

# **Technical Report on New Mexico CODES**

## ***Data Linkage for 1995***

**Carrie Rudd, Research Analyst**

**Division of Government Research**

University of New Mexico

*Produced Under*

**State of New Mexico Department of Health Contract No. 96.665.14.015**

for the Division of Epidemiology, Evaluation and Planning

**April 14th, 1997**

## **Acknowledgments**

**Many people helped to make the 1995 CODES Data Linkage possible. Sincere thanks go to Stuart Castle, Kathy Goodyear, Letty Rutledge, and Brian Woods for their friendly support and quick responses to my requests for information, to Keith Smith for his work on the Medicaid linkage, and to Jim Davis for his help in editing this document.**

**Special thanks go to Laura Ring for her excellent work on the 1994 CODES Data Linkage which set high standards for this and future CODES work.**

# Table of Contents

<b>1. INTRODUCTION</b> .....	<b>1</b>
<b>2. THE 1995 CODES FILES</b> .....	<b>1</b>
2.1 THE CRASH FILES .....	1
2.2 THE NEW MEXICO SYSTEMS TRAUMA REGISTRY .....	1
2.3 THE STATE OF NEW MEXICO OFFICE OF THE MEDICAL INVESTIGATOR DATA .....	2
2.4 NEW MEXICO HOSPITAL INPATIENT DISCHARGE DATA .....	2
2.5 THE NEW MEXICO MEDICAID CLAIMS FILE .....	2
<b>3. DATA PREPARATION</b> .....	<b>3</b>
3.1 THE CRASH DRIVER/OCCUPANT FILE .....	3
3.2 THE SYSTEMS TRAUMA REGISTRY DATA .....	4
3.3 OMI DATA .....	5
3.4 HIDD FILE .....	5
3.5 THE MEDICAID CLAIMS FILE .....	5
<b>4. QUALITY OF MATCHING VARIABLES</b> .....	<b>6</b>
<b>5. PROBABILISTIC MATCHING (LAURA RING)</b> .....	<b>7</b>
<b>6. UNDUPLICATING THE SYSTEM TRAUMA REGISTRY</b> .....	<b>8</b>
<b>7. GROUPING THE HOSPITAL INPATIENT DISCHARGE DATA</b> .....	<b>10</b>
7.1 DELAYING AGGREGATION OF HIDD .....	11
<b>8. MATCHING PEOPLE KILLED IN CRASHES TO THE OMI DATA</b>	
8.1 MATCHING DRIVERS, NON-DRIVER OCCUPANTS, PEDESTRIANS, AND PEDALCYCLISTS KILLED IN CRASHES TO OMI DATA .....	12
8.2 MATCHING INJURED DRIVERS, NON-DRIVER OCCUPANTS, PEDESTREIANS AND PEDALCYCLISTS TO THE RESIDUAL OMI DATA .....	15
8.3 SUMMARY .....	15
<b>9. MATCHING OMI DATA TO THE SYSTEMS TRAUMA REGISTRY DATA</b> .....	<b>16</b>
<b>10. MATCHING INDIVIDUALS INJURED OR KILLED IN CRASHES TO STR DATA</b> .....	<b>17</b>
10.1 ESTABLISHING A DENOMINATOR .....	17
<b>11. MATCHING HIDD TO THE SYSTEM TRAUMA REGISTRY DATA</b> .....	<b>19</b>
11.1 ESTABLISHING A DENOMINATOR .....	19
<b>12. MATCHING HOSPITAL INPATIENT DISCHARGE DATA TO PEOPLE IN CRASHES</b> .....	<b>21</b>
12.1 MATCHING HIDD TO DRIVERS, PEDISTRIANS AND PEDALCYCLISTS IN CRASHES .....	22
12.2 MATCHING HOSPITAL INPATIENT DISCHARGE DATA TO INJURED NON-DRIVER OCCUPANTS .....	24
12.3 AGGREGATING HIDD .....	25
<b>13. MATCHING HIDD RECORDS INDICATING MEDICAID CLAIMS TO THE MEDICAID DATA</b> .....	<b>25</b>
13.1 AGGREGATING MEDICAID CLAIMS .....	26
<b>14. THE 1995 CODES LINKED FILE AND FINAL MATCH RESULTS</b> .....	<b>27</b>
14.1 FINDING MATCHES NOT IDENTIFIED IN DIRECT FILE-TO-FILE MATCH PROCESSES .....	27
14.2 FINAL MATCH RESULTS AND FINAL MATCH RATES FOR THE 1995 CODES LINKED FILE .....	28
<b>15. PROJECT ASSESSMENT</b> .....	<b>30</b>
<i>References</i> .....	31

## **1. INTRODUCTION**

One aim of the CODES (Crash Outcome Data Evaluation System) project is to obtain and link together various New Mexico injury related data bases. The building of a CODES data base will be an important step in understanding and preventing injury in New Mexico and a valuable resource for researchers, policy makers, and law enforcement agencies throughout the state.

An important step in achieving that goal is the preparation and probabilistic linkage of the available data. Five 1995 data sources were obtained for preparation and linkage. These include: the State of New Mexico Crash Files, the New Mexico System Trauma Registry data, data from the New Mexico Office of the Medical Investigator, the Hospital Inpatient Discharge Data file, and the Medicaid claims file.

The Division of Government Research (DGR) processed and linked those data in calendar year 1996 and obtained encouraging results. What follows is a description of the data sources, preparation and matching of the data sources, and an assessment of the matching results. Further documentation, including detailed breakdowns of match rates and a data dictionary are included in appendices A and B, respectively.

## **2. THE 1995 CODES FILES**

### ***2.1 The Crash Files***

The Crash Files contain information from police crash reports submitted to and processed by the Transportation Statistics Bureau and reprocessed and maintained for statistical analysis purposes by DGR for the Traffic Safety Bureau. The information in the files are stored as three separate SAS data sets, which represent three different levels of information: the Accident Level, the Vehicle or Detail Level, and the Occupant Level. Selections from the Vehicle and Occupant Level Files were used in the creation of the CODES data base. The Vehicle Level File contains information on each motor vehicle and its driver, and on each pedestrian or pedalcyclist in most cases. When two or more pedestrians or pedalcyclists are struck together sometimes one individual is on the Vehicle Level File and the others are only on the Occupant Level File, for ease in discourse these few cases will also be referred to as non-driver occupants. Whether or not an individual is on the Vehicle Level File is important because as of 1995 the only information kept on non-driver occupants is age and sex. Last names and first initials are currently being collected for all injured non-driver occupants. That information will be available in 1997 on the 1996 Crash Files.

### ***2.2 The New Mexico System Trauma Registry***

The 1995 System Trauma Registry (STR) data contains trauma patient information submitted from 17 hospitals throughout the state of New Mexico. The participating hospitals in 1995 were: one level one trauma center: University Hospital in Albuquerque; five level two trauma centers: Lovelace Medical Center; Presbyterian Healthcare Systems, St. Joseph Medical Center in Albuquerque, St. Vincent Hospital in Santa Fe, and San Juan Regional Medical Center in

Farmington; one level three trauma center: Eastern New Mexico Medical Center in Roswell; and nine other hospitals: Guadalupe Medical Center - Columbia in Carlsbad, Dr. Dan C. Trigg Memorial Hospital in Tucumcari, Gallup Public Health Services, Lea Regional Hospital - Columbia in Hobbs, Memorial Medical Center in Las Cruces, Rehoboth McKinley Christian Hospital in Gallup, Guadalupe County Hospital in Santa Rosa, Nor-Lea General Hospital in Lovington, Northeastern Regional Hospital in Las Vegas, and Sierra Vista Hospital in Truth or Consequences. If available, field triage, or pre-hospital assessment of medical need, is used to determine patient inclusion in the STR, although sometimes inclusion must be based upon other criteria (Baulch and Rutledge, 1995). For more information see the Annual Trauma Systems Report for Calendar Year 1994, prepared by Baulch and Rutledge (1995) for the State of New Mexico Department of Health. Selected variables from the STR data were extracted for inclusion in the CODES data base.

### ***2.3 The State of New Mexico Office of the Medical Investigator Data***

The Office of the Medical Investigator (OMI) investigates and keeps data on any death occurring in New Mexico that is sudden, violent or untimely; or where a person is found dead and the cause of death is unknown (Office of the Medical Investigator, 1995). Therefore, theoretically all traffic crash fatalities should be found on the OMI data. All cases with manner of death judged to be an accident, homicide, suicide, undetermined, or resulting from unnatural causes were selected for inclusion in the CODES data set.

### ***2.4 New Mexico Hospital Inpatient Discharge Data (HIDD)***

The Hospital Inpatient Discharge Data (HIDD) consists of all hospital inpatient discharges from licensed, non-federal hospitals in New Mexico. Since 1992, the New Mexico Health Policy Commission has overseen the maintenance and data collection for HIDD. Thirty-four general hospitals and 21 specialty hospitals reported to HIDD in 1995, providing clinical, payer, and personal identifier information for each patient discharge. HIDD records with at least one injury diagnosis code (ICD-9 800-999.99 or any E-code) were selected for the CODES matching. Thirty-two hospitals were represented in the CODES data set. The extensive clinical, payer, and personal identifier information in HIDD made it an important link in the 1995 CODES project. The payer fields in HIDD help to bridge the gap between crash injuries and/or System Trauma Registry records and Medicaid claims. Additionally, HIDD diagnosis codes are critical in assessing the long term consequences and cost of treatment related to specific injuries.

### ***2.5 The New Mexico Medicaid Claims File***

The New Mexico Medicaid claims file contains records of all claims paid by Medicaid for services rendered in one calendar year. The claims originate with billing statements submitted by doctors, hospitals and others which, through electronic processing, become part of the Medicaid claims file. The Medicaid fiscal agent, under contract to the State of New Mexico, oversees and maintains the file. The CODES subset of the Medicaid file contained any records pertaining to an individual who had at least on injury during 1995. The CODES set contained almost two million records and 50,199 unique Medicaid ID numbers.

### 3. DATA PREPARATION

Before the five data bases could be matched several data preparation steps had to be accomplished. In order to comply with confidentiality agreements the data bases were divided into a matching file and a clinical file, then unique record identification numbers were assigned, allowing the files to be relinked after the matching process was carried out. All data preparation was carried out using the SAS Statistics System, Version 6.11 (SAS Institute, Inc. Cary, North Carolina). SAS was used to match the Medicaid claims file to HIDD. All other matching was done using *Automatch 3.0* (Matchware Technologies, Inc., Burtonsville, Maryland).

#### 3.1 *The Crash Driver/Occupant File*

For matching purposes the Vehicle Level File and the Occupant Level File were merged to form one occupant based file. The new Crash Driver/Occupant file included injured and non-injured drivers and non-drivers. A driver indicator variable was assigned so that drivers and non-driver occupants could be treated separately when necessary.

In order to obtain missing information concerning the driver personal identifiers: driver zip code, driver name and driver date of birth if missing, the New Mexico State Driver Master File, as of March 1996, was merged with the Crash Vehicle Level File by Social Security Number. The Driver Master File is maintained by the Motor Vehicle Division of the State of New Mexico Taxation and Revenue Department and pulled by the Traffic Safety Bureau and processed and kept by DGR for use in calculating traffic statistics and obtaining information on licensed drivers in New Mexico. Merging with the Driver Master File lowers the percentage of missing driver names and dates of birth.

The following data manipulations were carried out:

- The variable vehicle type (TYPEV) was used to create a new variable VTPE which discriminates between drivers and non-drivers of cars and trucks, pedestrians, pedalcyclists and motorcyclists.
- Errors in the date of birth variable were corrected: February 29 in non-leap years was coded to February 28, Days in September, April, May and November coded as 31 were changed to 30. Date of birth was then reformatted to be readable by Automatch.
- Any times with minutes over 59 were recoded into the next hour and any time over 2359 was recoded to begin at 0000.
- County in which the crash took place was recoded to be consistent with Federal Information Processing Standards (FIPS) county codes.
- Residence zipcodes were used to create county of residence FIPS codes.
- The character variable injury level (INJURY) was used to form a numeric injury level code. The numeric format was needed in order to use the Automatch 'prorated' matching algorithm.

### 3.2 The System Trauma Registry Data

The following data manipulations were carried out:

- The dashes in Social Security Numbers were removed.
- E-codes and place codes were used to determine motor vehicle crashes and whether or not the individual was a driver, pedestrian, pedalcyclist or non-driver occupant. If the E-code was less than 827 or if the place code was equal to street or highway (E849.5) the record was coded as a possible motor vehicle crash. See Table 1 for more information.
- A discharge date variable was calculated using the emergency department date (EDDATE) and number of hospital days (HOSPDAYS).
- If the injury date was coded as being later than the emergency department date, the injury date was set equal to the emergency department date.
- Missing values and dates were formatted to be compatible with Automatch.
- Errors in injury time and emergency department time were corrected. Times over 59 minutes were recoded into the next hour and times over 2359 were recoded starting at 0000.
- The FIPS codes for injury scene and residence of patient were broken up into a county component and a state component.
- Discharge disposition (DCDISPOS) was used to derive a new, less specific discharge disposition variable (DCDISP).

Table 1  
Type of Vehicle and Driver Information Assigned to E-codes in HIDD and STR Data

Type of Vehicle and Driver Information	E-Codes
Car or Truck Driver	E810.0,...., E827.0, E958.5, E988.5
Car or Truck Non-Driver Occupant	E810.1,...., E827.1
Motorcycle Driver	E810.2,...., E827.2
Motorcycle Non-Driver Occupant	E810.3,...., E827.3
Pedestrian	E814.8, E814.9, E810.7,...., E827.7
Pedalcyclist	E810.6,...., E827.6, E826.0, E826.1,....,E826.9

Table 2  
Type of Vehicle and Driver Information Assigned to Manner of Death Codes in the OMI Data

Type of Vehicle and Driver Information	Manner of Death Codes
Car or Truck Driver	A1, A3, A5, A13, A15, A17, A22, A65, A67, A69
Car or Truck Non-Driver Occupant	A2, A6, A14, A18, A23, A66, A70, A71
Motorcycle Driver	A7, A9
Motorcycle Non-Driver Occupant	A8, A10
Pedestrian	A19, A21, S16, U14
Pedalcyclist	A55, A56
Unknown Vehicle Driver	S10, U20

### ***3.3 OMI Data***

The following data manipulations were carried out:

- Manner of death codes were used to determine motor vehicle crashes and whether or not the individual was a driver, pedestrian, pedalcyclist or non-driver occupant. MDC codes: A1-A23, A55, A56, A65-A71, S10, S16, U14 and U20 represent motor vehicle crashes. See Table 2 for more information.
- The SAS scan function was used to divide the name into first name, last name and middle initial fields. "John Does" and "Jane Does" were set to missing.
- The dashes were removed from Social Security Number.
- Month of birth, day of birth and year of birth were extracted from Date of Birth.
- Age was truncated so that all ages were represented by integer values, e.g. an age coded as 2.5 years would equal 2 years.

### ***3.4 HIDD File***

The following data manipulations were carried out:

- Hospital license numbers were made compatible with STR hospital codes.
- County of hospitalization was determined using hospital zipcode (HOSPZIP).
- County of hospitalization was recoded to be consistent with Federal Information Processing Standards (FIPS) county codes.
- Age was truncated so that all ages were represented by integer values, e.g. an age coded as 2.5 years would equal 2 years.
- E-codes and place codes were used to determine motor vehicle crashes and whether or not the individual was a driver, pedestrian, pedalcyclist or non-driver occupant. If the E-code was less than 827 or if the place code was equal to street or highway(E849.5) the record was coded as a possible motor vehicle crash. Only a small percentage of HIDD records contain an E-Code. See Table 1 for more information on the use of E-Codes to determine motor vehicle crashes.
- Length of hospital stay was calculated from admit and discharge dates (DATE\_ADM and DATE\_DIS).
- Discharge disposition (STAT) was used to derive a new, less specific discharge disposition variable(DCDISP).
- Blanks, hyphens and other superfluous characters were removed from the Medicaid ID number (MCAIDID).

### ***3.5 The Medicaid Claims File***

Medicaid ID number (RECIPNO) was used to match the Medicaid file to HIDD; due to the high quality of the field no data manipulation was necessary.

## 4. QUALITY OF MATCHING VARIABLES

Table 3 contains information on the quality of the matching variables in the CODES data sets. If a variable is not available on a particular database it is indicated by an NA. All databases contain some sort of date information that is never missing, date of injury in Crash, emergency department date in STR data, hospital admit date on the HIDD file, date pronounced dead in OMI, and "from date" in Medicaid. The lack of personal identifiers for non-driver occupants is evident. Furthermore, there is not independence in the personal identifiers missing on The Crash file. If social security number is missing then name is always missing and for approximately 9% of drivers and 2% of injured drivers name, social security number, and date of birth are all missing. In the HIDD file, if a record contains a Medicaid ID number it does not contain Name or SSN. The absence of Name and SSN on Medicaid related records is required by the Health Policy Commission’s agreement with Medicaid.

Table 3.

**Matching Variables and Availability Information for 1995 CODES Files: OMI, Systems Trauma Registry, HIDD, Medicaid, and Records Coded as Having some Level of Injury in the Crash Files**

	CRASH		STR	OMI	HIDD	MEDICAID
	drivers	non-drivers				
<b>Location and Time Variables</b>						
Date of injury	100%		100%	91.9%	NA	NA
Time of Injury	100%		62.5%	62.5%	NA	NA
County of Injury	100%		97.2%	97.5%	NA	NA
State of Injury	100%		97.2%	100%	NA	NA
Date Pronounced Dead	NA		NA	100%	NA	NA
Emergency Department Date	NA		100%	NA	NA	NA
Emergency Department Time	NA		100%	NA	NA	NA
Hospital Admit Date (if applicable)	NA		NA	NA	100%	100%
Hospital Discharge Date (if applicable)	NA		NA	NA	100%	100%
Length of Hospital Stay	NA		100%	NA	100%	NA
<b>Personal Identifiers</b>						
Social Security Number	92.3%	NA	79.4%	15.1%	76.5%	NA
Name *	79.1%	NA	NA	99.7%	84.0%	100%
Date of Birth	96.5%	NA	100%	99.7%	100%	78.3%
Age	99.0%	96.4%	100%	100%	99.9%	NA
Sex	99.8%	99.8%	100%	100%	100%	98.7%
Residence County	99.4%	NA	99.1%	99.5%	98.7%	100%
Residence State	90.4%	NA	99.1%	100%	98.7%	100%
Medicaid Record Number(if applicable)	NA	NA	NA	NA	93.3%	100%

\* The Medicaid name fields sometimes contain numbers and special characters. A value was considered non-missing if it contained at least one letter.

## 5. PROBABILISTIC MATCHING

The data sets being used for CODES are large and the data are often collected under non-optimal conditions, thus errors are present. Deterministic matching requires variables to match perfectly and therefore requires perfect or nearly perfect identifiers. Probabilistic matching allows for errors in comparison of variables and is ideal for linking data sets with imperfect identifiers. Therefore, the software *Automatch 3.0* which applies a probabilistic matching algorithm developed by Mathew A. Jaro was used to carry out the record linkage. A brief description of the theory and methodology as proposed by Jaro (1989, 1995a, 1995b) are explained below. However, for more complete information see the referenced papers.

The goal of probabilistic matching is to link data set A to data set B. Set  $A \times B$  is the set of all possible pairs. Set  $A \times B$  needs to be divided into two mutually exclusive and exhaustive subsets: M which equals the set of all matches and U which equals the set of all non-matches. The cardinality of M (i.e.; the number of elements in M) will be much smaller than the cardinality of U, because given there are no duplicates in A and B, any record in A will only match to one record in B and vice-versa. The cardinality of  $A \times B$  can be very large even for data sets A and B of moderate size. For example, if there are 100 observations in both A and B then  $A \times B$  will have a cardinality of 10,000 (Jaro, 1989).

In order to cut down on the number of pairs in  $A \times B$  to be examined the set needs to be blocked. The smaller the blocks the fewer comparisons there are to be made. The best initial blocking variable is a unique identification number such as social security number or first and last name to keep the blocks as small as possible. To make up for errors in the blocking variables multiple passes are run. In later passes less stringent blocking criteria can be applied since many pairs will be classified as a match or a non-match in the first pass.

In order to decide whether or not a pair within a block is a match, a weight is calculated for each of the matching variables within a pair. In order to calculate the weights two probabilities need to be estimated. The **u**-probability equals the probability two records match on a certain variable accidentally. It can be estimated from the frequency distribution of the matching variables. The **m**-probability equals the probability the variable agrees on the two records when the two records are truly a match. It is functionally 1 minus the error rate in the field and can be assigned by the user so that low quality fields have low **m**-probabilities and important high quality fields have high **m**-probabilities.

Different comparison types can be implemented for different matching variables within a pair. For example the number of characters disagreeing can be counted or certain uncertainty comparisons can be used. For an exact explanation of the different type of comparison types available see the Automatch documentation (Jaro, 1995, p.34-41).

The weights are summed to form a composite weight for each pair and the "best matches" within

each block are selected. If a record doesn't match to any record in the other database but has a matching weight above the cutoff rate it is selected as a possible duplicate record. The user defines a cutoff weight for defining matched pairs and for defining clerical review cases. Specific variables can also be assigned as Clerical or Critical variables. If a variable is assigned to be Clerical and that variable does not agree within a user assigned margin of error the pair becomes a Clerical review case regardless of weight. If a variable is assigned to be Critical and that variable does not agree within a user assigned margin of error then that pair becomes a non-match regardless of weight. The cutoff weights are determined by examining a report of potential matches and frequency histograms of the weights.

## **6. UNDUPLICATING THE SYSTEM TRAUMA REGISTRY**

The 1995 System Trauma Registry contained three kinds of duplicates: transfer duplicates, repeat-visit duplicates, and same-visit duplicates. Transfer duplicates are expected in the Systems Trauma Registry data since it contains more than one record for people who are transferred from one STR participant to another. The method for identifying the transfer duplicates is explained in Table 4.

Repeat-visit duplicates result from a person visiting the hospital more than once. If the hospital visits are related to the same incident the visits should be, and usually are, consolidated into one STR record. However, a small number of repeat-visit duplicates have been present in both the 1994 and 1995 data (4 cases in 1994 and 1 case in 1995).

In addition to the expected duplicates, the 1995 data contained same-visit duplicates as a result of a file extraction error. The 1995 STR file was pulled more than once and at different times during the year. Therefore, it contained original and updated records that represented the same event. Because one record was intended to replace the other, the records were not identical. Same-visit duplicate pairs with nearly identical location/date/time information but different personal identifiers were tracked throughout the matching process. When one of the records matched to another CODES file via a personal identifier, its duplicate record was removed from the match file. In this way, the record with the correct identifiers was matched and the correct identifiers were carried through to further Automatch matchings. In all, 22 records were determined to be same-visit duplicates and were removed from the 1995 STR match file. The same-visit duplication was hard to detect and caused duplicates to surface during all phases of CODES matching. Future exact duplicate screenings will be more extensive. If same-visit duplicates are discovered the System Trauma Registry extraction will be repeated to eliminate the duplicates.

Table 4  
Blocking and Matching Strategies and Results for Unduplicating STR Data

Pass	Blocks	Matching Variables and Comparison Type	Results
1	Social Security Number	Age percentage difference 5% or less, Date of Injury within 1 day, Time of Injury within 120 minutes, Date of Birth within 7 days, Sex, Residence County, Residence State and County of Injury	107 duplicates found
2	Date of Birth	Difference in Social Security Number of 2 characters or less, Date of Injury within 1 day, Time of Injury within 120 minutes, Sex, Residence County, Residence State and County of Injury	77 duplicates found
3	Injury Date and Age	Date of Birth within 7 days, Difference in Social Security Number of 2 characters or less, Time of Injury within 120 minutes, Sex, Residence County, Residence State and County of Injury	5 duplicates found

Any pair of records with Social Security Numbers different in more than 2 characters or injury date more than one day apart were selected for Clerical review.

**Final Automatch Results:** 189 duplicate records were found.

Upon further examination four of the 189 duplicate records were thought to be false positives and not treated as a duplicates. One of the remaining records was an exact duplicate and could simply be deleted. Thirteen of the duplicates were found to be same-visit duplicates. The other 171 observations were cases where an individual was transferred to another facility or, in 1 case, where an individual left the hospital and returned later due to complications arising from the same injury.

In order to unduplicate the STR data the duplicate records had to be combined to make one record. Most of the pre-hospital information was taken from the first hospital and most of the discharge information was taken from the second hospital. If one hospital’s record contained a missing value for a given variable the information was obtained from the other hospital’s record. Some information from both hospitals was kept on the final analysis file. Those variables are included in the CODES Data Dictionary (see Appendix B). The number of hospital days and number of days spent in the ICU were obtained by summing across the two hospitals.

## 7. GROUPING THE HOSPITAL INPATIENT DISCHARGE DATA

The Hospital Inpatient Discharge Data is a hospital discharge based file; if an individual was admitted and discharged from more than one HIDD participant during 1995, the data contained multiple records for that individual. Automatch unduplication was used to assign a unique group identification number to records that refer to the same person. Grouping is like unduplicating except that the duplicates are not aggregated or removed from the output data set, but are flagged according to the Automatch specifications. After HIDD was matched to STR and Crash Driver/Occupant, the group identification number was used to aggregate the HIDD records to the appropriate level (see sections 7.1 and 12.3). The grouping strategy and results are summarized in Table 5.

Table 5

Grouping Strategy and Results for Grouping Hospital Inpatient Discharge Data

<b>Pass</b>	<b>Blocks</b>	<b>Matching Variables and Comparison Type</b>	<b>Results</b>
1	First Name and Last Name	Age percentage difference 5% or less, Social Security Number not almost certainly different, Date of Birth not almost certainly different, Sex, Ethnicity, Residence Zipcode and Hospital License Number (with no penalty for disagreement).	1,752 duplicates found
2	Social Security Number	Age percentage difference 5% or less, First and Last Name not almost certainly different, Date of Birth not almost certainly different, Sex, Ethnicity, Residence Zipcode and Hospital License Number (with no penalty for disagreement).	561 duplicates found.
3	Date of Birth	Age percentage difference 5% or less, Social Security Number not almost certainly different, First and Last Name not almost certainly different, Sex, Ethnicity, Residence Zipcode and Hospital License Number (with no penalty for disagreement).	68 duplicates found.

Table 5

Grouping Strategy and Results for Grouping Hospital Inpatient Discharge Data

4	Age and Sex	Social Security Number not almost certainly different, First Name probably the same, Last Name not almost certainly different, Ethnicity, Residence Zipcode and Hospital License Number (with no penalty for disagreement).	4 duplicates found.
---	-------------	---	---------------------

---

**Final Automatch Results:** 2,385 HIDD records were identified as corresponding to an individual who underwent injury related hospitalization more than once in 1995. 18,201 unique people are represented in the HIDD.

**7.1 Delaying Aggregation of HIDD**

The HIDD was grouped, but not aggregated, prior to matching. Since one individual may have generated multiple injury-related discharge records resulting from multiple *independent* injuries, aggregating the data to patient level prior to matching would have caused incorrect matches. Referring to Table 6, below, only record numbers 3-5 are crash related treatments. Aggregating the data to patient level (i.e. using record 1 as the primary record) would cause the crash related injury date (InjDate 2) to be lost as a matching variable. Using a date-based matching algorithm would assure a non-match since the only HIDD date available (HospDate 1) would occur prior to any crash date or crash-related STR entry date. Using a personal identifier-based matching algorithm would assure an overestimation of treatment time and, if matched to payer information, an overestimation of treatment costs. After the matching was completed and a crash and/or Trauma Registry date had been established, the HIDD was aggregated accordingly (see section 12.3).

Table 6

Hypothetical HIDD Records

Record Number	Type of Injury	Date of Injury	Hospital Admit Date	Admitted to Hospital (HIDD record generated upon discharge)
1	Non-Crash Injury	InjDate 1	HospDate 1	Treatment for Non-Crash Injury
2			HospDate 2	Continued Treatment for Non-Crash Injury
3	Crash Injury	InjDate 2	HospDate 3	Treatment for Crash Injury
4			HospDate 4	Continued Treatment for Crash Injury
5			HospDate 5	Rehabilitation for Crash Injury

## **8. MATCHING PEOPLE KILLED OR INJURED IN CRASHES TO THE OMI DATA**

The OMI/Crash matching was a two phase process. The first step was to match killed drivers, non-driver occupants, pedestrians and pedalcyclists to the OMI data. The second step was to match drivers, non-driver occupants, pedestrians, and pedalcyclists with incapacitating injuries (INJURY='A') to the OMI data. An injury is coded as 'K' in the Crash Driver/Occupant File if the individual died within 30 days of the crash date. The purpose of matching OMI against injured (not killed) individuals was to identify crash related deaths that occurred more than 30 days after the crash. A subset of the OMI data was used which included only those individuals indicated as being killed in a motor vehicle crash by the manner of death codes.

### ***8.1 Matching Drivers, Non-Driver Occupants, Pedestrians and Pedalcyclists Killed in Crashes to OMI Data***

Seven matching passes were performed to match the OMI data to killed drivers, non-driver occupants, pedestrians and pedalcyclists. The first three passes blocked on the indicator variable Driver. This subsets the OMI data so at first only those individuals in the OMI data expected to match to crash occupants coded as drivers, pedestrians and pedalcyclists (see Table 2 for more information), were included. The last four passes dropped this restriction to allow for a match if an individual was falsely coded as a non-driver.

Although there are fewer personal identifiers for non-driver occupants in the Crash Files than there are for drivers, pedestrians and pedalcyclists it was still possible to probabilisticly match to the OMI data since the probability of two non-driver occupants of the same age and sex being killed on the same day is low.

The data sets matched were: 483 individuals in OMI data with manner of death codes indicating they were killed in a motor vehicle crash on a NM street or highway in 1995; and 485 drivers, non-driver occupants, pedestrians and pedalcyclists in the Crash file who died within 30 days of a traffic crash as a result of that crash. There were fewer OMI records indicating highway/street deaths than there were Crash records with injury='K'. This discrepancy may have resulted from natural deaths that occurred in conjunction with a crash ( e.g. a heart attack while driving); such cases would be in the Crash File but not in the OMI CODES match set. See Table 7 for information on blocking and matching strategies and results.

Table 7

Blocking and Matching Strategies and Results for Matching OMI Data to Drivers, Pedestrians and Pedalcyclists Killed In Crashes

<b>Pass</b>	<b>Blocks</b>	<b>Matching Variables and Comparison Type</b>	<b>Results</b>
1	Driver Indicator Variable (i.e.; select only those in OMI indicated to be drivers, non-driver occupants pedestrians and pedalcyclists by Manner of Death Codes), Last Name and First Name	Age percentage difference 5% or less, Date Pronounced Dead within 30 days after accident date, Date of Injury within 2 days, Date of Birth within 7 days, Time of Injury within 120 minutes, Sex, Residence State, Residence County, County of Injury, and Type of Vehicle.	112 matched out of 485 killed drivers, non-driver occupants pedestrians and pedalcyclists ( 23%).  Leaving 373 unmatched killed drivers,non-driver occupants, pedestrians and pedalcyclists.
2	Driver Indicator Variable, Age and Sex	Last Name not almost certainly different, First Name not almost certainly different, Date Pronounced Dead within 30 days after accident date, Date of Injury within 2 days, Date of Birth within 7 days, Time of Injury within 120 minutes, Residence State, Residence County, County of Injury, Type of Vehicle and Residence Zipcode.	259 matched out of 373 killed drivers, pedestrians and pedalcyclists (69%).  Leaving 114 unmatched killed drivers, pedestrians and pedalcyclists.
3	Driver Indicator Variable, County of Injury and Month of Injury	Age percentage difference 5% or less, Last Name not almost certainly different, First Name not almost certainly different, Date Pronounced Dead within 30 days after Date of Injury, Date of Injury within 2 days, Date of Birth within 7 days, Time of Injury within 120 minutes, Sex, Residence State, Residence County, Type of Vehicle and Residence Zipcode.	66 matched out of 114 killed drivers, pedestrians and pedalcyclists (58%).  Leaving 48 unmatched killed drivers, pedestrians and pedalcyclists.
4	Last Name and First Name	Age percentage difference 5% or less, Date Pronounced Dead within 30 days after accident date, Date of Injury within 2 days, Date of Birth within 7 days, Time of Injury within 120 minutes, Sex, Residence State, Residence County, County of Injury, and Type of Vehicle.	9 matched out of 48 killed drivers, non-driver occupants pedestrians and pedalcyclists ( 19%).  Leaving 39 unmatched killed drivers, pedestrians and pedalcyclists.

Table 7 (cont.)

Blocking and Matching Strategies and Results for Matching OMI Data to Drivers, Pedestrians and Pedalcyclists Killed In Crashes

5	Age and Sex	Last Name not almost certainly different, First Name not almost certainly different, Date Pronounced Dead within 30 days after accident date, Date of Injury within 2 days, Date of Birth within 7 days, Time of Injury within 120 minutes, Residence State, Residence County, County of Injury, Type of Vehicle and Residence Zipcode.	9 matched out of 39 killed drivers, pedestrians and pedalcyclists (23%).  Leaving 30 unmatched killed drivers, pedestrians and pedalcyclists.
6	County of Injury and Month of Injury	Age percentage difference 5% or less, Last Name not almost certainly different, First Name not almost certainly different, Date Pronounced Dead within 30 days of Date of Injury, Date of Injury within 2 days, Date of Birth within 7 days, Time of Injury within 120 minutes, Sex, Residence State, Residence County, County of Injury, Type of Vehicle and Residence Zipcode.	2 matched out of 30 killed drivers, pedestrians and pedalcyclists (6%).  Leaving 28 unmatched killed drivers, pedestrians and pedalcyclists.
7	Street/Highway Indicator Variable (functionally, this served to escape the need for exact matches on identifiers)	Age percentage difference 5% or less, Last Name not almost certainly different, Date Pronounced Dead within 30 days of Date of Injury, Date of Injury within 2 days, Time of Injury within 120 minutes, Sex, County of Injury, Type of Vehicle and Residence Zipcode.	6 matched out of 28 killed drivers, non-driver occupants, pedestrians and pedalcyclists (21%).  Leaving 22 unmatched killed drivers, non-driver occupants, pedestrians and pedalcyclists.

---

**Final Automatch Results:** 463 out of 485 killed drivers, non-driver occupants, pedestrians and pedalcyclists (95.5%) were matched to OMI data.

### 8.2 Matching Injured Drivers, Non-Driver Occupants, Pedestrians, and Pedalcyclists to the Residual OMI Data

Matching motor vehicle and street/highway deaths in OMI with killed drivers, non-driver occupants, pedestrians and pedalcyclists left 20 unmatched OMI records. These residual records were matched against drivers, non-driver occupants, pedestrians and pedalcyclists coded as having incapacitating injuries (INJURY='A') in the Crash Driver/Occupant File.

The data sets matched were: 20 motor vehicle or street/highway death records in OMI that did not match to killed drivers, non-driver occupants, pedestrians and pedalcyclists in the first OMI/Crash match; and 5,649 drivers, non-driver occupants, pedestrians, and pedalcyclists with incapacitating injuries resulting from a traffic crash. See Table 8 for information on blocking and matching strategies and results.

Table 8  
Blocking and Matching Strategies and Results for Matching OMI Data to Drivers, Non-Driver Occupants, Pedestrians and Pedalcyclists Injured In Crashes

Pass	Blocks	Matching Variables and Comparison Type	Results
1	First Name and Last Name	Age within 5%, Date Pronounced Dead within 90 days after accident date, Date of Birth within 7 days, Date of Injury within 2 days, Time of Injury within 120 minutes, County of Injury ,Type of Vehicle, Sex, and County and State of Residence.	4 out of 20 motor vehicle or street/highway deaths in OMI (8%).
2	Age and Sex	Name not almost certainly different, Residence State, Residence County, Residence Zip, Date Pronounced Dead within 90 days after Date of Injury, Date of Injury within 2 days, Date of Birth within 7 days, Time of Injury within 120 minutes, Sex, County of Injury and Type of Vehicle.	2 out of 16 motor vehicle or street/highway deaths in OMI (4%).

**Final Automatch Results:** 6 out of 20 residual motor vehicle or street/highway deaths in OMI (30%) were matched to drivers, non-driver occupants, pedestrians and pedalcyclists coded as injured, but not killed, in the Crash File.

### 8.3 Summary

In conclusion, 95.5% of the individuals killed in traffic crashes were matched to the OMI data. Ninety-seven percent (97.1%) of the individuals in OMI with manner of death codes indicating a motor vehicle or street/highway death matched to killed or injured drivers, non-driver occupants, pedestrians, and pedalcyclists. Of the unmatched OMI records; one death resulted from a crash that occurred in 1963, another was that of a baby in the womb.

## 9. MATCHING OMI DATA TO THE SYSTEM TRAUMA REGISTRY DATA

The data sets matched were: all 1,498 records in OMI data; and 243 records with discharge dispositions of deceased in STR. The blocking and matching strategies and results are in Table 9.

Table 9  
Blocking and Matching Strategies and Results for Matching Deceased Individuals in STR Data to OMI Data

Pass	Blocks	Matching Variables and Comparison Type	Results
1	Place Pronounced Indicator (individual coded as pronounced dead at the hospital in OMI) and Date of Birth	Date Pronounced Dead within day after Injury date, Date of Injury within 2 days, Time of Injury within 120 minutes, Sex, Residence County and County of Injury.	215 out of 243 STR records with discharge dispositions of deceased matched (88%).  Leaving 28 unmatched STR records with discharge dispositions of deceased.
2	Place Pronounced Indicator (individual coded as pronounced dead at the hospital in OMI) and Date Pronounced Dead	Age percentage difference 5% or less, Date of Birth within 7 days, Date of Injury within 2 days, Time of Injury within 120 minutes, Sex, Residence County and County of Injury.	13 out of 28 STR records with discharge dispositions of deceased matched (46%).  Leaving 15 unmatched STR records with discharge dispositions of deceased.
3	Place Pronounced Indicator (individual coded as pronounced dead at the hospital in OMI) and County of Injury	Age percentage difference 5% or less, Date of Birth within 7 days, Date Pronounced Dead within 1 day after Injury date, Date of Injury within 2 days, Time of Injury within 120 minutes, Sex and Residence County.	1 out of 15 deceased individuals on STR matched (7%).  Leaving 14 unmatched STR records with discharge dispositions of deceased.

Table 9 (cont.)

Blocking and Matching Strategies and Results for Matching Deceased Individuals in STR Data to OMI Data

4	Place Pronounced Indicator (individual coded as pronounced dead at the hospital in OMI) and Sex	Age percentage difference 5% or less, Date of Birth within 7 days, Date Pronounced Dead within 1 day after Injury date, Date of Injury within 2 days, Time of Injury within 120 minutes, Residence County and County of Injury.	2 out of 14 deceased individuals on STR matched (14%). Leaving 12 unmatched STR records with discharge dispositions of deceased.
5	Date Pronounced Dead	Age percentage difference 5% or less and Sex	1 out of 12 deceased individuals on STR matched (8%) Leaving 11 unmatched STR records with discharge dispositions of deceased.

---

**Final Automatch Results:** 232 out of 243 deceased individuals on STR (95.5%) matched to OMI.

## 10. MATCHING INDIVIDUALS INJURED OR KILLED IN CRASHES TO THE SYSTEM TRAUMA REGISTRY DATA

The matching strategy for linking the Crash Driver/Occupant file and STR first blocked on whether or not the individual was coded as a driver and then eliminated this restriction. Injury level was used as a matching variable by coding all cases in STR with discharge dispositions equaling deceased as Killed and all others as having incapacitating injuries. Functionally, this caused the individual coded in the crash files as having the greatest level injury to be selected as a match to the STR data, if all other identifiers agreed equally.

### 10.1 *Establishing a Denominator*

Although those individuals in crashes with the highest level of coded injury are most likely to match to the STR data, no definitive subset of individuals in the Crash files are expected to match to the STR data. In order to get approximate match rates a suitable denominator had to be derived from the STR data. If a record on the STR data had an E-code indicating the patient was injured in a motor vehicle crash that person was included in the set used for matching (see Table 1 for a breakdown of E-codes into categories). If a record indicated the individual was involved in a motor vehicle crash on a street or highway in New Mexico they were counted as an expected match. The reported match rates represent the proportion that matched out of those expected to match, although 29 individuals coded as possible but not expected to match, matched to crash.

The data sets matched were: 1,673 records on STR data indicated as being in a motor vehicle crash; and 31,481 individuals in the Crash Driver/Occupant file with some injury reported by the police. See table 10 for information on blocking and matching strategies and results.

Table 10

Blocking and Matching Strategies and Results for Matching STR Data to Injured Individuals in Traffic Crashes

<b>Pass</b>	<b>Blocks</b>	<b>Matching Variables and Comparison Type</b>	<b>Results</b>
1	Driver Indicator Variable, Social Security Number	Age percentage difference 5% of less, Emergency Department Date within 7 days after accident date, Date of Injury within 2 days, Date of Birth not almost certainly different, Time of Injury within 120 minutes, Injury Level, Sex, Residence State, Residence County, County of Injury, Type of Vehicle.	472 STR records matched to injured crash occupants.
2	Driver Indicator Variable, County of Injury and Age	Social Security Number within 2 characters, Emergency Department Date within 7 days after accident date, Date of Injury within 2 days, Date of Birth not almost certainly different, Time of Injury within 120 minutes, Injury Level, Sex, Residence State, Residence County, Type of Vehicle.	462 STR records matched to injured crash occupants.
3	Driver Indicator Variable and Injury Date	Social Security Number within 2 characters, Emergency Department Date within 7 days after accident date, Date of Birth is not almost certainly different, Time of Injury within 120 minutes, Injury Level, Sex, Residence State, Residence County, County of Injury, Type of Vehicle.	109 STR records matched to injured crash occupants.
4	Social Security Number	Age percentage difference 5% of less, Emergency Department Date within 7 days after accident date, Date of Injury within 2 days, Date of Birth not almost certainly different, Time of Injury within 120 minutes, Injury Level, Sex, Residence State, Residence County, County of Injury, Type of Vehicle.	38 STR records matched to injured crash occupants.

Table 10 (cont.)

Blocking and Matching Strategies and Results for Matching STR Data to Injured Individuals in Traffic Crashes

5	County of Injury, Age and Sex	Social Security Number within 2 characters, Emergency Department Date within 7 days after accident date, Date of Injury within 2 days, Date of Birth not almost certainly different, Time of Injury within 120 minutes, Injury Level, Residence State, Residence County, Type of Vehicle	41 STR records matched to injured crash occupants.
6	Injury Date	Social Security Number within 2 characters, Age percentage difference 5% of less, Emergency Department Date within 7 days after accident date, Date of Injury within 2 days, Date of Birth not almost certainly different, Time of Injury within 120 minutes, Injury Level, Sex, Residence State, Residence County, County of Injury, Type of Vehicle	52 STR records matched to injured crash occupants.

---

**Final Automatch Results:** 1,174 STR (70.2%) records matched to injured drivers, non-driver occupants, pedestrians and pedalcyclists. Of the 1,550 expected matches, 1145 matched (73.9%).

## 11. MATCHING HIDD TO THE SYSTEM TRAUMA REGISTRY DATA

### 11.1 Establishing a Denominator

Initially, it was thought that a clear denominator could be established for the HIDD to STR match by excluding STR records with a hospital stay less than one day. This cut proved too restrictive in Automatch test passes. In response, a method similar to that used in the STR to Crash match was employed. The STR data was divided into expected and possible matches. An expected match was a record representing an individual admitted to or transferred to a HIDD participant with a hospital stay greater than one day. A possible match was any record representing an individual admitted to or transferred to a HIDD participant.

The data matched were: all 20,583 unique HIDD records; and 3,467 STR records representing a person admitted or transferred to a HIDD participant. Since the HIDD was not aggregated prior to

matching, it was important to find the correct *incident* match, as opposed to the same *patient* match (see section 7.1 for details). Therefore, the matching strategy used was predominantly date based. However, if a solid incident match was not found, personal identifiers were used to find patient matches with reasonable date discrepancies. STR was matched to Crash before being matched to HIDD, thus Name could be used as a matching variable in the HIDD /STR match. For blocking and matching strategies and results, please see Table 11.

Table 11.

Matching Strategy and Results for Matching Hospital Inpatient Discharge Data to the Systems Trauma Registry Data

<b>Pass</b>	<b>Blocks</b>	<b>Matching Variables and Comparison Type</b>	<b>Results</b>
1	Hospital and Emergency Department/Admit Date	Age percentage difference 5% or less, Date of Birth not almost certainly different, Social Security Number not almost certainly different, First and Last Names not almost certainly different, Discharge Dates within 30 days, Ethnicity, Sex and Discharge Disposition.	2,565 out of 3,476 records in STR matched to HIDD (74%). Leaving 911 unmatched STR records.
2	Hospital and Day After Emergency Department Date/Admit Date	Age percentage difference 5% or less, Date of Birth not almost certainly different, Social Security Number not almost certainly different, First and Last Names not almost certainly different, Discharge Dates within 30 days, Ethnicity, Sex and Discharge Disposition.	281 out of 911 records in STR matched to HIDD (31%). Leaving 630 unmatched STR records.
3	Hospital/'Transferred to' Hospital and Emergency Department/Admit Date	Age percentage difference 5% or less, Date of Birth not almost certainly different, Social Security Number not almost certainly different, First and Last Names not almost certainly different, Ethnicity and Sex.	28 out of 630 records in STR matched to HIDD (4%). Leaving 602 unmatched STR records.
4	Social Security Number	Hospital, Date of Birth not almost certainly different, First and Last Names not almost certainly different, Hospital Dates within 7 days, Length of Hospital Stay, Ethnicity, Sex and Discharge Disposition.	52 out of 602 records in STR matched to HIDD (9%). Leaving 550 unmatched STR records.

Table 11 (cont.)

Matching Strategy and Results for Matching Hospital Inpatient Discharge Data to the Systems Trauma Registry Data

5	Hospital and Date of Birth	Age percentage difference 5% or less, Social Security Number not almost certainly different, First and Last Names not almost certainly different, Hospital Dates within 35 days, Length of Hospital Stay, Ethnicity, Sex and Discharge Disposition.	12 out of 550 records in STR matched to HIDD (2%). Leaving 538 unmatched STR records.
---	----------------------------	---	--

---

**Final Automatch Results:** 2,938 out of 3,476 STR patients admitted to or transferred to a HIDD participant matched to HIDD records (85%).

## 12. MATCHING HOSPITAL INPATIENT DISCHARGE DATA TO PEOPLE IN CRASHES

The HIDD to Crash match posed unique challenges in terms of Automatch programming. Generally, "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 equivalent fields for any record on file B, and declare with reasonable certainty that the record pair examined is a match or a non-match" (Jaro, 1995, p. 11). Due to the lack of personal identifiers in the Crash file and the lack of motor vehicle information in the HIDD file, few Crash to HIDD matches were obviously solid. Therefore, the following guidelines were developed to increase the likelihood of correct matches:

1. If Name and/or Social Security Number were not used for blocking, injury level was used as a blocking variable so that only injured occupants from the crash file were candidates for matching in that pass.
2. If no personal identifiers were used for blocking, only seriously injured (killed, incapacitated, or visibly injured) occupants were candidates for matching in that pass.
3. If no personal identifiers or motor vehicle codes were available for blocking or matching, as was the case with most non-driver occupants, only incapacitated occupants were candidates for matching in that pass.

Additionally, a Quality of Match indicator was assigned to each match between HIDD and Crash. The scale assigned a value of 1 to the strongest matches, 2 to good matches, and 3 to weak matches. The quality of match assignments for each pass are noted in Tables 12 and 13, below.

**12.1 Matching Hospital Inpatient Discharge Data to Drivers, Pedestrians, and Pedalcyclists in Crashes**

The HIDD to Crash Driver\Occupant match was done in two steps. Step 1 aimed to match HIDD records to injured and uninjured drivers, pedestrians, and pedalcyclists. 13,376 HIDD records indicating injury were matched against 92,888 injured and uninjured drivers, pedestrians, and pedalcyclists (driver, pedestrian, and pedalcyclist residual from all previous matches). The blocking and matching strategy and results are summarized in Table 12. HIDD records not matched in step 1 were passed on to step 2.

Table 12.  
Blocking and Matching Strategy and Results for Matching Injured and Uninjured Drivers, Pedestrians, and Pedalcyclists to HIDD Records

Pass	Blocks	Matching Variables and Comparison Type	Results
1	First letter of First Name and first five letters of Last Name	First Name and Last Name not almost certainly different, Social Security number not almost certainly different, Date of Birth not almost certainly different, Date Admitted within 150* days of crash date, Age percentage difference 5% or less, Injury Level, Sex, County, Type of Vehicle, Driver indicator and Motor Vehicle Code indicator (with no penalty for disagreement).	347 HIDD records matched to drivers, pedestrians and pedalcyclists.  Quality of Match = 1
2	Social Security Number	First Name and Last Name not almost certainly different, Date of Birth not almost certainly different, Date Admitted within 150* days of crash date, Age percentage difference 5% or less, Injury Level, Sex, County, Type of Vehicle, Driver indicator and Motor Vehicle Code indicator (with no penalty for disagreement).	81 HIDD records matched to drivers, pedestrians and pedalcyclists.  Quality of Match = 1

Table 12 (cont.)

Blocking and Matching Strategy and Results for Matching Injured and Uninjured Drivers, Pedestrians, and Pedalcyclists to HIDD Records

3	Date of Birth and Injury Indicator (so that only injured drivers, pedestrians and pedalcyclists were included in this pass)	First Name and Last Name not almost certainly different, Social Security Number not almost certainly different, Date of Birth not almost certainly different, Date Admitted within 30* days of crash date, Age percentage difference 5% or less, Injury Level, Sex, County, Type of Vehicle, Driver indicator and Motor Vehicle Code indicator (with no penalty for disagreement).	66 HIDD records matched to drivers, pedestrians and pedalcyclists.  Quality of Match = 1
4	Crash Date/Date Admitted to Hospital, Age, Sex, and Seriously Injured Indicator (so that only seriously injured drivers, pedestrians and pedalcyclists were included in this pass)	Social Security Number not almost certainly different, Date of Birth not almost certainly different, Injury Level, County, Type of Vehicle, Driver indicator and Motor Vehicle Code indicator (with no penalty for disagreement).	9 HIDD records matched to drivers, pedestrians and pedalcyclists.  Quality of Match = 2
5	County, Age, Sex, and Seriously Injured Indicator (so that only seriously injured drivers, pedestrians and pedalcyclists were included in this pass)	Social Security Number not almost certainly different, Date of Birth not almost certainly different, Date Admitted within 7* days of crash date, Injury Level, Type of Vehicle, Driver indicator and Motor Vehicle Code indicator (with no penalty for disagreement).	2 HIDD records matched to drivers, pedestrians and pedalcyclists.  Quality of Match = 2
6	Crash Date/Date Admitted to Hospital, County, and Seriously Injured Indicator (so that only seriously injured drivers, pedestrians and pedalcyclists were included in this pass)	Social Security Number not almost certainly different, Date of Birth not almost certainly different, Age percentage difference 5% or less, Injury Level, Sex, Type of Vehicle, Driver indicator and Motor Vehicle Code indicator (with no penalty for disagreement).	1 HIDD record matched to drivers, pedestrians and pedalcyclists.  Quality of Match = 2

---

\*Note: Because the Crash Date/Date Admitted comparison was a Critical variable, the number of days required for a match is approximately one third that of the number stated.

---

**Final Automatch Results:** 506 HIDD records matched to drivers, pedestrians, and pedalcyclists in the Crash Driver\Occupant file.

**12.2 Matching Hospital Inpatient Discharge Data to Injured Non-Driver Occupants**

Step 2 of the HIDD/Crash matching process aimed to match HIDD records to injured non-driver occupants. 12,870 HIDD records (the residual from step 1) were matched against 4,009 seriously injured non-driver occupants (residual seriously injured non-drivers from all previous matches).

Uninjured non-driver occupants were excluded from the step 2 match set for the following reasons: the only matching variables available were county, age and sex; ninety-five percent of the HIDD records contained no information on whether or not the hospital stay was crash related; and the distribution of age, sex, and county on the two match files were not necessarily related due to trans-county emergency transports and non-crash-related injury reports in HIDD. By excluding uninjured non-driver occupants from the match set the probability of false positives was reduced. For information on blocking and matching strategies and results, see table 13.

Table 13.  
Blocking and Matching Strategy and Results for Matching Injured Non-Driver Occupants to HIDD Records

Pass	Blocks	Matching Variables and Comparison Type	Results
1	Motor Vehicle Code Indicator, Seriously Injured Indicator, Crash Date/Date Admitted to Hospital and County	Age percentage difference 5% or less, Sex, Injury Level, Type of Vehicle, Driver Indicator	38 HIDD records matched to non-driver occupants.  Quality of Match = 2
2	Motor Vehicle Code Indicator, Seriously Injured Indicator, Age, Sex and County	Date Admitted within 1 day of crash date, Injury Level, Type of Vehicle, Driver Indicator	25 HIDD records matched to non-driver occupants.  Quality of Match = 2
3	Outside Bernalillo County Indicator, Class A Injury Indicator, County, Age and Sex	Date Admitted within 1 day of crash date, Injury Level, Type of Vehicle, Driver Indicator and Motor Vehicle Code indicator (with no penalty for disagreement).	35 HIDD records matched to non-driver occupants.  Quality of Match = 3

Table 13 (cont.)

Blocking and Matching Strategy and Results for Matching Injured Non-Driver Occupants to HIDD Records

4	Class A Injury Indicator, County, Age and Sex	First Name and Last Name not almost certainly different, Social Security number not almost certainly different, Date of Birth not almost certainly different, Date Admitted within 50 days of crash date, Age percentage difference 5% or less, Injury Level, Sex, County, Type of Vehicle, Driver indicator and Motor Vehicle Code indicator (with no penalty for disagreement).	38 HIDD records matched to non-driver occupants.  Quality of Match = 3
---	---	---	--

**Final Automatch Results:** 136 HIDD records matched to non-drivers in the Crash Driver\Occupant file.

**12.3 Aggregating HIDD**

Once HIDD was matched to Crash and STR, the data was aggregated to an incident level. All hospital visits no more than 45 days apart following a crash or STR entry were attributed to the injury incident. Clinical data was taken from the first HIDD record and the number of days spent in General and Specialty hospitals were summed across the aggregated records. Four people in HIDD matched to two crashes each. These cases were reviewed, determined to be separate incidents, and aggregated to incident level.

**13. MATCHING HIDD RECORDS INDICATING MEDICAID CLAIMS TO THE MEDICAID DATA**

Due to time constraints, the 1995 Medicaid link was established using a straight SAS merge. The data matched were: 2,733 HIDD records with a unique Medicaid identification numbers; and 1,543,615 Medicaid records with a claim payment more than two dollars (payment>2).

Table 14

Blocking and Matching Strategy and Results for matching HIDD records with Unique Medicaid IDs to Medicaid Claim Records

Pass	Merged By	Results
1	Medicaid ID	1,357 HIDD records matched to Medicaid claims.  Leaving 1,376 unmatched records with Medicaid id numbers in the HIDD file.

**Final SAS Results:** 1,357 HIDD records with unique Medicaid ID numbers matched to Medicaid claims (50%).

**13.1 Aggregating Medicaid Claims**

1,357 HIDD records matched to 2,546 Medicaid claims. To preserve a consistent aggregation level in the linked file, the Medicaid had to be aggregated to the level of the already aggregated HIDD. As a result, only Medicaid claims that matched to a *linked* HIDD record were included in the final analysis file -- if a HIDD record was not matched to Crash or STR there was no available benchmark for aggregation. 380 aggregated HIDD records matched to Medicaid. Using HIDD aggregation markers, the Medicaid claims fields were combined so that payment and diagnosis information corresponded to one aggregated HIDD record. The numbers of inpatient, outpatient, physician, transportation, prescription and other claims were summed across the aggregated Medicaid claims. The total Medicaid inpatient claims payment was also calculated. Six Medicaid diagnosis codes were included in the aggregated record; beginning with the earliest claim, the first six unique diagnosis codes were assigned to variables in the aggregated record. Referring to Table 15, below, the aggregated Medicaid record corresponding to the primary HIDD record would contain information from claim 1 - claim 5.

Table 15

Medicaid Claims Corresponding to HIDD Records

Primary HIDD Record (Aggregated HIDD)	HIDD record 1	claim 1 claim 2 claim 3
	HIDD record 2	claim 3 claim 4 claim 5

## 14. THE 1995 CODES LINKED FILE AND FINAL MATCH RESULTS

The 1995 CODES analysis file contains a total of 5,439 records. 4,129 of these are linked records and the remaining 1,310 are unlinked STR or OMI records. What follows is a breakdown of the file contents.

- 3,612 STR Records: linked and unlinked STR records representing the complete, unduplicated 1995 STR file.
- 1,489 OMI Records: linked and unlinked OMI records representing the complete, unduplicated 1995 OMI file.
- 2,178 Crash Driver/Occupant records: any Crash record linked to OMI, STR, or HIDD.
- 3,499 HIDD Records: aggregated HIDD records linked to STR or Crash Driver/Occupant.
- 380 Medicaid Records: aggregated Medicaid records linked to a HIDD record containing STR or Crash information.

For analysis purposes, two supplemental CODES files were provided to the data users: the Crash Driver/Occupant residual file (n=142,985); and the HIDD residual file (n=17,084). These files provide crash and injury denominators, respectively. Note: The HIDD residual may contain records with "complications of care" injury codes so the denominator may be too inclusive.

### 14.1 *Finding Matches Not Identified in Direct File-to-File Match Processes*

The CODES linked file contains matched record pairs that were not identified by Automatch in direct file-to-file matches. These matched pairs surfaced during the building of the linked file, when matches from all processes were combined and cross checked. These matches are the natural result of a multiple process match strategy. Whenever n files (n>2) are matched in at least (n-1) separate match processes, the possibility of indirect matches is present. The simplest case is that of three files matched with two match processes. Given files A, B and C; if A is matched to B and B is matched to C, links will be established indirectly between files A and C. A more complicated example is that of three files matched in three processes. For example, if specific STR, OMI and Crash records were included in three separate match processes the following cases causing indirect matches may have occurred:

---

In a series of match processes including: STR record, **S**; OMI record, **O**; and Crash record, **C** ...

**Case #1:** **S** matched to **O**; **O** matched to **C**; **C** did not match to **S**

**Case #2:** **S** matched to **O**; **O** did not match to **C**; **C** matched to **S**

**Case #3:** **S** did not match to **O**; **O** matched to **C**; **C** matched to **S**

---

The respective indirect matched pairs are **S** to **O**, **O** to **C** and **S** to **O**.

When indirect matches surfaced they were reviewed, either by hand or using SAS programs. If the matches were solid they were added to the linked file.

**14.2 Final Match Results and Final Match Rates for the 1995 CODES Linked File**

The following tables present statistics for the final 1995 CODES analysis file: Table 16 presents the number of matches between the CODES files; Table 17 gives the percentage of disagreement for common variables within matched record pairs; Table 18 gives the final match rates for the 1995 CODES file. Due to the addition of indirect matches some of the final match numbers and rates given below are slightly higher than those given in previous sections of this report.

Table 16  
Number of Matches Between 1995 CODES Files

<b>STR</b>	<b>OMI</b>	<b>HIDD*</b>	<b>MEDICAID*</b>	
1,263	470	1,601	159	<b>CRASH</b>
	232	2,938	351	<b>STR</b>
		133	14	<b>OMI</b>
			380	<b>HIDD*</b>

\* aggregated records.

There were 889 records (24.6% of all STR records) that spanned the Crash, STR and HIDD data bases. There were eight (8) records that spanned all five 1995 CODES files.

Table 18  
Disagreement Rates between Common Variables within Matched Record Pairs

<b>Variable</b>	<b>STR/CRASH Matches</b>	<b>STR/OMI Matches</b>	<b>OMI/CRASH Matches</b>	<b>STR/HIDD Matches</b>
Age	3.4%	0.9%	5.1%	1.1%
Sex	0.8%	1.3%	0.0%	0.8%
Injury Date	1.3%	1.7%	0.4%	NA
County	11.3%	6.0%	6.0%	NA
Vehicle Type	2.7%	1.3%	0.2%	NA

**Table 18**  
**Final Match Rates for 1995 CODES Data Linkage**

<b>People Killed In Crashes</b>	with STR:	23.5%	(114 / 485)
	with HIDD:	13.8%	(67 / 485)
	with OMI:	<b>95.5%</b>	(470 / 485)
<b>People Injured (not killed) in Crashes</b>	with STR:	3.6%	(1129 / 30996)
	with HIDD:	4.7%	(1445 / 30996)
	with OMI:	~0.0%	(7 / 30996)
<b>Deceased People in STR</b>	with OMI:	<b>95.5%</b>	(232 / 243)
<b>STR records indicating a motor vehicle crash in New Mexico</b>	with Crash:	Expected <b>75.81%</b>	(1175 / 1550)
		Overall 75.5%	(1263 / 1673)
<b>STR records indicating admission or transfer to a HIDD facility</b>	with HIDD:	Expected <b>93.4%</b>	(2357 / 2528)
		Overall 84.5%	(2938 / 3476)
<b>HIDD records with unique Medicaid identification numbers</b>	with Medicaid:	<b>49.7%</b>	(1,357/2,733)

For more detailed information about the final match rates between STR and Crash and between STR and HIDD, please see Appendix A.

## **15. PROJECT ASSESSMENT**

The 1995 CODES project was successful in moving towards a more complete injury related database. For the first time, injured occupants were included in the match with the Office of the Medical Investigator data. Hospital Inpatient Discharge Data and Medicaid claims data were added to the linked file, providing new information about injury-related medical care and costs. Work is currently underway to match Medicaid claims to people in crashes, regardless of whether or not they were admitted to the hospital. Beginning with the 1996 CODES project, Emergency Medical Services (EMS) data will be added to the linked file. It will then be possible to track an injury from the injury scene, to the ambulance ride, entry in the trauma registry, admission to the hospital and, if a Medicaid claim was filed, payment for services rendered. Additionally, the 1996 Crash files will include Last Names and First Initials for all injured occupants, greatly improving the odds of matching these individuals to information in the other CODES files. The first two years of CODES work have made measurable progress towards meeting the core of the CODES data linkage objective set forth by the New Mexico Department of Health: "... to track injuries from the source through all phases of care (DOH, September, 1995)."

## REFERENCES

Baulch, S. & Rutledge, L.M. (1995). *Annual Trauma Systems Report: For Calendar Year 1994*. Santa Fe: State of New Mexico Department of Health, Community Health Systems Division, Emergency Medical Services Bureau.

Office of the Medical Investigator State of New Mexico (1995). *Annual Report 1994*. Albuquerque: University of New Mexico School of Medicine, Health Science Center.

Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *American Statistical Association Journal*, 84(406), 413-420.

Jaro, M. A. (1995). Probabilistic Linkage of Large Public Health Data Files. *Statistics in Medicine*, 14, 491-498.

Jaro, M.A. (1995). *Documentation on Automatch Generalized Record Linkage System, Version 3.0*. Burtonsville, MD: Matchware Technologies, Inc.

New Mexico Department of Health(1995). *CODES Approach, 2*. Santa Fe: State of New Mexico Department of Health, Division of Epidemiology, Evaluation & Planning.