

Technical Issues Related to the Probabilistic Linkage of Population-Based Crash and Injury Data

Sandra Johnson, National Highway Traffic Safety Administration

Abstract

NHTSA's Crash Outcome Data Evaluation System (CODES) project demonstrated the feasibility of using probabilistic linkage technology to link large volumes of frequently inaccurate state data for highway safety analyses. Hawaii, Maine, Missouri, New York, Pennsylvania, Utah, and Wisconsin were funded by NHTSA to generate population-based medical and financial outcome information from the scene to final disposition for persons involved in a motor vehicle crash. This presentation will focus on the technical issues related to the linkage of population-based person-specific state crash and injury data.

Data Sources and Access

Data for the CODES project included records for the same person and crash event located in multiple different files collected by different providers in different health care settings and insurance organizations at different points in time. Each data file had a different owner, was created for a specific use, and was not initially designed to be linked to other files. Crash data were more likely to be in the public domain. Injury data were protected to preserve patient confidentiality. Each data source added incremental information about the crash and the persons involved.

Six of the seven states linked person-specific crash data statewide to EMS and hospital data. The EMS data facilitated linkage of the crash to the hospital data because they included information about the scene (pick-up) location and the hospital destination. The seventh state was able to link directly to the hospital data without the EMS data because date of birth and zip code of residence were collected on the crash record for all injured persons. Other data files, such as vehicle registration, driver licensing, census, roadway/infrastructure, emergency department, nursing home, death certificate, trauma/spinal/head registries, insurance claims and provider specific data, were incorporated into the linkage when available and appropriate to meet the state's analytical needs.

Importance of Collaboration

Collaboration among the owners and users of the state data was necessary to facilitate access to the data. A CODES Advisory Committee was convened within each state to resolve issues related to data availability, patient confidentiality, and release of the linked data. The committee included the data owners such as the Departments of Public Safety, Health, Office of EMS, Vital Statistics, private and public insurers of health care and vehicles among others. Users included the owners, researchers, governmental entities, and others interested in injury control, improving medical care, reducing health care costs, and

improving highway safety.

File and Field Preparation

File preparation usually began with the creation of a person-specific crash file to match the person-specific injury data. Some of the data files only had one record per person; others, such as the EMS and hospital data files had more than one, reflecting the multiple agencies providing EMS care and the multiple hospital admissions for the same injury problem respectively. In some instances, all of the records were included in the linkage; at other times, the extra records were stored in a separate file for reference and analysis.

Except for Wisconsin which benefitted from state data which were extensively edited routinely, all of the states spent time, sometimes months, preparing their data for linkage. In most states, the hospital data required the least amount of editing. Preparation included converting the coding conventions for town/county codes, facility/provider, address, gender, and date in one file to match similar codes in the file being linked. Newborns were separated from unknown age. Date of birth and age discrepancies were resolved. Out of sequence times were corrected and minutes were added when only hour was documented. New variables were created to designate blocks of time, service areas for police, EMS and the hospital, probable admit date and others. Ancillary linkages to other data files were performed to beef up the discriminating power of the existing variables. Name and date of birth were the most common data added to the original files to improve the linkage.

Blocking and Linking Data Elements

Persons and events were identified using a combination of indirect identifiers and, in some linkages, unique personal identifiers, such as name, when they existed. Each of the CODES states used different data elements to block and link their files. Which variables were used for blocking and which for linkage depended upon both the reliability and availability of the data within the state, the linkage phase, and the files being linked. Most states used location, date, times, provider service area, and hospital destination to discriminate among the events. Age, date of birth, gender, and description of the injury were used most often to discriminate among persons. Hawaii, Missouri, New York, and Utah had access to name or initials for some of the linkages.

Linkage Results

Conditions of uncertainty govern the linkage of crash and injury state data. It is not certain which records should link. In the ideal world, records should exist for every crash and should designate an injury when one occurs; injury records should exist documenting the treatment for that injury and the crash as the cause; and the crash and injury records should be collected and computerized statewide. Linkage of a crash with an injury record should confirm and generate medical information about the injury. No linkage should confirm the absence of an injury. But that is the ideal world. In the real world, the crash record may not indicate an injury even though an injury occurred; the matching injury record may not indicate a crash or even be accessible; so it is difficult to know which records should link.

Linkage rates varied according to the type of data being linked. In each of the CODES states, about 10% of the person-specific police crash reports linked to an EMS record and slightly less than 1.8% linked to a hospital inpatient record, a reflection of the low rate of EMS transport and hospitalization for crash injuries. The linkage rates also varied by police designated severity level (KABCO). Linkage to the fatal injury records was not always 100%, but varied according to whether deaths at the scene were transported either by EMS or a non-medical provider. For the non-fatal injuries, linkage rates were higher for the more severe cases which by definition were likely to require treatment and thus to generate a medical record. About 76-87% of the drivers with incapacitating injuries linked to at least one injury or claims record (except

for Wisconsin, which had limited access to outpatient data and Pennsylvania which used 6 levels to designate severity). Linkage rates for persons with possible injuries varied widely among the seven states. Because of extensive insurance data resources, about two-thirds of the possible injuries linked in Hawaii and New York compared to a third or less in the other states. Many more records indicating “no injuries” matched in New York and Utah, again because of access to extensive computerized outpatient data for the minor injuries. Included in this group of not injured were people who appeared uninjured at the scene but who hours or days after the crash sought treatment for delayed symptoms, such as whiplash. Overall, the CODES states without access to the insurance data linked between 7-13% of the person-specific crash reports for crashes involving a car/light truck/van to at least one injury record compared to 35-55% for Hawaii and New York, the states with extensive outpatient data. Wisconsin linked 2% of its drivers to the hospital inpatient state data and this rate matched that for the seven states as a group.

Linkage of the records for the motorcycle riders was much higher than the car/light truck/van group, a reflection of the high injury rate for cyclists involved in police reported crashes. As expected the linkage rates were lower for the lower severities. Except for Pennsylvania and Wisconsin, more than 45 per cent of the person-specific motorcycle crash records linked to at least one injury record.

Validation of the Linkages

Causes of false negatives and false positives vary with each linkage because each injury data file is unique. Since it is unknown which records should link, validation of the linkage results is difficult. The absence of a record in the crash file prevents linkage to an injury record; the absence of a cause of injury code in the injury record risks a denominator inflated with non-motor vehicle crashes. The states assigned a high priority to preventing cases which should not match from matching and conservatively set the weight defining a match to a higher positive score. At the same time, they were careful not to set the weight defining a nonmatch too low so that fewer pairs would require manual review. The false positive rate ranged from 3.0 - 8.8 percent for the seven states and was viewed as not significant since the linked data included thousands of records estimated to represent at least half of all persons involved in motor vehicle crashes in the seven CODES states.

False positives were measured by identifying a random sample of crash and/or injury records and reviewing those that linked to verify that a motor vehicle crash was the cause of injury. Maine, Pennsylvania, and Wisconsin read the actual paper crash, EMS, and hospital records to validate the linkage. Missouri compared agreement on key linkage variables such as injury county, last initial, date of event, trafficway/trauma indicators, date of birth, or sex. Wisconsin determined that the false positive rate for the Medicaid linkage varied from that for hospitalizations generally since Medicaid cases were more likely to be found in urban areas.

False negatives were considered less serious than a false positive so the states adjusted the cut-off weight defining a nonmatch to give priority to minimizing the total matched pairs requiring manual review. A false negative represents an injury record with a motor vehicle crash designated as the cause which did not link to a crash report or a crash record with a designated severe injury (i.e., fatal, incapacitating) for which no match was found. The rates for false negatives varied from 4-30 percent depending on the linkage pass and the files being linked. The higher rates occurred when the power of the linkage variables to discriminate among the crashes and the persons involved was problematical. False negatives were measured by first identifying the records which should match. These included crash reports indicating ambulance transport, EMS records indicating motor vehicle crash as the cause of injury or hospital records listing an E code indicating a motor vehicle crash. These records were then compared to the linked records to identify those that did not link. False negatives were also identified by randomly selecting a group of crash reports and manually reviewing the paper records to identify those which did not link.

Crash and injury records failed to match when one or the other was never submitted, the linking criteria were too restrictive, key data linkage variables were in error or missing, the case selection criteria, such as the E-code, were in error or missing, the crash-related hospitalization occurred after several hours or days had passed, the crash or the treatment occurred out-of-state, etc. Lack of date of birth on the crash report for passengers was a major obstacle to linkage for all of the states except Wisconsin which included this information for all injured passengers. (As the result of the linkage process, Maine targeted the importance of including this data element on the crash report.) Among the total false negatives identified by Wisconsin, 12 percent occurred because the admission was not the initial admission for the crash and 10 percent occurred because key linkage variables were missing. Another 7.5 percent occurred because the linking criteria were too strict. About 7 percent were missing a crash report because the crash occurred out of state or the patient had been transferred from another institution. Twelve percent of the false negatives were admitted as inpatients initially for other reasons than the crash. It was not possible to determine the false negative rates when the key data linkage variables or E-code were in error, when out of state injuries were treated in Wisconsin Hospitals and when the crash record was not received at DOT.

In spite of the failure of some records to match, the estimates of matching among those that could be identified as “should match” was encouraging. Missouri estimated linkage rates of 65 percent of the hospital discharge, 75 percent of the EMS records, and 88 percent of the head and spinal cord injury registry records when motor vehicle crash as the cause of injury was designated on the record. Comparison of Missouri’s linked and unlinked records suggested that actual linkage rates were even higher, as unlinked records contained records not likely to be motor vehicle related injuries (such as gunshot, laceration, punctures, and stabs). The linked records showed higher rates of fractures and soft tissue injuries, which are typical of motor vehicle crashes. Seventy-nine percent of the fractures were linked, as were 78 percent of soft tissue injuries.

The comparison of linked and unlinked records does not suggest that significant numbers of important types of records are not being linked, though perhaps some less severely injured patients may be missed. Because ambulance linkage was used as an important intermediate link for the hospital discharge file, some individuals not injured severely enough to require an ambulance may have been missed, but they would also be less likely to require hospitalization. Any effect of this would be to erroneously raise slightly the estimate of average charges for hospitalized patients.

Significance of the False Positive and False Negative Rates

Although the rates for the false negatives and false positives were not significant for the belt and helmet analyses, they may be significant for other analyses using different outcome measures and smaller population units. For example, analyses of rural/urban patterns may be sensitive to missing data from specific geographic areas. Analyses of EMS effectiveness may be sensitive to missing data from specific EMS ambulance services or age groups. Another concern focuses on the definition of an injury link. Defining an injury to include linkage to any claim record that indicated medical treatment or payment increases the probability of including uninjured persons who go to the doctor for physical exams to rule out an injury. But this group also includes persons who are saved from a more serious injury by using a safety device, so although they inflate the number of total injuries, they are important to highway safety. When minor injuries are defined as injuries only if their existence is verified by linkage, then by definition the unlinked cases become non-injuries relative to the data sources used in the linkage. States using data sources covering the physician’s office through to tertiary care will have more linkages and thus more “injuries.” Estimates of the percentage injured, transported, admitted as inpatients, and the total charges will vary accordingly.

The Linkage Methodology is Robust and the Linked

Data Are Useful

Seven states with different routinely collected data that varied in quality and completeness were able to generate from the linkage process comparable results that could be combined to calculate effectiveness rates. The states also demonstrated the usefulness of the linked data. They developed state-specific applications to identify populations at risk and factors that increased the risk of high severity and health care costs. They used the linked data to identify issues related to roadway safety and EMS, to support safety legislation, to evaluate the quality of their state data and for other state specific purposes.