Public Safety
Vehicle Registration	Motor Vehicles
Driver Licensing	Motor Vehicles
Census	Health
Roadway	Dept. of Transportation
EMS	Health
Hospital	Health
Death Certificates	Vital Statistics
Vehicle Insurance Claims	Private Insurance Co/No Fault
Health Insurance Claims	Health/private insurers

STEP 3: Identifying and Resolving the Obstacles

Efforts to foster the development and utilization of injury data systems for highway safety must be based on a clear understanding of their potential value, their shortcomings, and the legal/institutional barriers to be overcome. Like police reported crash data, injury data systems offer tremendous analytic potential. However, the following technical and institutional issues may interfere with access to these data for linkage.

- **Confidentiality of Patient Information:** While police crash records are most often considered public records for the purposes of access, patient medical records are considered confidential and thus access to these data systems is greatly limited, and may even be restricted by legislation. New computer technology makes it possible to provide data security and protect patient confidentiality while still providing needed access to the patient identifiers for linkage.

- **Lack of Mutual Understanding by Various Data Owners:** Information useful to physicians in diagnosing and treating motor vehicle crash induced injuries is not necessarily that most wanted for highway safety applications and vice versa. Physicians study the relationship between the types of injuries and crash characteristics. Highway departments study the occurrence of injury with less concern about the type of injury, its survivability or cost of care. As a consequence, neither community is particularly inclined to support linkage activities, particularly if extra work is required to edit and reconfigure their data after it is collected. Under health care reform the medical community needs to know precisely which crash and vehicle characteristics have the potential to cause the most disabling and expensive injuries, so they can anticipate the need for early intervention. The highway safety community needs to know which roadside improvements and countermeasures will have the most impact on also reducing health care costs.

- **Interagency Politics:** Inter-agency politics may discourage the collaboration required among the different disciplines and multiple organizations for a successful linkage. Data owners initially may refuse to edit their data in advance or comply with the case selection guidelines. Sharing data may be viewed as a loss of power or a duplication of effort.

- **Communication Breakdowns:** Lack of information about the content of the data files may complicate the process of requesting the information you need for linkage. In some instances, even the data owner is not aware or forgets that a particular variable is no longer computerized or that a block of records is missing.

- **Price:** Not all data are free even when financed with public funds. You will have to pay the fee or negotiate a deal to generate needed information in exchange for access to the data.

- **Poor Data Quality:** The absence of identifiers with the power to disciminate among crashes and the occupants involved in the crash limits the ability to match records in two data files. Records with sufficient data for linkage still may be excluded because they are lacking information needed for the analyses.

  States that exclude information about the uninjured passengers or minor crashes/injuries limit the opportunities for linkage. This excluded population represents the potential "success" cases which should be included in any analysis of highway safety.

  Health data which do not record the cause of injury (E-code) complicate the process of defining the referent population to validate the linkage.

- **Data Not Computerized:** Linkage is impossible when the necessary records are not computerized. Health data may not be available in a centrally accessible computerized form. This is particularly true for physician office, emergency department, long term care data, and some types of claims data. Or the records may be computerized at each facility, but not merged into a statewide computerized file. When all but a small percentage of the records are computerized statewide, you may be forced to delay the linkage to obtain permission and funds to computerized the remaining records.

  Linkage also suffers when identifiers, important for linkage and previously computerized, are no longer computerized because of budget constraints.

- **Different Storage Media:** Problems may occur when translating EBCDIC to ASCII code, converting variable length records to fixed length, changing from DOS to UNIX, and defining the specifications for unlabeled tapes. Transfer of state data from a mainframe to a workstation or desktop computer requires an expert. It may be less expensive in the long run to sub-contract this task to an organization which specializes in such conversions.

STEP 4: Convening an Advisory Committee

Linkage requires collaboration among the owners and users of state data. Data owners are those entities responsible for initially collecting the data for their purposes. Users apply the data for purposes beyond that for which the data were originally collected. Convening an Advisory Committee consisting of the owners and users of state data provides a forum to discuss common issues and concerns. Initially the members should include, as a minimum, representatives with technical experience for the following major data systems being linked:

- Police crash report data system
- EMS data system
- Hospital Discharge data system
- Vital Statistics
- Insurance Claim data systems
- Other data systems being linked

Although the technical issues related to linkage are easier to resolve within a smaller group, expansion of the committee to also include the following decision-makers facilitates resolution of the political issues related to linkage.

- Governor's Highway Safety Representative
- Representative of Federal Highways
- EMS Director
- Director of Injury Control
- Medical professionals
- Researchers
- State Epidemiologist
- Law Enforcement
- Other Decision Makers
- Other Data Owners and Users

The purpose of the Advisory Committee is to promote collaboration among the owners who collect the data, the data managers who work with the data, the data interpreters who analyze the data, and the customers who use the data. Letters of commitment should be obtained from each member's organization indicating its willingness to support with shared services or actual funding the institutionalization of linked state data. The first tasks of the committee members are to determine what state data currently exist and to describe each data file according to the following criteria:

<u>Authority</u>

- Who collects the data and authorizes its use?
- Is the data collection mandated or voluntary?

Confidentiality Policies

- Are the data in the public domain or controlled to protect patient confidentiality?

Computerization of the Data

- Are the data computerized statewide?
- Who computerizes the data and what type of storage media is used?
- How many records are included in the data file?
- What is the size of each record?
- Is there more than one record for each person, for each event?
- Does the file include records for out-of-state residents?
- What media is used to store the data?
- Are all variables expected to be computerized actually computerized?
- How are the computerized data stored?
- What type of computer is used?

Accessibility

- Is there a fee to access the data?
- What data release policies must be followed to access the data?

Availability

- Are the data available quarterly, annually?
- Are the data stored for a calendar year or a snapshot in time?
- What is the current year of available data?
- When is data destroyed?

Reliability

- Are the data edited and, if so, when?
- Are the attribute value codes standardized among the different files?
- Are the data collectors/abstractors trained?
- What are the missing data rates for important variables?
- What types of records are missing from the data file?

Case Selection

- Are the records person-specific?
- Does the data file include unique person identifiers?
- Does the data file include E-codes, or other types of information to identify the cause (mechanism) of injury?
- Does the data file computerize a coded description of the injury?
- How many occupants are involved in crashes and how many are injured?
- How many occupants are treated only at a clinic or physician's office?
- How many occupants are transported by EMS?
- How many occupants bypass EMS and go directly to the emergency department or morgue?

- How many occupants are discharged home from the emergency department?
- How many occupants are admitted to a hospital?

Emergency Response Resources

- Obtain an inventory of public safety response resources.
- Obtain an inventory of EMS services and personnel by license level, hospital facilities trauma care facilities, etc.

Once the data sources are inventoried, their content is reviewed to determine what data need to be standardized and if the identifiers are sufficient for linkage.

The committee's functions also include:

- facilitating access to state data for linkage by encouraging timely processing and reasonable data access policies.
- advocating the development of statewide emergency department and other outpatient data
- improving data quality
- promoting and monitoring appropriate uses of the linked data
- developing data release policies to govern use of the linked data
- determining who should do the actual data linkage and on what computer
- developing a feedback process to routinely evaluate the data linkage resources, process, and outcome
- promoting the standardization of definitions
- monitoring the linkage results
- institutionalizing the linkage process and the analysis of linked data

STEP 5: Obtaining Other Resources for Linkage

- **Computer Software:** Two versions of the probabilistic software have been used for the linkage of crash and injury state data. MINICODES, probabilistic linkage software for a micro-computer, is preprogrammed for the crash/EMS/Hospital linkage and has been distributed free by the National Association of Governors' Highway Safety Representatives to the Governor's Highway Safety Representative in each state.

AUTOMATCH, a more comprehensive version of MINICODES, is capable of linking any two files and is available commercially through Matt Jaro, Match Ware Technologies, Inc. 301-384-3997. This version is more precise than MINICODES and also includes the capability to unduplicate files and perform geocoding linkages. The instruction manuals for both MINICODES and AUTOMATCH provide a detailed explanation of how to use the software. Neither MINICODES nor AUTOMATCH is a data file manager. Additional software will be required to manage the data files and perform the statistical analyses.

- **Computer Hardware:** The linkage process requires computer hardware with sufficient capability to process the expected volume of records. About 10,000 crash records can be linked to about 70,000 EMS records in less than 15 minutes using a 486 microcomputer and the MINICODES software. Larger files over 100,000 records are easier to link using the AUTOMATCH software and a computer workstation. Even a workstation computer may require creative file management when the linkage files include more than a million records. Some states store large state data files on a mainframe and then download selected records for linkage using a workstation or micro computer. Transfer of data between a mainframe and smaller computer might be facilitated by the purchase of a 9 track tape reader. A portable printer port access tape backup drive with the capability of storing 250MB using data compression allows the quick transfer of large data sets between computers. When the volume of records prevents all data files from being loaded simultaneously, and when tape access is slow, a solution is to perform the development and test work using small sample files.
- **Personnel:** Personnel with unique expertise working with the data files, knowledge about health outcomes and services, and expertise in data file maintenance and manipulation are necessary to perform the linkage and implement the linked data accurately. It is crucial to assign personnel who are familiar with the operation of the police and EMS emergency response system within the state to perform the clerical review of the unsure matches generated by the crash to EMS or hospital linkages.

- **Reference Materials:** Maps and code lists are necessary to facilitate the clerical review process. Useful references include the following:

    - Lists of county/town, Zip, census track, MCD, or other codes which your state uses to designate geographic locations

    - Provider code list for EMS services and hospitals

    - Police/ambulance/hospital service area definitions

    - Codes lists for severity levels (KABCO, ISS, AIS, trauma score)

    - Code lists for all categorical variables included in each data file: position in vehicle, ejection, type of EMS run, disposition (emergency department, hospital), etc.

SECTION II: PREPARING THE DATA

Section II is crucial for successful data linkage. **Your data files must contain sufficient information to discriminate among events and also among the multiple occupants involved in a specific event.**

The file and field preparation guidelines suggested below represent a beginning. You are encouraged to experiment with other edits and logic checks based on the unique characteristics of your data. Careful file and field preparation saves time, decreases the probability for invalid matches, and improves your chances for locating valid matches. This step is relatively simple to accomplish if your state data are routinely edited. Otherwise, it may take months, particularly if records must be computerized.

STEP 6: Editing the Data Files

Editing the data improves your linkage results. Thus, whenever possible, errors in the data should be corrected prior to linkage. Common errors and their corrections are listed below. Ideally they should be routinely implemented when the state data files are created. **Standardizing the coding of unknowns, not recorded, and not applicable in all of the data files facilitates linkage**.

Military time errors: Identify times out of sequence on both the crash and EMS data files to locate military time errors.

- Time of the crash should be no later than the earliest time indicated for the time of the report to the police or the time when police arrived at the scene.

- Time of the EMS call should be earlier than the time of EMS arrival at the scene or arrival at the destination.

Age errors: Compare date of birth and age for consistency.

- If date of birth is not collected for all occupants, use the driver information to validate the ages entered for the occupants identified as drivers.

- Check codes for unknown age, and newborns to make sure that zero is not used for both.

- Check the coding of ages greater than 99 and unknown age to make sure that 99 is not used for both.

- Check use of month versus year to code age for babies.

Location code errors: Identify incorrect county/town and provider service codes,

- Identify crash and EMS location codes which are inconsistent with the service areas for either the police or EMS responding unit.

- Identify hospital provider codes which are inconsistent with the designated crash location or EMS provider.

- Compare county/town and provider identification codes against master files to identify invalid codes.

Invalid or out of range coding errors: Identify attribute values which are invalid or out-of-range for each attribute.

Logic checks: Match values for two or more attributes to identify inconsistent data.

- Road character with crash location

- Crash type with crash location

- Alcohol related crash with no driver coded as under the influence

- Child restraint and age >5 years

- Weather coded as rain, snow, or sleet and road surface as dry

- Times between 0900-1200 and lighting conditions coded as dark

- Motorcyclist coded as wearing a safety belt; occupant of passenger car coded as wearing a helmet

- Date of birth coded for driver information compared to that coded for occupant indicated as being the driver.

Make friends with the data entry clerks or data collectors. They can help you understand the idiosyncrasies of your data.

STEP 7: Standardizing the File Structures

Both AUTOMATCH and MINICODES require that the file structure of each database be standardized:

- The files must be standard ASCII files. In other words, it should be possible to edit the files using a standard text editor. This means that each line is terminated by carriage-return, line-feed characters. Most files downloaded from other sources have this format.

- There is no limitation on line length (record size). Although if records are more than several hundred characters, they become difficult to examine.

- The records must be fixed size. (No variable length records are supported).

- All of the fields of a record must be fixed. No variations in record types are supported. Therefore, all records must have an identical format.

- Records in both files must be event, vehicle, or person specific.

- Crash to injury linkage requires that the crash file be converted to person-specific records.

- There should be one record per individual (event, or vehicle) in each file. More than one record may exist by mistake or because of multiple providers providing care. All but one record should be eliminated from the linkage process. The extra records can be deleted or reattached to the linked record during a separate linkage. In some instances, it may be possible to retain the multiple records and use the array matching feature of AUTOMATCH to perform the linkage.

- Default values vary depending on how a data element is defined, for example as a character or numerical field. Thus, definitions for the data elements chosen for blocking or linkage must be standardized.

- Data entry clerks and data collectors may not always follow the documentation instructions. It is important to talk to them to find out common practices which may affect your use of the data. Data quality improves when the data collectors/entry clerks receive feedback about the accuracy of the data.

- Information must be sufficient to discriminate among individuals. The following variables have been found to be useful:

  Age Name

Date of Birth
Year,Month, or Day of Birth
Gender
Address of Residence
Residence Code (town, city, county, state)
Residence (zip code)
First 3 digits of zip code
Last 2 digits of zip code

Initials
Soundex Name
Social Security Number

Transport: Yes/No
Run Report Number
Injury Yes/No
Types of Injury (Head, neck, etc.)
Injury Severity

Position in Vehicle
(driver/passenger)
Occurrence of Death
Date of Death

Hospital ID
Year, Month, Day of Admission
Probable Admit Date
Admit Hour

Month or Day of Discharge
Disposition
Pay Source
Diagnosis Codes

- Information must be sufficient to discriminate among crashes. The following variables have been found to be useful:

Date of Event (crash, EMS, Hospital, Claim)
Day of Event; Year of Event
Month of Event

Actual Time of Event
(Crash, EMS, Hospital, Claim)
Time Code

Location Code of Crash
(town, city, county, state)
Address

EMS Region
Hospital Service Area
Destination Hospital

Type of Event
Vehicle Type
MVA field
VIN

STEP 8: Standardizing the Data Elements

The data elements used to perform the blocking and linkage must conform to the following practices:

- **Standardize the codes used to represent categorical data in both files:**

Male designated as a 1 in the crash file should also be designated as a 1 in the injury files.

County/town codes designated in the crash files must also match the county/town codes used in the injury files.

Provider identification codes (ambulance, hospital, etc.) should be uniform on all data files to be linked. Watch out for variations in hospital codes caused when one data file assigns a single number to a group of hospitals and another data file assigns the same group of hospitals separate numbers for each facility.

- **Standardize coding for missing values, newborns, and unknown:** Missing values should be distinguishable from zero values for numeric fields. Particular care should be taken to distinguish an age of less than one year, for example zero to 6 months, from a missing age. Ages less than one year should be recorded as zero. Missing ages should be recorded as blanks.

  Because people are living beyond 100 years of age, age should be a 3 digit field in all of the data files to be linked.

- **Standardize person names to the extent possible:** Names should be separated into individual surname and given name fields. Use of a SOUNDEX algorithm is excellent for blocking purposes. SOUNDEX of surname provides a good blocking since many variations of a name will be included in a single block.

- **Standardize the coding of dates:** Dates should be in year-month-day order, if possible (e.g. 19920304 = March 4, 1992), since this places date fields in ascending collating sequence.

- **Standardize the coding of time as military time:** Time should be represented as hour-minute (or hour), (e.g. 1230) is 12:30 PM. Remember that 0000 is a valid time (Midnight). When actual hour is coded and minutes are blank, you might consider the feasibility of entering 30 for the minutes.

STEP 9: Coding Nonuniform Data

Because probabilistic linkage has the capability of linking many variables simultaneously, you are encouraged to use all available information. Nonuniform information can be recoded into uniform variables for participation in the linkage. The new variables may be binary (yes/no), date (e.g., date+1), or categorical (e.g., area 1-4). They facilitate matching by helping to reduce the need for manual review. The variables presented below are examples of the types of variables which can be created. The types and quantity of new variables you create will depend on the information which is computerized and available in

your state data files. If your data system does not collect the information described below, you will not be able to create the variables. On the other hand, if your data system collects other unique information, you will be able to create your own unique variables.

- **Gender:** When gender is not available, but name is, convert common female names to female; make the rest male. Code gender using numeric codes.

- **Classification of actual times into time blocks:** Convert actual times into time block codes. Other time blocks (6 hour, etc.) may also be created if appropriate for your emergency response system. Time blocks are useful for linking crash records directly to the hospital data.

  4 hour time blocks: Convert the following times into 4 hour time blocks. The values for 6 new time blocks will range from 1-6. The time blocks are usually defined as 0000 to 0359, 0400 to 0759, 0800 to 1159, 1200 to 1559, 1600 to 1959, 2000 2359.

  Crash time
  Time police arrive at scene
  Call to EMS
  EMS arrival at scene
  EMS arrival at destination

  8 hour time blocks: Convert the following times into 8 hour time blocks. The values for the new time blocks will range from 1-3. The time blocks are usually defined as 0000 to 0759, 0800 to 1559, 1600 to 2359.

  Crash hour
  Admit hour

- **Classification of type and area of injury information:** Convert nonuniform information describing type and area of injury into uniform binary variables on each data file. The number of new variables you create is limited only by the limitations of the information computerized on your data files.

  Codes indicating area of injury: Create new variables using the area of injury fields designated on the crash record. These fields usually describe the area of injury such as the head, neck, back, etc. The purpose of this type of variable is to minimize the need for manual review of unsure records. If this type of information is limited in the data file, the variables will not contribute much to the matching decision.

  Create new binary variables on the EMS data file to match the new crash variables. Thus, EMS data describing the area of injury (head, spinal, burns, etc.) and/or the treatment (cervical immobilization, long board, splinting, bleeding controlled, etc.) should be used to create new EMS variables to match the crash variables for head, neck, back, etc.

Create new binary variables on the hospital data file to match those created on the crash data file. Identify records with an ICD-9-CM or procedure code which when recoded match the new binary variables on the crash data file; for example cases with an ICD-9-CM discharge or procedure code related to head injuries should be coded as head.

Codes indicating type of injury: Create new variables using the type of injury fields designated on the crash record. These fields usually describe the type using terms such as bleeding, broken bones, shock, etc. The purpose of this type of variable is to minimize the need for manual review of unsure records. If this type of information is limited in the data file, the variables will not contribute much to the matching decision.

Create new binary variables on the EMS data file to match the new crash variables. Thus, if the EMS data file indicates the type of injury as bleeding, broken bones, shock, etc. and/or the treatment (cervical immobilization, long board, splinting, bleeding controlled, etc.) then use one or a combination of these variables to create new EMS variables to match the new crash data file variables for bleeding, bones, shock, etc..

Create new binary variables on the hospital data file to match those created on the crash data file. Identify records with an ICD-9-CM or procedure code which when recoded match the new binary variables on the crash data file; for example ICD-9-CM discharge or procedure codes related to bleeding, broken bones, shock.

Conversion of both area and type of injury information to a standardized code is complicated if the data are computerized as free form text.

INJURY

Create an injury variable using any information on the crash record indicating that the occupant suffered some type of injury.

Create the same variable on the EMS file using any information indicating that the occupant suffered some type of injury.

Create the same variable on the hospital file using any information indicating that the occupant suffered some type of injury.

- **Standardize Identification Codes:** Identification codes, assigned, for example, to indicate the geographical location of the crash, the identification of the hospital or an EMS service agency, should be standardized using the same numerical coding system. Codes which vary should be standardized prior to linkage.

- **Reclassification of geographic location information:** Create a new variable which converts geographic locations into a categorical variable that will be uniform on each of the data files. This variable indicates the service area for the crash.

  HOSPITAL SERVICE AREA

  Create a new categorical variable in the crash data file which converts the county/town code of the geographic location of the crash into a code to identify the hospital area where the crash victim is most likely to be treated. Modify the definition of the hospital service areas so that EMS regions are subsets within the areas. This variable is particularly useful in the absence of EMS data indicating the hospital destination. The codes for this variable should be defined to also include the closest trauma center to which the victim is most likely to be transported.

  Create the same hospital service area variable in both the EMS and hospital data files by converting the town/county code of the EMS pickup and the town/county code of the location of the hospital to the hospital service area of the location for the crash.

  EMS REGION

  Create a new categorical variable on the crash data file to indicate the EMS Region in which the crash occurred. Modify the definition of EMS region so that very few ambulance runs have the potential of crossing regional boundaries. In ambiguous cases, the state EMS Director should indicate the correct EMS Region for the geographic location of the crash.

  Create the same EMS Region variable in the EMS data file to indicate the EMS Region in which the EMS service is located. In ambiguous cases, the state EMS Director should indicate the correct EMS Region for the service.

  Create the same EMS Region variable in the hospital data file to indicate the EMS Region in which the hospital is located. In ambiguous cases, the state EMS Director should indicate the correct EMS Region for the hospital.

- **Classification of crash date information:** Convert crash date into a new date variable recording the crash date plus one day.

  PROBABLE ADMIT DATE

  Create a new date variable, probable admit date, to record crash date plus 1 day on the crash record for those crashes which occur after 8 p.m.. For crashes which occur before 8 p.m., enter the current date. Use this variable instead of date when linking directly to the hospital file.

- **Classification of disposition information:** Convert disposition information into a new binary variable,

  DEATH

  Create a new binary variable, DEATH, on the crash data file to indicate all fatal crash injuries.

  Create the same new binary variable, DEATH, to record any information from the EMS run record which might indicate that the patient died at the scene, enroute, or at the emergency department.

  Create the same new binary variable, DEATH, on the hospital data file to indicate all admissions for which disposition from the hospital is indicated as death.

- **Reclassification of driver date of birth information:** Date of birth is a unique identifier for linkage to medical records. When date of birth is not routinely recorded on the crash record for each occupant, this information can be obtained for the occupants who are drivers from the driver specific information recorded on the crash record.

  DATE OF BIRTH

  This variable should apply to all occupants. Record date of birth from the driver information for vehicle one. Repeat for drivers of vehicle two, etc. Perform ancillary linkages to other data files, if possible, to add date of birth for passengers. Date of birth is an important variable for linkage to injury records.

- **Classification of cause of injury information:** Convert medical record information indicating motor vehicle crash as the cause of injury into a new binary variable,

  MVA

  Create a new binary variable on the crash data file and enter a value of 1 for each record since by definition all crash records are MVA related.

  Create the same new binary variable on the EMS data file and convert information on the run record indicating a motor vehicle crash as the cause of injury.

  Create the same new binary variable on the hospital data file and convert discharge information indicating an E-code for a motor vehicle crash as the cause of injury.

- **Classification of vehicle characteristics:** Most crash records include the make, model, year, and registration number of the vehicles involved in the crash. Some crash records also record the vehicle identification number (VIN) which when decoded

provides additional information about the vehicle's characteristics (type restraint, wheel base, weight, etc.). If the crash record does not include the VIN, a new categorical variable should be created for the transfer of the VIN from the vehicle registration file.

VIN

Obtain missing vehicle characteristics by linking the crash and vehicle registration data files via the vehicle identification number (VIN). Once the linkage is complete, transfer the VIN from the vehicle registration data file to the crash data file.

- **Classification of vehicle damage information:** Routinely collected crash data do not include information to calculate the actual force of a crash. Instead surrogate measures, alone or in combination, are used. These measures include towed, number of vehicles in crash, speed, damage estimate, etc. and can be used to generate useful definitions for the crash severity level.

  CRASH SEVERITY

  Create a new variable in the crash data file to transfer coded information generated from one or more surrogate measures of crash severity.

- **Standardize insurance claim numbers:** Standardize insurance claim numbers to eliminate special characters plus suffix and prefixes from the root claim number or to eliminate number, date, location, plus suffixes from claimant names.

STEP 10: Creating New Variables to Support the Analyses

New variables may be created from existing data to expand the usefulness of the linked data for analytical purposes. As an example, the variables below are required to calculate an analytical measure called the Sensitivity Index (Step 21). But these same variables are also useful for other types of highway safety analyses.

- **Classification of type of vehicle:** Convert type of vehicle into two new variables.

  MOTOR VEHICLE

  Motor vehicle should be defined to exclude motorcycles, ATV, motorbikes, plus other types of vehicles in which safety belt utilization is invalid.

  MOTORCYCLE

  Motorcycle should be defined to include motorcycles, ATV, motorbikes, and other vehicles for which helmet utilization is valid.

- **Classification of occupant protection devices:** Convert use of occupant protection devices into three new binary variables.

  SAFETY BELTS

  Safety belt utilization describes if the crash record indicated that the occupant was wearing a safety belt at the time of the crash. Safety belt is defined to include all of the types (lap belt, shoulder harness, etc.) listed on the crash record but does not include air bag. It should be recorded only for occupants of motor vehicles regardless of age. Use of child restraints could be considered the same as safety belt utilization for this field. Belts should be recorded for occupants of vehicles required to have safety belts.

  HELMETS

  Helmet utilization describes if the crash record indicated that the driver or passenger of the motorcycle was wearing a helmet. It should be recorded only for occupants of motorcycles.

  AIR BAG

  Air bag utilization describes if the air bag was inflated during the crash. It may be difficult to determine belt usage for those crashes when the air bag inflates.

- **Classification of crash as alcohol related:** Add a variable to each occupant specific record to indicate if the crash was alcohol related.

  ALCOHOL

  Alcohol is defined as a binary variable to indicate if the occupant was involved in a crash in which either driver was under the influence of alcohol/drugs. A "yes" indicates that at least one driver was under the influence. **A "no" indicates that neither driver was under the influence of alcohol/drugs.**

- **Classification of census information:** Create a new variable on the crash data file to indicate the population density for the geographic location of the crash. This information is crucial for the Sensitivity Index calculations.

  POPULATION PER SQUARE MILE

  Create a new categorical variable in the crash record to record the population per square mile for the geographic location of the county/town in which the crash occurred. Population per square mile is a calculated field generated from census data. A reference file should be created indicating all the county/town codes in the state. For each code, the file should indicate the total population, total square miles, and the population per square mile. This reference file should then be linked via the county/town code to the crash record file. Once linked, the appropriate population per square mile should then be transferred to match the county/town code for the crash location.

- **Classification of severity:** Different measures of severity may be recorded directly or calculated from component information on the crash, and injury data files. Each of the potentially useful severity measures is discussed below.

  Crash Record KABCO Indicators: Most police crash records use the KABCO scale (or similar) to record a functional level of severity as fatal, incapacitating (needs help from the scene), nonincapacitating (obviously injured but ambulatory), possible (no injury apparent but might exist), or none/unknown. The police designate the level of injury for each occupant based on visual information available at the scene. This information does not and is not intended to predict survivability.

  EMS Trauma Score: In some states, the EMT at the scene records a severity score (i.e., revised trauma score, CRAMS score) directly on the run record; in other locations, the score is generated retrospectively by the computer from the components of the score. The EMS record lists the victim's first set of vital signs recorded at the scene. This information along with the Glasgow Coma Score are used to generate the Revised Trauma Score.

  (Optional) Hospital AIS and ISS scores: Create two new categorical variables on the hospital data file to record the Abbreviated Injury Score (AIS) and Injury Severity Score (ISS) severity scores. The AIS and ISS scores are not part of the Sensitivity Index. But they are important for different types of medical outcome analyses. Special software must be purchased to calculate both the AIS and ISS using anatomical information described by the ICD-9-CM codes assigned to each inpatient at the time of discharge from an acute care hospital. These scores correlate with survivability and are useful for medical outcomes research. Your trauma center may have a copy of this software.

  Vital Signs: Create new variables combining indicators for systolic, diastolic, respirations, and/or other vital signs to indicate levels of severity.

- **Calculated time variables:** Access, response, destination and other times can be calculated automatically by the computer from the actual times as recorded.

  ACCESS TIME

  Access time is defined as the time of the crash on the crash record subtracted from the time of the call to EMS on the run record.

  RESPONSE TIME

  Response time is defined as the time of the call to EMS on the EMS record subtracted from the time of EMS arrival at the scene also on the EMS record.

  AT SCENE TIME

At scene time is defined as the time of EMS arrival at the scene subtracted from the time EMS left scene.

TO HOSPITAL

To hospital time is defined as the time EMS left scene subtracted from the time of EMS arrival at the destination.

DESTINATION TIME

Destination time is defined as the time EMS arrives at the scene subtracted from the time EMS arrives at the hospital destination.

TOTAL TIME

Total time is defined as the time of arrival at the hospital destination recorded on the EMS record minus the time of the crash recorded on the crash report.

- **Occupant Information:** The Sensitivity Index requires access to information describing the identity of the vehicle (first, second, etc.) and the occupant's position (driver, passenger, front seat, back seat) and the severity of injury, if any, in order to construct the correct denominators to calculate the utilization rates for protective devices (air bag, safety belts, helmets, or no alcohol).

SECTION III: CASE SELECTION FOR LINKAGE

STEP 11: Ancillary Linkages

Existing state data can be improved by ancillary linkages to other data files to add date of birth, name, zip code, the vehicle identification number, and other identifiers for linkage. Ancillary files include injury registries, the driver licensing file, vehicle registration file, regional EMS data systems, national insurance index, and others. When two files vary in the power of their identifier information, ancillary linkages can fill in the gaps. Ancillary linkages are usually easier since, frequently, direct identifiers exist to perform the linkage. These identifiers include the driver license number, the vehicle registration or VIN number, etc. It should be noted that ancillary data files themselves often require file preparation before they can be linked to the primary files.

STEP 12: Choosing the Records for Linkage

Reducing the state data files to include only those records with the potential for linkage improves the efficiency and effectiveness of the linkage process. For example, out-of-state-residents should be eliminated if they are not included on both files being linked. The general rule should be to include records on the sub-file which you think may have the potential of linking. If you are unsure about a record's potential for linkage, include the record. It is easier to eliminate records and variables than to go back and add them.

The characteristics of your data files determine which records should be selected for linkage. To be selected, the record must be computerized. Records that are mandated for licensure or regulation are more likely to be complete and accurate, and thus more likely to be selected. Cause of injury codes, if available, also can be used to select records for linkage. For example, the EMS record may include a box to indicate a motor vehicle injury. The emergency department or hospital records may include an E-code indicating cause of injury as a motor vehicle crash. It is important to ensure the accuracy of the cause of injury information before using it for case selection. In some states, E-codes are not mandated and training may not be provided to ensure uniform documentation.

Case Selection Criteria: The following case selection criteria are examples for creating the sub-files for the actual linkage.

> **Crash records:** To minimize systematic bias, select all crash records. The police may have neglected to indicate the injury.

EMS records: Select only EMS records indicating an emergency (unscheduled) transport, refused transport or treatment, or a death at the scene. Eliminate the routine (scheduled) transports.

Hospital records: Select hospital records which have at least one ICD-9-CM code between 800-959 indicating an injury. From this group, consider eliminating records which have a cause of injury E-code not related to a motor vehicle crash or which represent multiple admissions for the same injury. The best match to the hospital record usually means the first hospital to which the patient is taken. Consider including records with injury related ICD-9-CM codes outside of the 800-959 range, such as V15.5, V71.3, V71.4, and codes for lumbago. Once the initial linkage has been completed, it may be useful, depending on the type of analysis, to match the hospital patient identification numbers back to the hospital files to locate subsequent admissions which may not have a discharge code in the originally selected ranges.

In some states it may be necessary because of patient confidentiality concerns to also restrict the file to only those variables necessary for the blocking and linkage. Once the reduced hospital file is linked, then the hospital records can be returned to the owner of the hospital data to obtain the rest of the data. This method eliminates concern that an outside party may have access to information about hospital utilization above and beyond the needs of the linkage project.

Data Quality: It is important to evaluate the completeness of the crash, EMS, and hospital data files and determine if specific geographical areas, providers (crash, EMS, hospital), occupants (front or back seat passengers) or other characteristics are under reported. Missing records prevent valid matches and thus skew the results.

Time Frame: Linkage is more efficient when performed using complete, statewide data for a finite period, usually 12 months. Data linkage is complicated when the records included in one data file represent a snapshot of current records that span a period of time greater than 12 months and the other data file includes records only for a calendar year. Crashes which occur before the first day of the twelve month period should be eliminated. EMS and hospital records for two days following the end of the 12 months should be added for possible linkage to those crashes which occur after 9 p.m. on the last day of the 12th month.

Hierarchical Linkage: Linkage should be performed in a hierarchical manner when the linkage variables are weak. Thus, sub-files should be created so that the records with the highest probability of linkage should be linked first and the records with the lowest probability of linkage should be linked last. For example, you may find it more efficient to link the records with a designated motor vehicle injury first and the records without a designated injury last. Hierarchical linkage is not necessary when the identifiers are sufficient to discriminate among the crashes and the occupants of a crash.

SECTION IV: PERFORMING THE DATA LINKAGE

Prior to the use of computers, small groups paper records could be linked manually after consideration of all of the information included in the record. With the advent of computers, it became feasible to link data files with a large volume of records. However, linkage required an exact match among the linkage variables and only the computerized information could be considered. To achieve an exact match, the attribute values for the linkage variables would have to be adjusted to compensate for the inevitable errors. As a result, many passes through the entire data file were necessary to handle all of the adjustments.

Probabilistic linkage techniques became available for highway safety data linkage in the form of new software, (AUTOMATCH/MINICODES), which focuses on the probability of a match without requiring the attribute values to match exactly. This methodology was extensively tested by matching a sample of individuals counted in the U.S.Census to a Post Enumeration Survey with the object of determining which individuals and households were present in both the census and survey. For this type of application, very high precision matching was required. Probabilistic linkage performed the task successfully. Section IV discusses the linkage process and theory of probabilistic linkage as implemented by the AUTOMATCH and MINICODES software. The instruction manuals provide more details about how to actually implement the software.

STEP 13: About Probabilistic Linkage

- **Purpose of Record Linkage**

The purpose of a record linkage application is to locate in different files records pertaining to the same crash victim despite the fact that the records may contain missing or incorrect information. Individual record linkage involves two files: **file A** and **file B**. For the crash and injury linkage, file A is usually the crash record and file B the injury record being linked.

Each file consists of fixed **fields**, which contain the information to be matched. The fields must be in fixed locations and have a uniform size (see Step 7).

One or more fields on file A must have equivalent fields on file B. For example, in order to match on date and age, both files A and B must include fields containing this information. The location and length of a field on file A may be different from its equivalent field on file B.

To link the two files, one could create a set of all possible record pairs. The first pair would be record number one from file A matched to record number 1 from file B. The next

pair would be record one from file A and record 2 from file B, until **n X m** pairs were formed (where **n** is the number of records on file A and **m** is the number of records on file B).

The objective of the record linkage process is to classify each pair as belonging to one of two sets: the set of matched record pairs **M(matched)**, and the set of unmatched record pairs, **U(unmatched).** For example, if we were to inspect, say the pair created from record 123 on file A with record 217 on file B, we must be able to say it is not a match (and belongs in set U) or it is a match and belongs in set M.

Thus, many more pairs are unmatched than matched. To illustrate this, consider two files with 1000 records each. There are 1,000,000 possible record pairs, but only 1000 possible matches (if there are no duplicates on the files). Thus, set M will include at most 1000 pairs and set U will include the remaining 999,000 pairs.

- **Feasibility of Record Linkage**

In order for a record linkage application to be feasible, it should be possible for a human to examine the match fields for any record on file A and the equivalent fields for any record on file B, and declare with reasonable certainty that the record pair examined is a match or a nonmatch.

For example, if the only field in common between two files were gender, then no one would say that if the gender agreed then the pair represented the same individual. However, if both files contained a field such as Social Security Number, then one could claim that if there was a match that it represented the same individual.

A rule of thumb for determining if a record linkage application is feasible is to multiply the number of values in each variable used for linkage and then to compare this product with the number of records in both files. If the product is much greater than the number of records, the application is probably feasible.

For example, if gender, age and initial were the only fields that could serve as matching fields, then the following calculation can be made: gender has two possible values, age has one hundred and initial has twenty-six. (2 x 100 x 26 = 5200). Since there are only 5200 possible values for the fields only very small data sets can be matched with any confidence. The probability that more than one record is an exact duplicate and does not represent the same individual is very high with a file size of 5200. The actual probabilities would depend on the distribution of the values in the fields.

STEP 14: Blocking the Data Files

- **Concept of Blocking**

For any files of reasonable size it is not feasible to compare all record pairs since the number of possible pairs is the product of the number of records on each file. Even two small files of 1000 records each generate 1,000,000 possible pairs to examine. Of this million, a maximum of 1000 will be matches. The other 999,000 are unmatched pairs. If there were a way to look at pairs of records having a high probability of being matches and ignoring all pairs with very low probabilities, then it would be feasible to conduct the linkage with large files.

Fortunately, the concept of **blocking** provides a method of limiting the number of pairs being examined. If one were to partition both files into mutually-exclusive and exhaustive subsets and only search for matches within a subset, then the process of linkage becomes manageable.

To understand the concept of blocking, consider a field such as **age.** If there are 100 possible ages, then this variable partitions a file into 100 subsets. The first subset is all people with an age of zero, the next is those with an age of 1, etc. These subsets are called **blocks** (or **pockets** in some systems). Suppose, for example, that the age values were uniformly distributed. If this were so, then out of a sample file consisting of 1000 records, there would be ten records for people of age zero on each file, ten records for people of age 1, etc.

The pairs of records to be compared are taken from records in the same block. The first block would consist of all persons of age zero on files A and B. This would be 10 x 10 or 100 record pairs. The second block would consist of all persons on files A and B with an age of 1. When the process is complete, we would have compared 100 (blocks) x 100 (pairs in a block) = 10,000 pairs, rather than the 1,000,000 record pairs required without blocking.

Multiple Passes: Blocking causes all records having the same value in the blocking variables to be matched. One consequence is that records failing to match on the blocking variables do not participate in the matching and thus are automatically classified as nonmatched. For example, if our blocking variable was age, and age was missing on some of the records, than those records would be unmatched.

To get around this problem, **multiple passes** are used. Any records that do not match can be rematched using another blocking scheme, say crash location. Only those cases that have errors on both age and crash location will not participate in the matching, unless additional passes are performed.

- **Selecting Blocking Strategies**

It should be obvious from the example above that smaller blocks are many times more efficient than large blocks. It is much better to use very restrictive blocking schemes (especially in the first pass). Since most of the records will match on the first pass, a second pass has much fewer records to process, and can be less restrictive.

A variable such as age alone is not a good blocking strategy since age is generally unevenly distributed (some ages may be much more prevalent in the files than others). In addition, partitioning a large file into 100 age categories still leaves many records in each block.

More than one variable may be chosen as a blocking variable in a pass. For example, if date (365), crash 4 hour time code (6), and gender (2) were blocking variables, and gender is coded as M or F, then the first block is female injured January 1 during the first 4 hours, the second is male injured January 1 during the first 4 hours, etc. This would now partition the files into 4,380 subsets for pass 1. Records with an error in gender, crash time code, or date would be unmatched and available for the second pass.

The unmatched records from pass 1 are called **residuals.** These **residuals** become the input files for the pass 2 match.

In general, 2 passes are sufficient to match almost all the cases included in each of the linkage phases. Blocking variables should be chosen so that as many records as possible have an opportunity to match during one of the two passes.

The blocks should be as small as possible. Less than 10-20 records per file is a good block size. Blocks should never exceed 100 records per file, or efficiency will be quite poor. The largest square block permitted would have 180 records on file A compared to 180 records on file B (32,400 pairs). The largest nonsquare block would have 12,000 records on file A and 2 records on file B (or vice versa). Any combination of sizes that does not exceed 32,400 pairs or 12,000 records on any single file are permissible for block sizes. If the maximum block size is exceeded, then all of the records in the block are skipped. They become residuals to be processed in pass 2.

The variables that are the best blocking variables are those with the most number of values possible and the highest reliability. For example, gender alone is a poor choice, since it only divides the file into 2 subsets. Similarly, fields subject to a great probability of error should be avoided. Apartment number is generally misreported or omitted, and hence would not make a good blocking variable. **In mathematical terms, the fields with the highest weights make the best blocking variables. Variables which can match exactly without allowances for errors also should be used as blocking variables.** Exhibit 5 indicates the types of information which are useful for blocking and the availability of this information in the major data files being linked.

In summary, probabilistic linkage reduces the scope of the linkage problem to small blocks of matched records within which the linkage occurs. Variables used for blocking should be reliable and applicable to all records. Blocks should be as small as possible (generally less than 20 on each file) to speed the matching. The number of records to be compared must not exceed memory on the computer. For a microcomputer, the total number of records on either file in one block must not exceed 1000 and the total number of records to be compared (A X B) in one block is limited to 12000. However, a block that large usually indicates that you need more or better blocking variables.

Exhibit 5: Types of Information Useful for Blocking

BLOCKING VARIABLES	Crash	Medical Record	Vehicle Insurance	Health Insurance
Date of Event: (Onset of Injury, EMS Pickup Admit to Hospital)	X	X	X	X
Age	X	X	X	X
Date of Birth	X	X		X
Gender	X	X	X	X
County of Event	X	EMS only	X	
Town of Event	X	EMS only	X	
Time Code: (Onset of Injury; Call to EMS EMS at scene; Arrival at Hospital)	X	X		
Probable Admit Date	X			X
Admit Hour to Hospital		X		
Destination / Hospital		X		X

STEP 15: Assigning the Weights

Weights are assigned only to the linkage variables. They are assigned based on the frequency of the variable value, meaning that rare values have higher weights. The frequency information allows the matcher to vary the weights according to the particular values of a field. Reviewing the frequency results helps to identify the following errors:

Unknown/Newborn age error: Frequency totals should exist for the blanks (unknown age), and age 0 (newborns). Failure to show totals for the unknown ages indicates a problem with the field designation for age. Age should be a character field so that the unknown ages are not recorded as a zero.

Elderly age error: Large totals for age 99 may indicate that your data file also records unknown ages as 99.

Times: If frequencies are calculated for times, the totals listed by the computer may include only times ending with a zero or a five. This pattern indicates that times are being rounded by the data collectors. Times in between are reported "outside" the computer table and thus will have a different set of weights. Although the variation in weights is not expected to be significant, they should be reviewed to ensure reasonableness.

The information contained in the variables to be matched helps the matcher (or a human) determine which record pairs are matches and which are nonmatches. Each field provides some information. Taken together, all the fields should determine with little equivocation, the status of the pair being examined.

Discriminating power: Some fields provide more information more reliably than others. For example, it would be absurd to sort both files on the gender variable, and assert that if the gender agrees, the record pair represents the same individual. However, it would not be so silly to sort both files on Social Security Number, and assert that if this number agrees then the record pair represents the same individual. This is because the probability of chance agreement on a rare event is relatively low compared to chance agreements on the other values. This section discusses how the **discriminating power** of each variable can be measured.

Each field has two probabilities associated with it. These are called the **m** and **u** probabilities. The **m** probability is the probability that a field agrees given that the record pair being examined is a matched pair. This is effectively one minus the error rate of the field. For example, in a sample of matched records, if gender disagrees 10 percent of the time due to a transcription error, or being misreported, then the m probability for this variable is 0.9 (1 - 0.1). The more reliable a field is, the greater the **m** probability will be.

The **u** probability is the probability that a field agrees given that the record pair being examined is an unmatched pair. Since there are so many more unmatched pairs possible than matched pairs, this probability is effectively the probability that the field agrees at random. For example, the probability that the gender variable agrees at random is about 0.5. Given a uniform distribution, there are four possibilities:

File A	File B
M	F
M	M
F	M
F	F

The gender agrees in two of the four combinations (thus, 0.5 **u** probability).

The weight for a field is computed as the logarithm to the base two of the ratio of **m** and **u**. To see how this translates into actual values, let's examine our example of the gender and the Social Security Number variables. Assume that gender has a 10 percent error rate and Social Security Number has a 40 percent error rate. The **m** probability for gender is 0.9. The **u** probability is 0.5 (from the above table). Thus, the weight for gender is $\log_2 (m/u) = \ln(m/u)/\ln(2) = \ln(0.9/0.5)/\ln(2) = 0.85$.

Conservatively, assume that the probability of chance agreement of Social Security Number is one in ten million. Given **m** as 0.6 (40 percent error rate in matched pairs), then the weight for Social Security is $\ln(0.6/0.0000001) = 22.51$. Thus, the weight for a match on the gender variable is 0.85 and a match on SSN is worth 22.51. The **weights have captured what we know intuitively about the variables.**

Composite Weights: For each record pair compared, a composite weight is computed and stored. The composite weight is the sum of the individual weights for all attribute comparisons. If an attribute value agrees in the pair being compared, the agreement weight, as computed above, is used. If an attribute value disagrees in the pair being compared, the disagreement weight is computed as: $\log_2 [(1-m)/(1-u)]$. This results in field disagreements receiving negative weights. **Thus, agreements add to the composite weight and disagreements subtract from the composite weight. Obviously, the higher the score, the greater the agreement.**

STEP 16: Linking the Files

In order to link crash and injury data, the blocking and linkage variables must be able to discriminate among the events and among multiple persons involved in the same event. Information to discriminate among events includes descriptors of the event. Information to discriminate among persons includes personal descriptors.

Data files are linked using direct and indirect identifiers capable of identifying a specific person involved in a specific event. Direct identifiers include name, social security number, unique patient identification number, or other types of identifiers which alone have the capability to identify a specific person or event. Exhibit 6 presents a list of direct variables useful for linkage.

Exhibit 6: Types of Direct Identifiers Information Useful for Linkage
Name
Initials
Last Name
Social Security Number
Unique Number: EMS Run Record Crash Record Unique Patient ID
Vehicle Identification Number
Driver License Number
Geographical Location: Global Positioning Latitude/longitude Node Marker

Exhibit 7 presents examples of indirect variables useful for linkage. Indirect variables must be combined in order to identify a specific person or event. Event identifiers are particularly important in the absence of strong personal identifiers. The indirect variables have been classified as either person or event identifiers. Person identifiers include personal descriptors such as age, date of birth, gender, residence, and type/area of injury. Event identifiers include descriptors of the event such as date, geographic location, times, providers, service area. Both direct and indirect variables are used in the linkage since there is the potential for errors to exist in both. In those instances when the variables available for blocking and linking are weak, it is permissible to use the same information for both blocking and linking. However, the variables should be modified slightly. For example, instead of date of crash, use year and month of crash to block, and the complete date to link.

Some states will be able to implement all of the variables listed in Exhibits 6 and 7; other states will use only a few. It is not necessary to use all of the variables presented to perform the linkage. Additional variables unique to your state should be included whenever they contribute to your capability to discriminate among events and occupants. Use all of the information you have even though it may apply to only a small percentage of the cases. Each bit of information contributes to the linkage of valid pairs.

Exhibit 7: Types of Indirect Identifiers Useful for Linkage

PERSON IDENTIFIERS	Crash	Medical Records	Vehicle Insurance	Health Insurance
Age	X	X	X	X
Date of Birth Birth Year Birth Month Birth Day	Usually available only for drivers	X		X
Gender	X	X	X	X
Area of Injury Codes	X	X		X
Type of Injury Codes	X	X		X
Residence Town Code		X	X	X
Residence Zip Code First 3 digits Last 2 digits		X	X	X
Probable Admit Date	X			
EVENT IDENTIFIERS				
Date of Event	X	X	X	X
County of Event	X	EMS		
Town/municipality of Event	X	EMS		
Mechanism of Injury	X	EMS Hospital		
Indication of Injury	X	X	X	X
Actual Times: (Onset Call to Police/EMS Arrival at Scene Arrival at Hospital Admit to Hospital)	X	X		
Time Codes: (Report to Police Arrival time)	X			
EMS Service	X	X		
Destination/Hospital		X		X
Service Area	X	X		
EMS Region	X	EMS		

STEP 17: Match, Nonmatch, Almost/Suspect Match

Probabilistic linkage generates matched pairs, clerical review pairs, and residuals in each of the two files being linked. The composite weight score for each pair is used to define a match, a nonmatch, or an almost/suspect match. Lowering the score to define a match may increase the likelihood of increasing your rate of false positive matches. Increasing the score to define an unsure match may increase the likelihood of increasing your rate of false negative matches. Initially, the range of scores to identify the almost/suspect matches should be fairly wide to avoid generating too many false negatives.

The clerical review process: During the clerical review process, the user designates if the almost/suspect matched pair is a match or a nonmatch. When duplicates occur, they may be swapped with the matched record. The clerical review screen displays the linkage variables and other information to assist the decision making. Exhibit 8 presents a list of variables which are useful during the clerical review process. These variables provide information for resolving the almost/suspect matches but do not participate in the blocking or linkage. A benefit of the clerical review process is that the review of the almost/suspect matches highlights more precisely how the matches and nonmatches should be defined to minimize the false positives and the need for clerical review.

Summary of the linkage process: The matching algorithm can be summarized as follows:

- A block of records is read on both files.
- For each possible record pair in the block, all fields are compared and a composite weight is computed. A matrix of composite weights results. The matrix size is nXm where n is the number of A records in the block and m is the number of B records in the block. The elements of the matrix are the composite weights.
- A linear sum assignment program is used to optimally assign the best matches.
- The assigned elements are examined. If they have a weight greater than the cutoff values, the pair becomes a match or clerical review pair.
- Duplicates are detected on both files by examining the row and column of an assigned pair. If there is more than one element whose weight is greater than the match cutoff weight, the pair is a potential duplicate.
- The assignments are written out to the pointer files.
- The residual pointers are updated to indicate which records did not match.

Exhibit 8: Types of Information Useful for Clerical Review

CLERICAL REVIEW	Crash	Medical Records	Vehicle Insurance	Health Insurance
PERSON IDENTIFIERS				
Ejected	X			
Actual Injury Type	X	X		X
Actual Injury Area	X	X		X
Actual Severity (KABCO, Trauma Score ISS)	X	X		
Position: Driver Passenger	X		X	
Vehicle Occupied	X		X	
Diagnosis Code DX1, DX2, etc.		X		
Disposition		X		
EVENT IDENTIFIER				
Number of Persons Involved	X		X	
Number of Vehicles Involved	X		X	
Times: (Report to Police Arrival at Scene Arrival at Hospital)	X	X		
EMS Service	X	X		X

STEP 18: Resolving Problems

Failure to link: Common problems prevent the linkage of records regardless of the characteristics of the data file. These problems include:

- Match parameters are too strict to allow linkage when errors exist
- Key data linkage variables were in error
- Key data linkage variables were missing
- The potential match record was not included in files being linked
- An out-of-state event is not documented on in-state data files with subsequent treatment by an in-state provider or vice versa
- Motor vehicle E-code was in error

Some linkage problems can be corrected using one or more of the following suggestions.

Edit the data files: Linkage problems may justify additional editing in areas previously thought insignificant. For example, the existence of multiple records per occupant or confusion over the coding of newborns and unknown age may generate confusing linkage results. If the problem is significant, the data files should be edited and relinked.

Change the dictionaries: Adjusting the dictionaries may be necessary to reflect the unique characteristics of your data files or to correct errors in the variable listings for column, length, or missing value indicator.

Change the match specifications: Adjustments to the match specifications for the blocking variables or the variable types may be necessary to make the linkage more efficient or to restrict the numbers of records selected for clerical review.

Correct the comparative variable designations: Increasing the m probabilities will result in a higher penalty being assigned if the value does not match. Adjust the definitions for matches and nonmatches to minimize the number of false positive linkages and the number of matched pairs requiring clerical review.

SECTION V: ANALYZING THE DATA

STEP 19: Reviewing the Linkage Results

Histogram: The matcher program produces a histogram to display the distribution of the composite weights for all comparisons. Records that do not match generally have high negative weights since most fields should disagree. Records that match have high positive weights. This produces a bimodal distribution that allows the matcher to discriminate between matches and nonmatches. The clerical review cases should be defined as the weight where the "bump" in the histogram reaches near the axis. A portion of the histogram generated by the computer looks like this:

```
*  H I S T O G R A M
*
*  Distribution of observed weights for all possible comparisons
*  Scale based on mean frequency of: 8

*  For weights with a frequency greater than the mean -
*  The histogram shows an arrow in the last column
*
*     WGT    Freq
*  -10.00     100  ************************************************************ >
*   -9.50       0
*   -9.00       0
*   -8.50       1 *
*   -8.00       7 *******
*   -7.50       0
*   -7.00       7 *******
.....................................................................
*    4.00       0
*    4.50       0
*    5.00       2 **
*    5.50       2 **
*    6.00       1 *
*    6.50       1 *
*    7.00       1 *
*    7.50       1 *
*    8.00       1 *
*    8.50       2 **
*    9.00       2 **
*    9.50       1 *
*   10.00       7 *******
*   10.50       3 ***
*   11.00       8 ********
*   11.50      11 ***********
*   12.00       9 *********
*   12.50       2 **
*   13.00       7 *******
*  TOTAL COMPARISONS: 0000307
```

The histogram above presents an example of how the weights for all possible pairs are likely to be distributed. Matches are included among the higher positive weights. Each line indicates the weight (in 0.5 increments), the frequency and a graphic representation. Lines ending with > exceed the range. There are always many more unmatched pairs, negative weights, than matched pairs so the ranges in the beginning of the chart are high. The histogram is essential for deciding the cutoff values for defining a match and an unsure match.

Notice in the example above the match cases trail off around 6. Below this there are some "bumps." Consequently, in this case we made the clerical review cutoff 6 and the unmatched cutoff 4. **It should be noted that the program arranges all possible pairs by weight into a large matrix. These same pairs are presented in the histogram.** The frequencies for unmatched cases trail off as the weights go higher and the frequencies for matched cases trail off as the weights go lower. This forms two curves (or modes). These represent the unmatched and the matched cases. The farther apart these modes are from each other, the better the discrimination between the matched and unmatched records. Try to draw a continuous curve from the histogram chart, and examine the tails of the curve to decide where to make the cutoff points. Exhibit 9 translates the histogram into a graphic representation of the bimodal distribution of matched and unmatched pairs.

Exhibit 9: Distribution of Weights

The bimodal characteristic of the histogram is affected by the number and types of linkage variables. The more variables used for linkage, the wider the range of weights and the more pronounced the bimodal distribution. An insufficient number of variables will decrease the range of rates and obscure the bimodal distribution. Use of variables for linkage which correlate with each other, for example date and year/month of crash, will increase the range of weights. In some instances, the increased weight assigned to date information may obscure the value of other variables. If the correlation is known in advance, it might be advisable to assign a lower **m**prob (for example .5 instead of .9) to the date information. Another solution would be to use year and month for blocking and the complete date for linkage. Since no weight is assigned to the blocking variables, year/month information can be used for blocking and the normal **m**prob for date (for example .9) can be used for the linkage. Summary output

statistics are generated for each linkage pass and are reported below using the following format:

```
*  OUTPUT STATISTICS FOR MATCH:
*  PASS: 1
*     80  Records on File A
*     75  Records on File B
*           30  A residuals from previous pass
*           30  B residuals from previous pass
*           78  A records read
*           72  B records read
*     20  Blocks processed
*      0  OVERFLOW blocks
*            8  Maximum A block size (including overflow)
*            5  Average A block size (including overflow)
*            9  Maximum B block size (including overflow)
*            5  Average B block size (including overflow)
*     52  Matched pairs
*      2  Exact matched pairs
*      3  Clerical pairs
*      3  A duplicates
*            1  Exact A duplicates
*      2  B duplicates
*            0  Exact B duplicates
*      9  A residuals (including skips & missing)
*      9  B residuals (including skips & missing)
*     13  A records skipped
*      6  B records skipped
```

The number of records read on each file appears first. The number of blocks processed is the number of times the blocking keys agreed. Residuals are those records on each file that remain unmatched. A record is skipped whenever the blocking keys do not agree. **Only records included in a block participate in the linkage.**

(In MINICODES, records are also skipped whenever there is a block overflow. A block overflow occurs whenever the number of comparisons exceeds the maximum matrix size. In this case, the affected records on both files are skipped. Block sizes should be kept small enough so that this does not happen. However, if it does, then subsequent passes should match these records. Skipped records do not participate in the linkage and are assigned as residuals for the next pass.)

STEP 20: Validating the Linkage Results

The process of validating your linkage results helps you to understand how the linkage works. It also highlights problems with the quality of your data. This process focuses on answering the questions:

What records linked which should have linked?
What records linked which should not have linked?

What records did not link which should not have linked?
What records did not link which should have linked?

Is there a difference analytically between different samples of linked data?

Potential systematic bias in the linked data: Systematic bias skews the results. Frequency distributions of each variable to be included in the analyses should be prepared for both the linked and unlinked data. These distributions will highlight out of range data, outliers, and unexpected patterns. Bias may be caused by different reporting thresholds, definitional inconsistencies, incomplete computerization, missing data, single option data fields, etc.

False positive cases (records which should not have matched but did match): False positives overstate the results. They usually result when the identifiers available in the two files being linked are sufficient to cause a match but not sufficient to discriminate among different individuals. False positives introduce a bias if they occur nonrandomly since they are more likely to occur among populations that are common in each data file, such as the young men in zip code areas with large populations.

False positives can be identified by selecting a random sample of the actual paper crash records and manually linking them to the actual medical records to identify records which should not have matched.

False negative cases (records which should have matched but did not): False negatives understate the results by failing to link all possible matches. False negatives usually are not randomly distributed. They include people in counties adjacent to neighboring states, the less serious acute injuries, and the populations (i.e., uninjured passengers) for which information is not collected on the crash record. The false negatives may cause under representation of specific types of injuries common to passengers or to the less seriously injured. When working with hospital charges, the false negative rate is likely to cause average charges to be understated.

The under reporting of crash or injury records for specific areas or populations or the omission of the motor vehicle crash as the cause of injury in the injury records contribute to the linkage of false positive or false negative pairs.

False negatives can be estimated by determining the linkage rate of the injury records for which the motor vehicle cause was designated and then manually reviewing the records which did not link.

Identify EMS records with an MVA code: EMS records frequently designate if the injury was the result of a motor vehicle crash. This information may be designated by a box for MVA, trafficway, etc. Although not always recorded by the EMS personnel at the scene, when indication of a crash is recorded, these records are useful as a tracer group to identify records which should have matched but did not.

Identify hospital records with an E-code indicating motor vehicle crash as the cause of injury: Hospital records may include an E-code to indicate if the injury was the result of a motor vehicle crash. Although not always recorded by the hospital personnel, when the E-code is recorded, these records are useful as a tracer group to identify records which should have matched but did not. This population is particularly important because it represents persons with severe injuries requiring hospitalization.

Identify the analytical impact of the case mix generated by the linkage process: By definition, the linked records should include the injured and the unlinked records should include the uninjured and the characteristics of these two populations should differ. Since we do not expect all crash injuries to link to an injury record, it is important that the sample of linked records for the injured represent the unlinked records for the injured. Implementing an analysis (such as the Sensitivity Index described below) on different samples of linked data may be useful to highlight potential bias in the results. All phases of the linkage (Crash to EMS, linked crash/EMS to hospital, unlinked crash to hospital, etc.) must be evaluated for bias since systematic bias may be more prevalent with some data files then with others. This is very important when ancillary sources of data have been used to expand the information available for linkage. Records receiving the additional information may have a higher probability for matching and thus of being included in the analyses. The analytical results may be skewed accordingly.

Analytical results are affected by outliers and the number of cases in the cell being analyzed. A single patient may incur costs of more than $900,000 while the majority of the other crash victims incur costs of less than $100,0000. Deaths frequently have lower costs than victims who live. Administrative practices in the health care system may have more influence on charge variations than crash severity. Locating the actual total charges is complicated by the presence of multiple payers. Averaging the charges when the number of cases is small may produce questionable results.

STEP 21: Applying the Linked Data

Descriptive Statistics: Linked state data provide access to a wealth of computerized information connecting the event with medical and financial outcomes. The analytical potential is limited only by the completeness and quality of the computerized data.

CODES Mandated Model: As part of NHTSA's Crash Outcome Data Evaluation System (CODES) project, seven states have successfully implemented linked data to study the benefits of safety belts and helmets on mortality, morbidity, severity, and inpatient hospital charges. The results were reported to Congress in February 1996. Each state linked crash data to injury and claims data as described in Exhibit 2. The CODES model defined injury using a combination of the KABCO designations and linkage to a medical or claim record indicating

an injury. The logistic regressions controlled for the type of crash, urban/rural location, age, gender, crash severity, roadway surface conditions, time of day, type of vehicle, seating position, and intersection related. In addition to the mandated model, each of the CODES states performed other state specific analyses.

Sensitivity Index: The Sensitivity Index was designed to compare EMS performance within and between states. It is generated from linked state crash and EMS data. Index criteria were chosen to address the following questions:

1. How significant is the problem of motor vehicle injuries?
2. How quickly are motor vehicle injuries responded to?
3. How severe are motor vehicle injuries?
4. How quickly are motor vehicle injuries treated and transported?
5. How well are serious motor vehicle injuries treated?
6. Could the injuries have been prevented?

The Index was originally developed as a statewide measure to generate data that are standardized by police designated severity level or by population per square mile for interstate comparisons. The chosen criteria met tests of feasibility, simplicity, availability, and usefulness. Each is defined below.

1. **Injuries Per Hundred Million Vehicle Miles (HMVM)**
2. **Average Revised Trauma Score**
3. **Average EMS Transport Rate**
   (Variables 1-3 reported for each police severity level: fatal, incapacitating, nonincapacitating)

4. **Average, Standard Deviation Access Time**
5. **Average, Standard Deviation Response Time**
6. **Average, Standard Deviation Treatment/Transport Time**
   (Variables 4-6 reported for each location type defined as metro, urban, suburban, rural, wilderness statewide)

7. **Survivability**
8. **Prevention: safety belts, helmets, no alcohol**
   (Variables 7-8 reported as a single statewide percent)

STEP 22: Documenting the Linkage Process

Replication of the linkage process is facilitated if you document what you do as you go along. For example, documenting the editing process makes it possible for you to save time when you perform the next linkage.

Your documentation files should specify the following:

- Names of contacts for the state data and the agreements controlling use of the state data
- Complete description of the files and the file preparation required for linkage
- Notes describing problems associated with the linkage process and suggestions for improving the linkage when it is repeated

APPENDIX A: TECHNICAL ASSISTANCE FOR DATA LINKAGE

- Provided to States On-Site by Teams of Experts from the CODES States

NHTSA funds a team of data linkage experts to provide customized on-site technical assistance to states upon request. This assistance provides information on how to obtain state data, prepare the files for linkage, and establish an Advisory Committee to institutionalize data linkage. Experts also are available to come to your state to help you implement the probabilistic linkage software, link your data, and validate the linkage results. And the experts are available to assist in developing analytical uses for your linked state data. Requests for technical assistance should be directed to your NHTSA Regional Administrator.

George A. Luciano
Region I - CT, ME, NH, RI, VT
617-494-3427

Thomas M. Louizou
Region 2 - NY, NJ, PR, VI
914-682-6162

Eugene Peterson
Region 3 - DE, DC, MD, PA, VA, WV
410-768-7111

Thomas J. Enright
Region 4 - AL, FL, GA, KY, MS, NC, SC, TN
404-347-4537

Donald J. McNamara
Region 5 - IL, IN, MI, MN, OH, WI
708-503-8892

Georgia S. Chakiris
Region 6 - AR, LA, NM, OK, TX
817-334-4300

Troy R. Ayers
Region 7 - IA, KS, MO, NE
816-822-7233

Louis R. DeCarolis
Region 8 - CO, MT, ND, SD, UT, WY
303-969-6917

Joseph Cindrich
Region 9 - AZ, CA, HI, NV
415-744-3089

Curtis A. Winston
Region 10 - AK, ID, OR, WA
206-220-7640

- Technical Assistance Provided by Telephone

Information about CODES and data linkage also is available by telephone from CODES experts at NHTSA and each of the CODES states. Contact can be made directly to the persons listed below or through your state's Governor's Highway Safety Representative.

Dennis Utter, CODES COTR
202-366-5351

Sandy Johnson, CODES Consultant
202-366-5364

NHTSA
400 Seventh St., SW Room 6125
Washington, DC 20590

Karl Kim
Hawaii CODES
808-956-7381

Karl Finison
Maine CODES
207-623-2555

Mark Van Tuinen
Missouri CODES
314-751-6274

Richard Guerin
New York CODES
518-474-2219

Hank Weiss
Pennsylvania CODES
412-647-1110

Pat Nechodom
Utah CODES
801-581-6410

Martha Florey
Wisconsin CODES
608-266-3557

APPENDIX B: INTERNET SITES FOR CODES INFORMATION

NHTSA

World Wide Web at: http://www.nhtsa.dot.gov/nrd/nrd30/reports.html

PENNSYLVANIA

World Wide Web at: http://www.pitt.edu/~icrin (CODES page to be added in near future)

UTAH

World Wide Web at: http://www-CODES.med.utah.edu

WISCONSIN

World Wide Web at: http://linear.chsra.wisc.edu/chip/linkinfo

GLOSSARY OF TERMS

AIS: The Abbreviated Injury Score is an anatomical measure of severity, derived from narrative descriptions of the injury or International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes, frequently used in highway safety analyses.

AUTOMATCH: This software, developed by Matt Jaro of MatchWare Technologies Inc. (301-384-3997), implements probabilistic linkage techniques. The linkage process generates matches, nonmatches, and almost/suspect matches. The CODES project demonstrated the feasibility of using this software to link crash and injury state data.

CODES: The Crash Outcome Data Evaluation System (CODES) refers to a National Highway Traffic Safety Administration (NHTSA) project which funded Hawaii, Maine, Missouri, New York, Pennsylvania, Utah, and Wisconsin to implement probabilistic linkage techniques to link computerized state data. The state data include computerized crash, EMS, emergency department, hospital, claims, and death certificate records. Records were linked using direct and combinations of indirect identifiers to identify person-specific information located in different data files.

Discriminating Power: Information which is used to link for the same person records located in different files must be able to locate the person and the event in which the person was involved. For the crash and injury linkage, the data needed for linkage must be able to identify the occupant and the specific crash in which the occupant was involved.

Duplicate: AUTOMATCH uses a linear sum assignment algorithm to designate the pairs as matches within the block. The algorithm chooses those matched pairs which maximize the total score for the block. Pairs not chosen as matches but which have similar composite weights as those chosen are considered as duplicates which should be reviewed during the clerical review process.

E-Code: The E-Code is used to indicate the external cause of the injury. These codes are part of the International Classification of Diseases, 9th Revision, Clinical Modification.

EMS: EMS refers to the Emergency Medical Services System which provides prehospital care to victims of emergency illness and injury.

Glasgow Coma Score (GCS): The Glasgow Coma Score (GCS) consists of measures for eye opening, motor and verbal response to indicate the level of consciousness for patients suffering an injury.

Histogram: A histogram is a chart or graph which presents a frequency distribution. For the crash and injury linkage, the computer produces a histogram to present the frequency of composite weights for the pairs generated by the linkage. The weights range from negative scores to high positive scores to create a bimodal distribution.

Identifier: An identifier may be a direct identifier such as a name or number which is unique to the individual. Uniqueness can also be achieved by combining indirect identifiers such as date of birth, gender, initials, zip code of residence, type and area of injury, date and hour of admission, etc.

MINICODES: The MINICODES software is preprogrammed to implement the probabilistic algorithms to link crash to EMS, linked crash/EMS to hospital, and unlinked crash to hospital data files using a microcomputer. Compared to AUTOMATCH, this software is useful for experimenting with the linkage process, but has fewer capabilities and is not suitable for very large data files. A copy of MINICODES plus an instruction manual were distributed to the Governor's Highway Safety Representative in each state.

Outlier: An outlier is an exception which greatly exceeds the average and which will skew the average when included in its calculation. For example, the average of one $900,000 inpatient charge and nine $100,000 inpatient charges equals $180,000 when the $900,000 is included but only $100,000 when the outlier ($900,000) is excluded.

Residual: A residual represents a record which is not matched either because it was excluded from the block or did not match during the linkage process.

State Data: State data are computerized and include all records statewide. Common state data include the crash, EMS, and hospital discharge data files.

Exhibit 3. Characteristics of the state data useful for highway safety and injury control

Data Source	Data Collector	Statewide	Population Based	Edited	Record Unit	Indicates CrasNh	
Nonmedical Data Sources							
Crash **Report**	Department of Transportation/ Public Safety / Motor Vehicles	✔	✔		Crash	✔	
Vehicle Registration	Department of Motor Vehicles	✔	✔		Vehicle		
Driver Licensing	Department of Motor Vehicles	✔	✔		Driver		
Census	Department of Health	✔	✔		Person		
Roadway/Infrastructure	Department of Transportation	✔	✔		Road	✔	
Medical Data Sources							
EMS	Depts of Health or Public Safety		✔		Event	✔	
Emergency outpatient	Hospital/Claims				Event		
Hospital discharge	Dept. of Health	✔	✔	✔	Event		
Registries: Trauma, Head & Spinal Cord, Poison	Hospital or Dept. of Health			✔	Person	✔	
Death Certificates	Dept of Vital Statistics	✔	✔	✔	Person	✔	
Insurance Claims Data							
Medicaid, Medicare	Dept of Health	✔		✔	Claim		
Private Health Insurance	Health Insurance Co				Claim		
Worker's Compensation	Dept. of Labor	✔		✔	Claim		
Private Vehicle Insurance	Vehicle Insurance Co				Claim	✔	
National Auto Insurance Files	Association of Insurance Co				Claim	✔	