U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

NHTSA
www.nhtsa.gov

DOT HS 812 179

July 2015

Crash Outcome Data Evaluation System (CODES): An Examination Of Methodologies and Multi-State Traffic Safety Applications

DISCLAIMER

This publication is distributed by the U.S. Department of Transportation, National Highway Traffic Safety Administration, in the interest of information exchange. The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the Department of Transportation or the National Highway Traffic Safety Administration. The United States Government assumes no liability for its contents or use thereof. If trade or manufacturers' names or products are mentioned, it is because they are considered essential to the object of the publication and should not be construed as an endorsement. The United States Government does not endorse products or manufacturers.

Suggested APA Format Citation:

Cook, L. J., Thomas, A., Olson, C., Funai, T., & Simmons, T. (2015, July). *Crash Outcome Data Evaluation System (CODES): An examination of methodologies and multi-state traffic safety applications.* (Report No. DOT HS 812 179). Washington, DC: U.S. Department of Transportation, National Highway Traffic Safety Administration.

This report was partially funded by the U.S. Centers for Disease Control and Prevention.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT HS 812 179		

4. Title and Subtitle	5. Report Date
Crash Outcome Data Evaluation System (CODES): An Examination of Methodologies and Multi-State Traffic Safety Applications	July 2015
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Lawrence J. Cook, Ph.D.; Andrea Thomas; Cody Olson; Tomohiko Funai; Timothy Simmons	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
University of Utah Intermountain Injury Control Research Center 615 Arapeen Dr., Suite 202 Salt Lake City, UT 84108-1226	11. Contract or Grant No. DTNH22-08-H-00304

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
National Highway Traffic Safety Administration National Center for Statistics and Analysis 1200 New Jersey Avenue SE. Washington, DC 20590	NHTSA Technical Report
	14. Sponsoring Agency Code NVS-412

15. Supplementary Notes

John Kindelberger was the Contracting Officer's Technical Representative for the CODES cooperative agreements involved in this project.

16. Abstract

This report provides a summary of recent technical work in the Crash Outcome Data Evaluation System (CODES), a program facilitated by the National Highway Traffic Safety Administration (NHTSA). CODES involves a statistical methodology to augment State crash data with medical outcome data using probabilistic linkage. In 2013, NHTSA transitioned the CODES program to full State autonomy. This two-part report comprises the final technical report from the CODES Technical Resource Center at the Utah CODES project. Part one provides information on the probabilistic linkage methodology employed by CODES, and addresses related topics including other types of linkage, alternative linked data sets, match probabilities, and missing data imputation; and part two reports on demonstration projects pooling multi-state standardized data for four topics relevant to traffic safety.

17. Key Words	18. Distribution Statement
NHTSA; Crash Outcome Data Evaluation System; CODES; CODES Data Network; data linkage; probabilistic linkage; multiple imputation; linked data applications	Document is available to the public from the National Technical Information Service www.ntis.gov.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22
Unclassified	Unclassified	100	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Table of Contents

Executive Summary

This report provides a summary of recent technical work in the Crash Outcome Data Evaluation System (CODES), a program facilitated by the National Highway Traffic Safety Administration (NHTSA). CODES involves a statistical methodology to augment State crash data with medical outcome data using probabilistic linkage. In 2013, NHTSA transitioned the CODES program to full State autonomy. This two-part report comprises the final technical report from the CODES Technical Resource Center.

Probabilistic linkage is a powerful method for combining information from different databases into a single dataset for analysis. Desired information about study subjects is often contained in two or more databases, and if a unique key does not exist between these databases, it is not possible to combine the information directly. Rather than relying on a unique key to combine records, probabilistic linkage makes use of fields that are common to each database. CODES uses probabilistic linkage to combine information from motor vehicle crash (MVC) reports and hospital records, sometimes also adding databases such as EMS, death certificates, and others.

Part 1 of this report provides detailed information on the probabilistic linkage methodology employed by CODES, as well as other types of linkage, alternative linked data sets, match probabilities, and questions about imputation of missing data. Topics and findings include:

- CODES linkage uses multiple imputations to avoid clerical matching by selecting a weighted sample, based on match probability, from the set of all candidates. Other ways to create a matched set include high-probability links and maximum a posteriori (MAP) data sets. Some other available software packages use a variety of features and methodologies.
- In a comparison of linkage types, when minimum potential match probability was high, high-probability, multiply imputed, and MAP matched sets were not significantly different from the true match population; when minimum potential match probability was low, multiply imputed and MAP matched sets were representative of the true match population with multiply imputed matched sets performing slightly better than MAP matched sets when modeling the binary outcome of hospitalization status. High-probability matched sets were not representative of the true match population when minimum potential match probability was low. High-probability matched sets underrepresented common values thereby overrepresenting rare values; however, the matched pairs identified were likely correct.
- In a study of identifiers and match probabilities when linking records between crash and hospital databases, typical identifiers are incident date, sex, age, date of birth, first name, last name, and seat position; these identifiers and other identifiers such as emergency department and hospital flags, social security number, and longitude/latitude tend to produce high match probabilities. There appears to be a negative relationship between the crash file size and the match probabilities. Therefore, careful selection of identifiers is crucial when linking bigger crash files; crash records with high match probabilities appear in a higher number of imputed datasets.
- In a study of imputation of missing data for analysis, demonstrations found that odds-ratio estimates from complete case analyses are usually very close to those from multiple-imputation analyses, especially when rates of missing data are low; that multiple-imputation-

based estimates tend to be more powerful (i.e., tend to identify more predictors as significant, and have shorter confidence intervals for those estimates) compared to estimates based on complete-cases only; that in some cases multiple-imputation methods give very different estimates compared to complete case methods and that in these situations, missing-at-random assumptions appear to be violated; and that despite the many variables involved in specifying the imputation, two different procedures of imputing missing data in a CODES dataset resulted in similar results.

Part 2 explains the design and implementation of a general-use data standardization model and provides demonstration projects using standardized data from multiple CODES States on four topics relevant to traffic safety: older occupants, child safety restraints, motorcycle helmets, and graduated driver licensing. Findings included:

- The percentage of occupants sustaining chest injuries and fractures tripled as the age of occupants increased from 21 to 64, to 85 and older.
- The odds of sustaining an injury to the neck, back, or abdomen among children reported as using child restraint systems were almost half the odds of reported unrestrained children (OR: 0.64; 95% CI: 0.59, 0.70). This reduction was less evident among seat-belt reported children (OR: 0.91; 95% CI: 0.83, 1.00).
- After adjusting for other factors, the relative risk of motorcycle head and face injuries was higher when no helmet was worn: not wearing a helmet was associated with a 201-percent increase in the risk of head injuries and 263-percent increase in the risk of facial injuries in single-vehicle crashes in partial law States.
- Insurance Institute for Highway Safety graduated driver licensing (GDL) programs rated "good" were associated with lower rates of teenage driver involvement in injury motor vehicle crashes compared to teens driving under GDL programs rated "poor".

The demonstration projects showed that CODES methodology is not only feasible within a single State, but when combined, linked multi-State data analyses can produce sensible, meaningful results. Combined data can be used to study populations that may be too small to analyze in a single-State study. An additional benefit is the ability to compare crash outcomes in relation to the type of legislation that has been enacted in the different States. These efforts provide an example for how future multi-State projects may be carried out.

Preface

This report provides a summary of recent technical work in the Crash Outcome Data Evaluation System (CODES), a program facilitated by the National Highway Traffic Safety Administration (NHTSA). It is the final report from the CODES Technical Resource Center, a function provided by the Utah CODES project at the University of Utah Intermountain Injury Control Research Center. It is in two parts: the first providing information on the probabilistic linkage methodology employed by CODES as well as addressing related topics including other types of linkage, alternative linked data sets, match probabilities, and missing data imputation; and the second reporting on demonstration projects pooling multi-state standardized data for four topics relevant to traffic safety.

CODES analyses are accomplished by combining information from police crash reports with data from medical records such as emergency department, hospital discharge and Emergency Medical Services (EMS) databases. Since these databases are collected by different State agencies and at different times during the injury event, there is rarely a common key among the databases that would allow a direct join. In the absence of common keys, CODES projects use probabilistic linkage, a method that determines the probability that two records refer to the same person and event, to bring these data sets together. Once combined, an analyst may then examine the relation of event factors from the crash report with medical outcomes from the hospital and EMS databases.

CODES began in response to the Intermodal Surface Transportation Efficiency Act of 1991 (ISTEA), which required NHTSA to conduct a study to determine the efficacy of seat belts and motorcycle helmets for preventing injuries in motor vehicle crashes. After the initial study (NHTSA, 1996), NHTSA set up and coordinated the CODES Data Network, an ongoing collection of CODES grantee States that worked together to share methodological and analytical findings and study common traffic safety issues. Since then CODES has been implemented in up to 30 States. At the State level, CODES projects have used their data to support traffic safety legislation regarding primary seat belt enforcement, motorcycle helmet usage, GDL, booster seats, and many other topics, as well as to support State Traffic Records Coordinating Committees (TRCC) and other decision-makers, to conduct problem identification, and to provide traffic safety facts to the public (NHTSA, 2010a; NHTSA, 2015).

In 2013, the CODES program went through a transition in which NHTSA gave State grantees full responsibility for their programs, including funding responsibility. NHTSA had been encouraging CODES grantees to prepare for such a transition for some years and had encouraged States to plan to continue the programs. In the years leading up to the transition, the CODES Data Network worked with NHTSA on several analyses and data requests and also provided support to the U.S. Centers for Disease Control and Prevention (CDC) and the National Transportation Safety Board (NTSB). For more information on the program transition and recent activities, see *CODES: Program Transition and Promising Practices* (Report No. DOT HS 812 178) (NHTSA, 2015).

One of the challenges to conducting studies that utilize data from multiple States is that each State's crash file, while similar, does not always capture the same fields, and common fields are

frequently coded differently. NHTSA estimates that it would cost nearly a billion dollars to collect and code all crashes in the Nation into a uniform format (NHTSA, 2010b). In an attempt to harmonize data collected from multiple States, the CODES Technical Resource Center worked with NHTSA's State Data System and the CODES Data Network to develop a standardization model that mapped State-specific crash files onto a standardized format called the General Use Model (GUM). The second part of this report covers the assembling of the GUM and reports on findings using GUM data from 11 CODES States for demonstration analyses of 4 traffic safety topics: older occupants, child safety restraints, motorcycle helmets, and GDL.

References

National Highway Traffic Safety Administration. (1996). *Report to Congress: Benefits of safety belts and motorcycle helmets* (Report No. DOT HS 808 347). Washington, DC: U.S. Department of Transportation.

National Highway Traffic Safety Administration. (2010a). *The Crash Outcome Data Evaluation System and applications to improve traffic safety decision-making* (Report No. DOT HS 811 181). Washington, DC: U.S. Department of Transportation.

National Highway Traffic Safety Administration. (2010b). *Report to Congress: NHTSA's crash data collection programs.* (Report No. DOT HS 811 337). Washington, DC: U.S. Department of Transportation.

National Highway Traffic Safety Administration. (2015*). CODES: Program transition and promising practices*. (Report No. DOT HS 812 178). Washington, DC: U.S. Department of Transportation.

Part 1: Probabilistic Record Linkage

Chapter 1: Linkage Methodology

Introduction

Often the information required to answer a research question resides in multiple databases. When faced with this issue researchers must find a way to join or link the databases before the study can continue. For instance, in order to determine the outcome for patients referred to the emergency department (ED) by poison control, one would need to combine information from poison control, ED, possibly hospital admission, and vital records databases. Similarly, one would need to combine information from an emergency medical services (EMS) database with hospitalization and death databases to determine if patients undergoing a specific treatment have better outcomes depending on the provider's experience with the treatment.

Questions requiring the combination of multiple databases are ubiquitous in motor vehicle crash (MVC) research (Gonzalez et al., 2007; Mango & Garthe, 2007; Senserrick et al., 2009; Thygerson, Merrill, Cook, Thomas, & Wu, 2011a; Thygerson et al., 2011b; Newgard et al., 2012; Thomas, –Thygerson, Merrill, & Cook, 2012; Brubacher, Chan, Fang, Brown, & Purssell, 2013; Vladutiu et al., 2013). Researchers might be interested in knowing if passenger vehicle occupants using safety restraints are at less risk of being hospitalized following a MVC compared to occupants not using safety restraints. In an effort to support universal helmet legislation, researchers might be interested in determining the risk of sustaining a traumatic brain injury (TBI) for motorcyclists wearing helmets compared to riders not wearing helmets. More recently, there has been increased interest in better understanding the correlation between police-reported injury severity compared to what is reported by trained medical professionals. In all of these cases, combining information from MVC records with ED and hospital admission data is required to answer the research question on a population level.

There are three broad methods for combining databases: direct linkage or interface through a common identifier, deterministic linkage, and probabilistic linkage. We proceed through this chapter by briefly describing direct and deterministic linkage. The main focus will be on probabilistic methods used by CODES funded by NHTSA. We then conclude by examining a number of probabilistic linkage software options currently available.

Types of Linkages

Direct Linkage or Interface

An interface describes a situation where two data sources are able to seamlessly interact with each other in real time. Such a situation might arise when a police officer scans a driver license and is immediately provided with information regarding the driver and vehicle. Different hospitals under the same ownership may have their databases interfaced through the use of a medical record number. Interfaces should be highly reliable and usually support critical business practices. Interfaces between MVC and health care data rarely exist due to the complex nature of data ownership and how the data are compiled. MVC data and hospital data are collected by different entities: MVC by public agencies such as police departments and hospital information by private health care companies. Similarly, compilation of data from multiple law enforcement

agencies or hospital systems is often done well after the event and by different State agencies, Department of Public Safety or Department of Transportation for MVC records, and the Department of Health or State Hospital Association for hospital data.

Direct linkage between two data sources occurs when there is a unique single identifier, or combination of identifiers, that allow records from the databases to be joined using a simple query. This type of linkage requires the identifier (or set of identifiers) to be collected and coded the same way on both data sources. If both the MVC and hospital databases contained a person's Social Security number (SSN), then a direct linkage between the two would be possible. Unfortunately, in our experience, SSN rarely exists in either the MVC or hospital databases, making direct linkage in CODES impossible. One situation where the opportunity for a direct linkage frequently arises is when combining information from a driver's MVC record and his or her driver license file where the driver license number can be used to directly join the two databases. Even though a direct linkage may be feasible, researchers need to remember that the success of the direct linkage is highly dependent on the quality of the identifiers being used. If driver license numbers are frequently missing or poorly coded in crash files, then a direct linkage with the driver license file would result in many drivers failing to link to their license information. In such a situation researchers should turn to another linkage method to augment the direct linkage.

Deterministic Linkage

Rather than relying on a single identifier or set of identifiers that uniquely specifies an individual within a database, deterministic linkage relies on using multiple quasi-unique fields. Examples of variables that are frequently used in deterministic linkages include dates of events, dates of birth, names, and county or ZIP codes. The simplest form of a deterministic linkage requires all quasi-unique fields to agree on a pair of records for it to be considered a match. Researchers will often construct scoring schemes for their deterministic linkages giving higher point totals to variables that are considered to be more reliable or specific and fewer points to more general or less reliable fields. In order to be considered as a match, a pair of records must exhibit a combination of agreements and disagreements that lead to a point total above a determined threshold. While deterministic linkages can lead to successful results, they do have some shortcomings. Typically the point totals for weighting matching fields is determined by a given researcher. Two researchers are likely to arrive at different weighting schemes and thus different linking results. Another shortcoming is that the relative rarity or commonness of a specific value in a field is not considered. Agreement on a value that occurs in half of the records in a database is weighted exactly the same as agreement on a value that occurs on only a single record. Finally, there is no way to assess the researcher's confidence in a given pair of records beyond the fact that the pair achieved the threshold. The remainder of this section addresses how probabilistic linkage—the methodology used within CODES—overcomes these limitations.

Probabilistic Linkage

Probabilistic linkage is the methodology used by the CODES data network; it addresses the limitations noted above with deterministic linkage. Rather than relying on a researcher determined threshold, probabilistic linkage generates the probability that a pair of records refers to the same person and event. Below we provide a summary of how probabilistic linkage identifies potential matches; summarizes the probability that two records refer to the same person

and event; and ways to identify sets of matches to use in analysis based on these probabilities. For a more detailed description of the probabilistic linkage process please see Jaro (1995), Cook, Olson, and Dean (2001), Dean et al. (2001), and McGlincy, (2004, 2006).

Operationalizing Comparisons of Records in a Probabilistic Setting

Probabilistic record linkage is accomplished by comparing data fields in two different files, such as the date of birth or the gender of a patient. The comparisons of numerous data fields lead to a judgment that two records refer to the same MVC patient event (and should be linked) or that the records do not refer to the same MVC patient event (and should not be linked). As with deterministic linkage, this judgment is based on the cumulative agreement and disagreement of field values. Data fields that are compared have differing impacts on a judgment that two records should be linked. For instance, agreement of the gender field alone would not suffice to conclude that two records refer to the same patient, while agreement on SSN alone greatly enhances the probability that two records refer to the same individual. By assigning the log-odds to field comparisons, it is possible to calculate match weights and computerize the judgment process. The calculation of match weights rely on two probabilities: reliability, the probability that field *i* will agree given two records are known to be a true match, and discriminating power, the probability that field *i* will agree given that two records are known not to match. It is customary to represent the reliability and discriminating power of field *i* as m_i and u_i, respectively. For a given pair of records, if field *i* agrees, the odds the records match are m_i/u_i and there will be an agreement weight of $w_i = \log_2(m_i/u_i)$. If field *i* disagrees, then the odds that the records match are $(1-m_i)/(1-u_i)$ and a disagreement weight of $w_i = \log_2((1 - m_i)/(1 - u_i))$ is assigned. The composite weight (w_t) for a record pair will be the sum of agreement and disagreement weights for all the data fields that are available for comparison. As w_t increases, the likelihood that two records refer to the same MVC patient event increases. As w_t decreases there is decreased likelihood that the records refer to the same MVC patient event. Another benefit of this procedure is that value specific *u* probabilities within a field can be assigned resulting in common values of field, such as a last name of Smith, have lower match weights compared to rarer values, such as a last name of Funai.

Both m_i and u_i are theoretical quantities and are rarely known prior to conducting a linkage. To estimate m_i for fields from a database CODES analysts use multiple techniques. First is historical knowledge of the databases being linked. Most analysts have linked their MVC and hospital databases many times. If no major modifications have been made to either database since the previous year's linkage then the m_i's from that linkage can be carried forward as an estimate for the new linkage. If faced with a major modification to one of the regular databases or linking to a new database then the past year's linkage will be of little use. In this case, one can usually get good estimates of the reliability, m_i*s*, of data fields from the data owners. In the absence of any information regarding reliability, CODES analysts are encouraged to link one month of data to obtain initial reliability estimates to inform the full year linkage. The discriminating power is estimated from the data being linked. Typically the set of expected matched pairs is negligible in size compared to the set of all possible matched pairs between two databases. Thus, u_i for a given field can be estimated by taking a sample of random pairs and determining how often field *i* agrees.

Identifying True and False Matches

Once all candidate match pairs have been identified and received a match weight the next step is to determine which pairs are matches and which ones are not. We will first begin by describing two traditional methods for selecting pairs by using cut points and then describe the current methodology implemented in the software used throughout the CODES Data Network. Graphs are one method to identify cut points for selecting which pairs are true or false matches. To use this method, one must first sort the candidate matches by weight. The first cut point is a value at which all pairs with a weight above will be considered true matches. The second cut point determines the value at which all pairs with a weight below will be considered to be non-matches. All pairs with weights between the two cut points are manually reviewed and classified as a match or non-match.

Figure 1.1.1 shows the distribution of match weights from a MVC and EMS linkage. To choose our cut points we want to identify a point at the high end of the distribution where we can feel comfortable that all pairs of records above that weight are true matches; in this case we have selected a match weight of 33. The second cut point will be the one at which all pairs below are false matches; in this example we have chosen a match weight of 20. Thus, all pairs between match weights of 20 and 33 will need to be reviewed by hand, which is more than 1,000 pairs.

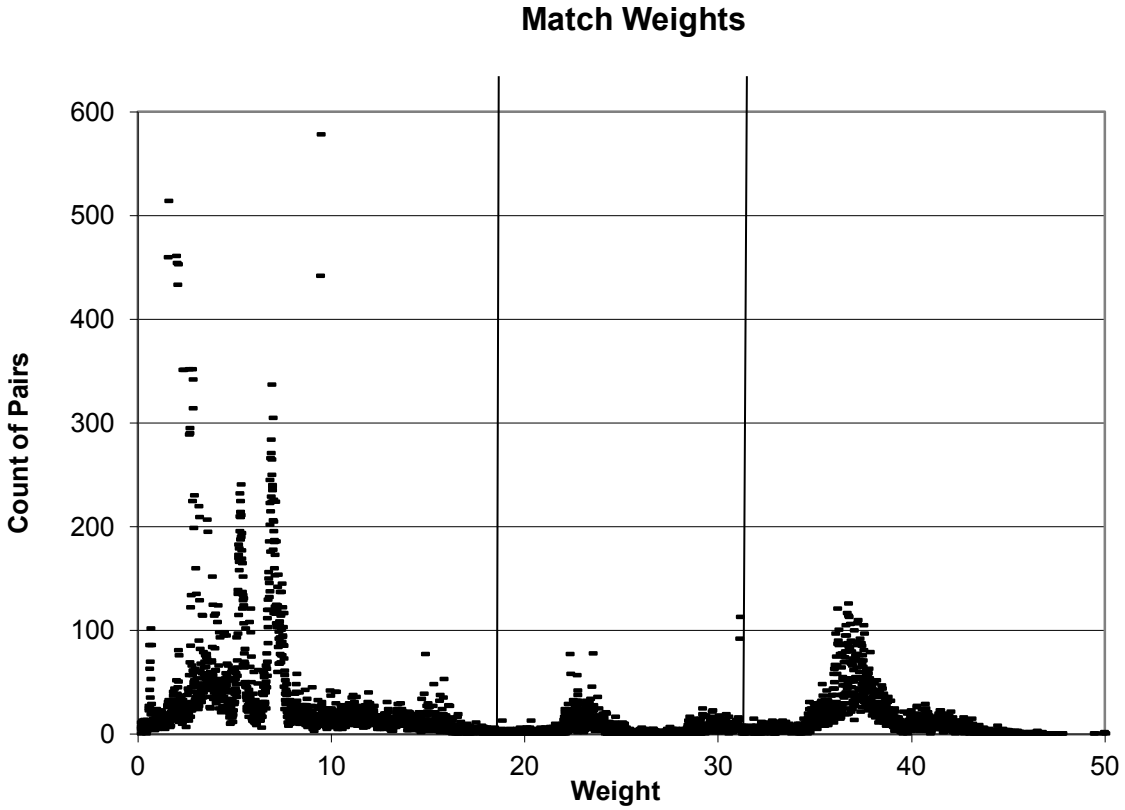


Figure 1.1.1. Distribution of match weights from an EMS and MVC probabilistic linkage

Figure 1.1.2 shows the results of the clerical review and the final determination of all pairs. The large red circles represent pairs that were rejected as matches and the blue dashes are pairs that

were considered true matches. The plot shows that the distribution of false match weights and true match weights slightly overlap. However, the burden of manually reviewing thousands of pairs and the need for reproducible research has led some researchers to abandon clerical review altogether. Under this methodology we would pick a single cut point, say a match weight of 30, and all pairs with weights less than 30 would be considered false matches, while all pairs with weights above 30 would be true matches.

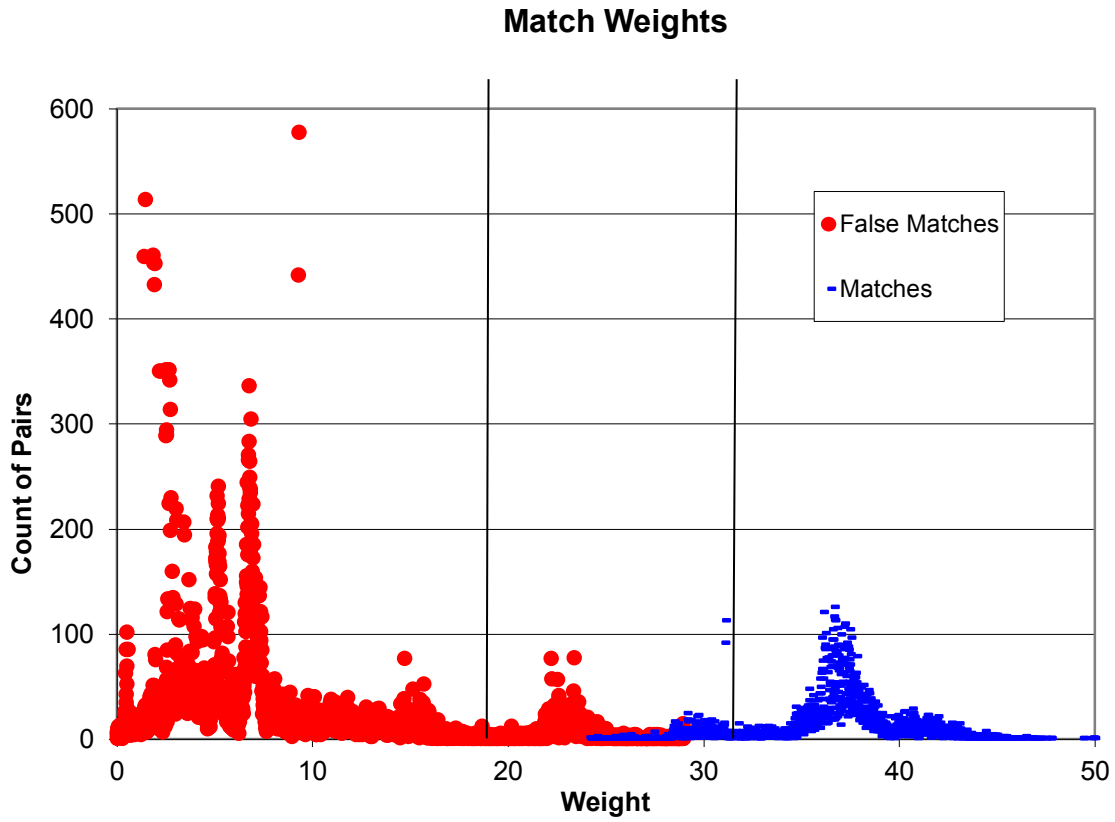


Figure 1.1.2. Distribution of match weights from an EMS and MVC probabilistic linkage by determination of final match status

Match Probabilities

An alternative to choosing cut points based on graphs is to calculate probability based cut points. Since match weights are derived from the logarithm of the odds (log odds) it is possible to determine the weight needed to achieve a specified probability that two linked records are a true match. The following section was adapted from formulas developed by McGlincy (2004) and supplies the necessary background needed to make the above determination. The odds of an event A are defined as

$$\frac{\text{Probability that A occurs}}{\text{Probability that A does not occur}} \text{ or } \frac{P(A)}{1-P(A)}.$$

By rearrangement one obtains

$$P(A) = \frac{\text{Odds of A}}{1 + \text{Odds of A}} . \quad (1)$$

Using Equation 1 it is possible to calculate the odds and probability of picking a matched pair at random. Given two files, 1 and 2, with number of records A and B, respectively, the number of possible record pairings is A x B. If E of the A x B pairings are true matches (note that E must be less than both A and B, since the number of true matches cannot exceed the minimum of the two file sizes) the probability of picking a true match at random is

$$P(E) = \frac{\text{number record pairings that are true matches}}{\text{total number of record pairings}} = \frac{E}{AxB} . \quad (2)$$

Therefore, the odds of picking a true match at random is

$$\frac{P(E)}{1 - P(E)} = \frac{\frac{E}{AxB}}{1 - \frac{E}{AxB}}$$

$$= \frac{E}{AxB - E} . \quad (3)$$

This equation will produce a small numeric value since the number of possible record pairings greatly exceeds the number of valid matches. For instance, in an ideal case where we have 1,000 records in File 1 and 1,000 records in File 2 and every record in File 1 is known to match uniquely to a record in File 2, there are 1,000 expected matches. Using Equation 3, one can calculate the odds of picking a true match at random to be $\frac{1000}{1000x1000 - 1000} = 0.001$ or 1 in 1,000 tries. Using equation 1 it can be shown that the probability of picking a true match at random is also approximately 0.001.

How much information is needed to improve the probability of selecting true matches to 0.90? The odds are calculated from equation 3, $\frac{0.90}{1 - 0.90} = 9.0$. The ratio of the desired odds to the current odds will reveal how much the odds must improve to obtain the desired probability, 0.90. The ratio of the desired odds and the current odds is $\frac{9}{0.001} = 9{,}000$, so the current odds must increase by a factor of 9,000 to improve the probability of picking correct matches from 0.001 to 0.90. If we take the $\log_2(9{,}000)$ we can express the needed improvement in odds as an improvement in match weight. To increase the probability of selecting a true match, the match weight must increase from its current value of $\log_2(0.001) = -9.97$ to $\log_2(9{,}000)$ or 13.14. Therefore, only accepting matched pairs that have a match weight of 13.14 will yield a probability of selecting correct matches of at least 0.90.

Using the same notation as above for A, B, and E and denoting the desired probability of selecting a correct match as p, the match weight (w_t) that corresponds to probability p of a true match can be expressed as

$$w_t = \log_2\left(\frac{\text{desired odds}}{\text{current odds}}\right) = \log_2\left(\frac{\frac{p}{1-p}}{\frac{E}{AxB - E}}\right) . \quad (4)$$

By holding file sizes (A and B) and the number of expected matches (E) constant, one can examine the effect of increasing p, the desired probability. If the value of p is increased to p' the necessary match weight will increase. The increase in necessary weight, I_w, can be quantified.

$$I_w = \log_2\left(\frac{\frac{p'}{1-p'}}{\frac{E}{AxB-E}}\right) - \log_2\left(\frac{\frac{p}{1-p}}{\frac{E}{AxB-E}}\right) \quad (5)$$

$$I_w = \log_2\left(\frac{p'}{1-p'}\right) - \log_2\left(\frac{E}{AxB-E}\right) - \left(\log_2\left(\frac{p}{1-p}\right) - \log_2\left(\frac{E}{AxB-E}\right)\right) \quad (6)$$

$$I_w = \log_2\left(\frac{p'}{1-p'}\right) - \log_2\left(\frac{p}{1-p}\right) \quad (7)$$

$$I_w = \log_2\left(\frac{\frac{p'}{1-p'}}{\frac{p}{1-p}}\right) \quad (8)$$

$$I_w = \log_2\left(\frac{\text{desired odds of a true match}}{\text{current odds of a true match}}\right) \quad (9)$$

Similarly, if we held the file sizes (A and B) and probability (p) constant, it is easily demonstrated (from equation 4) that an increase in expected matches (E) will decrease the match weight needed to achieve a linkage with probability p. If we hold everything constant except for the size of file 1, then Equation 4 shows that more match weight is necessary.

Equation 4 can now be used to determine cut points for true matches, false matches, and the clerical review region. One option, for instance, would be to use Equation 4 to determine the weights associated with probabilities of correct matches equal to 0.9 and 0.5, $w_{0.9}$ and $w_{0.5}$ respectively. All pairs with a weight above $w_{0.9}$ would then be considered true matches, all pairs with a weight below $w_{0.5}$ would be considered false matches, and all pairs between $w_{0.9}$ and $w_{0.5}$ would be manually reviewed. To eliminate the human element of clerical review, another option is to choose to select a single cut point, such as a match probability of 0.9, and consider all pairs

of records above the threshold to be true matches and all pairs of records below the threshold to be false matches. It is common to refer to matches generated this way as 'high-probability matches' to indicate they have passed the high threshold cut point.

While being able to quantify a specific cut point in terms of match probability is a desirable property, there is no guarantee that a given probabilistic match algorithm will have pairs that can achieve the cut point. As mentioned above, the ability to achieve a pre-specified probability is related to the sizes of the files being linked, the number of expected true matches, the desired probability that needs to be achieved, and, of course, the quality of the matching variables. One of the properties of probabilistic linkage is that pairs of records that agree on rarer values of the matching variables will have higher weights, and therefore higher probabilities, than pairs of records agreeing on common values of the matching variables. If a researcher finds him or herself in a situation where only a portion of the matches can achieve the specified upper cut point then there is a risk of producing a set of matches which are biased on the matching variables. For example, the motor vehicle crash population tends to be much younger than the general population, with the largest portion coming from the 16 to 25 year-old age group. Also, more crashes tend to occur in urban areas than in rural areas. Therefore, matched pairs that agree on older ages and rural counties typically result in higher match weights and probabilities. Thus, in constructing a high-probability matched set, if the cut point for determining true and false matches falls in the middle of the true match weight distribution then the resulting match results are likely to over represent older and rural crashes.

Current CODES Methodology

Multiple Imputation of Match Status

To avoid potential biases introduced by focusing on a subset of high-probability pairs, McGlincy (2004) has proposed using multiple imputation to select matched sets. Briefly, the process involves setting a single cut point associated with a very small probability, such as 0.01. All pairs that are able to achieve this low cut point are then considered as potential candidate pairs. Rather than keeping all pairs as true matches, the procedure continues by selecting a weighted sample, based on match probability, from the set of all candidates. In other words, pairs that have a probability of 0.90 of being correct are selected in about 90 percent of all samples, pairs with a probability of 0.50 of being correct are selected in about half of all samples, and pair with a probability of 0.01 are selected in only about 1 out of 100 samples. To account for the uncertainty introduced by taking a random sample of candidate match pairs, the process is repeated multiple times; standard CODES practice is to take five samples. The software incorporates an additional Markov chain Monte Carlo (MCMC) step between samples to ensure that successive samples are not dependent on the initial starting values of the field agreement probabilities (m_i) and that the five sets of matched pairs are not related. To analyze the resulting matched sets, special methods are needed, which have been described elsewhere and are discussed in Chapter 4.

Alternatives to Multiple Matched Sets

There are some instances when a researcher needs a single data set, such as when providing data for a public website. In these instances one can still use the results of the Markov chain Monte Carlo process. Before describing these methods, it needs to be emphasized that while a single

imputation can be used to obtain point estimates, standard errors from an analysis of a single imputation will underestimate the true variability in the data. Thus, there is a risk of obtaining false significance from hypothesis tests and artificially small confidence interval lengths.

One method of obtaining a single data set is to construct a set of linked pairs while controlling the false error rate in the matched set. To begin this process select a single MCMC matched set. Construct the matched set by using the following algorithm.

1. Order all matches from the largest probability to the smallest.
2. Calculate each pair's false match probability as 1 – true match probability.
3. Determine the desired false positive rate (FP), usually 1 percent, 5 percent, or 10 percent.
4. Repeat the following steps until the false positive rate (FP) is obtained
    a. Remove the linked pair at the top of the list (the pair with the highest true match probability and the smallest false match probability) from the pool of potential matches and insert it into the set of selected matches.
    b. Calculate the estimated false positive rate (EFP) in the set of selected matches as (the sum of false match probabilities)/(total number of selected matches).
    c. If EFP < FP repeat steps a and b.
    d. If EFP $\geq$ FP stop.
5. Use the resulting set of selected matches as the set of linked pairs with corresponding false positive rate FP.

Another option for generating a single imputed set for analysis is via maximum a posteriori estimation. Maximum a posteriori estimation is the Bayesian analogue to maximum likelihood estimation with the exception that rather than finding the maximum of the likelihood, we are finding the maximum of the posterior distribution of match probabilities. To generate a maximum a posteriori set, the software will generate many, at least 50, imputed sets. The maximum a posteriori set is the sample that makes the most correct decisions, defined as deciding pairs with match probabilities above 0.5 are true matches and pairs with match probabilities below 0.5 are false matches.

While the MCMC and multiple imputation processes outlined above help overcome many pitfalls that traditional probabilistic linkages are susceptible to, they do add an increased burden to the analyst. While much of the process is handled by the software, there is an extra level of statistical sophistication required by the user to understand all of the parameters and options available to guide the process. Additionally, five data sets are inherently more complicated to analyze than a single data set. Analysts must be familiar with additional SAS procedures (PROC MI and PROC MIANALYZE) and plug-ins (IVEware) in order to use the linked results.

Survey of Linkage Software

To help understand various available products for data linkage, and how they compare with CODES methodology, we searched online for products that were described as enabling linkage. Selected products are profiled in this section, but except for LinkSolv, the profile only extends to knowledge as gleaned from online descriptions. Alternatives to LinkSolv were not purchased nor tested for this overview.

LinkSolv

LinkSolv (http://www.strategicmatching.com/) is the commercial version of the linkage software used by the CODES Data Network. (MH 2004, 2006, Matching 2013) LinkSolv runs in Microsoft Access but has the ability to connect to SQLServer in order to accommodate larger data files. LinkSolv uses the methodology described in the above sections to derive match probabilities and generate imputed matched sets. The software provides the ability to compare any type of matching variable. Character strings, numbers, dates, and latitude/longitude coordinates can all be compared appropriately within LinkSolv. Each of the comparison methods allows the user to require an exact match on the two fields or for the fields to agree within a specified window. For example, depending on the user's preference, two numeric values can be given an agreement weight if they match exactly, or if one is within a specified numeric distance from the other, or if one is within a specified percentage of the other. Similar utilities exist for strings, dates, and geographic location fields. There are also many standardization routines available to the data user to facilitate preparing multiple files for linkage within the program, rather than having to do this prior to loading the data. The software also provides the ability for users to define their own data standardization and comparison routines.

To aid users in constructing their linkage models, LinkSolv offers the ability to simulate databases. The advantage of simulating data is that the user knows which pairs of records are true matches and which ones are not. After linking the simulated data, methods can be used to determine the overall fit of the linkage and the percent of true matches identified. Once a researcher has achieved a good linkage result with the simulated data, the linkage algorithm can then be applied to the actual data.

LinkSolv also provides a number of tools to assess the quality of the linkage once it has been completed. One can assess the impact of any comparison tolerances that were incorporated into the match and then can determine the effect of widening or shortening the tolerance. One can also evaluate whether or not match fields have dependent agreements or disagreements. Finally, all match and testing information is collated onto a final report that allows the user to make modifications to the following year's linkage.

LinkSolv provides the ability to run a self-match or unduplication (link a database against itself searching for multiple records for the same individual). LinkSolv has also extended the probabilistic linkage algorithms to matching three files at once.

LinkageWiz

LinkageWiz (http://www.linkagewiz.net) is a commercial linkage product that is available for purchase online. The number of records that a file can hold is dependent on your licensing agreement and purchasing fee. A user may purchase an agreement that allows unlimited records (defined as 4 – 5 million records). A standard amount of maintenance and technical support is included with a license but additional support and upgrades can also be purchased. LinkageWiz is a standalone piece of software and does not run within another program.

Data from a number of sources can be imported into LinkageWiz. The software also provides the ability to standardize and clean. LinkageWiz allows for matching, either exactly or within a specified tolerance, on first and last name, date of birth, date of death, date of event, sex, address,

ZIP code, SSN, medical record number, business name, email address, medical diagnoses, Medicare Number, event type, and up to five user defined data fields. LinkageWiz provides the ability to conduct deterministic or probabilistic linkages. Probabilistic weights can be calculated as described above or user defined.

LinkageWiz does not appear to provide built in tools to simulate data or evaluate the fit of a linkage model. There is also no mention of the use of MCMC to refine the linkage model. There is not a built in mechanism for creating imputed matched sets.

LinkageWiz does appear to provide the ability to deduplicate files.

The Link King

The Link King (http://the-link-king.com/) is a free (public domain) SAS/AF product. There is a limitation of 99,999,999 records on the size of files that can be used. Users interact with The Link King through a GUI so one does not need to be an experienced SAS programmer to use it. The Link King adapted its linkage algorithms from MEDSTAT which was used by the Substance Abuse and Mental Health Services Administration's (SAMHSA) Integrated Database.

The Link King can use files stored as SAS data sets, SPSS portable files, comma delimited files, and Excel spreadsheets. It appears that The Link King only allows for very specific variables: SSN, date of birth, first name, middle name, last name, maiden name, gender, race, and client ID. First and last name and either date of birth or SSN are required to run a linkage in Link King. Matching variables can either be required to match exactly or within a specified tolerance. Since The Link King restricts the fields that are available, it tries to guide the user through the variable standardization process based on the matching variable types. The software also allows users to specify certain values of fields to be designated as missing during the standardization process. The Link King provides the ability to conduct either probabilistic or deterministic linkage.

The Link King does not appear to provide built in tools to simulate data or evaluate the fit of a linkage model. While not having the same level of validation tools as LinkSolv, The Link King does allow users to generate random samples of linked records to review the quality of the linkage model. There is not a built in mechanism for creating imputed matched sets.

The Link King does appear to provide the ability to de-duplicate files.

Link Plus

Link Plus (http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm) is a free probabilistic linkage program developed by the Centers for Disease Control and Injury Prevention's (CDC) Division of Cancer Prevention and Control to support the development and maintenance of cancer registries. Link Plus is a standalone application that can be used to detect duplicates in a single registry or link the registry to external files. Link Plus requires data files to be fixed width text or delimited. Link Plus can support files as large as 4.5 – 4.8 million records.

Link Plus allows the following data fields: last and first names, middle names, dates, SSN, generic strings, and ZIP codes. User defined variables can also be included. Algorithms are incorporated into the software such that a user can require an exact match or fuzzy match on

each of the data types. Link Plus calculates probabilistic match weights. There does not appear to be built in functions to help users standardize or clean data.

Link Plus does not appear to provide built in tools to simulate data or evaluate the fit of a linkage model. There is also no mention of the use of MCMC to refine the linkage model. There is not a built in mechanism for creating imputed matched sets.

FRIL

FRIL (http://fril.sourceforge.net/), or the Fine-Grained Records Integration and Linkage Tool, is a free open source tool that enables record linkage. FRIL was developed as a joint project between Emory University and the CDC. Data from text files, Excel, and JDBC databases can be imported into FRIL.

FRIL appears to be very flexible with the types of matching fields and standardization routines available. Users can input dates, names, character strings, and numeric values for matching fields. During the standardization process, users can concatenate two matching fields or split a single field into multiple matching variables. Match variables can either be required to agree exactly or within a certain distance using a fuzzy match. Unlike other programs described above, users can specify an upper tolerance for which pairs receive a full agreement weight and a lower tolerance for which pairs receive a full disagreement weight. Pairs that have a comparison which fall between the two tolerances receive a partial agreement weight. Another feature that is unique to FRIL is that in addition to allowing users to define typical blocking schemes used in normal probabilistic linkages, users can also look for pairs within a neighborhood, a specified mathematical distance measure, of each other.

FRIL does not appear to provide built in tools to simulate data or evaluate the fit of a linkage model. There is also no mention of the use of MCMC to refine the linkage model. There is not a built in mechanism for creating imputed matched sets.

FRIL does appear to provide the ability to de-duplicate files.

Table 1.1.1 contains summary information regarding the reviewed software.

Table 1.1.1. Summary of Linkage Software Capabilities.					
	LinkSolv	LinkageWiz	The Link King	Link Plus	FRIL
Commercial Versus Free	Commercial	Commercial	Free	Free	Free
Platform	Access	Stand Alone	SAS	Stand Alone	Stand Alone
Standardization/ Data Cleaning Available	Yes	Yes	Yes	No	Yes
Probabilistic Match Weights	Yes	Yes	Yes	Yes	Yes
Deterministic Match Weights	No	No	Yes	Yes	No
Customizable Match Weights	No	Yes	No	Yes	Yes
Custom Variable Types	Yes	Up to 5	No	Yes	Yes
Fuzzy Matching Comparisons	Yes	Yes	Yes	Yes	Yes
Model Evaluation Tools	Yes	No	Limited	No	No
MCMC and Imputation of Missing Links	Yes	No	No	No	No
Deduplication/ Self Match	Yes	Yes	Yes	Yes	Yes

References

Brubacher, J. R., Chan, H., Fang, M., Brown, D., & Purssell, R.(2013). Police documentation of alcohol involvement in hospitalized injured drivers. *Traffic Injury Prevention,14*(5), 453-460.

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

Dean, J. M., Vernon, D. D., Cook, L., Nechodom, P., Reading, J., & Suruda, A. (2001). Probabilistic linkage of computerized ambulance and inpatient hospital discharge records: A potential tool for evaluation of emergency medical services. *Annals of Emergency Medicine, 37*(6), 616-626.

Gonzalez, R. P., Cummings, G. R., Phelan, H. A., Harlin, S., Mulekar, M., Rodning, C. B.(2007). Increased rural vehicular mortality rates: Roadways with higher speed limits or excessive vehicular speed? *The Journal of Trauma, 63*(6), 1360-1363.

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

Mango, N. & Garthe, E. (2007). Statewide tracking of crash victims' medical system utilization and outcomes. *The Journal of Trauma, 62(2)*, 436-460.

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

McGlincy, M. (2006). Using test databases to evaluate record linkage models and train linkage practitioners. Joint Statistical Meetings. Seattle, WA, 3404-3410.

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

Senserrick, T. Ivers, R., Boufous, S., Chen, H.-Y., Norton, R., Stevenson, M., et al. (2009). Young driver education programs that build resilience have potential to reduce road crashes. *Pediatrics, 124*(5), 1287-1292.

Strategic Matching. (2013).

Thomas, A. M., Thygerson, S. M., Merrill, R. M., & Cook, L. J., (2012). Identifying work-related motor vehicle crashes in multiple databases. *Traffic Injury Prevention 13*(4), 348-354.

Thygerson, S. M., Merrill, R. M., Cook, L. J., & Thomas, A. M. (2011a). Comparison of factors influencing emergency department visits and hospitalization among drivers in work and nonwork-related motor vehicle crashes in Utah, 1999-2005. *Accident Analysis & Prevention 43*(1), 209-213.

Thygerson, S. M., Merrill, R. M., Cook, L. J., Thomas, A. M., Wu, A. C. (2011b). Epidemiology of motor vehicle crashes in Utah. *Traffic Injury Prevention, 12*(1), 39-47.

Vladutiu, C. J., Poole, C., Marshall, S. W., Casteel, C., Menard, M. K., & Weiss, H. B. (2013). Pregnant driver-associated motor vehicle crashes in north carolina, 2001-2008. *Accident Analysis & Prevention, 55*, 165-71.

Chapter 2: A Comparison of High-Probability, Multiply Imputed, and Maximum a Posteriori Matched Sets

Introduction

As large databases often containing thousands of records have become more available, computers more powerful, and probabilistic linkage software more prevalent, studies using probabilistic linkage have become more widespread. As described in Chapter 1, probabilistic linkage is a method for combining information from different databases into a single dataset for analysis by comparing multiple fields common to each database (Newcombe, 1988; Jaro, 1989;, Roos and Wajda, 1991; Bell et al., 1993; Jaro, 1995; Cook et al., 2001). Comparisons of multiple fields lead to the determination of the probability that two records refer to the same person and event and should therefore be linked. High probabilities assigned to pairs can be achieved by using specific and accurate matching fields (Jaro, 1995). Databases used for probabilistic linkage are usually administratively collected for purposes other than linkage study.(Jaro 1989, Bell et al., 1993; Chamberlayne et al., 1998; & Dean et al., 2001) As a result, the quality or completeness of the information is outside of the researchers control and can be variable. If linkage fields are frequently missing or erroneous, then the linkage may fail to identify many true matches. If missing or erroneous data are related to some mechanism, such as injury severity, biases may be unintentionally introduced into the linked dataset. These issues should be taken into account whenever a probabilistic linkage method is being considered.

High-probability, multiply imputed, and maximum a posteriori (MAP) matched sets can each result from probabilistic linkage. High-probability and MAP methods result in one matched set, whereas the imputation method results in multiple matched sets (see Chapter 1 for more details). The goal of this chapter is to compare high-probability, imputed, and MAP matched sets across differing levels of information available to the linkage.

Methods

In order to compare the impact of analyzing high-probability, multiply imputed, and MAP matched sets, we performed a series of linkages and analyses on two simulated datasets in which we could identify true matches.

Data source

We used the Utah motor vehicle crash (MVC) database obtained from the Utah Department of Transportation, Division of Traffic and Safety for years 1992 (n=114,016 people), 1993 (n=126,276 people), 1997 (n=143,213 people), and 2003 (n=133,327 people). This database contains information on all reported MVCs in Utah. A MVC is reportable if it occurs on public roadways and results in at least one injury or fatality or at least $1,500 in property damage. The data are collected on reports completed by the responding police officer at the scene of the MVC and include identifying information on persons involved (i.e., name, birth date, sex) as well as details about the MVC: each person's injury status, seating position, and restraint use; as well as detailed information about the time, location, and type of MVC; and vehicles involved.

From the Utah MVC database, we created two simulated datasets (referred to as File A and File B) with a set of known true matches. This was accomplished by first selecting 10,000 records from the MVC database and placing these records in both File A and File B. We selected an additional 180,000 records from the MVC database that were different from the first selection and inserted 90,000 records in File A and the remaining 90,000 records in File B. The result was two simulated datasets, each with 100,000 records, 10,000 of which should match exactly to a single record and 90,000 that should not match to any records in the other dataset.

A typical CODES probabilistic study links MVC databases to hospital databases in order to associate hospital outcomes with MVC characteristics.(Smith 1984, Cook *et al.* 2000, Conner *et al.* 2010, Olsen *et al.* 2010, Thygerson *et al.* 2011, Thomas *et al.* 2012) To replicate a typical study of CODES data (linked MVC and hospital databases), we generated a hospital outcome for File B: simulated log hospital charges. We also considered hospitalization status, which indicated whether or not a case went to the hospital (linkage status). Cases that linked from File A to File B were considered hospitalized cases; otherwise, cases were considered non-hospitalized. Simulated log hospital charges were obtained from a linear regression model with normally distributed errors. The values of the parameters and variance of the error term used to simulate log hospital charges were derived from a linear regression model of observed log hospital charges from previously linked Utah CODES data. The model is summarized by Equation (1) below:

$$\begin{aligned} simulated\ log\ hospital\ charges = {} & 6.2993 + 0.0379 * occupant\ sex + 0.0057 \\ & * occupant\ age + 0.0875 * MVC\ location + 0.4795 \\ & * police\ suspicion\ of\ alcohol\ or\ drug\ use + \varepsilon \sim N(0, 0.9527) \end{aligned} \quad (1)$$

The following covariates were binary: occupant sex (male = 1, female = 0), MVC location (rural = 1, urban = 0) and police suspicion of alcohol or drug use (suspected = 1, not suspected = 0). Occupant age was a continuous covariate.

Linkages

We performed three linkages with File A and File B, varying the amount of available information in each linkage. High information variables, such as name and birth date, may be unavailable for the linkage process for a variety of reasons. Therefore, we used some variable combinations that virtually guaranteed a perfect linkage and other combinations that made use of very little information, making identification of correct matched pairs less certain. We calculated the minimum potential match probability to quantify the quality of linkage variable combinations using methods outlined elsewhere.(Cook *et al.,* 2001). The linkage variable combinations used in this study, along with the corresponding minimum potential match probabilities, are summarized in Table 1.2.1. All linkages were conducted using Strategic Matching LinkSolv version 8.3.0328.(McGlincy, 2000)

Table 1.2.1: Summary of linkage variable combinations and corresponding minimum potential match probability for each linkage performed.			
	Linkages		
	A	**B**	**C**
Minimum potential match probability	0.999	0.470	0.027
Occupant first name	X		
Occupant first initial		X	
Occupant last name	X		
Occupant last initial			
Occupant birth date	X		
Occupant birth month and day		X	
Occupant age			X
MVC date	X	X	X
MVC time	X		
MVC hour		X	X
Occupant sex	X	X	X
MVC county	X		X

For each of the three linkages, we produced high-probability, multiply imputed, and MAP matched sets. For high-probability matched sets, we only retained matched pairs that achieved a probability of 0.90 or greater. Multiply imputed matched sets are designed to include high and low probability matched pairs (McGlincy, 2004). All matched pairs that achieved a probability of 0.01 or greater were used to generate a distribution of candidate matched pairs from which to sample. For this analysis, we created five imputed matched sets. Like multiply imputed matched sets, all matched pairs that achieved a probability of 0.01 or greater were considered for the MAP matched sets. This method resulted in only one matched set per linkage.

We conducted analyses in SAS and used PROC MIANALYZE to combine results from different imputations.(Schafer, 1997; SAS Institute Inc., 2002)

Analysis

We used sensitivity and specificity to evaluate each linkage across high-probability, multiply imputed, and MAP matched sets. Sensitivity is how well each linkage identified true matched pairs and specificity is the ability of the linkage to exclude false matched pairs. In addition to accounting for the percent of correct and false matches obtained, it is important to compare the resulting distribution of the match variables to determine if the sample of matched records accurately reflects distributions from the underlying population. Thus, we compared distributions of high-probability, multiply imputed, and MAP matched sets against the distribution of correct matched pairs for each of the three linkages across three linkage variables: occupant age, MVC county, and MVC hour. These variables were selected for the analysis because they have many different values and are commonly used in CODES-type analyses. We used Kolmogorov-Smirnov tests to determine statistical differences between distributions.

We fit regression models to both a continuous and binary outcome to study differences in statistical inference between high-probability and imputed matched sets. We used simulated log hospital charges as the continuous outcome and hospitalization status as the binary outcome. For both outcomes, we applied the same model to each linkage for high probability, multiply imputed, and MAP matched sets. Covariates used to model these outcomes were the same as the covariates shown in Equation (1). Estimated coefficients and corresponding 95-percent confidence intervals from the continuous outcome were compared to the coefficients used to generate simulated log hospital charges [see Equation (1)]. Because hospitalization status was not simulated but derived from linkage status, we fit a model to hospitalization status using the 10,000 true matched pairs as the hospitalized cases. This result would have been achieved if the linkage perfectly identified the 10,000 true matched pairs without including any false matches. The coefficients derived from this model are considered the true value of the parameter when comparing the coefficients and 95-percent confidence intervals generated from a model fit to the results of each of the five linkages.

Results

Sensitivity and Specificity

The total matches identified, as well as the sensitivity and specificity, for high-probability, multiply imputed, and MAP matched sets are displayed in Table 1.2.2. For high-probability, multiply imputed, and MAP matched sets, linkages with more information (linkage A) have higher sensitivity and specificity than linkages with less information (linkage C).

Across all linkages, high-probability matched sets had specificity above 99.0 percent, indicating that most of the matched pairs identified are correct. Sensitivity for high-probability matched sets decreases substantially as the minimum potential match probability decreases with linkage C identifying only 27.3% of correct matched pairs. Interpreting sensitivity in conjunction with specificity indicates that while the majority of matched pairs identified are true matched pairs, linkages with poor information, as illustrated by linkage C, produce very few matched pairs.

For multiply imputed and MAP matched sets, at least 98.6 percent of matched pairs identified by linkages A and B are true matched pairs. Linkage C identifies fewer true matched pairs, with 89.6 percent and 89.2 percent identified by multiply imputed and MAP matched sets, respectively. Multiply imputed and MAP matched sets had higher sensitivity than the high-probability matched sets. Linkages A and B identify 98.0 percent or more of true matched pairs for both multiply imputed and MAP matched sets. Multiply imputed matched sets from linkage C identified 73.1 percent of true matched pairs while MAP matched sets from linkage C identified 75.0 percent of true matched pairs. These findings indicate that while multiply imputed and MAP matched sets identify more matched pairs, many of which are true matched pairs, these matched sets also include false matched pairs.

Table 1.2.2: Total matches, sensitivity, and specificity for high-probability and imputed match sets.				
		Linkages		
		A	**B**	**C**
Minimum potential match probability		0.999	0.470	0.027
High-probability matched sets	Total matches	10,001	10,305	3,265
	Sensitivity	0.998	0.995	0.273
	Specificity	> 0.999	0.996	0.994
Multiply imputed matched sets	Total matches	10,028	11,121	16,653
	Sensitivity	> 0.999	0.983	0.731
	Specificity	> 0.999	0.986	0.896
MAP matched sets	Total matches	10,027	11,039	17,235
	Sensitivity	> 0.999	0.980	0.750
	Specificity	> 0.999	0.986	0.892

Linkage Variables

Figures 1.2.1 to 1.2.3 compare the distribution of high-probability, multiply imputed, and MAP matched sets from linkages A through C with the distribution of true matched pairs for occupant age, MVC county, and MVC hour, respectively. Across the three linkage variables, high-probability, multiply imputed, and MAP matched sets are nearly identical to the distribution of correct matched pairs for linkages A and B (all: $p>0.690$). These distributions show that in moderate to high information settings, there is no difference between high-probability, multiply imputed, and MAP matched sets for these three linkage variables. Under linkage C, the distribution of high-probability matched sets is significantly different from the distribution of true matched pairs for occupant age ($p<0.001$) and MVC county ($p=0.027$); however, it is not significantly different for MVC hour ($p=0.476$). The distributions of multiply imputed and MAP matched sets is not significantly different from the distributions of true matched pairs across these three linkage variables (all: $p>0.346$).

Figure 1.2.1: Distribution of high-probability, imputed, and MAP matched sets for occupant age.

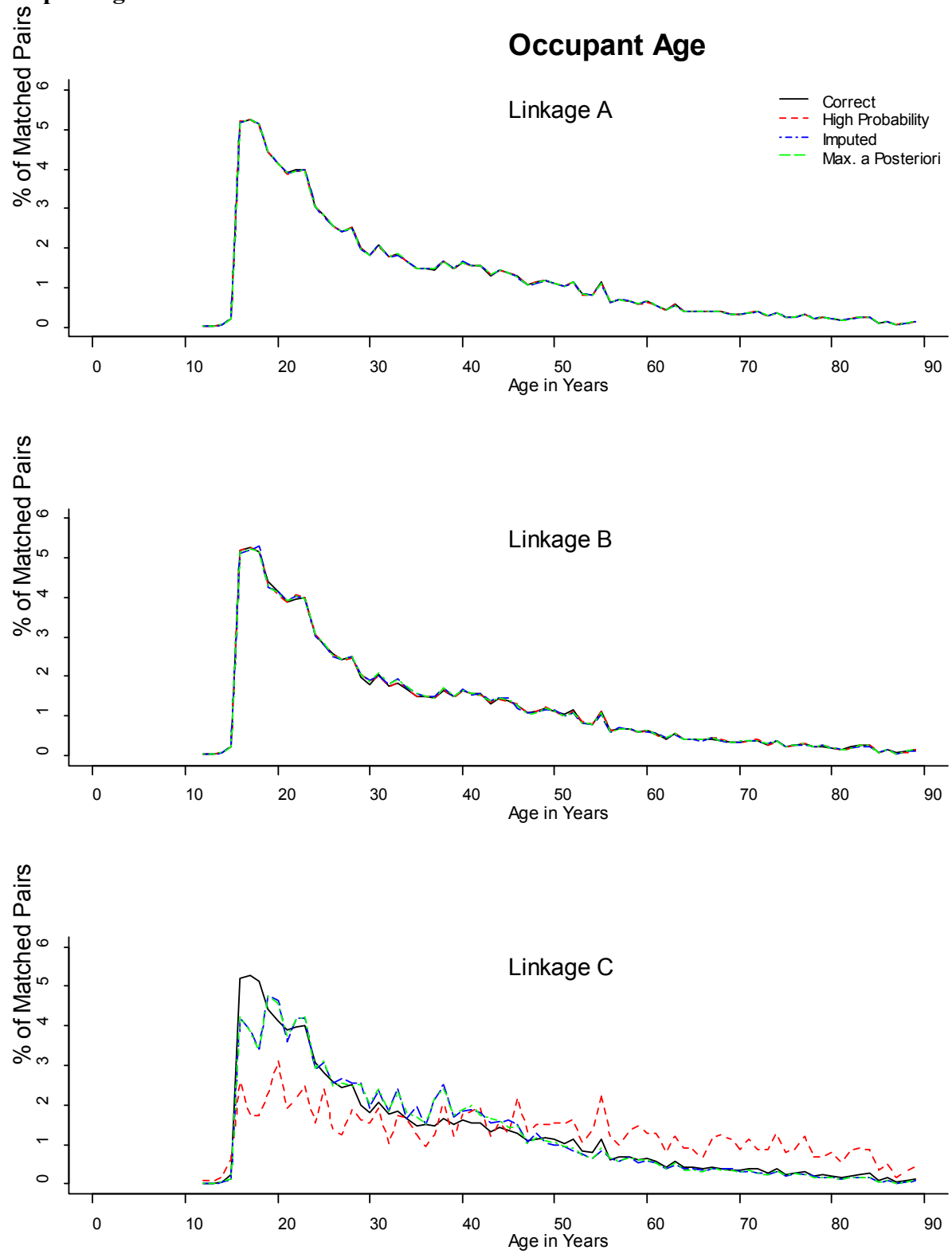


Figure 1.2.2: Distribution of high-probability, imputed, and MAP matched sets for MVC county.

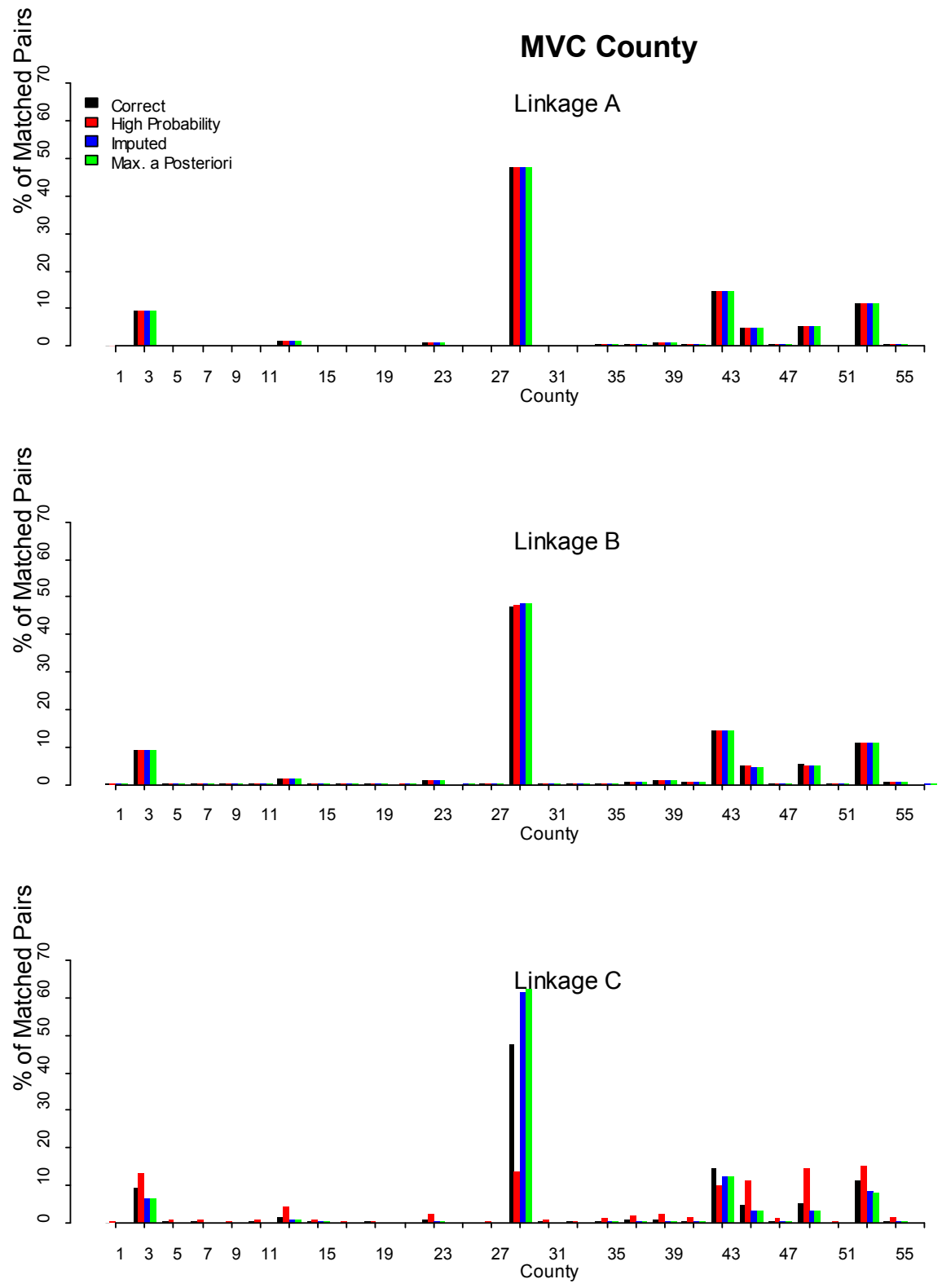
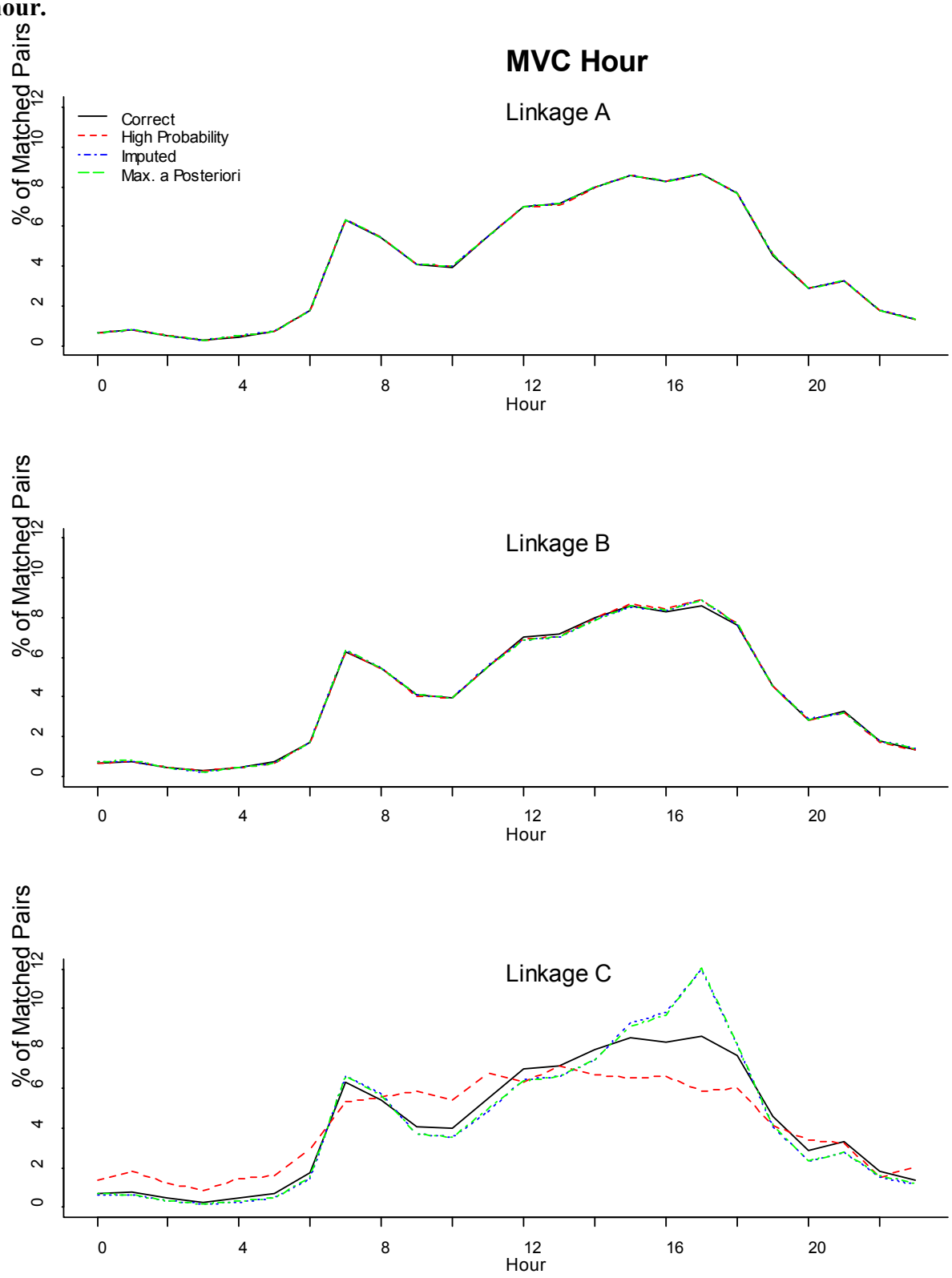


Figure 1.2.3: Distribution of high-probability, imputed, and MAP matched sets for MVC hour.

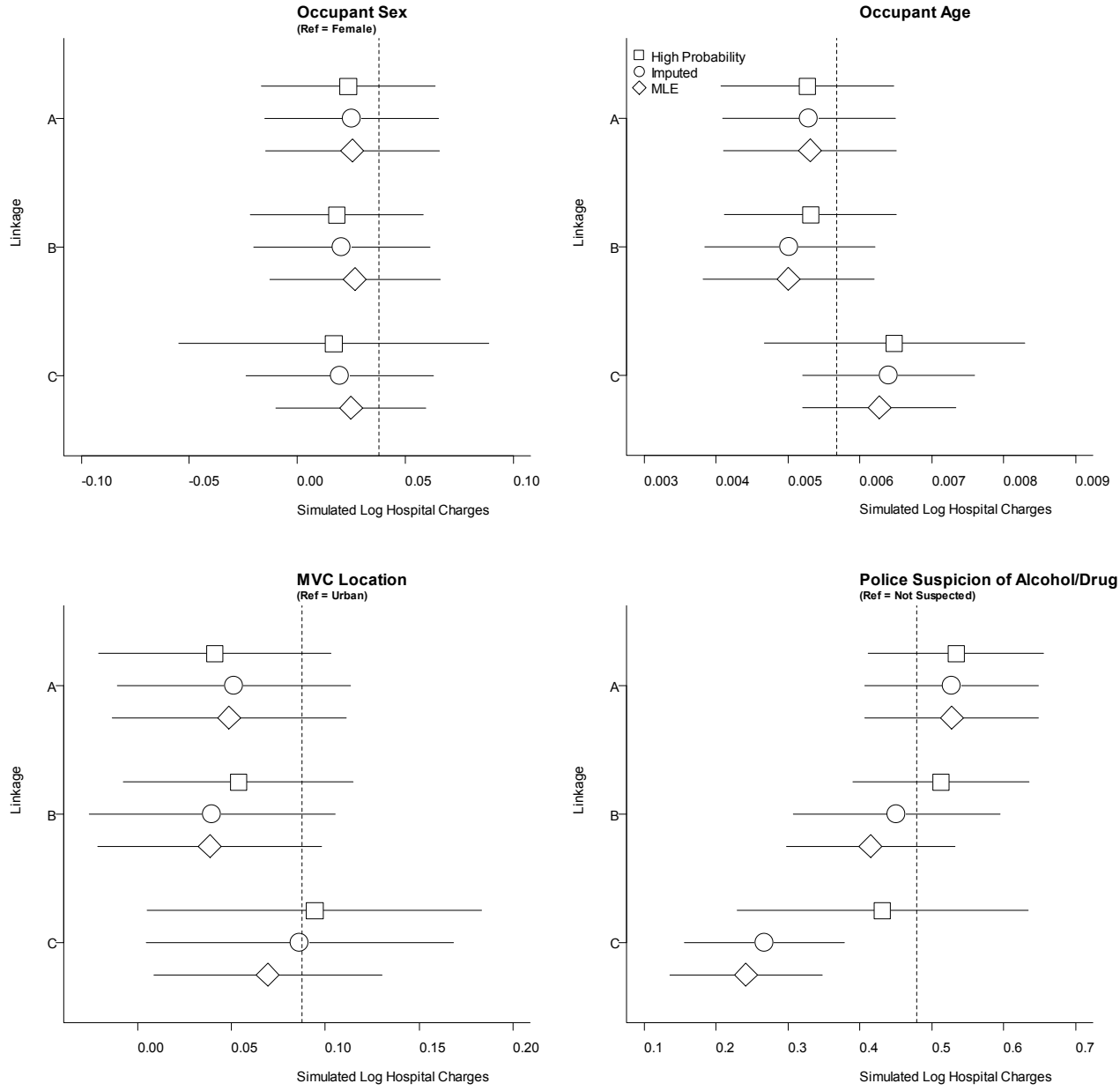


The differences between the distribution of high-probability matched sets and the distribution of true matched pairs for occupant age and MVC hour, as observed in linkage C, are apparent in Figures 1.2.1 and 1.2.3. Occupants ages 80 to 89 account for 1.63 percent (n=163) of correct matched pairs in the true match population. High-probability matched sets estimate the proportion of occupants ages 80 to 89 at over three times (5.79%, n=189) the true proportion. An underestimate of occupants ages 80 to 89 years exists in multiply imputed and MAP matched sets, to a much smaller magnitude than the overestimate seen among high-probability matched sets (1.16%, n=193 for imputed; 1.12%, n=193 for MAP). Additionally, nearly 50% (n=4,753) of all correct matched pairs are from county 35; however, high-probability matched sets estimate that only 13.5% (n=442) of matched pairs are from county 35. The least populous counties are overestimated by high-probability matched sets. Multiply imputed and MAP matched sets do not show this same pattern. High-probability matched sets from linkage C tend to overstate rare values, as seen by the large percent of matched pairs that are older and have MVCs occurring in less populous counties. Multiply imputed and MAP matched sets follow the distribution of correct matched pairs more closely.

Estimating simulated log hospital charges

We modeled simulated log hospital charges to understand the impact of high-probability, multiply imputed, and MAP matched sets on an analysis of a hospital outcome commonly studied by CODES projects. Figure 1.2.4 summarizes the coefficients and corresponding 95 percent confidence intervals of high-probability, multiply imputed, and MAP matched sets from linkages A through C. The coefficient estimates of high-probability, multiply imputed, and MAP matched sets from linkages A and B are not significantly different from the coefficients used to generate simulated log hospital charges, as assessed through 95 percent confidence intervals. The coefficients from high-probability matched sets are not significantly different from the coefficients used to generate simulated log hospital charges for linkage C; however, the confidence intervals associated with high-probability matched sets in linkage C are nearly twice as wide (>1.5 times) for occupant sex, occupant age, and police suspicion of alcohol or drug use compared to the same confidence intervals from multiply imputed and MAP matched sets. The coefficient for police suspicion of alcohol or drug use from multiply imputed and MAP matched sets is significantly different from the coefficient used to generate simulated log hospital charges for linkage C.

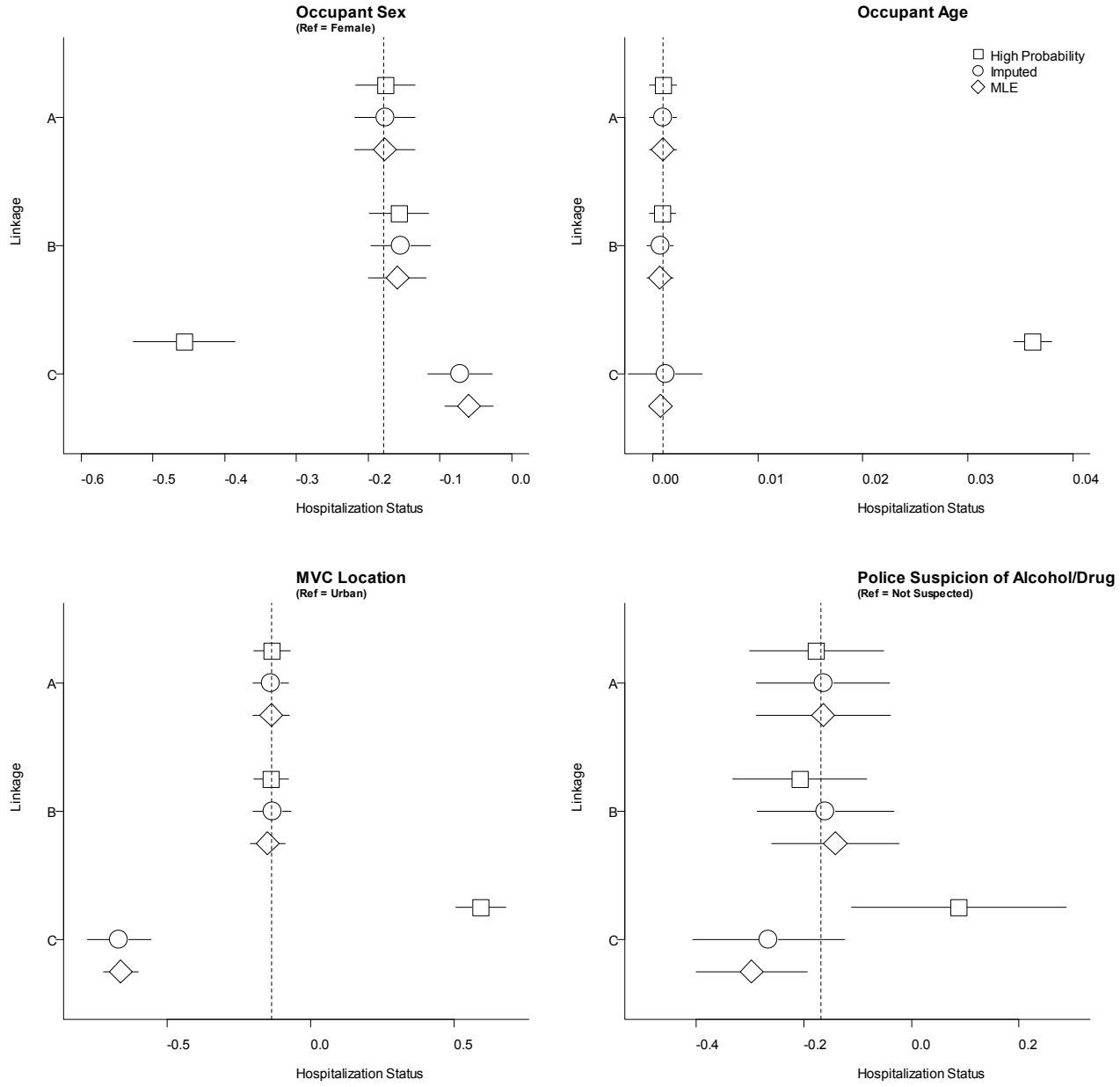
Figure 1.2.4: Coefficient estimates and corresponding 95-percent confidence intervals used to model simulated log hospital charges for high-probability, imputed, and MAP matched sets from linkages A through C compared to coefficients used to generate simulated log hospital charges (dashed line)



Estimating hospitalization status

To understand the impact of high-probability, multiply imputed, and MAP matched sets on an analysis of a binary outcome, we modeled hospitalization status. Figure 1.2.5 contains the coefficients and corresponding 95-percent confidence intervals of high-probability, multiply imputed, and MAP matched sets from linkages A through C. The coefficients for occupant sex, occupant age, MVC location, and police suspicion of alcohol or drug use from high-probability, multiply imputed, and MAP matched sets are not significantly different from the true coefficients across linkages A and B. These coefficients are significantly different from the true coefficients for high-probability matched sets in linkage C. The coefficients from multiply imputed matched sets are not significantly different from the true coefficients across all linkages with one exception: the coefficient for police suspicion of alcohol or drug use in linkage C is significantly different from the true coefficient. The coefficients for occupant sex and MVC location from MAP matched sets in linkage C are significantly different from the true coefficients. When considering the binary outcome of hospitalization status, multiply imputed matched sets more accurately modeled the true match population compared to high-probability and MAP matched sets.

Figure 1.2.5: Coefficient estimates and corresponding 95-percent confidence intervals used to model hospitalization status for high-probability, imputed, and MAP matched sets from linkages A through C compared to the true coefficients (dashed line)



Other Considerations

Although multiply imputed and MAP matched sets performed similarly well in this study, this may not necessarily be the case for other studies where researchers have access to different variables or qualities of variables. Another consideration is that the standard error of estimates made from multiply imputed matched sets, when properly analyzed, takes into account the variability between imputations (Rubin, 1987). MAP matched sets only use a single imputed set, so standard variance calculation will ignore the variation that may have been seen in different imputations. For studies involving statistical significance, this difference could be important. Newgard et al. (2012) also found that probabilistic linkage using multiple imputations minimized bias and better preserved the sample size when compared to probabilistic linkage using a single matched set.

2.4 Conclusions

We compared high-probability, multiply imputed, and MAP matched sets across different minimum potential match probabilities. We conducted three linkages on simulated data with varying amounts of available information in each linkage and compared the resulting distributions and inference between high-probability, multiply imputed, and MAP matched sets and the true match population. We found that when minimum potential match probability was high, high-probability, multiply imputed, and MAP matched sets were not significantly different from the true match population. When minimum potential match probability was low, multiply imputed and MAP matched sets were representative of the true match population with multiply imputed matched sets performing slightly better than MAP matched sets when modeling the binary outcome of hospitalization status. High-probability matched sets were not representative of the true match population when minimum potential match probability was low. High-probability matched sets underrepresented common values so thereby overrepresented rare values; however, the matched pairs identified were likely correct.

The type of linkage methodology that a researcher chooses to employ will likely depend on the goal of the linkage. When it is important to follow an individual, as in health registries, the matched pairs from high-probability matched sets will likely be true matches and follow the correct individual, but may still exclude many lower probability true matches. When conducting population level analyses, the resulting linked dataset from multiply imputed matched sets, which are more representative of the true match population, are desirable.

References

Bell, R. M., Kessey, J., & Richards, T. (1993). The urge to merge: A computational method for linking datasets with no unique identifier. *Proceedings of the SAS Users Group 18 Conference,* New York, 1-6.

Chamberlayne, R., Green, B., Barer, M. L., Hertzman, C., Lawrence, W. J., & Sheps, S. B. (1998). Creating a population-based linked health database: A new resource for health services research. *Canadian Journal of Public Health, 89*(4), 270-273.

Conner, K. A., Xiang, H., Smith, G. A. (2010). The impact of a standard enforcement safety belt law on fatalities and hospital charges in Ohio. *Journal of Safety Research, 41*(1), 17-23.

Cook, L. J., Knight, S., Olson, L. M., Nechodom, P. J., & Dean, J. M. (2000). Motor vehicle crash characteristics and medical outcomes among older drivers in Utah, 1992 - 1995. *Annals of Emergency Medicine, 35*(6), 585-591.

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

Dean, J. M., Vernon, D. D., Cook, L., Nechodom, P., Reading, J., Suruda, A. (2001). Probabilistic linkage of computerized ambulance and inpatient hospital discharge records: A potential tool for evaluation of emergency medical services. *Annals of Emergency Medicine, 37*(6), 616-626.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *American Statistical Association, 84*(406), 414-420.

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

McGlincy, M. H. (2000). Linksolv. Strategic Matching. Morrisonville, New York.

McGlincy, M. H. (2004). A bayesian record linkage methodology for multiple imputation of missing links. *Proceedings of the Survey Research Methods Section, American Statistical Association,* 4001-4008.

Newcombe, H. B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration, and business*. New York: Oxford University Press.

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

Olsen, C. S., Cook, L. J., Keenan, H. T., & Olson, L. M. (2010). Driver seat belt use indicates decreased risk for child passengers in a motor vehicle crash. *Accident Analysis & Prevention, 42*(2), 771-777.

Roos, L. L., & Wajda, A. (1991). Record linkage strategies. Part I: Estimating information and evaluating approaches. *Methods of Information in Medicine, 30*(2), 117-123.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

SAS Institute Inc. (2002). SAS Software. 9.2 ed. Cary, NC: SAS Institute Inc..

Schafer, J. L. (1997). Analysis of incomplete multivariate data. Boca Raton, FL: Chapman & Hall/CRC.

Smith, M. E. (1984). Record linkage: Present status and methodology. *Journal Clinical Monitoring and Computing, 13*(2-3), 52-71.

Thomas, A. M., Thygerson, S. M., Merrill, R. M., & Cook, L. J. (2012). Identifying work-related motor vehicle crashes in multiple databases. *Traffic Injury Prevention, 13*(4), 348-354.

Thygerson, S. M., Merrill, R. M., Cook, L. J., & Thomas, A. M. (2011). Comparison of factors influencing emergency department visits and hospitalization among drivers in work and nonwork-related motor vehicle crashes in Utah, 1999-2005. *Accident Analysis and Prevention, 43*(1), 209-13.

Chapter 3: Analysis of Match Probability in Probabilistic Linkage

Introduction

Record linkage allows the combination of different databases into one extensive dataset for analysis. For example, linking the records from motor vehicle crashes (MVC), the emergency department (ED), and the hospital inpatient database allows us to perform the analysis of crash with regards to the medical outcomes. Probabilistic record linkage uses properties of variables common to databases to determine the probability that two records refer to the same person. This probability is called match probability. When linking the records between the MVC and hospital records, variables such as name, date of birth, date of incident, and county of the MVC and the hospital can be used in linkage modeling. Linkage as performed in CODES results in multiply imputed datasets, each with the possibility of different links between MVC and hospital records. Proper analysis of multiply imputed datasets accounts for the uncertainty inherent in the linkage process (Cook et al., 2001). Variability in the linkage modeling across multiple CODES States can arise due to the presence or absence of useful identifiers in a State's databases. It is suspected that the size of a State's MVC and hospital file also have an effect on linkage results. States with larger databases need more information to identify the correct matches. It is expected that pairs of records with high match probabilities are more likely to appear in all imputed datasets. The purposes of the chapter are: (1) identify the commonly used identifiers used to link the MVC and the hospital records, (2) evaluate the effects of commonly and uncommonly used identifiers on match probabilities and on match weights, (3) examine how MVC file size can affect match probabilities, and (4) ascertain whether or not MVC-hospital pairs of records with high match probabilities are more likely to appear in all imputations.

Definitions

Matched pairs or paired records signify any MVC records being linked to the hospital records. Identifiers are the variables that are common to both of the databases that are being linked. Match probability is the probability that two records refer to the same person given the values of the identifiers that are used in the linkage modeling. Match weight is the paired record's sum of all weights assigned to each identifier used in the linkage model. The weight for each identifier is determined as a function of the odds of a match. A more complete discussion of probabilistic linkage is given by Cook et al. (2001), Crash Outcome Data Evaluation System (2010), Jaro (1995), and McGlincy (2004).

Hypothesis

When linking records between the MVC and hospital databases, the common identifiers can include date of birth (DOB), first and last names, date of incidents, county of incident, sex, and age. We predict that linkage models with many common identifiers will tend to have more matched pairs with high match probabilities than the models with few common identifiers. We also expect matched pairs that were linked with commonly used identifiers to have higher match weights than matched pairs that were linked with rarely used identifiers. In addition, we hypothesize that the large MVC file sizes are negatively correlated with match probabilities. That is, the larger the MVC file, the more the distribution of the high match probabilities will shift

towards the left. Lastly, we speculate that paired records with high match probabilities will appear in many, if not all, imputed datasets.

Methods

We used a total of 15 probabilistically linked MVC and hospital records from 8 CODES States for the MVC years 2005–2008. Each linkage result has five imputed datasets. Individual CODES analysts were responsible for linking data from their own State MVC and hospital files and used the same probabilistic linkage software, CODES2000 (McGlincy, 2000). CODES2000 generates diagnostic reports which can be used to evaluate the linkage model. Linkage diagnostic reports for each State were used to identify the widely used (> 6 States), moderately used (3-5 States), and rarely used (1-2 States) identifiers. We obtained matched pairs and their match probabilities and match weights from the General Use Model (GUM) data prepared by eight CODES States. The GUM was developed through a joint effort between CODES and the NHTSA State Data System and is a data mapping of a standardized set of data elements routinely collected on police MVC reports and in hospital databases. Graphical methods are used to represent the distributions of match probabilities in eight CODES States. We evaluated performance of each identifier using the match probability and match weights distributions. We categorized MVC file size into four groups (0 – 99,000; 100,000-199,999; 200,000-299,999; 300,000-399,999; and 400,000 – 799,999), and evaluated their match probability percentiles. We use graphical methods to illustrate the likelihood of linked pairs appearing in multiply imputed datasets.

Results and Discussion

Identify common/uncommon identifiers

Table 1.3.1 shows the list of variables used in the linkage model by eight CODES States. This Table also categorizes each identifier based on its frequency of use. It also totals the number of identifier used by each CODES State. Incident date, sex, age, DOB, first name, last name, and seating position were most widely used identifier in the model. The rarely used identifiers include Hospital ZIP, Injury flag, Latitude/Longitude, SSN, and Race. Though used by only few CODES States, identifiers such as SSN and latitude / longitude may have a high discriminative power. However, they may suffer from the lack of reliability since they are prone to data entry error. Home State, middle name, and race are only used by one State, though this information may be readily available in the other CODES States. It could be argued that, since most of the occupants in the MVC and patients at the hospital usually are resident of the same State as where the MVC and hospitals are, home State may not have high discriminative power, especially in geographically large States. Similarly, middle name can lack the reliability and the discriminative power that first and last name can have because not everybody has the middle name and often only the initials are recorded. Also, race is typically determined by police officers at the scene without asking the person directly, which can possibly lead to measurement errors.

Table 1.3.1: List of variables used in the linkage model by eight CODES States											
Identifiers	**CODES States**										
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**N used**	**% Used**	**Identifier type**
Incident Date	X	X	X	X	X	X	X	X	8	100	Widely used
Sex	X	X	X	X	X	X	X	X	8	100	
Age	X	X	X		X	X	X	X	7	88	
DOB*	X	X	X	X		X	X	X	7	88	
First Name**	X	X	X	X	X	X			6	75	
Last Name**	X	X	X	X	X	X			6	75	
Seat position	X	X		X		X	X	X	6	75	Moderately used
Crash flag			X	X	X	X			4	50	
Hour of incident		X		X	X			X	4	50	
Vehicle Type	X	X					X	X	4	50	
Home ZIP				X		X	X		3	38	
Hospital Flag	X	X		X					3	38	
Incident County***					X	X	X		3	38	
Collide with			X				X		2	25	Rarely used
Hospital ZIP		X		X					2	25	
Injury flag					X	X			2	25	
Latitude / Longitude	X							X	2	25	
SSN			X			X			2	25	
Fatal injury				X					1	13	
Home State						X			1	13	
Injury code	X								1	13	
Middle Name			X						1	13	
Race							X		1	13	
RPC			X						1	13	
N identifiers used	11	11	11	12	9	13	10	8			
* Full date, just month, or just year ** Full name, initial, or soundex *** defined as counties of MVC and hospital admission											

Figure 1.3.1 shows the heat map that graphically represents the distribution of match probabilities for each contributing CODES State project. The columns represent eight CODES States who had submitted applicable GUM data with match weights and probabilities at the time of this study. The rows represent the percentages of match probabilities separated into groups every 10-percent increment. The columns are sorted by the percentages on the top row (match probability ≥ 0.9) and CODES States labeling for Table 1 and Figure 1 correspond to each other. Percentages of matches that have all match probabilities are greater than 0.9 ranges from 78.99 percent to 96.90 percent. Percentages of matches whose match probabilities were less than 0.1 ranged from 0.04 percent to 3.22 percent.

Comparing Table 1.3.1 against Figure 1.3.1, it is interesting to see that the CODES States that did not use first and last name have the lowest percentages of matches that had match probability ≥ 0.9, which could be an indication that first and last names are powerful identifiers. Three CODES States that used a hospital flag identifier also achieved higher percentage of matches with high match probabilities. The figure also suggests that the linkage model with too few identifiers could lose the model's potential to produce linkage results with high match probabilities. For example, CODES State H only used eight identifiers. This CODES State has the lowest percentages of match probabilities ≥ 0.9 and highest percentages of match probabilities < 0.1.

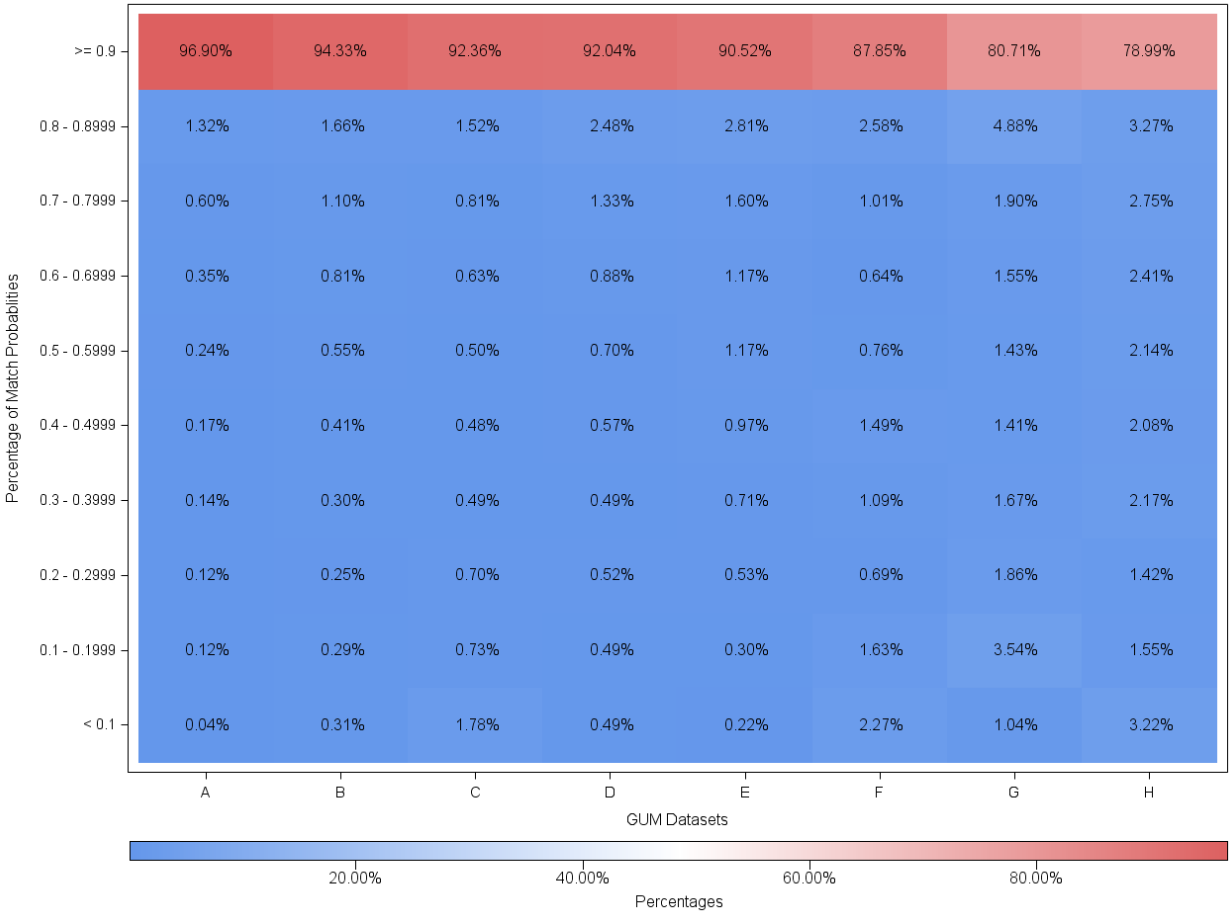


Figure 1.3.1: Percentage of Match Probabilities for each of eight CODES States

Identifier performance with respect to match probabilities and match weights

Table 1.3.2 shows, for each identifier, the median percentage of match probabilities across all CODES States that used that identifier. For example, among all CODES States that used incident date identifier, the median percentage of matches that achieved match probability ≥ 0.9 was 91.28 percent. For ease of view, each identifier is color coded into widely used, moderately used, and rarely used identifiers, as previously categorized.

Table 1.3.2: Median percentage of match probabilities across all CODES States

		Incident Date	Sex	Age	DOB	First Name	Last Name	Seat Position	Crash Flag	Hour of Incident	Vehicle Type	Home Zip	Hospital Flag
	N States	8	8	7	7	6	6	6	4	4	4	3	3
Match Probability	≥ 0.9	91.28	91.28	90.52	92.04	92.20	92.20	89.94	91.28	91.28	87.52	87.85	94.33
	0.8 - 0.8999	2.53	2.53	2.58	2.48	2.07	2.07	2.53	2.53	2.65	2.47	2.58	1.66
	0.7 - 0.7999	1.21	1.21	1.10	1.10	1.05	1.05	1.21	1.17	1.47	1.50	1.33	1.10
	0.6 - 0.6999	0.84	0.84	0.81	0.81	0.72	0.72	0.84	0.76	1.03	1.18	0.88	0.81
	0.5 - 0.5999	0.73	0.73	0.76	0.70	0.63	0.63	0.73	0.73	0.93	0.99	0.76	0.55
	0.4 - 0.4999	0.77	0.77	0.97	0.57	0.53	0.53	0.99	0.77	0.77	0.91	1.41	0.41
	0.3 - 0.3999	0.60	0.60	0.71	0.49	0.49	0.49	0.79	0.60	0.60	0.98	1.09	0.30
	0.2 - 0.2999	0.61	0.61	0.69	0.69	0.53	0.53	0.61	0.61	0.53	0.83	0.69	0.25
	0.1 - 0.1999	0.61	0.61	0.73	0.73	0.39	0.39	1.02	0.61	0.39	0.92	1.63	0.29
	< 0.1	0.76	0.76	1.04	1.04	0.40	0.40	0.76	1.14	0.40	0.67	1.04	0.31

		Incident County	Collide with	Hospital Zip	Injury Flag	Latitude / Longitude	SSN	Fatal Injury flag	Home State	Injury Code	Middle Name	Race	RPC
	N States	3	2	2	2	2	2	1	1	1	1	1	1
Match Probability	≥ 0.9	87.85	86.54	93.18	89.19	87.95	90.11	92.04	87.85	96.90	92.36	80.71	92.36
	0.8 - 0.8999	2.81	3.20	2.07	2.70	2.29	2.05	2.48	2.58	1.32	1.52	4.88	1.52
	0.7 - 0.7999	1.60	1.35	1.21	1.30	1.67	0.91	1.33	1.01	0.60	0.81	1.90	0.81
	0.6 - 0.6999	1.17	1.09	0.84	0.90	1.38	0.63	0.88	0.64	0.35	0.63	1.55	0.63
	0.5 - 0.5999	1.17	0.97	0.63	0.96	1.19	0.63	0.70	0.76	0.24	0.50	1.43	0.5
	0.4 - 0.4999	1.41	0.95	0.49	1.23	1.12	0.98	0.57	1.49	0.17	0.48	1.41	0.48
	0.3 - 0.3999	1.09	1.08	0.39	0.90	1.16	0.79	0.49	1.09	0.14	0.49	1.67	0.49
	0.2 - 0.2999	0.69	1.28	0.39	0.61	0.77	0.69	0.53	0.69	0.12	0.70	1.86	0.7
	0.1 - 0.1999	1.63	2.14	0.39	0.96	0.83	1.18	0.49	1.63	0.12	0.73	3.54	0.73
	< 0.1	1.04	1.41	0.40	1.24	1.63	2.03	0.49	2.27	0.04	1.78	1.04	1.78

Widely used identifiers | **Moderately used identifiers** | **Rarely used identifiers**

All identifiers categorized as widely used identifiers achieve high median percentages for match probabilities ≥ 0.9 (range 89.94-92.20) and low median percentages for match probabilities < 0.1 (range 0.4 to 1.04). However, identifiers such as crash flag, hour of incident, hospital flag, hospital ZIP, and middle names still achieve comparable match probability results. Clearly, widely used identifiers are not always the only identifiers that produce favorable results. There clearly is a difference between common/uncommonly used identifiers and informative/uninformative identifiers. We can see that informative identifiers are generally accurately recorded, have clearly distinguishable outcomes (e.g., ZIP codes), and have many possible outcomes, such as birth dates (see also Cook et al., 2001).

The identifier that achieves the lowest mean percentage was race (80.71%). Though SSN performed relatively well (90.11%), the latitude and longitude did not perform as well (87.95%). This could be due to numerous reasons. One is that latitude and longitude are two separate numbers, which increases the risk of data errors. Also, if one recording the MVC information is unfamiliar with the latitude and longitude, the two may be switched by accident, which would lead to mismatch with the latitude and longitude for the hospital.

It should be noted that the limitation of this analysis is that, for a particular CODES State, when non-informative identifiers are used with informative identifiers, effect of non-informative identifiers can be nullified by the informative identifiers in the linkage model. For example, injury code on the MVC side may not always be accurately recorded, and thus may not always be an informative identifier. However, CODES State A still achieves median percentage of 96.90 percent for match probability ≥ 0.9, the highest median of all CODES States. This is probably due to the fact that CODES State A uses all of widely used identifiers (see Table 1).

Figure 1.3.2 shows the boxplots used to evaluate the effect of each identifier on match weights. Almost all identifiers floor at approximately 14.00. All widely used identifiers cap around 100.00. Some moderately used identifiers and rarely used identifiers such as home ZIP, incident county, injury flag, and SSN achieve results similar to that of widely used identifiers. Race, unsurprisingly, received the lowest maximum and median match weights of all identifiers. This analysis still suffers the same shortcomings as Table 1.3.2 where the effects of informative and non-informative identifiers may not be distinguishable.

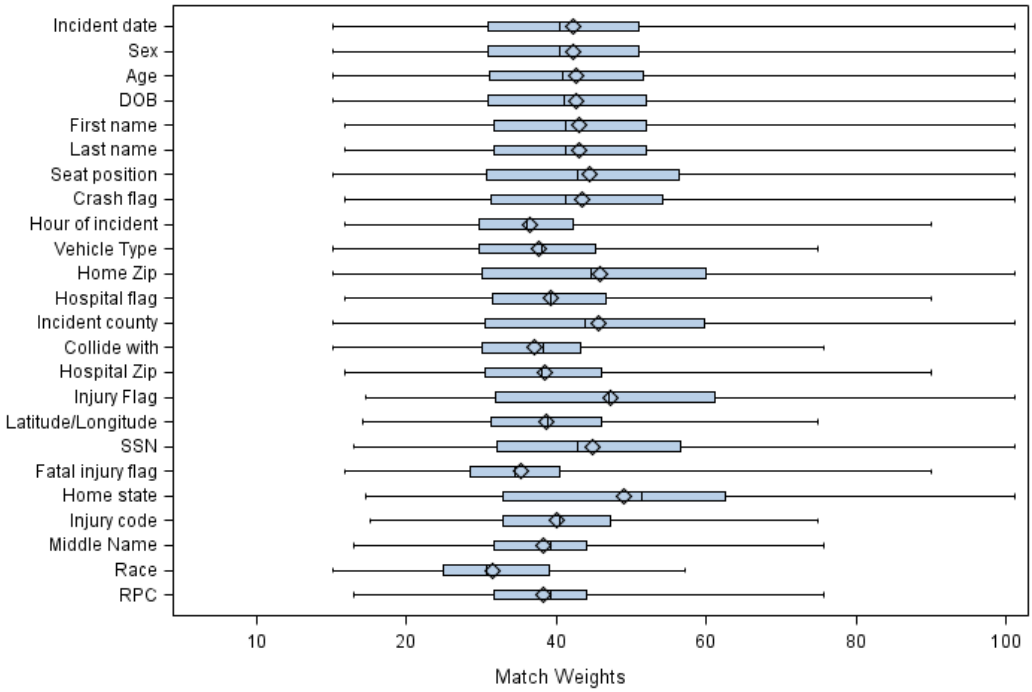


Figure 1.3.2: Boxplot of match weights for each linkage identifier

Effect of MVC file size on match probability

In this section of the chapter, we treat 15 yearly GUM datasets from each contributing State CODES project separately. There are three GUM datasets where the MVC file sizes were 0 – 99,999; three between 100,000 and 199,999; four between 200,000 and 299,999; three between 300,000 and 399,999; and two between 400,000 and 799,999. Match probabilities in each MVC file size group were evaluated using the percentiles. Figure 1.3.3 shows the relationship between percentiles and the MVC file sizes. As the MVC file size increases, the match probabilities decrease for 5th to 25th percentile. The match probabilities converge to 1.0 for 25th percentile and higher. When examining the line for MVC file size 0 – 99,999 and MVC file size 100,000 – 199,999, the match probabilities are very close even at 5th percentile. A limitation of this analysis is that there may not be a sufficient number of GUM datasets in each MVC file size group to determine whether there is a significant difference between MVC file size groups. Additionally, we cannot extrapolate outside of MVC file sizes greater than 800,000.

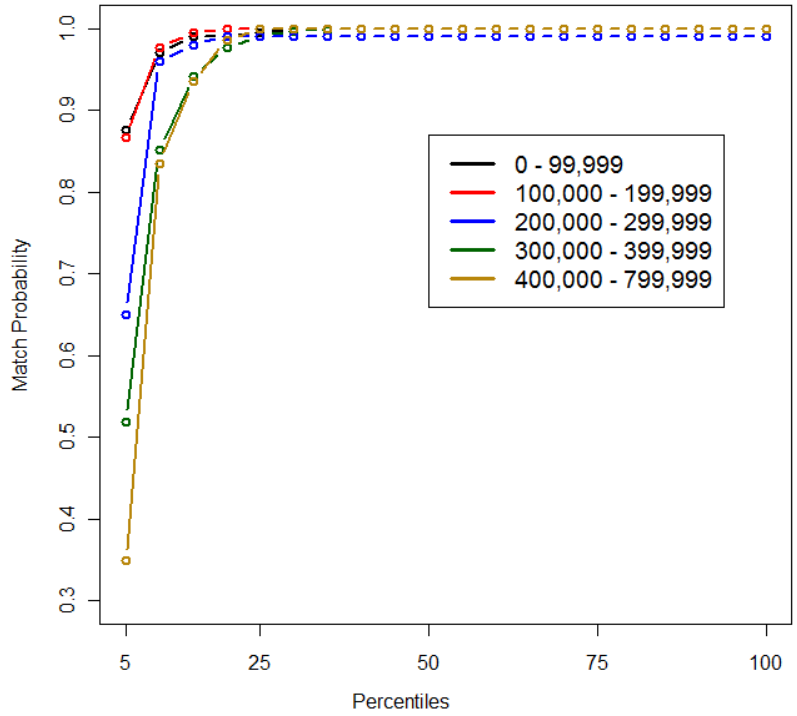


Figure 1.3.3: Match Probability Percentile for each MVC File Size Group

Imputations

In order to investigate whether an MVC-hospital pair with high match probability is likely to appear in all imputations, the number of times each MVC occupant linked to the hospital was obtained. Each MVC occupant's match probability was categorized into every 0.1 increment. Since all MVC occupants are assigned differing match probabilities across imputed datasets as the result of Markov chains in the CODES2000 software, the maximum match probability was used to categorize the match probabilities. The heat map on Figure 1.3.4 summarizes these results. The structure of the heat map is similar to that of Figure 1.3.1 except the columns represent the number of imputations in which each record appears. Each grid represents the percentages of records that appear in one, two, three, four, or five imputations. For example, for match probabilities ≥ 0.9, 0.45 percent of records appear in only one imputation. Similarly, for match probabilities ≥ 0.9, 93.86 percent of pairs appear in all five imputations.

There is a clear trend where pairs with high match probabilities are very likely to be in all five imputations and pairs with low match probabilities are likely to be in only one imputation.

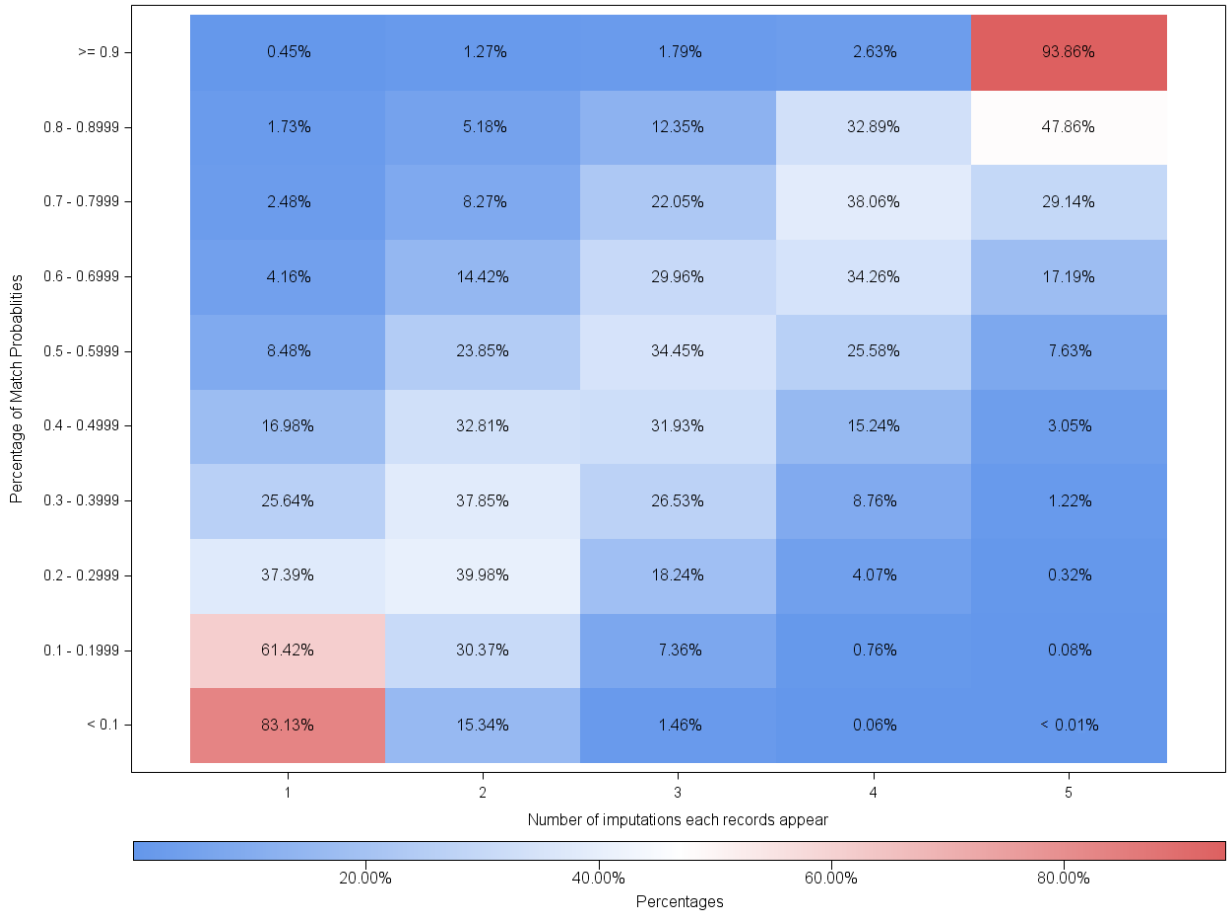


Figure 1.3.4: Match probability and number of imputations each record appears

Conclusion

When linking the records between the MVC and hospital databases, the typical identifiers are incident date, sex, age, DOB, first name, last name, and seat position. These identifiers and other identifiers such as ED, and hospital flags, SSN and longitude and latitude tend to produce high match probabilities. We suggest that when requesting MVC and hospital files to be linked, identifiers that yield high match probabilities should be considered first. There is negative relationship between the MVC file size and the match probabilities. Therefore, careful selection of identifiers is crucial when linking bigger MVC files. MVC records with high match probabilities appear in a higher number of imputed datasets.

References

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

Crash Outcome Data Evaluation System. (2010). *The crash outcome data evaluation system (CODES) and applications to improve traffic safety decision-making.* Washington, DC: National Highway Traffic Safety Administration.

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

McGlincy, M. H. (2000). *CODES 2000*. Morrisonville, NY: Strategic Matching.

McGlincy, M. H. (2004). A bayesian record linkage methodology for multiple imputation of missing links. *Proceedings of the Americian Statistical Association*, 4001-4008.

Chapter 4: A Comparison and Demonstration of Multiple Imputation of Missing Data

Introduction

Previous chapters have described data linkage, and in particular the probabilistic linkage method used in the CODES 2000 and LinkSolv software. These methods allow the linkage of databases that have no primary keys for joining databases. If you consider the link between datasets to be missing data, then probabilistic linkage is in essence imputation of missing data. This chapter discusses the next step of imputation: imputing unknown or missing values within linked datasets.

Missing Data

There are many ways that missing data can occur within a dataset, and it is important to try to understand the mechanism that causes the data to be missing before one handles it through imputation or another tactic. Missing data are commonly classified by the type of mechanism that gave rise to the missingness: Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR) (See Table 1.4.1). While there are no statistical methods that can be used to distinguish MCAR, MAR, and MNAR mechanisms through analysis of the data, a careful investigation of the data collection process can give insight into the mechanisms of missing data.

Table 1.4.1. Descriptions and Examples of Missingness Types

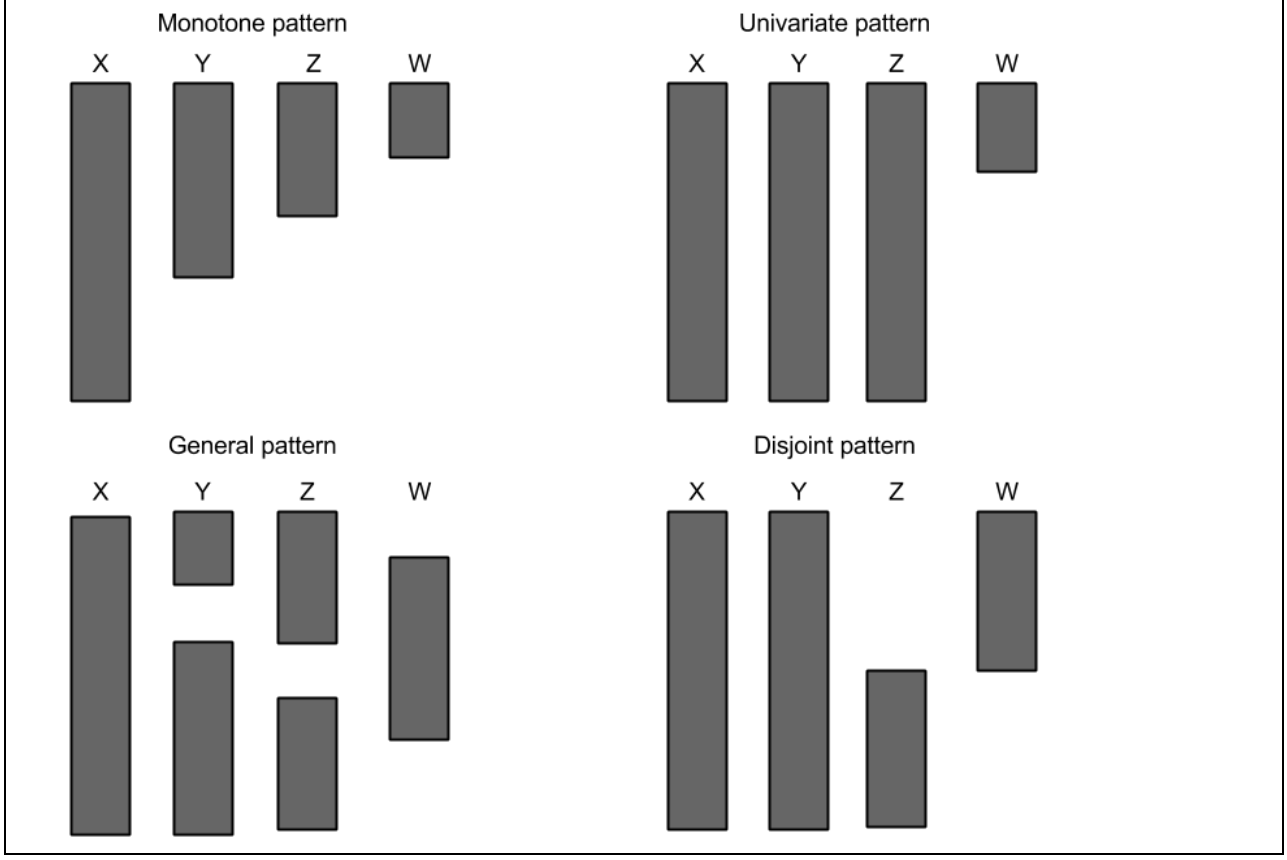
Mechanism of Missing Data	Description	Examples
Missing Completely at Random (MCAR)	The probability that a data point is missing is independent of all other observed and unobserved characteristics of the study sample. In other words, subjects with missing data are a random sample of the study population.	• In an EMS dataset, transport time was deleted for the top *n* patients from a file sorted in a random order.
Missing at Random (MAR)	The probability that a value is missing depends on the observed values in the sample, but is independent of any unobserved or missing values. In other words, the observed data contain information that explain the mechanism of missingness up to an element of randomness.	• Transport time is missing from a dataset more often for children than for adults. • One hospital failed to report charges. • Birth date is missing more often for passengers.
Missing Not at Random (MNAR)	The probability that a value is missing depends on unobserved variables or the missing value itself. As a consequence, it is impossible to estimate missing values using other variables in the dataset.	• Transport time in an EMS dataset is missing more often for patients transported by a certain agency and the data about which agency transported the patient is unavailable. • Hospitals did not report disposition for patients who were transferred or those who died. • Birth date was not collected for children under 10 years old.

Patterns of Missing Data

The pattern of missing data is important to consider when choosing how to handle missing data. Figure 1.4.1 below shows a graphical representation of 4 missing data patterns, where the columns represent data variables, the vertical axis represents observations, and a gray pattern represents the presence of observed data values. In a monotone pattern, observations and variables can be arranged so there is sequential censoring by variable. For example, in the figure, variable X is always observed when Y is observed, Y is observed when Z is observed, and Z is observed when W is observed. There are no cases where W is observed and Z is not. This pattern might be expected in a longitudinal study where all future observations of a case are missing after that case is censored. A univariate missing data pattern is a special case of monotone missing, where only one variable is missing values.

A general pattern of missing data cannot be arranged into a monotone pattern. There are cases with missing values for W and observed values for Z, and vice versa. In a disjoint pattern, there are variables that are never observed at the same time. In the figure, Z is never observed with W and vice versa.

Figure 1.4.1. Patterns of missingness



Methods to Handle Missing Data

Imputation is the process of replacing missing values with plausible values so that all observed data may be included in the analysis. Multiple imputation is a special case where values are imputed or drawn multiple times from a derived distribution, resulting in multiple datasets to analyze instead of the single original dataset. There are many methods of deriving plausible values to impute, but all methods result in a set of imputed datasets where the imputed values are randomly assigned, conditional on observed data.

An alternative to imputation is exclusion of all cases with at least one missing data-point, also known as complete case analysis. Complete case analysis is the default of many procedures in statistical software. Depending on the size of your dataset and the amount of missing data, deletion of cases can result in a serious reduction of statistical power and possible bias. If the data are MCAR, then estimates will be unbiased despite the loss of power. If the data are MAR or MNAR, then results may be biased for failure to account for the missingness process. Alternatives to complete case analysis exist, including single-imputation methods. Single imputation methods replace missing values with estimated values, and result in a single imputed dataset. The estimated value may be a conditional mean or median, a predicted value from a regression model, a value sampled from a similar case in the dataset, a value sampled from a similar case in an external dataset, or the last observed value carried forward in a longitudinal dataset. Single imputation methods generally do not account for the uncertainty inherent in the imputed values, may result in underestimates of variances and inflated type I error, and may introduce additional bias into the dataset. In many cases, multiple imputation can overcome limitations of complete-case analysis and single imputation methods. Multiply imputed datasets allow all cases to be included in analyses, and account for the uncertainty inherent in the imputed values. This leads to unbiased results in the case of MCAR or MAR mechanisms, and at least as much, if not more, statistical power than exclusion of cases with missing data. Multiple imputation does not overcome bias introduced by MNAR mechanisms, but have been shown to be less biased when data are MNAR than other methods (Haukoos, 2007).

Producing Multiply Imputed Datasets

While there are many methods to produce multiple imputed datasets, these generally fall into two categories: (1) Model-based imputation, where a statistical model is developed that predicts the value of a missing data point. A random component is included when filling in the missing data to account for the uncertainty of that prediction. (2) Cell imputation, where each case with missing values is matched with a similar case and all missing data from the first case are replaced with data from the second case. The matched case is usually randomly selected from a pool of candidate cases in order to account for the uncertainty of the match. There are many methods of identifying a pool of candidate cases to draw from including predictive models, propensity scores, and distance measures.

Some variations of model based imputation, and most variations of cell imputation demand that the data be monotone-missing. In these methods, the pool of candidate cases, or the pool of cases to be included in data models, is made up of cases with fully-observed or previously-imputed data. For example, if one were using one of these methods to impute the monotone pattern shown

in the figure above, values of W would be imputed first, using only candidates with observed values of X,Y, and Z. Next, values of Z and W would be imputed using only candidates with observed values of X, and Y, and observed or imputed values of W. The method would progress through the dataset until it imputed all values of all observations. If the data can't be arranged into a monotone missing pattern, a naïve imputation method may be used to fill in enough data to make the pattern monotone missing before applying the more sophisticated imputation model.

Multiple Imputation Using a Sequence of Regression Models

The method of multiple imputation of missing data used for many CODES analyses is model based, and does not require a monotone pattern. This method uses chained regression models, and is described by Raghunathan et al. (2001) and implemented in IVEware software (University of Michigan, 2002).

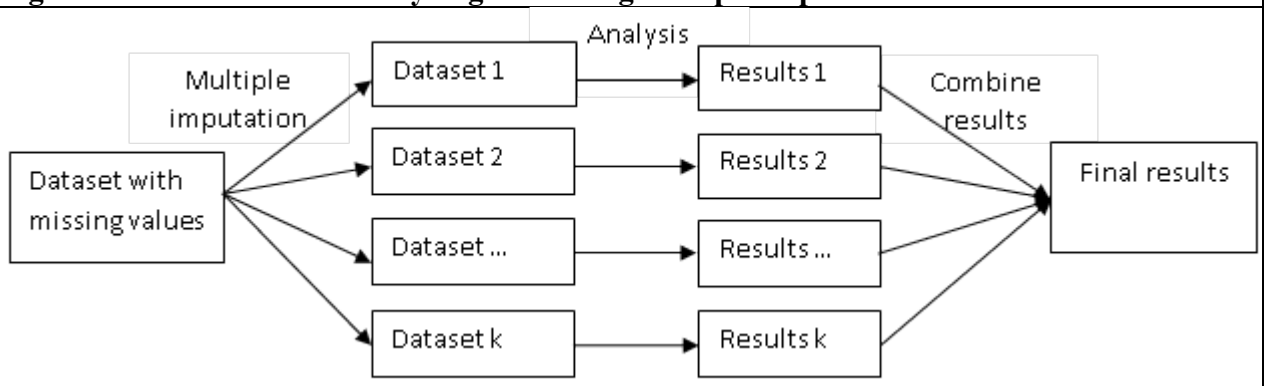
Briefly, this method fits a regression model to each variable in a dataset, and creates a predictive distribution from which to impute missing values. If this method were applied to the general missing pattern in the figure above, a model would be built using only cases with observed values of all other variables (complete cases) to predict and fill-in missing values for the least-missing variable, Y. Those imputed values and all observed values would then be used to impute the next-most observed variable, Z. These imputed values would be used in the next model, and so-on. Once all values were imputed, a second model for Y would be built, this time using all cases. Previously imputed values for Y would be replaced with newly imputed values. A second model would then be built for Z and so on until all variables were imputed several times and the models were stable.

In this method, regression models may be linear, logistic, Poisson, generalized logit, or a mixture of these. Data are not required to be monotone missing, conditional models may be created to apply to only a subset of the data, and imputed values can be restricted using lower and upper bounds.

Analyzing Multiply Imputed Datasets

The methods of analyzing multiply imputed datasets are described by Rubin (1987). Briefly, once the original dataset is imputed multiple times, analyses are applied to each imputed dataset individually, and results are combined (see Figure 1.4.2).

Figure 1.4.2. Overview of analyzing data using multiple imputation methods



The analysis depends on the goals of the study, but most univariate and multivariate statistical procedures may be used. Once the same analysis has been applied to the multiple datasets, and multiple results are obtained, parameter estimates (Q_i) from the *m* imputed datasets are averaged to obtain a final parameter estimate ($\bar{Q}_m$).

$$\bar{Q}_m = \frac{1}{m}\sum Q_i$$

The variance estimate of the final parameter estimate is a combination of within-imputation and between-imputation variance components. Within-imputation variance ($\bar{U}_m$) is calculated as the average of the variances, (U_i), obtained in the *m* multiple analyses.

$$\bar{U}_m = \frac{1}{m}\sum U_i$$

Between-imputation variance (B_m) is calculated as the sample variance of the *m* parameter estimates.

$$B_m = \frac{1}{(m-1)}\sum (Q_i - \bar{Q}_m)^2$$

The total variance (T_m) is a combination of the within and between-imputation variances.

$$T_m = \bar{U}_m + \left(1 + \frac{1}{m}\right) B_m$$

This total variance can be used with the overall parameter estimate to construct confidence intervals, or to test hypotheses using a Student's t-distribution with υ degrees of freedom.

$$\upsilon = (m-1)\left[1 + \frac{\bar{U}_m}{\left(1 + \frac{1}{m}\right) B_m}\right]^2$$

Analyses can be combined manually or using software. The MIANALYZE procedure incorporated in SAS software (SAS Institute, Cary, NC), for example, accepts multiple sets of results and produces overall estimates, standard errors, confidence intervals, and hypothesis test results as well as estimates of the relative increase in variance due to missing values, the fraction of missing information, and the relative efficiency for each estimate.

Multiple Imputation Demonstration

Study Population

Crash data linked with emergency department (ED) and hospital data were provided by 11 States as part of the CODES General Use Model (GUM). These States submitted up to four years of data. Each State performed probabilistic linkage using multiple imputation for missing links prior to submission. After submission, the CODES Technical Resource Center imputed missing crash, ED, and hospital data in IVEware.

We selected one year of GUM data from four States for use in this demonstration. General attributes of the 4 datasets are described in table 1.4.2 below. We chose 4 States varying in size from 73,563 to 337,986 crash occupants. Each State GUM dataset had similar variables for use in the imputation process, though the availability of these variables differed slightly between States. The rate of linkage to ED and hospital records also varied between States. We excluded crashes occurring outside a traffic-way and those involving parked vehicles or non-motor vehicles (boats, trains, etc.) since not all States reported these crashes. We also excluded non-occupants (pedestrians, bicyclists, etc.) and occupants of non-passenger vehicles (large trucks, buses, motorcycles).

Table 1.4.2. Description of datasets.

Attribute	State A	State B	State C	State D
Number of crashes	73,563	337,986	126,547	231,809
Number of variables in imputation models	55	53	56	52
Percent Linking to ED record	6.7%	15.6%	7.0%	15.6%
Percent Linking to Inpatient Record	0.5%	1.2%	0.5%	0.8%

Methods

First, we described the rate of missingness for each variable and the variability of missingness between States. We then fit logistic regression models to each dataset separately. The following models were fit:

1. *Demonstration 1: Complete-Case Crash Data vs. Multiply Imputed Crash Data.* Using only crash data (not hospital or ED data), we estimated the odds of incapacitating or fatal injury according to the crash report. Up to 20 predictors were included in the model. Predictors that were missing 100 percent for a dataset were not imputed or included in the logistic model. We fit the logistic models using a complete-case dataset and repeated it using multiply imputed datasets. We combined results from multiply imputed data using methods explained in the previous section. Odds-ratios and 95-percent confidence intervals from each method were compared for each predictor.

2. *Demonstration 2: Multiply Imputed Links Only vs. Multiply Imputed Links and Data.* Using probabilistically linked crash, ED, and hospital data, we estimated the odds of a moderate or more severe injury (overall Maximum Abbreviated Injury Scale score of 3-6) from the linked hospital record dependent on the same 20 predictors from demonstration 1. Unlinked crash records were considered to not have a moderate or severe injury. We fit the logistic models to multiple datasets obtained from probabilistic linkage, excluding cases with missing data for the outcome or predictors. We repeated the analysis using datasets with missing data multiply imputed. Odds-ratios and 95-percent confidence intervals from each method were compared for each predictor.

3. *Demonstration 3: Using Only Linked Cases: Multiply Imputed Links Only vs. Multiple Imputed Links and Data.* We repeated demonstration 2 using only crash occupants linking to an ED or hospital record; unlinked occupants were excluded. Again, logistic models were fit to linked cases with non-missing data for the outcome and predictors in one model, and again to all cases with missing data multiply imputed.

Results

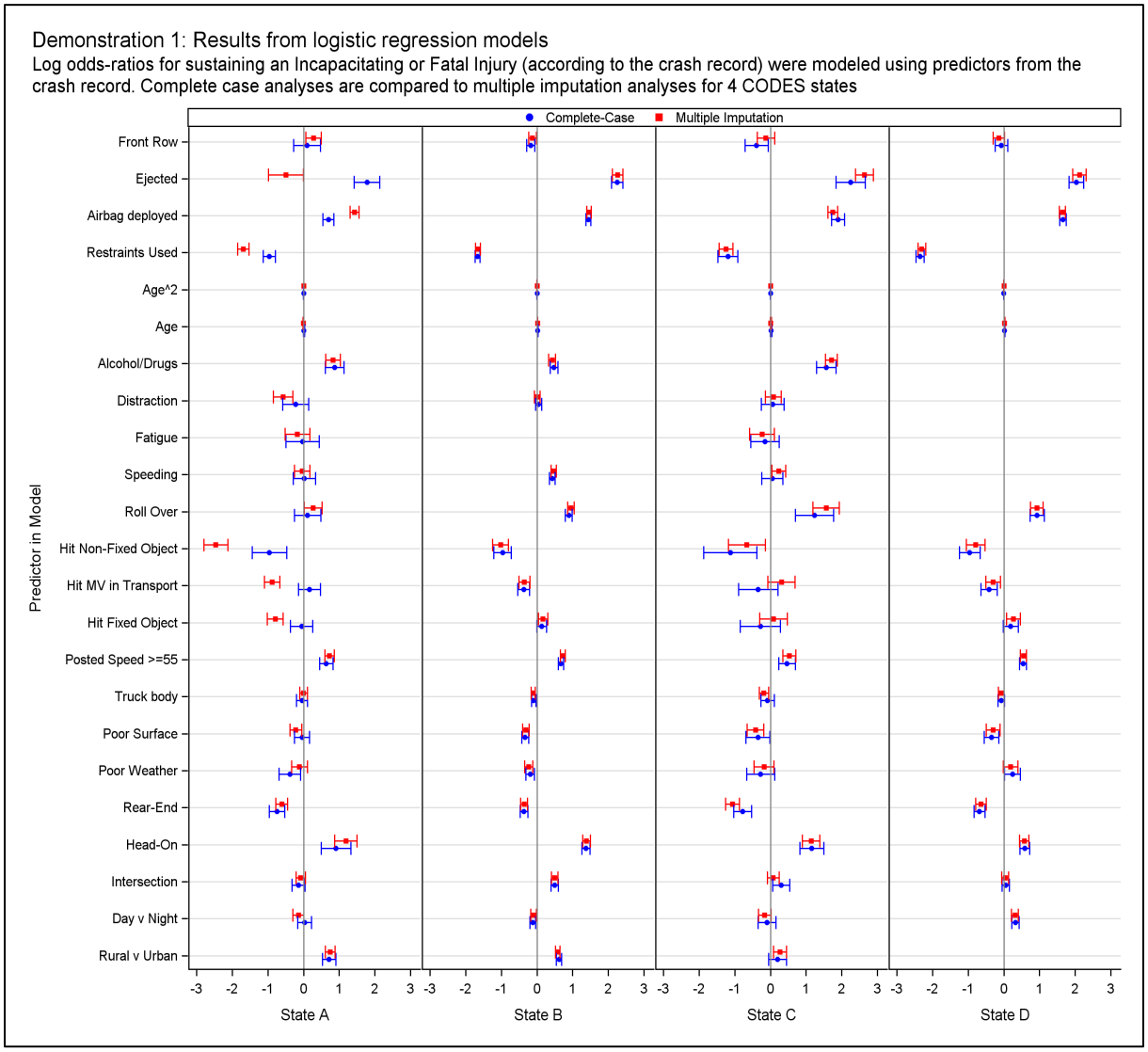
Rates of missing data varied within and across datasets. Table 1.4.3 describes rates of missing data for variables to be included in logistic regression models. Rates vary from 0 percent to more than 25 percent. Some variables were not collected on the State specific crash report (missing 100% of the time). State A had nine variables coded as missing more than 5 percent of the time, with information about being/not being ejected missing for 79.2 percent of crash occupants. State B had more than 5 percent of the observations coded as missing for five variables data. Information about being/not being ejected was missing in 16.6 percent of cases in State B. State C had six variables with at least 5 percent of observations coded as missing, with speed limit information being missing the most frequently (15.5%). State D did not collect data for five variables, but had relatively low rates of missingness among variables that were included. The most frequently missing variable in State D data was restraint use information (6.5%).

Table 1.4.3. Missingness Rates of Variables					
Variables included in logistic regression models		***Percent Missing***			
Model Variables	**Description/Levels**	**State A**	**State B**	**State C**	**State D**
Rural Location	Rural vs. Urban Location	0.0%	0.0%	0.3%	100.0%
Night time	Crash occurred between 8:00 pm and 5:59 am	0.6%	0.4%	0.0%	0.0%
Intersection related	Crash was related to travel through an intersection	0.5%	0.0%	3.4%	0.8%
Manner of Crash	Crash classified as head-on vs. rear-end vs. other (Angle, Sideswipe, not collision with MV in transport, or Other)	0.0%	0.0%	2.0%	0.3%
Adverse weather conditions	Adverse (rain, snow, sleet, hail, fog, smog, smoke, severe cross winds, other) vs. Not adverse (clear or cloudy)	1.0%	1.5%	1.0%	0.1%
Poor surface conditions	Poor (wet, snow, slush, ice, frost, other) vs. Dry	0.8%	0.8%	0.8%	0.1%
Truck Body	Light Truck, passenger van, SUV vs. Passenger car	0.0%	0.0%	0.0%	0.0%
Posted Speed limit >=55	Posted speed limit >=55 mph	0.0%	5.3%	15.5%	1.0%
Most harmful event	Collision with MV in transport vs. collision with non-fixed object vs. collision with fixed object vs. other (non-collision or no harmful event)	3.7%	0.0%	6.7%	0.0%
Roll over	Vehicle rolled over	0.1%	0.0%	4.3%	0.0%
Speed Related	Vehicle's speed was a contributing factor in the crash	5.4%	1.8%	3.2%	100.0%
Fatigue Related	Driver fatigue/drowsiness was a contributing factor in the crash	1.9%	100.0%	4.1%	100.0%
Distraction Related	Driver distraction was a contributing factor in the crash	5.4%	0.0%	12.9%	100.0%
Alcohol or Drugs Suspected	Alcohol or Drug use was suspected of the driver of the vehicle	4.4%	1.4%	0.6%	100.0%
Age	Age in years	28.0%	5.7%	2.6%	2.7%
Age-squared	Age in years, squared	28.0%	5.7%	2.6%	2.7%
Restrained	Restraints reported used vs. not used	15.9%	12.4%	8.6%	6.5%
Airbag deployed	Airbag deployed vs. not deployed or not applicable	11.1%	4.0%	6.5%	2.5%
Ejected	Ejected, partially ejected, ejected—unknown degree vs. Not ejected	79.2%	16.6%	6.4%	1.4%
Front Seat	Front seat vs. Back seat or exterior	0.1%	0.5%	0.5%	0.9%
Model Outcomes	**Description/Levels**	**State A**	**State B**	**State C**	**State D**
Injury (>3)	Injury level of Incapacitating, or Fatal Injury vs. Not injured, possible injury, non-incapacitating injury, injured-severity unknown.	21.0%	4.6%	0.1%	0.0%
Highest level of care	Not Linked vs. Emergency Department/Outpatient vs. Inpatient	0.0%	0.0%	0.0%	0.0%
Discharged home	Discharge status of Home vs. Died/long term care/rehab/left against medical advice	11.1%	1.8%	1.6%	1.5%
Total hospital charges	Total hospital charges unadjusted	0.0%	0.1%	0.0%	0.0%
MAIS over 2	MAIS of Serious, Severe, Critical, or Maximum vs. Not injured, Minor, and Moderate	0.0%	0.0%	0.0%	0.0%
Cells are colored: gray if missing 100%, red if missing >25%-99.9%, orange if missing >10%-25%, and yellow if missing >5%-10%.					

Demonstration 1: Complete-Case Crash Data vs. Multiply Imputed Crash Data

For demonstration 1, we fit a logistic regression model to crash data from each State. Hospital data were not included in this analysis and the imputation of hospital-crash record links was not used. The outcome was the odds of incapacitating or fatal injury according to the crash report. Model predictors are listed in the top portion of the preceding table. Variables missing 100% were not included in the model for that State.

In one set of analyses, we only included cases that included observed data for the outcome and all predictors (complete cases) in the model. This was a single dataset, and represented a complete-case analysis applied to non-imputed crash data. In another set of analyses, we used multiply imputed datasets to retain all cases in the analysis. The figure below shows results from demonstration 1. Each panel of the plot shows the log odds-ratios from the logistic regression for a State, with State A model results in the left panel, State B in the second and so on.



Results from Demonstration 1.						
State	Number of observations		Number of Significant Predictors		Mean (SD) length of 95% confidence intervals	
	Complete Case	Multiple Imputation	Complete Case	Multiple Imputation	Complete Case	Multiple Imputation
A	10,846	73,563	10 of 23	16 of 23	0.52 (0.25)	0.40 (0.22)
B	237,089	337,986	19 of 22	20 of 22	0.20 (0.10)	0.18 (0.09)
C	74,380	126,547	12 of 23	13 of 23	0.64 (0.33)	0.46 (0.23)
D	206,647	231,809	13 of 18	13 of 18	0.29 (0.15)	0.27 (0.14)

State A had a considerable amount of missing data; only 15 percent of their 73,563 cases were included in the complete-case analysis. This lead to some differences in parameter estimates. For example, the odds-ratio for being ejected was 5.9 (95% CI 4.1, 8.4) for the complete-case analysis, but 0.61 (95% CI 0.37, 1.00) for the multiple imputation analysis. The same odds-ratio in other States analyses ranged from 8 to 14. The estimate using imputed data in this case is clearly incorrect, and illustrates a limitation of imputing large amounts of data. The effect of an airbag deploying differed between methods as well, with the complete case odds-ratio being 2.01 (95% CI 1.72, 2.34) and the multiple imputation analysis odds-ratio being 4.18 (95% CI 3.69, 4.74). Odds-ratios for the effect of airbags ranged from 4 to 6 in the other analyses, suggesting that the estimate from the complete-case analysis may be incorrect.

These two examples show the potential differences in results that can be obtained from using complete case or multiply imputed data. Other log odds-ratio estimates from multiple imputation analyses were similar to complete case results, with a tendency for imputed data to result in tighter confidence intervals.

Overall in State A, 10 of the 23 predictors were significant (log-odds ratio confidence interval did not cross 0) in the complete case analysis, while 16 were significant in the multiple imputation analysis. Confidence intervals in the multiple imputation analysis were shorter on average compared to the complete case analysis (mean log-odds ratio confidence interval length: 0.40 vs. 0.52).

Complete-case methods utilized 70 percent of the original sample from State B, and complete-case and multiply imputed methods resulted in similar estimates and confidence intervals. This State had the largest sample size of the 4 included, and as a result, had the tightest confidence intervals of the 4 States. The confidence intervals from multiple imputation analyses tended to be slightly smaller (mean interval length=0.18) than the matching complete-case intervals (mean interval length=0.20). Complete case analysis resulted in 19 of 22 predictors being significant; multiple imputation analysis resulted in 20 of 22 predictors being significant.

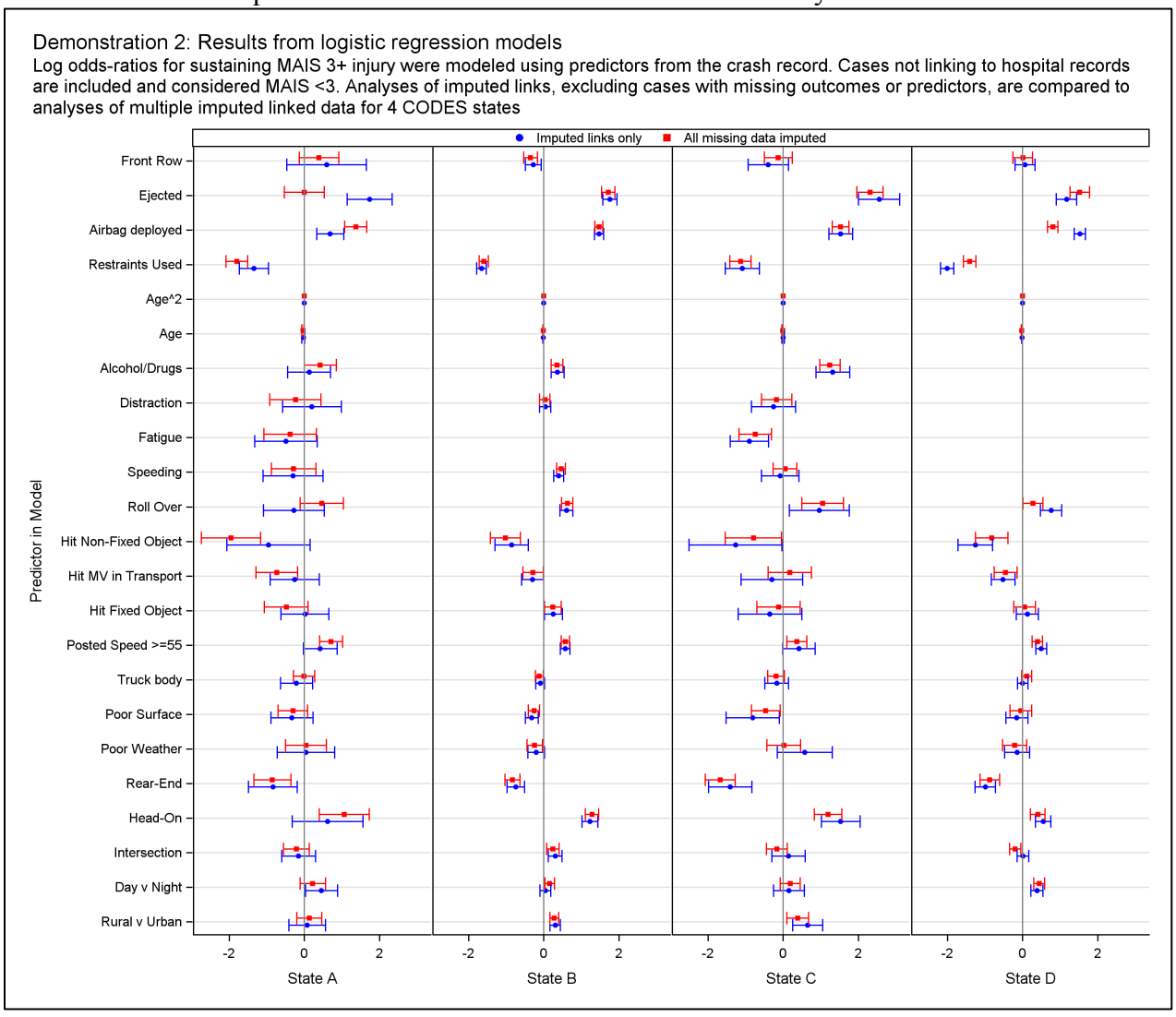
Only 59% of cases from State C were included in complete-case analyses. While there were no reversing of effects, confidence intervals were tighter when multiply imputed data were used (mean interval length=0.46) compared to the complete case analysis (mean interval length=0.64). The complete case analysis resulted in 12 significant predictors compared to 13 in the multiple imputation analysis our of 23 model predictors. The effects of Truck body vs. Passenger car body, Rural vs. Urban location, and Speeding became significant using multiply imputed data, while the effects of Intersection and Front Row became non-significant.

State D had the least amount of missing data, with only 11% of cases being excluded in the complete-case analysis. This resulted in very similar results between the 2 methods. Mean confidence interval lengths were slightly shorter for the multiple imputation analysis (0.27) vs. the complete case analysis (0.29), and both analyses resulted in the same 13 predictors being significant.

Demonstration 2: Multiply Imputed Links Only vs. Multiply Imputed Links and Data

For demonstration 2, we fit logistic regression models to linked crash and hospital data. The outcome was the odds of a moderate or more severe injury (overall Maximum Abbreviated Injury Scale score of 3-6) from the linked hospital record. Unlinked crash records were included in this analysis, and were considered to not have a moderate or severe injury.

In one set of models, we excluded cases missing the outcome or any of the model predictors. This set of models was like running a complete-case analysis using multiply imputed links and therefore methods for analyzing multiply imputed data, but without imputing missing data. We compared this analysis to an analysis using imputed links and multiply imputed missing outcomes and model predictors in order to retain all cases in the analysis.



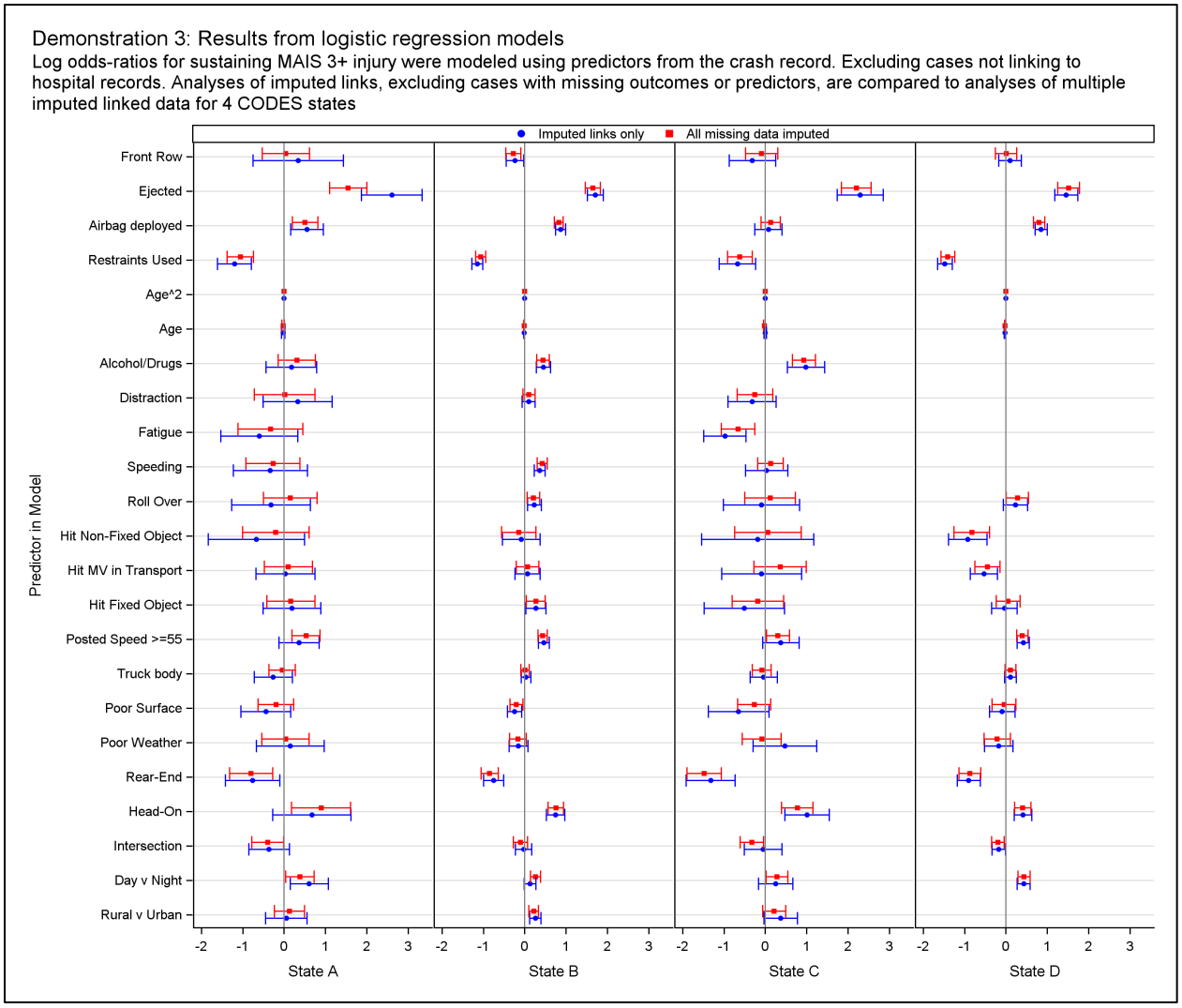
Results from Demonstration 2.						
State	Number of observations		Number of Significant Predictors		Mean (SD) length of 95% confidence intervals	
	Imputed links only	Multiple Imputation	Imputed links only	Multiple Imputation	Imputed links only	Multiple Imputation
A	10,846	73,563	6 of 23	9 of 23	1.20 (0.55)	0.89 (0.40)
B	237,089	337,986	17 of 22	21 of 22	0.35 (0.18)	0.31 (0.16)
C	74,380	126,547	12 of 23	13 of 23	1.06 (0.52)	0.70 (0.34)
D	206,647	231,809	11 of 18	13 of 18	0.44 (0.23)	0.42 (0.22)

The numbers of cases included in complete case models were very similar to demonstration 1, with 15% of cases included from State A, 70 percent from State B, 59 percent from State C, and 89 percent from State D.

Results from this demonstration were similar to those from demonstration 1. Multiple imputation analyses resulted in a greater number of significant predictors in each State's regression model. Confidence intervals using imputed missing data were typically shorter than the comparable intervals using imputed links only.

Demonstration 3: Using Only Linked Cases: Multiply Imputed Links Only Versus Multiply Imputed Links and Data

Demonstration 3 was similar to demonstration 2, in that we modeled the odds of a moderate or more severe injury (overall Maximum Abbreviated Injury Scale score of 3-6) from the linked hospital record. However, unlike demonstration 2, demonstration 3 did not include unlinked crash records. These models focused only on cases linking to ED or hospital records. In one set of analyses, we only included cases with completely observed data for the outcome and predictors (linked complete cases). In another set of analyses, we imputed missing data in order to retain all cases in the analysis.



Results from Demonstration 3.						
State	Number of observations		Number of Significant Predictors		Mean (SD) length of 95% confidence intervals	
	Imputed links only	Multiple Imputation	Imputed links only	Multiple Imputation	Imputed links only	Multiple Imputation
A	3,088	5,288	6 of 23	9 of 23	1.30 (0.59)	0.95 (0.42)
B	43,336	56,737	14 of 22	16 of 22	0.35 (0.19)	0.32 (0.17)
C	5,575	9,574	7 of 23	11 of 23	1.12 (0.60)	0.73 (0.37)
D	34,577	37,876	12 of 18	13 of 18	0.45 (0.24)	0.42 (0.22)

Models with imputed links only included 58 percent of linked cases for State A, 76 percent for State B, 58 percent for State C, and 91 percent for State D.

Results from this demonstration were similar to those from demonstration 1 and 2. Analyses using imputed missing data typically resulted in shorter 95 percent confidence intervals and a greater number of significant predictors than the comparable analyses using imputed links only.

Summary of Demonstration Findings

These demonstrations illustrate three general themes. First, odds-ratio estimates from complete case analyses are usually very close to those from multiple imputation analyses, especially when rates of missingness are low. Second, multiple imputation-based estimates tend to be more powerful (i.e., tend to identify more predictors as significant, and have shorter confidence intervals for those estimates) compared to estimates based on complete-cases only. Third, in some cases multiple imputation methods give very different estimates compared to complete case methods. In these situations, MAR assumptions appear to be violated.

Odds-ratio estimates from multiple imputation analyses were generally very similar to those from complete case analyses. For State D, which had very little data missing, confidence intervals from one method overlapped those from the other method for all parameters estimated in demonstrations 1 and 3; in demonstration 2, intervals for Restraints Used and Airbag Deployed differed enough for confidence intervals to disagree. For States C and B, which also had relatively little missing data, confidence intervals from complete-case analyses overlapped those from multiple imputation analyses in all demonstrations. Despite State B missing restraint use data for 12 percent and ejected data for 16 percent, those estimates were similar between methods for all three demonstrations. State C was missing posted speed limit data for 16 percent and distraction data for 13 percent. Estimates for posted speed limit > 50 mph were similar between methods for all demonstrations, but were significant using multiply imputed data and non-significant using imputed links only in demonstrations 2 and 3, showing the increased power gained by using multiply imputed data. Estimates for distraction in State C were not significant using any method in any demonstration. Confidence intervals in analyses of the State with the most missing data (A) overlapped for 17/23 parameters in demonstration 1, 21/23 in demonstration 2, and all 23 in demonstration 3. Though the odds-ratio estimates were similar between methods, this would not necessarily be the case if we were estimating counts (e.g. total number of hospitalizations) or sums (total charges or costs). In those types of analyses, multiple imputation analyses would include more observations and result in higher estimates. However, we assume that when estimating other measures, such as means, medians, rates, etc., multiple imputation analyses would result in similar estimates compared to complete-case analyses, especially in the case of MAR.

Multiple imputation methods increase standard error estimates of the final estimate compared to the estimates obtained from the analyses of the imputed datasets separately. The magnitude of that increase depends on the variability of the individual estimates from those imputed datasets. Despite this built-in increase in variability, multiple imputation confidence intervals were consistently tighter than the complete case estimates. The decrease in overall standard error compared to a complete-case analysis results from additional cases that are included (i.e., larger sample sizes) This demonstrates that the increase in the sample size due to multiple imputation outweighed the loss of precision due to variability between multiple datasets in all demonstrations and States.

Despite the general agreement between methods, there were a few parameters within State models that were quite different between complete-case and multiple imputation methods. The estimate for Ejected in State A was the greatest offender. In demonstration 1, multiple imputation methods estimated a protective effect of being ejected, while complete-case analysis showed a

large risk associated with ejection. The complete-case analysis of State A only used 15 percent of available cases due to eject being missing in 79 percent of cases, the outcome being missing in 21 percent and other predictors missing values . An investigation into the distribution of ejection status between States shows a similar percentage of ejected occupants in each of the other 3 States, assuming missing at random: 0.46 percent ejected in State B, 0.59 percent in State C, and 0.34 percent in State D. The percentage in State A was 2.21 percent. However, if we assume that all occupants missing data about ejection status were not ejected, the percentage of ejected occupants would be 0.46 percent—much more in line with other States. This suggests that ejection status is not likely missing at random; the probability of the value being missing seems to be much higher if the occupant was not ejected. The violation of the MAR assumption seems to have affected the imputation models and resulted in an incorrect estimate of the effect of ejection on injury in the logistic regression analyses. This result re-iterates the importance of understanding the mechanism of missing data in the original dataset.

Comparison of Multiple Imputation Methods

Many methods exist to impute missing data. CODES currently uses a series of regression models for multiple imputation implemented in IVEware, called from within SAS software. The same general method is implemented in other packages (MICE in S-plus and R for example). As of version 9.3, SAS software includes an experimental option for imputation using sequential regression models within the MI Procedure.

Comparison

We re-imputed the GUM data for State A using SAS software and the MI Procedure and compared it to data for State A used in the demonstrations 1-3, which was imputed using IVEware.

There are several technical differences between IVEware and the MI Procedure. One such difference is that IVEware allows you to restrict the imputation of a variable to a subset of the data. For example, we restrict the imputation of hospital data (charges, length of stay, etc.) to linked cases. Similarly, we restricted the imputation of alcohol/drug use to drivers. There is no analogous mechanism in the MI Procedure to restrict imputation, so we imputed the data in steps: First, we imputed the hospital variables for cases that linked to the hospital; second, we imputed suspicion of alcohol/drug use for drivers only; and third, we merged those two imputed datasets with the remaining crash variables and imputed those in a third set of models.

Another difference between procedures is that IVEware uses model selection to select the ‘best’ model to predict each variable. Models are re-evaluated in each iteration, so the ‘best’ model for a variable may change over the course of the imputation process. This is convenient when there is no reason to believe one set of predictors to be better than another set. Options exist to limit the number of predictors to make the process more efficient. However, you cannot force IVEware to fit a certain model other than the fully saturated one. The MI Procedure requires that you specify the model to impute categorical variables. This is inconvenient when you have many variables to impute and little idea of what the best models are.

In our case, we initially specified models in the MI Procedure that we thought made scientific sense for each variable. However, these models did not always converge. When a model doesn't converge, the procedure terminates, leaving one to wonder which model has an issue and how to fix it. Eventually we used the 'best' models selected from IVEware as the specified models in the MI Procedure. This may have the effect of making these two methods more similar then they otherwise would be, but it enabled us to avoid early termination of the procedure. Once the data were re-imputed using the MI Procedure, we compared the distributions of the 2 sets of multiply imputed datasets.

Table 1.4.4. Comparison of IVEware and PROC MI

Characteristics of State A data before imputation (complete case) and after imputation using IVEware and the MI Procedure (PROC MI) in SAS Software. Only variables missing more than 5% are shown. Relative frequencies or means (age) are shown.

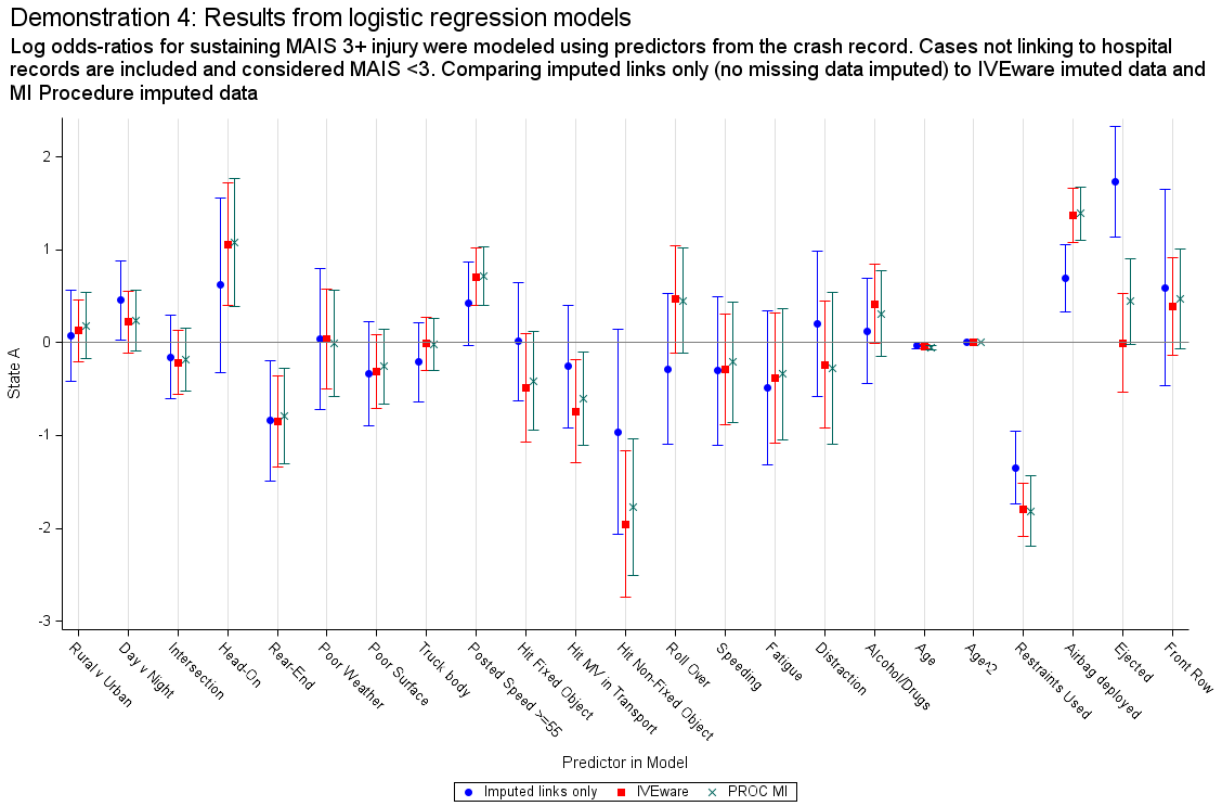
Variable	Percent Missing	Complete Case	IVEware	PROC MI
Speed contributed	5.4%	4.7%	4.8%	4.8%
Distraction contributed	5.4%	6.7%	6.6%	6.9%
Age (Mean)	28.0%	36.4%	35.0%	35.2%
Restraints Used	15.9%	92.9%	92.6%	92.6%
Airbag Deployed	11.1%	12.2%	11.8%	11.7%
Ejected	79.2%	2.2%	9.3%	10.3%
Discharged Home	11.1%	97.4%	97.4%	97.3%
Incapacitating or Fatal Injury	21.0%	2.7%	2.1%	2.6%

Data imputed using the SAS software was similar to data imputed using IVEware. Most variables also had a similar distribution to the raw (complete-case) dataset. For example, even though age was missing in 28 percent of cases, the mean age among those with non-missing data was 36.4 years old. Data imputed using IVEware had a mean age of 35.0 years, and data imputed using SAS Software had a mean age of 35.2 years. The Ejected variable, which we have discussed previously and which we suspect to be Missing Not At Random, was the most commonly missing in this dataset and varied the most between the imputed. Whereas 2.2 percent of occupants with non-missing ejection status were ejected, IVEware estimated that 9.3 percent of all occupants were ejected, and SAS Software estimated 10.3 percent.

Logistic Regression Models

We compared the results from logistic models applied to data with imputed links only, data imputed using IVEware, and data imputed using the MI Procedure. These models used the same outcome and predictors as the models in Demonstration 2 above. We modeled the odds of sustaining a moderate or more severe injury (overall Maximum Abbreviated Injury Scale score of 3-6) from the linked hospital record. Unlinked crash records were included in this analysis, and were considered to not have a moderate or severe injury.

Results from the 3 sets of models are shown below.



The two logistic regression models using multiply imputed data identified the same 9 predictors as significant (95% confidence interval did not include 0). Lengths of the confidence intervals for parameter estimates were similar: IVEware mean (SD) interval length: 0.89 (0.40); PROC MI mean (SD) interval length: 0.91 (0.41).

Comparison Conclusions

The two sets of multiply imputed datasets compared in this demonstration were similar in distribution and resulted in similar analytic results. The two datasets were created with the same basic methodology of using sequential regression models to estimate the values of missing data. Other imputation methods may produce different distributions and/or analytic results. Also, because there are many decisions that must be made, and settings that may be adjusted when specifying the models that impute the missing data, these results may differ from results obtained by models that were specified differently. However, despite the many variables involved in specifying the imputation, we have shown that two different procedures of imputing missing data in a CODES dataset resulted in similar results.

References

Haukoos, J. S., & Newgard, C. D. (2007, July). Advanced statistics: missing data in clinical research--part 1: An introduction and conceptual framework. *Academic Emergency Medicine, 14*(7), 662-668.

Newgard, C. D., & Haukoos, J. S. (2007, July). Advanced statistics: missing data in clinical research--part 2: Multiple imputation. *Academic Emergency Medicine, 14*(7):669-678.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001, June). A multivariate technique for multiple imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1):89-95.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Part 2: Applications From Multiple State CODES Data

CODES General Use Model Overview

As shown in Part 1, probabilistic linkage is a powerful method for combining information from different databases into a single dataset for analysis. Desired information about study subjects is often contained in two or more databases and if a unique key does not exist between these databases, it is not possible to combine the information directly. Rather than relying on a unique key to combine records, probabilistic linkage makes use of fields that are common to each database. The CODES methodology uses probabilistic linkage to combine information from motor vehicle crash (MVC) reports and hospital records, sometimes also adding databases such as EMS, death certificate, and others. As a network, CODES has studied hospital outcomes of motorcycle crashes (Cook et al., 2009), and internally addressed standardization modes for other topics such as older motor vehicles occupant, maximum abbreviated injury scale (MAIS) by MVC characteristics, and hospital outcomes of distracted driving. These projects each required a new data model tailored to the topic. In 2010, the CODES network set out to develop and implement one data model that would be used to study a variety of different MVC topics.

This report incorporates probabilistically linked MVC and hospital records from 11 CODES States for the crash years 2005-2008. States that participated include Connecticut, Georgia, Kentucky, Maryland, Minnesota, Missouri, Nebraska, New York, Ohio, South Carolina, and Utah. Probabilistic linkage is a method that uses personal and event information common to a pair of records to estimate the probability a MVC record and hospital record describe the same person and event. Linking information may include names, date of birth, sex, date and time of MVC/hospital treatment, MVC/hospital location, and the roles of people and vehicle involved. This method results in multiply imputed datasets, each with the possibility of different links between MVC and hospital records (Wang et al., 2010). Analysis of multiply imputed datasets accounts for the uncertainty inherent in the linkage process. A more complete discussion of probabilistic linkage is given by Cook et al. (2001), Crash Outcome Data Evaluation System (2010), Jaro (1995), and McGlincy (2004). Individual CODES analysts were responsible for linking data from their own State MVC and hospital files. CODES analysts have extensive training with probabilistic linkage and use the same probabilistic linkage software, CODES2000 (McGlincy, 2000, 2006).

In order to combine all MVC datasets into a single analytical database, each State's linked MVC and hospital files were mapped onto a standardized set of data elements known as the 'General Use Model' or GUM. The GUM, developed through a joint effort between CODES and the NHTSA State Data System (SDS), provides a standardized set of data elements routinely collected on police MVC report. Standardized medical outcome variables are also added. All standardized crash variables were designed to conform as closely as possible to previously published NHTSA guidelines and data systems, such as the Model Minimum Uniform Crash Criteria (MMUCC) Guideline (NHTSA, 2008), Fatality Analysis Reporting System (FARS), and National Automotive Sampling System General Estimates System (NASS GES). CODES analysts were provided with a detailed coding manual to aid in the creation of the standardized dataset. State analysts also submitted proposed mappings of State data elements to the GUM elements for approval prior to data submission. SDS analysts independently mapped State data elements to the GUM and the CODES Technical Resource Center (TRC) compared the mappings from the State analysts to the mappings from the SDS analysts. If any discrepancies

were found, the CODES TRC met with the State analyst and SDS analyst until a consensus was reached. Once mappings were approved, CODES State analysts submitted five imputations of linked, mapped data to the CODES TRC. Datasets were then subjected to validity and consistency checks prior to incorporation into the full GUM database.

As missing data are a frequent occurrence in administrative data such as the GUM, missing values were imputed using sequences of regression models implemented in IVEware (Raghunathan et al., 2001). Analysts at the TRC developed State and year specific imputation models as data were approved, and imputed missing data from each of the five submitted datasets separately. Analyses were conducted separately on each imputation and the results were combined using methods presented by Schafer (1997) using SAS PROC MIANALYZE (SAS Institute Inc., 2002).

The GUM contains 50 MVC variables and 18 medical outcome variables. The MVC variables include information about the time, location, and circumstances of the MVC; vehicles involved and vehicle characteristics; and details about injured and uninjured MVC participants. It is important to note that safety restraint use and helmet use are self-reported and are often over-reported. Medical outcomes were derived from emergency department and hospital discharge databases and include billing information related to the visit, such as billed charges, length of stay, and discharge status. Also included were injury severity scores and the Barell Injury Diagnosis Matrix derived from *International Classification of Diseases, 9th Revision Clinical Modification* (ICD-9-CM) codes and external causes of injury codes (Centers for Disease Control and Prevention, 2010). Billed hospital charges were adjusted for inflation to 2008 dollars and for State difference, making them comparable across States. Charges represent the total hospital charges accumulated while being treated in the hospital. These charges do not represent what payment the hospital actually received or actual costs to the hospital.

The GUM is experimental and was created to understand what types of analyses were possible with a large, multiple-State CODES dataset. This report section profiles four analyses designed to demonstrate the utility of the GUM and potential uses of combined linked data. Each analysis compares medical outcomes by MVC circumstances. The first analysis looks at older occupants, the second children by safety restraint use, the third motorcyclists by helmet laws, and the fourth teen drivers by graduated driver licensing characteristics.

References

Burch, C., & Cook, L. J. (2013, November). A comparison of KABCO and AIS injury severity metrics using CODES linked data. *Traffic Injury Prevention*.

Centers for Disease Control and Prevention. (2010). *The Barell injury diagnosis matrix, classification by body region and nature of the injury*. Atlanta, GA: National Center for Health Statistics.

Cook, L. J., Olson, L. M., Dean, J. M. (2001). Probabilistic record linkage: Relationships between file sizes, identifiers and match weights. *Methods and Information in Medicine, 40*(3), 196-203.

Cook, L. J., Kerns, T., Burch, C., Thomas, A. M., & Bell, E. (2009). *Associations between helmet use and motorcyclist head and facial crash outcomes in CODES linked data.* (Report No. DOT HS 811 208). Washington, DC: National Highway Traffic Safety Administration, Washington, DC.

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

McGlincy, M. A. (2000). LinkSolv. Morrisonville, NY: Strategic Matching.

McGlincy, M. A. (2004). A bayesian record linkage methodology for multiple imputation of missing links. *Proceedings of the Americian Statistical Association*, 4001-4008.

McGlincy, M. A. (2004). Using test databases to evaluate linkage models and train linkage practitioners. *Proceedings of the American Statistical Association*, 3404-3410.

National Highway Traffic Safety Administration. (2008*). Model Minimum Uniform Crash Criteria Guideline, 3rd edition*. (Report No. DOT HS 810 957). Washington, DC: Author.

National Highway Traffic Safety Administration. (2009). *The crash outcome data evaluation system (CODES) and applications to improve traffic safety decision-making*. (Report No, DOT HS 811 181). Washington, DC: Author.

National Highway Traffic Safety Administration. (2012). *Model Minimum Uniform Crash Criteria Guideline, 4th edition*. (Report No. DOT 811 631). Washington, DC: Author.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85-95.

SAS Institute Inc. (2002). SAS Software. Cary, NC: SAS Institute Inc.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. Boca Raton, FL: Chapman & Hall/CRC.

Wang, H. E., Balasubramani, G. K., Cook, L. J., Lave, J. R., & Yealy, D. M. (2010). Out-of-hospital endotracheal intubation experience and patient outcomes. *Annals of Emergency Medicine, 55*(6), 527-537.

Analysis 1. Comparison of Medical Consequences of Motor Vehicle Crashes among Older Occupants

Abstract

Objective: To examine differences in injury patterns between younger (ages 21–64) and older (65+) occupants in motor vehicle crashes.

Methods: Probabilistically linked crash and hospital data from years 2005-2008 were collected from eleven States in the CODES Network. State data were mapped onto common elements and submitted for combined analysis. State and year specific multiple imputation models were developed to replace missing data with estimated values. AIS, MAIS, ISS, Barell Matrix nature of injury, and body region were calculated from ICD-9-CM codes. Means, medians, and percentages are used to describe differences between older and younger occupants.

Results: There were 7,131,628 persons aged 21 years or older available for analysis. The majority (90%) were under 65 years-of-age. More than 54,000 persons were over age 85. Safety restraint use was positively correlated with age. Nearly 97% of those 85+ were reported as using safety restraints whereas 90% of those 21-64 were reported restrained. Injury severity and the percent of persons treated at the hospital or dying increased with age. Thirteen percent of hospital treated persons 21–64 incurred MAIS 2+ injury compared to 30 percent of occupants 85+. While more than half of 21- to 64-year-olds killed were not reported restrained, three-quarters of those occupants 85+ were killed. Over 25 percent of hospitalized 85+ occupants were discharged dead, to rehab, or long-term care, while 98 percent of those 21–64 were discharged home. The percent of cervical vertebral column injuries decreased from 27 percent in 21- to 64-year-old hospital treated occupants to 9 percent in those 85+. Conversely, chest injuries account for 6 percent of injuries in those 21–64 but the percentage increases to 20 percent for those 85 and older. The percent of fractures increased from 8 percent of all injuries in 21- to 64-year-olds to almost 25 percent in those 85+. Internal injuries became more prevalent as age increased.

Conclusions: Despite having higher safety restraint usage rates, older occupants sustain more severe injuries and are more likely to suffer fractures, internal, and chest injuries. This study demonstrates that safety restraints may not have the same protective impact on older drivers as they do on younger drivers. Results of this study can be used to improve trauma care for older occupants.

Introduction

Older persons are the fastest growing age group in the United States. The number of persons over the age of 85 is projected to increase by 350 percent between the years 2000 and 2050. By 2050, one in five people in the United States is expected to be age 65 or older (Wiener & Tilly, 2002). As the population becomes older, potential impacts to traffic safety and resulting motor vehicle crash (MVC) outcomes should be studied. Previous studies of older drivers have shown that they are more likely to be involved in intersection and T-bone MVCs (Cook et al., 2000; Chen et al., 2012; Friedman et al., 2013) MVCs involving older drivers are also more likely to result in hospitalization or death (Cook et al., 2000). As medical care costs increase along with

the age of the population, a better understanding of the types of injuries that older occupants sustain in MVCs is needed. The purpose of this study is to examine injuries and medical outcomes of MVC-involved older occupants.

Methods

To study injuries and medical outcomes of older occupants involved in MVCs, this study uses probabilistically linked MVC and hospital data from participants in the National Highway Traffic Safety Administration's (NHTSA) Crash Outcome Data Evaluation System (CODES) Network. Use of this database was approved by the University of Utah Institutional Review Board.

Data Source

This project uses the GUM for crash years 2005-2008. The GUM incorporates probabilistically linked MVC and hospital records from 11 CODES States.

Population

This study was restricted to drivers and passengers of passenger vehicles and light trucks in transport who were age 21 or older at the time of the MVC. Only MVCs occurring in the traffic way were included in the data set. Older occupants are defined to be age 65 or older at the time of the MVC.

Analysis

We used counts and percentages to summarize MVC and hospital characteristics. We also used medians to compare hospital charges across reported safety restraint use. All analyses were completed in SAS 9.3 (SAS Institute Inc., 2002).

Results

In 2005-2008, there were 7,131,628 occupants age 21 or older who were riding in a passenger vehicle or light truck at the time of the MVC identified in the GUM. The majority (90%) were age 21 to 64, and 10 percent were age 65 or older. More than 54,000 occupants were age 85 years or older.

Occupant Characteristics

Table 2.1.1 displays occupant characteristics of the MVC population. The likelihood of being a driver was negatively associated with age. While nearly 83 percent of those age 21 to 64 years were drivers, this percentage dropped to 75 percent for those age 85 years or older. Safety restraint use, conversely, was positively correlated with age. Nearly 97 percent of those age 85 or older were coded as using safety restraints whereas 90 percent of those age 21 to 64 were similarly coded. MVC reported severity shows that older occupants were less likely to be coded as uninjured (77% for 85+ versus 90% for 21 – 64) and more likely to be killed (1% for 85+ versus 0.2% 21 – 64).

Table 2.1.1. Occupant characteristics by age group.									
Age Group	Total (%)	Driver	Restraint Use	MVC Reported Injury Severity*					
				O	C	B	A	K	U
21 – 64	6,441,215 (90.3%)	82.7%	90.2%	90.4%	13.7%	4.3%	1.4%	0.2%	0.1%
65 – 69	232,910 (3.3%)	82.4%	97.5%	80.1%	13.1%	4.2%	1.4%	0.3%	0.1%
70 – 74	171,968 (2.4%)	81.1%	97.5%	80.2%	13.1%	4.7%	2.7%	0.4%	0.1%
75 – 79	137,363 (1.9%)	80.6%	97.4%	79.7%	12.7%	5.2%	1.7%	0.5%	0.1%
80 – 84	93,576 (1.3%)	79.5%	97.2%	78.7%	12.9%	5.9%	1.9%	0.6%	0.1%
85 +	54,596 (0.8%)	75.2%	96.9%	77.4%	12.8%	6.7%	2.1%	1.0%	0.1%
Total	7,131,628 (100%)	82.5%	90.9%	80.3%	13.6%	4.3%	1.4%	0.2%	0.1%

- O = Not Injured, C = Possible Injury, B = Non-incapacitating Injury, A = Incapacitating Injury, K = Fatal Injury, U = Injury Severity Unknown

Driver MVC Characteristics

MVC characteristics by driver age are shown in Table 2.1.2. Older driver MVCs were more likely to occur at an intersection (52% for 85+ versus 44% for 21 - 64) and have traffic controls present. Surprisingly, distraction was more likely to be listed as a contributing factor in older driver MVCs compared to younger driver MVCs (19% for 85+ versus 12% for 21-64). Younger driver MVCs were more likely to be speed- or fatigue-related and occur on interstates. Younger drivers were more likely to be involved in single vehicle (11% for 85+ versus 15% for 21 – 64) (Table 2.1.3) and rear-end (23% for 85+ versus 37% for 21-64) MVCs. Angle MVCs, those most likely to be associated with an intersection, were highest among older drivers (43% for 85+ vs. 26% for 21 – 64).

Table 2.1.2: MVC characteristics by driver age group.						
Age Group	Daylight Hours	Intersection Related	Traffic Control Present	Speed Related	Fatigue Related	Distraction Related
21 – 64	73.8%	44.2%	41.7%	5.9%	0.8%	12.0%
65 – 69	81.3%	45.9%	44.6%	3.2%	0.8%	12.1%
70 – 74	83.5%	47.4%	45.9%	2.9%	0.9%	13.4%
75 – 79	85.5%	48.7%	47.0%	2.6%	1.0%	15.1%
80 – 84	87.3%	49.5%	47.8%	2.4%	1.0%	17.2%
85 +	88.7%	51.5%	48.5%	2.3%	0.9%	19.2%
Total	74.7%	44.5%	42.2%	5.6%	0.8%	12.2%

Table 2.1.3. MVC configuration by driver age group.			
Age Group	MVC Type		
	Single Vehicle	Rear-end	Angle
21 – 64	15.3%	36.8%	26.3%
65 – 69	12.9%	32.6%	31.7%
70 – 74	12.2%	30.2%	34.9%
75 – 79	11.7%	27.5%	38.1%
80 – 84	11.0%	25.1%	40.4%
85 +	10.8%	22.5%	43.3%
Total	15.0%	36.1%	27.2%

Medical Outcomes

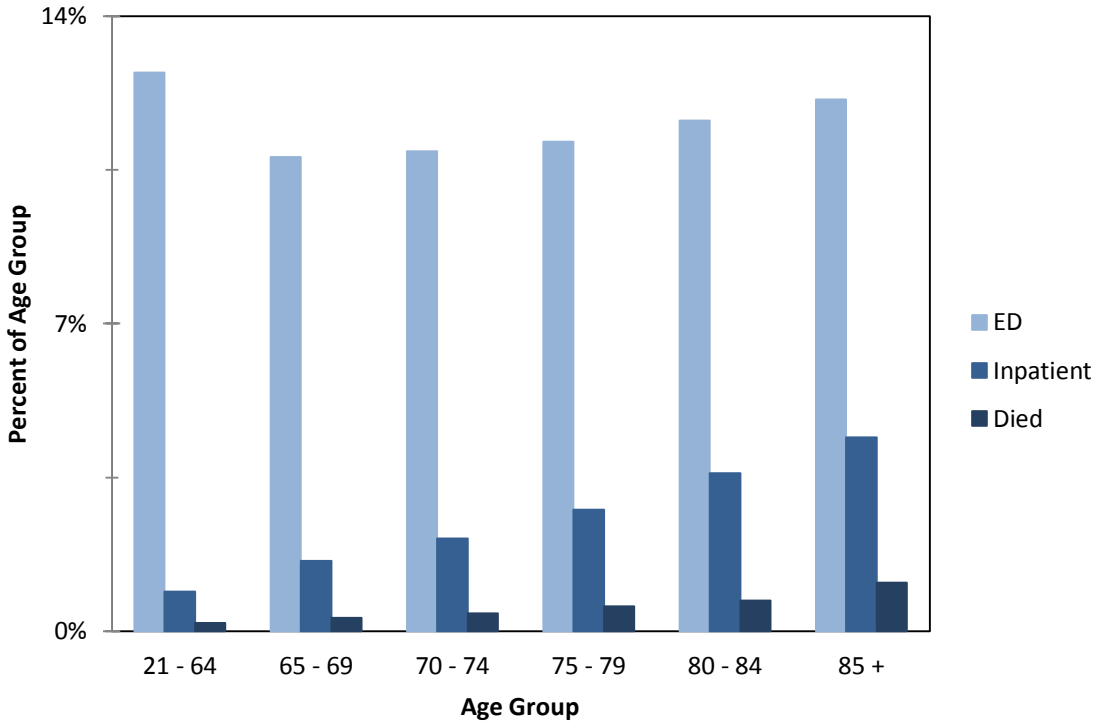
As shown in Figure 2.1.1, the percentage of occupants that were treated at the hospital or died increased with age. While emergency department treatment was highest for occupants age 21 to 64 years the reverse was true for being admitted to the hospital and dying in the hospital or at the scene of the MVC. We examined the percent of occupants in each level of care that were using safety restraints by age group. Not surprisingly, well over 90 percent of occupants that were not treated at the hospital or killed were using safety restraints. Also, not surprising is that the percent of occupants using safety restraints decreased as the level of care increased. While more than half of those age 21 to 64 that were killed were not restrained, three-quarters of those age 80 or older that were killed were using safety restraints (Figure 2.1.2).

As expected, the percent of hospital treated MVC occupants using public insurance greatly increased with age from a low of 7.9 percent for those age 21 to 64, to a high of 36.8 percent for those age 85 years or older. Hospital treated older occupants were also more likely to incur at least moderate injuries as measured by the Maximum Abbreviated Injury Scale (MAIS) compared to younger occupants (Table 2.1.4). Only 13 percent of occupants age 21 to 64 years incurred at least moderate injuries compared to 19 percent of those age 65 to 69 years, 21 percent of those age 70 to 74 years, 25 percent of those age 75 to 79 years, 27 percent of those age 80 to 84 years, and 30 percent of those age 85 years or older.

Table 2.1.4. Maximum Abbreviated Injury Scale (MAIS) by age group.		
Age Group	MAIS	
	No Injury Code to Minor Severity	Moderate to Maximum Severity
21 – 64	87.0%	13.0%
65 – 69	81.2%	18.8%
70 – 74	78.7%	22.3%
75 – 79	75.2%	24.8%
80 – 84	72.9%	27.1%
85 +	69.7%	30.3%
Total	86.0%	14.0%

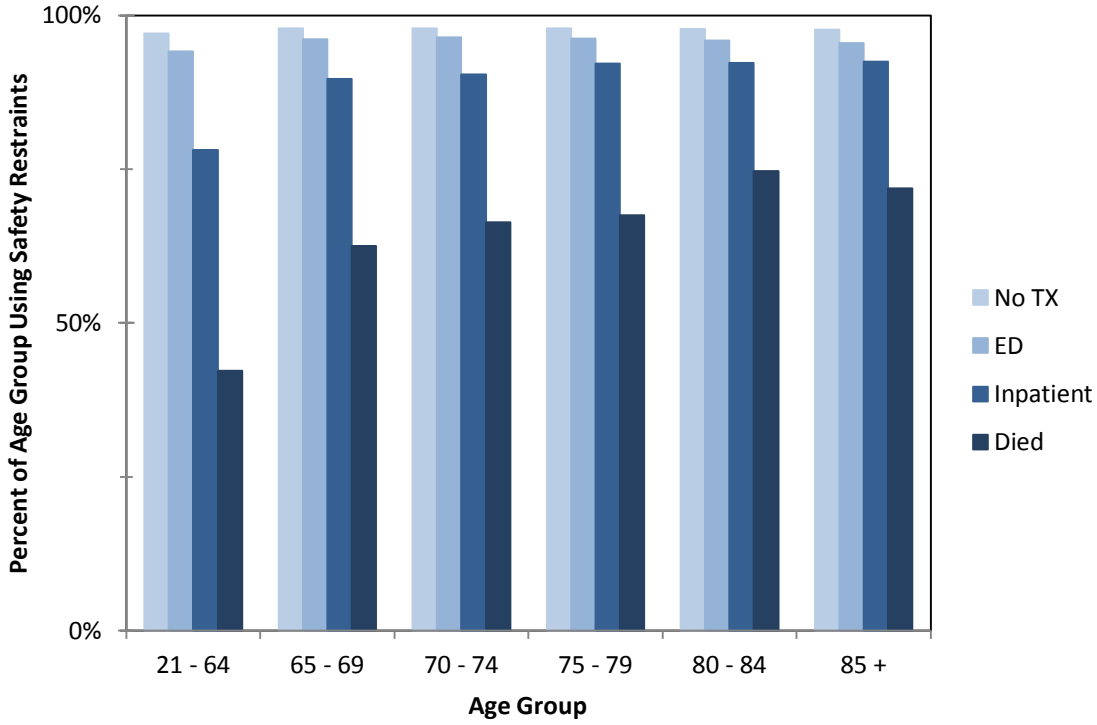
While almost all hospital treated occupants age 21 to 64 years were discharged home (98%) the same was not true for the older age groups: 65 to 69 (95%), 70 to 74 (93%), 75 to 79 (90%), 80 to 84 (86%), and 85 (81%) years or older. Figure 2.1.3 shows the discharge status for all hospitalized occupants not discharged home by age group. The percent of occupants discharged to longer term care (LTC) and rehab greatly increases with age.

Figure 2.1.1. Highest level of care by age group.



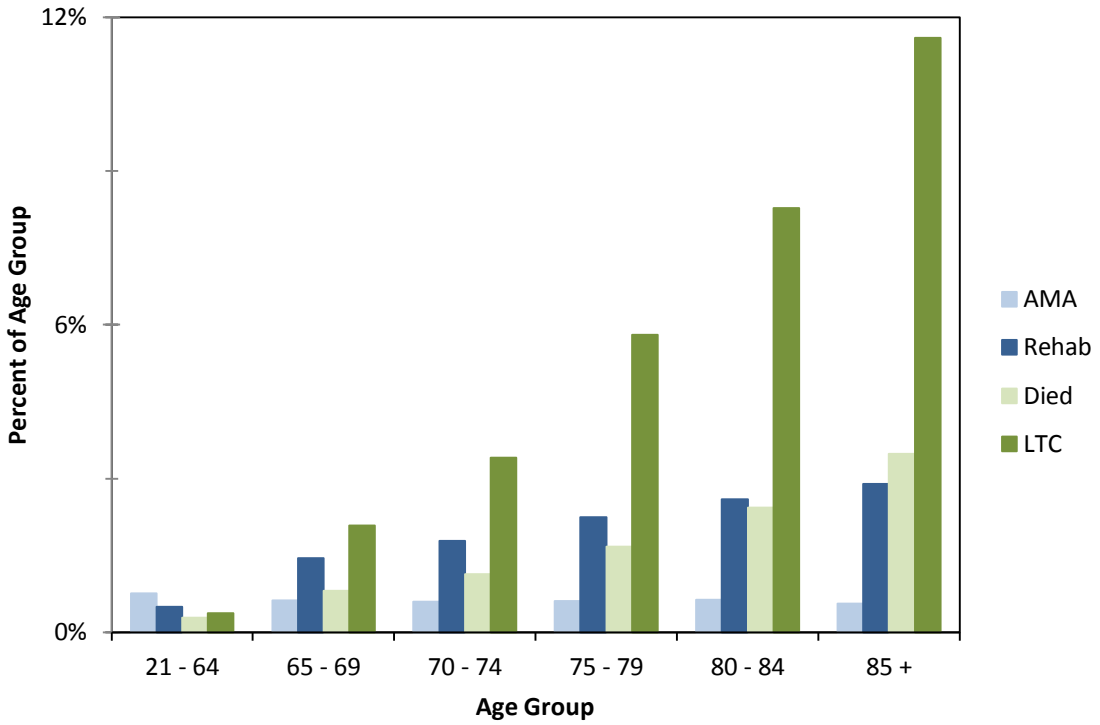
* ED = Emegency Department

Figure 2.1.2. Safety restraint usage by highest level of care by age group*.



* No TX= No Treatment, ED = Emergency Department

Figure 2.1.3. Discharge status by age group of those not discharged home*.



*AMA = Left Against Medical Advice, Rehab = rehabilitation, LTC = Long Term Care

The top five injured body regions, based on the Barell Matrix, by age group are displayed in Table 2.1.5. Cervical vertebral column injuries, commonly associated with whiplash, are the most common injury for those age 21 to 64, 65 to 69, and 70 to 74 years. After age 75 years chest injuries are the most common. Across all age groups, cervical vertebral column injuries decrease from a high of 27 percent in occupants age 21 to 64 years to a low of 9 percent in those age 85 years or older. Conversely, chest injuries account for only 6 percent of injuries in occupants age 21 to 64 years but increase to 20 percent for those age 85 years or older.

Table 2.1.5. Top five injured body regions by age group.

	Age Groups					
Ranking	21 – 64	65 – 69	70 – 74	75 – 79	80 – 84	85 +
1	Cervical VCI (26.9%)	Cervical VCI (21.1%)	Cervical VCI (17.4%)	Chest (18.1%)	Chest (20.0%)	Chest (19.8%)
2	Chest (6.3%)	Chest (13.5%)	Chest (16.4%)	Cervical VCI (14.3%)	Cervical VCI (11.1%)	Cervical VCI (9.2%)
3	Wrist/Hand (6.2%)	HFN Unsp (6.3%)	HFN Unsp (6.5%)	HFN Unsp (6.7%)	HFN Unsp (7.0%)	HFN Unsp (7.2%)
4	HFN Unsp (6.1%)	Shoulder & Arm (5.9%)	Shoulder & Arm (5.9%)	Wrist/Hand (5.8%)	Wrist/Hand (6.1%)	Wrist/Hand (5.9%)
5	Shoulder & Arm (5.7%)	Wrist/Hand (5.7%)	Wrist/Hand (5.8%)	Shoulder & Arm (5.5%)	Shoulder & Arm (5.2%)	Other Head (5.7%)

VCI = Vertebral Column Injury
HFN Unsp = Head, Face, Neck Injury Unspecified

Fractures increased from only 8 percent of all injuries in occupants age 21 to 64 to almost 25 percent in those age 85 years or older (Table 2.1.6). Additionally, internal injuries became more prevalent as age increased.

Table 2.1.6. Top five natures of injury by age group.

	Age Group					
Ranking	21 – 64	65 – 69	70 – 74	75 – 79	80 – 84	85 +
1	Spr & Str (47.2%)	Spr & Str (35.9%)	Superficial (31.5%)	Superficial (32.8%)	Superficial (33.6%)	Superficial (32.6%)
2	Superficial (24.3%)	Superficial (29.5%)	Spr & Str (29.8%)	Spr & Str (24.1%)	Fracture (21.9%)	Fracture (24.4%)
3	Fracture (8.1%)	Fracture (14.0%)	Fracture (16.3%)	Fracture (19.1%)	Spr & Str (18.7%)	Spr & Str (14.4%)
4	Unspecified (7.2%)	Unspecified (6.9%)	Open Wound (7.3%)	Open Wound (8.2%)	Open Wounds (9.9%)	Open Wounds (11.2%)
5	Open Wound (7.1%)	Open Wound (6.7%)	Unspecified (7.0%)	Internal (7.0%)	Internal (7.6%)	Internal (8.4%)

Spr & Str = Sprains and Strains

The pattern of injured body regions for occupants using safety restraints was similar to the overall pattern in Table 2.1.7. Different injury patterns were observed in the unrestrained occupant group however, with traumatic brain injuries (TBI) becoming more prevalent in all age groups (Table 2.1.7).

Table 2.1.7. Top five injured body regions by age group among unrestrained occupants.

	Age Group					
Rating	21 – 64	65 – 69	70 – 74	75 – 79	80 – 84	85 +
1	Cervical VCI (14.8%)	Cervical VCI (12.5%)	Chest (13.1%)	Chest (16.1%)	Chest (16.2%)	Chest (14.9%)
2	TBI (9.2%)	Chest (13.5%)	TBI (13.0%)	TBI (12.9%)	TBI (12.8%)	TBI (11.8%)
3	HFN Unsp (8.4%)	TBI (9.8%)	Cervical VCI (11.5%)	HFN Unsp (8.9%)	HFN Unsp (8.8%)	HFN Unsp (8.4%)
4	Face (7.8%)	Other Head (8.2%)	HFN Unsp (7.7%)	Cervical CVI (8.4%)	Cervical CVI (7.3%)	Other Head (7.8%)
5	Other Head (7.7%)	HFN Unsp (8.2%)	Face (6.2%)	Face (6.4%)	Other Head (6.6%)	Cervical CVI (7.7%)

VCI = Vertebral Column Injury
TBI = Traumatic Brain Injury
HFN Unsp = Head, Face, Neck Injury Unspecified

Median hospital charges for all hospital treated occupants are summarized in Table 2.1.8. Older occupants tended to have higher ED and hospital admission charges compared to younger occupants. Median ED charges increase by just over $200 between occupants age 21 to 64 years and those age 85 years or older. Hospital admission charges see a nearly $900 increase between the same two age groups.

Table 2.1.8. Median hospital charges and 95% confidence intervals in 2008 dollars by age group.		
ED		**Charges (95% CI)**
Age Group	21 – 64	808 (439, 1,550)
	65 – 69	912 (482, 1,854)
	70 – 74	959 (506, 1,988)
	75 – 79	995 (521, 2,077)
	80 – 84	1,000 (526, 2,081)
	85 +	1,016 (528, 2, 162)
Hospital admission		
Age Group	21 – 64	17,711 (9,105, 36,734)
	65 – 69	18,325 (9,455, 37,961)
	70 – 74	18,059 (9,263, 36,338)
	75 – 79	19,198 (9,983, 38,983)
	80 – 84	18,184 (9,478, 36,056)
	85 +	18,584 (9,349, 36,409)

Conclusions

Our analysis used combined CODES data from eleven States and showed that older drivers have distinctive MVC patterns as seen in other studies.(Cook et al., 2000, Chen et al., 2012, Friedman et al., 2013) Additionally, older occupants were more likely to be treated at the hospital following the MVC compared to younger occupants. In particular, those age 85 year or older were nearly five times more likely to be admitted to the hospital compared to those age 21 to 64 years. Older occupants were much more likely to be discharged from the hospital to long term care and rehab compared to younger occupants, where almost all were discharged home. For those occupants treated at the hospital, we found that older occupants had distinctive injury patterns. Older occupants were more likely to have chest injuries, whereas younger occupants were more likely to sustain cervical vertebral column injuries. Finally, older occupants tended to have higher median hospital charges compared to younger occupants. These results can be used to advocate for safety programs and support safety improvements targeted at older occupants.

References

Chen, H., Cao, L., & Logan, D. B. (2012). Analysis of risk factors affecting the severity off intersection crashes by logistic regression. *Traffic Injury Prevention, 13*(3), 300-307.

Cook, L. J., Knight, S., Olson, L. M., Nechodom, P. J., Dean, & J. M. (2000). Motor vehicle crash characteristics and medical outcomes among older drivers in Utah, 1992 - 1995. *Annals of Emergency Medicine 35*(6), 585-91.

Friedman, C., Mcgwin, G. J., Ball, K. ., Owsley, C. (2013). Associations between higher order visual processing abilities and a history of motor vehicle collision involvement by drivers ages 70 and over. *Investigative Opthalmology & Visual Science,* 54(1), 778-82.

SAS Institute Inc. (2002). SAS software. Cary, NC: SAS Institute Inc.

Wiener, J. M., & Tilly, J. (2002). Population aging in the United States of America: Implications for public programmes. *International Journal of Epidemiology, 31*(4), 776-781.

Analysis 2. Comparison of Medical Outcomes by Reported Safety Restraint Use among Children Ages 1 to 7 Years

Abstract

Objective: Compare medical outcomes of motor vehicle crash (MVC)-involved children between the ages 1 to 7 years reported as using child safety restraints, only seat belts, and no safety restraints.

Methods: We used the Crash Outcome Data Evaluation System's (CODES) General Use Model for the crash years 2005-2008. This dataset contains probabilistically linked motor vehicle crash and hospital records from 11 CODES States. Only children between the ages 1 to 7 years riding in passenger vehicles or light trucks that were involved in MVCs occurring in the trafficway were included. Reported safety restraint use was classified into three groups: CRS (child safety restraints used), seat-belt (only seat belts used), and unrestrained (no restraints used). We used summary statistics to compare medical outcomes by safety restraint use.

Results: There were 390,920 children in the dataset: 57.0 percent of children were CRS restrained, 40.3 percent were seat-belt restrained, and 2.7 percent were unrestrained. Less than 10 percent of CRS and seat-belt restrained children went to the hospital or died at the scene following a MVC compared to 21.9 percent among unrestrained children. The odds of sustaining an injury to the neck, back, or abdomen among CRS restrained children were almost half the odds among unrestrained children (OR: 0.64; 95% CI: 0.59, 0.70). This reduction was less evident among seat-belt restrained children (OR: 0.91; 95% CI: 0.83, 1.00). CRS restrained children had the lowest median hospital charges with seat-belt restrained children having the second lowest median hospital charges.

Conclusions: Among children ages 1 to 7 years, CRS restraint use was associated with the best medical outcomes compared to seat-belt only and no restraint use and seat-belt only use had better medical outcomes than no restraint use. These findings suggest that both CRS and seat-belt restraint use among children ages 1 to 7 years offer some protection; however, CRS restraint use is preferred.

Introduction

Motor vehicle crashes (MVCs) are a leading cause of injury-related deaths among children in the United States (Centers for Disease Control and Prevention, 2011). Child safety restraints, when used properly, are an effective way to reduce injuries and deaths among children involved in MVCs (NHTSA, 2008). Fatalities are reduced by over 70 percent for infants and over 50 percent for toddlers using child safety restraints in passenger vehicles (National Highway Traffic Safety Administration, 2008). Child safety restraint laws have been shown to increase child safety restraint use and reduce injuries and fatalities among children involved in MVCs (Zaza et al., 2001). A recent study found that when child restraint laws were extended to include older children ages 6 to 8 years, more children were placed in child safety restraints and were seated in rear rows, contributing to a reduction of injuries and fatalities (Eichelberger et al., 2012). Although using no restraints results in injuries and fatalities among children involved in MVCs,

using only seat belts as opposed to child safety restraints can also be harmful to children because seat belts do not fit children properly and can result in injuries to the neck, spine, back, or abdomen (Durbin et al., 2003; NHTSA, 2008). The goal of this study is to compare medical outcomes of MVC-involved children between the ages 1 and 7 years using child safety restraints, only seat belts, and no safety restraints.

Methods

To compare medical outcomes of children between the ages 1 to 7 years involved in MVCs by reported safety restraint use, we used the Crash Outcome Data Evaluation System's (CODES) standardized (CODES) General Use Model (GUM) data from contributing States for 2005-2008. The University of Utah Institutional Review Board approved this study.

Data Source

This project uses the GUM for crash years 2005-2008. The GUM incorporates probabilistically linked MVC and hospital records from 11 CODES States. Only children between the ages 1 to 7 years riding in passenger vehicles or light trucks that were involved in MVCs occurring in the trafficway were included in this study. Passenger vehicles and light trucks include sedans, station wagons, pickup trucks, SUVs, and minivans.

Definitions

Safety restraint use as reported on the police crash report was classified into three groups: CRS, seat-belt, and none. CRS (child restraint system) use is defined as the use of child safety restraints (car or booster seats) for children ages 1 to 7 years. Seat-belt restraint use includes children ages 1 to 7 years using only seat belts. Restraint use was classified as none when no safety restraints were reported used. The classification of safety restraint use is based in part on the definitions outlined by Olsen et al. (2010).

Analysis

We used counts and percentages to summarize crash and hospital characteristics by safety restraint use for children ages 1 to 7 years. We also used medians to compare hospital charges across safety restraint use. All analyses were performed in SAS 9.2 (SAS Institute Inc. 2002).

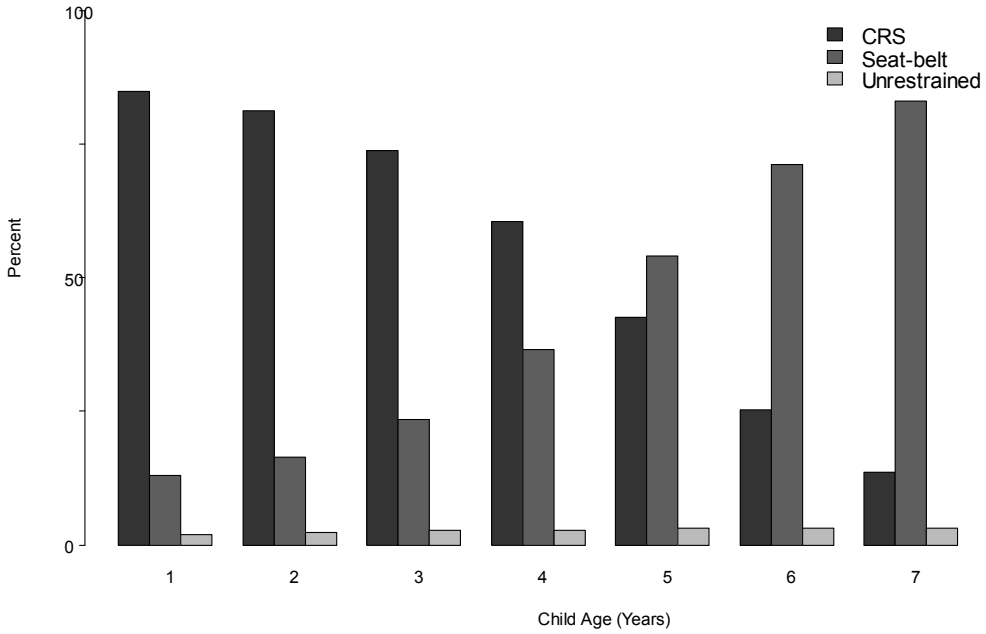
Results

There were 390,920 children ages 1 to 7 years in the GUM dataset: 57.0 percent were CRS restrained, 40.3 percent were seat-belt restrained, and 2.7 percent were unrestrained.

Crash Characteristics

Figure 2.2.1 summarizes child age by child safety restraint use. Younger children ages 1 to 4 years were more likely to be CRS restrained. As child age increased, CRS restraint use decreased such that the majority of children ages 5 to 7 years were seat-belt restrained. In this study, the percent of unrestrained children did not substantially change across child age.

Figure 2.2.1. Child age by reported child safety restraint use.



Almost all (95.6%) of CRS restrained children were riding in rear rows at the time of the MVC. Only 81.3 percent of seat-belt restrained and 79.2 percent of unrestrained children were riding in rear rows at the time of the MVC, with a fifth of these children riding in the front row.

Table 2.2.1 contains driver safety restraint use by child safety restraint use. Nearly all of children that were CRS or seat-belt restrained were riding with drivers that were using safety restraints. About a third (29.4%) of unrestrained children were riding with drivers that were also unrestrained.

Table 2.2.1. Driver safety restraint use by reported child safety restraint use.

	CRS *n=222,668*		**Seat-belt** *n=157,614*		**Unrestrained** *n=10,638*	
	#	**%**	**#**	**%**	**#**	**%**
Driver restrained	219,144	98.4	155,241	98.5	7,514	70.6
Driver unrestrained	3,524	1.6	2,373	1.5	3,124	29.4

Medical Outcomes

Less than 10 percent of CRS and seat-belt restrained children went to the hospital or died at the scene or hospital following a MVC compared to over a fifth (21.9%) among unrestrained children. Of all children involved in MVCs, 18.7 percent of unrestrained children were treated in the emergency department (ED) and 2.3 percent were admitted to the hospital. Only 8.3 percent of CRS and 9.1 percent of seat-belt restrained children were treated in the ED while 0.2 percent of CRS and 0.3 percent of seat-belt restrained children were admitted to the hospital. The distribution of highest level of care received by child safety restraint use is summarized in Table 2.2.2.

Table 2.2.2. Highest level of care by reported child safety restraint use.						
	CRS *n=222,668*		**Seat-belt** *n=157,614*		**Unrestrained** *n=10,638*	
	#	**%**	**#**	**%**	**#**	**%**
No hospital record	203,684	91.5	142,670	90.5	8,304	78.1
Emergency department	18,375	8.3	14,334	9.1	1,991	18.7
Hospital admission	495	0.2	515	0.3	248	2.3
Died at scene or hospital	114	0.1	95	0.1	95	0.9

Unrestrained children had more severe injuries compared to CRS restrained children, with 16.5 percent of unrestrained children sustaining at least minor injuries on the Maximum Abbreviated Injury Scale (MAIS) compared to 5.0 percent of CRS restrained children. Seat-belt restrained children also tended to be more likely to have at least minor injuries compared to CRS restrained children. Injury severity by child safety restraint use is compared in Table 2.2.3.

Table 2.2.3. MAIS by reported child safety restraint use.						
	CRS *n=222,668*		**Seat-belt** *n=157,614*		**Unrestrained** *n=10,638*	
	#	**%**	**#**	**%**	**#**	**%**
No hospital record	203,684	91.5	142,670	90.5	8,304	78.1
Not injured	7,855	3.5	4,709	3.0	574	5.4
Minor	10,001	4.5	9,074	5.8	1,341	12.6
Moderate	751	0.3	799	0.5	191	1.8
Serious	172	0.1	187	0.1	96	0.9
Severe	102	0.0	86	0.1	42	0.4
Critical	24	0.0	18	0.0	11	0.1
Maximum/died at scene	79	0.0	71	0.0	79	0.7

Table 2.2.4 summarizes the number of children that went to the hospital and had neck, back, or abdomen injuries and traumatic brain injury (TBI) by child safety restraint use. The odds of sustaining an injury to the neck, back, or abdomen among CRS restrained children were almost half the odds among unrestrained children (OR: 0.64; 95% CI: 0.59, 0.70). A similar result is seen in comparing CRS to seat-belt restrained where the odds of sustaining an injury to the neck, back, or abdomen were also significantly lower (OR: 0.71; 95% CI: 0.68, 0.74). Seat-belt restrained children were also less likely to have injuries to the neck, back, or abdomen compared to unrestrained children; however, this difference was much smaller (OR: 0.91; 95% CI: 0.83, 1.00). The odds of TBI were also significantly lower among CRS and seat-belt restrained children compared to unrestrained children: CRS restrained children had a 75-percent reduction in the odds (OR: 0.26; 95% CI: 0.21, 0.32) of receiving a TBI and seat-belt restrained children had a 70% reduction (OR: 0.30; 95% CI: 0.24, 0.38).

Table 2.2.4. Neck, back, or abdomen injuries and TBI of children that went to the hospital by reported child safety restraint use.						
	CRS *n=18,907*		**Seat-belt** *n=14,874*		**Unrestrained** *n=2,257*	
	#	**%**	**#**	**%**	**#**	**%**
Neck, back, abdomen						
No	12,387	65.5	8,529	57.3	1,241	55.0
Yes	6,520	34.5	6,345	42.7	1,016	45.0
Traumatic brain injury						
No	18,425	97.5	14,435	97.0	2,056	91.1
Yes	482	2.5	439	3.0	201	8.9

Of children that went to the hospital following a MVC, over 99.0 percent of CRS and seat-belt restrained children were discharged home. Less than 1.0 percent of these children left against medical advice, died, or were discharged to long term care or rehab. Fewer unrestrained children were discharged home (98.4%), with 1.6 percent leaving against medical advice, dying, or continuing onto long term care or rehab.

CRS restrained children had the lowest median ED and hospital admission charges compared to unrestrained and seat-belt restrained children (see Table 2.2.5). Median hospital charges for unrestrained children were $196 and $3687 more for ED and hospital admissions, respectively, than the median hospital charges for CRS restrained children. The difference in medians between seat-belt and CRS restrained children was smaller, with an $88 difference in ED charges and a $246 difference in hospital admissions.

Table 2.2.5. Median hospital charges in 2008 dollars by reported child safety restraint use.			
	CRS *n=222,668*	**Seat-belt** *n=157,614*	**Unrestrained** *n=10,638*
Emergency department	$355.40	$443.36	$551.72
Hospital admission	$13,496.54	$13,742.24	$17,183.36

Conclusions

This study has two main findings. First, among children ages 1 to 7 years, CRS restraint use was associated with the best medical outcomes compared to seat-belt restraint use and no restraint use. Second, children that were seat-belt restrained tended to have better medical outcomes than unrestrained children.

CRS restrained children had the lowest rate of ED visits or hospital admissions compared to seat-belt restrained and unrestrained children. CRS restrained children also sustained the lowest rates of injuries as measured by MAIS and had the lowest median ED and hospital admission charges. The majority of CRS restrained children were discharged home following a trip to the hospital. This group also saw the largest reduction in the odds of receiving a TBI. CRS restrained children were far less likely to sustain an injury to the neck, back, or abdomen compared to seat-belt restrained and unrestrained children, providing additional evidence that child safety restraints fit children better than seat belts (Durbin et al., 2003; NHTSA, 2008).

Although CRS restrained children had the best medical outcomes compared to children in the other two safety restraint groups, seat-belt restraint use also resulted in lower ED visits and hospital admission compared to unrestrained children. Seat-belt restrained children were less injured than unrestrained children and were discharged home at a similar rate as CRS restrained children. Seat-belt restrained children also saw a large reduction in the odds of receiving a TBI compared to unrestrained children.

These findings suggest that both CRS restraint and seat-belt restraint use among children ages 1 to 7 years offer some protection; however, CRS restraint use is preferred because it is associated with the best medical outcomes, including fewer injuries to the neck, back, or abdomen.

References

Centers for Disease Control and Prevention. (2011). Web-based injury statistics query and reporting system (WISQARS). Atlanta, GA: Author.

Durbin, D. R., Elliott, M. R., & Winston, F. K. (2003). Belt-positioning booster seats and reduction in risk of injury among children in vehicle crashes. *JAMA, 289*(21), 2835-2840.

Eichelberger, A. H., Chouinard, A. O., Jermakian, J. S. (2012). Effects of booster seat laws on injury risk among children in crashes. *Traffic Injury Prevention, 13*(6), 631-639.

National Highway Traffic Safety Administration. (2008). 2006 motor vehicle occupant protection facts. (Report No. DOT HS 810 654). Washington, DC: Author. Olsen, C. S., Cook, L. J., Keenan, H. T., & Olson, L. M. (2010). Driver seat belt use indicates decreased risk for child passengers in a motor vehicle crash. *Accident Analysis and Prevention, 42*(2), 771-777.

SAS Institute Inc. (2002). SAS software.Cary, NC: SAS Institute Inc..

Zaza, S., Sleet, D. A., Thompson, R. S., Sosin, D. M., & Bolen, J. C. (2001). Reviews of evidence regarding interventions to increase use of child safety seats. *American Journal of Preventative Medicine, 21*(4 Suppl), 31-47.

Analysis 3. Comparing Medical Outcomes by Helmet Use Laws in 11 States Using CODES Data

Abstract

Objective: Compare medical outcomes of motorcycle crashes between States with partial and universal helmet laws, and describe injuries related to motorcycle crashes.

Methods: We used the Crash Outcome Data Evaluation System's General Use Model's five States with universal helmet laws and six States with partial helmet laws for the years 2005-2008. The dataset consisted of motorcycle operators involved in crashes according to motor vehicle records probabilistically linked to hospital records. We described and compared medical outcomes between States with partial laws to those with universal laws. We estimated relative risks of medical outcomes and tested for differences in injury patterns using likelihood ratio tests.

Results: Reported helmet use was higher in universal law States (88% versus 42%). Billed emergency charges were higher in partial law States, as was the proportion of patients using public or government insurance (12% versus 9%). Injuries to the head and face were more common in partial law States. After adjusting for other factors, the relative risk of head and face injuries was higher when no helmet was worn: not wearing a helmet was associated with a 201-percent increase in the risk of head injuries and 263-percent increase in the risk of facial injuries in single-vehicle crashes in partial law States. Body regions and the nature of injuries sustained differed slightly by helmet law type, with more extremity injuries, sprains/strains, and contusion/superficial injuries observed in universal law States, and more head/neck injuries, fractures, and open wounds observed in partial law States.

Conclusions: Helmet use was effective in reducing head and facial injuries regardless of the type of helmet law or type of crash. The risk of head and facial injuries was higher in States with partial helmet laws compared to those with universal helmet laws.

Introduction

While motor vehicle crash (MVC) rates have declined in recent years, motorcycle crash injuries and fatalities have increased in the United States. (Centers for Disease Control and Prevention, 2012) Motorcycle helmets are effective in preventing head and brain injuries in MVCs and universal motorcycle helmet use laws have been shown to be effective at increasing the use of motorcycle helmets (NHTSA, 2011). However, helmet use laws vary state-to-state and have been difficult to pass and retain historically (Homer & French, 2009; NHTSA, 2011). The goal of this study is to compare rates of helmet use, medical care provided, head injuries, facial injuries, traumatic brain injury (TBI) rates, and resource utilization in terms of charges and length of stay between motorcyclists who crashed in universal law States to those who crashed in partial helmet use States. We also looked at the distribution of body regions injured and the nature of injuries of those motorcyclists involved in crashes in the 11 States and 4 years included in the study.

Methods

We used the Crash Outcome Data Evaluation System's (CODES) General Use Model (GUM) for 2005-2008 in this analysis. The University of Utah Institutional Review Board approved this study.

Data Source

This project uses the GUM for crash years 2005-2008. The GUM incorporates probabilistically linked MVC and hospital records from 11 CODES States. Only operators of motorcycles involved in a MVC were included in this study, excluding parked vehicles and crashes occurring outside of the traffic way.

Study Population

Of the 11 participating CODES States, five had a universal helmet law and six States had a partial helmet law during the study period. A total of 31 State/years were included in this analysis, with 10 States contributing data for 2005; 7 for 2006; 6 for 2007; and 8 for 2008. Universal and partial law States were represented in the data for each year.

The six partial helmet laws represented in this study vary, with age restrictions for un-helmeted riders ranging from 17 to 20 years old. Two laws require helmets for only those with instructional/learner's permits. One has provisions that require proof of medical insurance for unhelmeted riders over an age limit.

Analysis Methods

Helmet use rates are compared between universal and partial law States. We further describe the helmet use rates of those motorcycle riders covered by partial laws according to their age. We describe medical care provided and rates of injuries using counts and relative frequencies. Charges and length of stay are described using means, medians, and other descriptive statistics. Charges were adjusted for yearly inflation and differences between State incomes.

We examined the effect of helmet use on the rate of head injuries, facial injuries, traumatic brain injuries, and moderate to severe head or facial injuries (including fatalities) using generalized logistic regression models applied to motorcycle operators. Helmet use was as reported in State crash data from police accident reports and does not distinguish between different types of helmets. We excluded those who died at the scene since the specific injuries are unknown. We estimated the relative risk of sustaining each medical outcome of interest after adjusting for the following: gender, age, intersection related, night-time (9 pm to 5:59 am), poor surface conditions, year, type of crash (single vs. multi-vehicle), helmet law, and helmet use. We included interactions between helmet use and type of crash, and between helmet use and helmet law. The estimate of interest was the relative risk for the outcome when a helmet was not used compared to when a helmet was used.

We used multiply imputed datasets in all analyses, and combined results using appropriate methods (Schafer, 1999).

Results

Description of the Study Population

This study included 79,917 motorcycle operators, with 34,364 (43%) records submitted from partial helmet law States, and 45,552 (57%) from universal helmet law States. Reported helmet use was 42 percent in partial law States and 88 percent in universal law States. Among those operators covered by a partial helmet law according to their age (N=1,660), helmet use was 44 percent. In comparison, operators under age 21 in universal law States (N=4,166) showed a helmet use rate of 81 percent. Crash and operator characteristics are given in Table 2.3.1.

Table 2.3.1. Description of the study population.

Characteristic	Partial Law	Universal Law	Total
	N=34,364	N=45,552	N=79917
Helmet used	42%	88%	70%
Age (Median)	37	36	37
Male	93%	93%	93%
Single Vehicle Crash	39%	45%	43%
Crash at Intersection	36%	40%	39%
Night time	18%	17%	18%
Speed related*	10%	17%	15%
Suspicion of alcohol or drugs*	9%	5%	6%
Rural location*	17%	31%	28%
Poor Surface Conditions	6%	7%	7%
Medical Care Not linked to a hospital or emergency department visit and not dead at the scene	43%	40%	41%
Emergency department visit	39%	39%	39%
Median Charges	$1,986	$1,443	$1,618
Mean Charges	$3,688	$3,217	$3,398
Hospitalized	14%	18%	17%
Median Charges **	$32,287	$25,950	$28,389
Mean Charges**	$59,032	$56,325	$57,223
Mean Length of stay in days**	6.7	7.1	7.0
Discharged home**	81%	83%	82%
Died during medical treatment**	3%	3%	3%
Died at the scene	3%	3%	3%
Payer source for those with Emergency Department visit or hospital admission			
Public/Government	12%	9%	10%
Private Insurance	65%	69%	67%
Self/Uninsured	21%	20%	20%
Other	2%	3%	3%

* Available for only a subset of States: speed related 10, suspicion of alcohol or drugs 9, rural location 8 out of 11 included States

** Percentages are out of the number linking to a hospital visit.

Analysis of Medical Outcomes

Rates of linking crash records to ED and hospital medical records varied between States. Overall, 57% of the study population linked to an ED (40%) or hospital (17%) record. Median and mean ED and hospital charges were higher in partial law States among motorcyclists linking to a medical record. Although mean length of stay was longer in universal helmet law States, the number discharged home was also higher. Public insurance was responsible for costs incurred by 12 percent of the motorcycle operators in partial law States compared to 9 percent of those in universal law States who received medical care.

Head injuries, facial injuries, traumatic brain injuries, and moderate to severe head or facial injuries or death, were more frequent among motorcyclists in partial law States compared to universal law States. Table 2.3.2 gives the rates of these outcomes along with risk ratios comparing partial law States to universal law States regardless of helmet use. The rate of each outcome was higher in the partial law States. The largest difference was for facial injuries, which were 1.60 times more prevalent in partial law States compared to universal law States.

Table 2.3.2. Rates and Relative Risks of Medical Outcomes with 95% Confidence Intervals[1]

Statistic	Head Injury	Facial Injury	Traumatic Brain Injury	Moderate to Severe Head or Facial Injury or Death
Rate in Partial Law States	16.49% (16.03,16.95)	12.54% (12.13,12.95)	7.77% (7.43,8.10)	7.13% (6.82,7.45)
Rate in Universal Law States	12.18% (11.86,12.50)	7.83% (7.57,8.08)	7.13% (6.88,7.38)	6.16% (5.92,6.39)
Relative Risk for Partial vs. Universal Helmet Law	1.35 (1.30, 1.41)	1.60 (1.53, 1.68)	1.09 (1.03, 1.15)	1.16 (1.09, 1.23)

[1] Excluding operators who died at the scene

We estimated the adjusted relative risk of each outcome with multivariable models adjusting for age, type of crash (single vs. multi-vehicle), intersection, time of day, surface conditions, helmet law, helmet use and interactions for helmet use by type of crash, and helmet use by helmet law. Adjusted relative risks of each outcome are given in Table 2.3.3 comparing no helmet use vs. helmet use. Each model included significant interactions, meaning that the effect of helmet use depended on the type of crash and the type of helmet law. Four conditional adjusted relative risks are given for the four combinations of helmet law and type of crash. In all cases, helmets were protective. Effects ranged from a 42 percent increase in the risk of head injury for non-helmeted operators in multi-vehicle crashes in universal law States, to a 263-percent increase in the risk of facial injuries for non-helmeted operators in single-vehicle crashes in partial law States. The largest effects of helmets were seen in partial law States and single-vehicle crashes. However the risk of each outcome was higher among non-helmeted operators in universal law States and multi-vehicle crashes as well.

Table 2.3.3. Adjusted Relative Risks of Medical Outcomes for No Helmet Versus Helmet with 95% Confidence Intervals[1,2]				
Helmet law and type of crash	**Head Injury**	**Facial Injury**	**Traumatic Brain Injury**	**Moderate to Severe Head or Facial Injury or Death**
Partial Law, Multi-Vehicle	2.00 (1.85 ,2.16)	2.44 (2.20 ,2.69)	1.72 (1.52 ,1.93)	1.93 (1.70 ,2.19)
Partial Law, Single-Vehicle	3.01 (2.79 ,3.24)	3.63 (3.30 ,3.98)	2.63 (2.34 ,2.95)	2.94 (2.59 ,3.34)
Universal Law, Multi-Vehicle	1.42 (1.29 ,1.57)	1.53 (1.35 ,1.72)	1.57 (1.36 ,1.80)	1.65 (1.43 ,1.90)
Universal Law, Single-Vehicle	2.14 (1.95 ,2.35)	2.27 (2.02 ,2.55)	2.40 (2.09 ,2.76)	2.51 (2.17 ,2.90)
[1] Excluding operators who died at the scene [2] Adjusting for: gender, age, intersection, night-time, and poor surface conditions				

Description of Injuries

We describe the body regions injured among motorcycle operators receiving medical care in Table 2.3.4. In table 2.3.5, we look specifically look at motorcyclists who were, or would have been, covered by a partial helmet law according to State and age. For this comparison, we included motorcyclists less than 21 years old in universal helmet law States, and motorcyclists meeting age restrictions (ranging from 17 to 21 years old) in partial helmet law States.

Table 2.3.4. Body regions injured among motorcycle operators seen in the emergency department or admitted to the hospital.[1] P-values compare partial to universal law States.[2,3]						
	Emergency Department			**Hospital**		
Body Region	**Partial N=11,261**	**Universal N=18,057**	**P-value**	**Partial N=4,197**	**Universal N=8,456**	**P-value**
Head and Neck	**32%**	18%	0.00	**51%**	41%	0.00
Traumatic Brain Injury	**8%**	5%	0.00	**34%**	29%	0.00
Spine and Back	10%	9%	0.17	18%	18%	0.36
Torso	22%	20%	0.32	48%	47%	0.24
Extremities	72%	**74%**	0.00	76%	**79%**	0.00
Other and Unspecified	**25%**	23%	0.00	11%	11%	0.22
[1] Multiple body regions per motorcyclist were included. [2] P-values are testing universal vs. partial law States using the likelihood ratio test. [3] Bolded Percentages are those that are significantly higher when comparing partial to universal (P<0.01).						

The most prevalent body region injured among all groups was extremities. Head and neck injuries were the second most prevalent injuries for partial law States, with torso injuries being the second most prevalent injuries for hospital admissions in universal law States. Universal law States saw significantly fewer head and neck injuries, including traumatic brain injury, in the ED and hospital according to likelihood ratio tests. Universal law States saw slightly more extremity injuries in the ED and hospital, and slightly fewer other and unspecified body region injuries. The majority (80%) of injuries to other and unspecified body regions were contusion/superficial injuries.

Motorcyclists covered by partial helmet laws according to State and age had significantly more head and neck injuries, including traumatic brain injury, in the ED compared to riders under age 21 years covered by universal helmet laws (Table 2.3.5).

Table 2.3.5. Body regions injured among motorcycle operators covered by a helmet law according to age and State (partial law State), or under 21 years-old (universal law State) seen in the emergency department or admitted to the hospital.[1] P-values compare partial to universal law States.[2,3]

	Emergency Department			Hospital		
Body Region	**Partial N=686**	**Universal N=1716**	**P-value**	**Partial N=171**	**Universal N=618**	**P-value**
Head and Neck	**32%**	20%	0.00	50%	45%	0.29
Traumatic Brain Injury	10%	7%	0.02	39%	36%	0.46
Spine and Back	9%	7%	0.06	11%	17%	0.07
Torso	18%	16%	0.23	37%	45%	0.06
Extremities	74%	74%	0.85	71%	81%	0.01
Other and Unspecified	30%	28%	0.30	19%	15%	0.18

[1] Multiple body regions per motorcyclist were included.
[2] P-values are testing universal vs. partial law States using the likelihood ratio test.
[3] Bolded Percentages are those that are significantly higher when comparing partial to universal ($P<0.01$).

Conclusions

This study has three main findings. First, rates of head and facial injuries, including traumatic brain injury, are higher in States with partial helmet laws. Second, after adjusting for other factors, helmets are associated with decreased risk of head and facial injuries regardless of the type of crash or helmet use law. Third, head and neck injuries, including traumatic brain injuries, were more frequent in partial law States among motorcycle operators seen in the ED or hospital. The results indicate that, while many factors vary between States, including the implementation of helmet use laws and the compliance with those laws, helmets are consistently effective in preventing head and facial injuries.

References

Centers for Disease Control and Prevention. (2012). *Motorcycle safety guide*. Atlanta, GA: U.S. Department of Health and Human Services.

Homer, J., & French, M. (2009). Motorcycle helmet laws in the United States from 1990 to 2005: Politics and public health. *American Journal of Public Health, 99*(3), 415-423.

National Highway Traffic Safety Administration. (2011*). Countermeasures that work: A highway safety countermeasure guide for State highway safety offices. 6th ed.* Washington, DC: Author.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*(1), 3-15.

Analysis 4. Graduated Driver Licensing and Teenage Driver Involvement in Injury Crashes

Abstract
Objective: Describe the motor vehicle crash (MVC) characteristics of teenage drivers and compare the crude rates of teenage driver involvement in injury MVCs by ratings of graduated driver licensing (GDL) programs defined by the Insurance Institute for Highway Safety (IIHS). *Methods:* We used 2005-2008 data from the Crash Outcome Data Evaluation System (CODES) network's General Use Model to identify 16- to 18-year-old drivers in passenger cars and light trucks in transport that were involved in an injury MVC occurring in traffic ways. We computed IIHS ratings of GDL programs and used census data to derive per capita rates of teenage driver involvement in injury MVCs. We described MVC characteristics and medical outcomes of teenage drivers in injury MVCs and we estimated rates and rate ratios using Poisson regression models estimated by GEE. *Results:* "Good" GDL programs are associated with lower rates of teenage driver involvement in injury MVCs.

Introduction

Motor vehicle crashes (MVCs) are a leading cause of death and nonfatal injuries treated in hospital emergency departments (ED) (Centers for Disease Control and Prevention, 2011). Graduated driver licensing (GDL) programs aim to reduce exposure to dangerous situations to young and novice drivers by restricting their driving privileges, thus reducing their risk of being in MVCs while allowing them to gain driving experience in low-risk environments before obtaining full driving privileges. Although programs and restrictions vary from State to State, many include nighttime driving and passenger restrictions. MVCs among teenage drivers have declined in States that have adopted GDL programs (Ulmer et al., 2000; Foss et al., 2001; Mayhew et al., 2001; Shope and Molnar, 2004; Hallmark et al., 2008; Zhu et al., 2009), and the adoption of GDL programs has been associated with a reduction in the rate of fatal MVCs among young drivers (Baker et al. 2006). In 2000, the Insurance Institute for Highway Safety (IIHS) developed an algorithm for classifying State GDL systems into poor, "Fairl", "Marginal", and "Good" ratings. The algorithm has been used to evaluate the strength of GDL ratings, taking into account required practice hours, restrictions on nighttime driving, and restrictions on passengers allowed in the vehicle (McCartt et al., 2010).

Although GDL programs can be examined using MVC databases alone, such analyses are unable to include hospital outcomes of these MVCs. In this respect, the Crash Outcome Data Evaluation System (CODES) provides a unique platform for analyzing results in different GDL programs since it not only provides information about the MVC itself, but also provides data from the emergency department (ED) and hospital admission records. The purpose of this paper is to utilize the CODES data from multiple States to describe MVC characteristics and medical outcomes of teenage drivers in MVCs and to compare the rates of teenage driver involvement in injury MVCs across different IIHS ratings of GDL programs.

Methods

To study MVC characteristics and medical outcomes of teenage drivers involved in MVCs, this study uses probabilistically linked MVC and hospital data from participants in NHTSA CODES Network. Use of the standardized CODES data from various States was approved by the University of Utah Institutional Review Board.

Data Sources

This project uses the GUM for crash years 2005-2008. The GUM incorporates probabilistically linked MVC and hospital records from 11 CODES States.

This study uses data provided by IIHS (2012) to determine for each State and month in the GUM the main components of GDL programs that applied to teenage drivers. These components include: minimum permit age, permit holding period, minimum required number of practice hours, restrictions on nighttime driving and restrictions on passengers. This study also uses the algorithm described by McCartt et al. (2010) to rate GDL programs as either "Marginal/Fair" or "Good".

To compute the rate of teenage driver involvement in injury MVCs, this project relies on intercensal estimates of the resident population of 16- to 18-year-olds within each State and year represented in the GUM (US Census Bureau 2012).

Definitions

The primary outcome of this study is the rate of teenage driver involvement in injury MVCs that occur in traffic ways. Teenage drivers are defined to be 16 to 18 year-old drivers of passenger cars or light trucks in transport. Injury MVCs are defined as MVCs in which at least one occupant sustained a moderate to critical injury (MAIS $\geq$ 2) or died following the MVC. Within any given set of States and months, the rate of teenage driver involvement in injury MVCs is defined as the total number of teenage drivers reported in injury MVCs divided by the total number of teenage-years. Total teenager-years are estimated from the intercensal estimates of the resident population of 16- to 18-year-olds within States by dividing the estimated number of resident 16- to 18-year-olds for each State and year by 12 and summing over the given set of States and months.

Analysis

We used percentages to summarize the MVC and medical characteristics of teenage drivers. To analyze the crude rates of teenage driver involvement in injury MVCs, we used single Poisson regressions of the counts of teenage drivers with the natural logarithm of the total number of teenager-years as an offset. To account for clustering of teenage driver counts within States, the regression models were estimated by generalized estimating equations (GEE) using an independence working correlation structure to ensure numerical stability and achieve unbiased estimates. All analyses were completed in SAS software, version 9.3 (SAS Institute Inc., 2002). We used multiply imputed datasets in all analyses and combined results using appropriate methods (Schafer, 1997).

Results

Teenage Driver MVC Characteristics and Medical Outcomes

A total of 519,094 teenage drivers were involved in MVCs in the 2005-2008 GUM dataset. Of these teenage drivers, 45.5 percent, and 54.5 percent were exposed to "Marginal/Fair" and "Good" GDL programs. Table 2.4.1 shows the description of the teenage driver population in the 2005-2008 GUM combined data. Of the teenage drivers that were involved in MVCs, 22.0 percent were 16 years old, 36.6 percent were 17 years old and 41.4 percent were 18 years old. Compared to "Good" GDL programs, "Marginal/Fair" GDL programs showed higher percentages among teenage drivers of failure to use a seatbelt. However, other risk factors of MVCs and injuries were more prevalent under "Good" GDL programs than "Marginal/Fair" GDL programs: under "Good" GDL programs, we saw a higher prevalence of nighttime MVC, speed-relatedness, and suspicion of alcohol/drug use than under "Marginal/Fair" GDL programs. Moreover, "Marginal/Fair" GDL programs had lower percentages of teenage drivers linking to ED or hospital records (12.5%, 0.7%) than "Good" GDL programs (18.2%, 1.0%). On the other hand, among teenage drivers, death at the scene of the MVC is less prevalent under "Good" GDL programs than "Marginal/Fair" GDL programs. Some of these results may seem counterintuitive, but also may reflect the different age distributions between the GDL programs, and the tendency for fewer younger teenagers to be licensed in "Good" States due to the restrictions of the programs; we use exposure by population to further examine these findings in the next section.

Table 2.4.1. Description of the teenage driver study population: Teen driver characteristics by IIHS GDL rating in State.

IIHS GDL Rating	"Marginal/Fair"	"Good"	Total
Teen driver characteristics	N = 236,139 (45.5%)	N = 282,955 (54.5%)	N = 519,094
Age			
16	28.5%	16.5%	22.0%
17	34.2%	38.5%	36.6%
18	37.3%	44.9%	41.4%
Male teenage driver	53.7%	55.1%	54.5%
MVC information			
Nighttime MVC	13.0%	13.4%	13.2%
No seatbelt use	5.1%	2.7%	3.8%
Speed related [1]	11.7%	14.4%	13.0%
Suspicion of alcohol/drug use [1]	1.9%	2.0%	1.9%
Percent of teenage drivers by worst medical outcome			
Not linked to ED or hospital admission record	86.8%	80.8%	83.5%
ED visit	12.5%	18.2%	15.6%
Hospitalized	0.7%	1.0%	0.9%
Died [2]	0.2%	0.1%	0.2%
Percentage of highest level of care for all occupants involved in the MVC involving teenage drivers			
Not linked to ED or hospital admission record	74.5%	65.5%	69.6%
ED visit	23.6%	32.2%	28.3%
Hospitalized	1.9%	2.4%	2.1%

[1] Out of 11 States included in the GUM data, only available for 9 States for speed related and suspicion of alcohol and drugs.
[2] Defined as died at the scene of the MVC.

Teenage Driver Involvement in Injury MVCs

To control for the different distributions of age between "Good" and "Marginal/Fair" GDL programs, the rates of teenage driver involvement in injury MVCs per 1000 teenager-year were examined. Table 2.4.2 shows the crude rates and relative risks of teenage driver involvement in injury MVCs in each GDL category. Rates were computed via single Poisson regression on teenage driver counts from all injury MVC as discussed in the methods sections. Overall, the rate of teenage driver involvement in injury MVCs under "Good" GDL programs is 38 percent lower than under "Marginal/Fair" GDL programs. The teenage drivers most affected by "Good" GDL programs (age 16 drivers) show the largest risk reduction between "Good" and "Marginal/Fair" GDL programs (0.31). "Good" GDL programs are also associated with lower rates of injury MVC involvement by male teenage drivers, lower rates of involvement in nighttime injury MVCs, and lower rates of involvement by unbelted teenage drivers and teenage drivers suspected of alcohol/drug use. The lowest relative risks between "Marginal/Fair" and "Good" GDL programs are found in no seatbelt use (0.28), and suspicion of alcohol/drug use (0.45).

Table 2.4.2. Per capita rates and rate ratios of teen driver involvement in injury MVCs by driver characteristic and IIHS GDL rating

IIHS GDL Rating	"Marginal/Fair"	"Good"	"Good" vs. "Marginal/Fair"
Driver Characteristic			
Overall *	4.53 (3.44, 5.96)	2.81 (2.65, 2.99)	0.62 (0.48, 0.81)
Age			
16 *	1.37 (1.01, 1.87)	0.43 (0.33, 0.56)	0.31 (0.21, 0.47)
17 *	1.50 (1.16, 1.94)	1.10 (1.04, 1.17)	0.73 (0.57, 0.94)
18	1.65 (1.26, 2.16)	1.28 (1.16, 1.42)	0.78 (0.59, 1.02)
Male driver *	2.55 (1.91, 3.40)	1.68 (1.54, 1.82)	0.66 (0.49, 0.88)
Nighttime MVC *	0.83 (0.58, 1.19)	0.50 (0.48, 0.52)	0.60 (0.42, 0.87)
No seatbelt use *	0.85 (0.49, 1.47)	0.24 (0.15, 0.39)	0.28 (0.14, 0.58)
Speed-related	0.92 (0.44, 1.89)	0.54 (0.47, 0.61)	0.59 (0.28, 1.22)
Suspicion of alcohol/drug use *	0.28 (0.21, 0.38)	0.13 (0.09, 0.19)	0.45 (0.28, 0.72)
* Significant difference between "Marginal/Fair" GDL ratings and "Good" GDL rating.			

Table 2.4.3 shows the rate ratios comparing the risk of teenage driver involvement in injury MVCs under the GDL component listed in the table to the risk in the absence of these components. For all ages, restrictions beginning at or before 12 AM are associated with significant reductions in the crude rate of teenage driver involvement in injury MVCs. Restrictions on driving at night and carrying unrelated peer passengers are also associated with significant risk reductions for 16-year-old and 17-year-old drivers. Overall, the largest risk reductions are associated with 16-year-old drivers who are directly affected by GDL programs, and the risk reductions diminish with increasing age.

Table 2.4.3. Age-specific per capita rate ratios of teenage driver involvement in injury MVCs comparing ideal GDL components to non-ideal.			
	Age-specific rate ratios (95% CI)		
GDL component	16	17	18
Minimum permit age 16 or older	0.34 (0.23, 0.52)	0.85 (0.66, 1.08)	0.91 (0.71, 1.17)
Permit holding period 6+ months	0.64 (0.34, 1.18)	1.02 (0.77, 1.37)	1.14 (0.79, 1.65)
Required to practice for 30+ hours	1.48 (0.59, 3.72)	0.96 (0.74, 1.25)	0.93 (0.72, 1.19)
Night driving restricted after 12am	0.34 (0.20, 0.58)	0.66 (0.48, 0.89)	0.67 (0.50, 0.90)
No more than 3 teenage passengers	0.35 (0.20, 0.60)	0.71 (0.52, 0.97)	0.75 (0.54, 1.04)

Conclusion

Among teenage drivers, "Good" GDL programs are associated with a lower prevalence of failure to use seatbelts and death at the scene of the MVC. Though "Marginal/Fair" GDL programs appear to be associated with a lower prevalence of other MVC risk factors and with adverse medical outcomes other than death among teenage drivers and other participants in MVCs, the per capita rate of teenage driver involvement in injury MVCs is lower under "Good" GDL programs than under "Marginal/Fair" GDL programs. In particular, "Good" GDL programs are associated with lower rates of injury MVC involvement by male teenage drivers, lower rates of involvement in nighttime injury MVCs, lower rates of involvement by unbelted teenage drivers and teenage drivers suspected of alcohol/drug use. Compared to "Marginal/Fair" GDL programs, "Good" GDL programs are associated with a reduction in the crude rate of involvement in injury MVCs for younger teenage drivers. The typical components of GDL programs are also associated with risk reductions of teenage driver involvement in injury MVCs.

References

Baker, S. P., Chen, L. H., & Li, G. (2006). *National evaluation of graduated driver licensing programs.* (Report No. DOT HS 810 614). Washington, DC: National Highway Traffic Safety Administration.

Centers for Disease Control and Prevention (2011). *Web-based injury statistics query and reporting system (WISQARS).* Atlanta, GA: National Center for Injury Prevention and Control.

Foss, R. D., Feaganes, J. R., & Rodgman, E. A. (2001). Initial effects of graduated driver licensing on 16-year-old driver crashes in North Carolina. *Journal of the American Medical Association, 286*, 1588-1592.

Hallmark, S. L., Veneziano, D. A., Falb, S., Pawlovich, M., Witt, D. (2008). Evaluation of Iowa's graduated driver's licensing program. *Accident Analysis and Prevention 40*, 1401-1405.

IIHS (2012). *Effective dates of graduated driver license laws*. Arlington, VA: Insurance Institute for Highway Safety. Accessed on August 21, 2013. www.iihs.org/laws/pdf/gdl_effective_dates.pdf.

Mayhew, D. R., Simpson, H. M., Des Groseilliers, M., & Williams, A. F. (2001). Impact of the graduated driver licensing program in Nova Scotia. *Journal of Crash Prevention and Injury Control 2*, 179-192.

McCartt, A. T., Teoh, E. R., Fields, M., Braitman, K. A., & Hellinga, L. A. (2010). Graduated licensing laws and fatal crashes of teenage drivers: A national study. *Traffic Injury Prevention, 11*(3), 240-248.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Boca Raton, FL: Chapman & Hall/CRC.

Shope, J. T., Molnar, L. J. (2004). Michigan's graduated driver licensing program: Evaluation of the first four years. *Journal of Safety and Research, 35*, 337-344.

Ulmer, R. G., Preusser, D. F., Williams, A. F., Ferguson, S. A., & Farmer, C. M. (2000). Effect of Florida's graduated licensing program on the crash rate of teenage drivers. *Accident Analysis and Prevention, 32*, 527-532.

U.S. Census Bureau. (2012). Intercensal Estimates of the Resident Population by Sex and Age for States: April 1, 2000 to July 1, 2010. Accessed on August 21, 2013. www.census.gov/popest/data/intercensal/state/ST-EST00INT-02.html.

Zhu, M., Chu, H., & Li, G. (2009). Effects of graduated driver licensing on licensure and traffic injury rates in upstate New York. *Accident Analysis and Prevention, 41*, 531-535.

Part 2 Summary

One of the major challenges in conducting multi-State studies is combining information from crash reports that may not use the same definitions for similar data elements or collect variables at a different crash levels. In order to overcome this hurdle the CODES Data Network and State Data System (SDS) worked together to produce individualized State mappings onto a common set of variables in the General Use Model (GUM). These demonstration projects show that CODES methodology is not only feasible within a single State, but when combined, linked multi-state data analyses can produce sensible, meaningful results. As shown, combined data can be used to study populations that may be too small to analyze in a single-State study such as with abdominal injuries associated with seat belt misuse in the younger population or specific nature of injury and injured body regions in older occupants. An additional benefit of multi-state studies is the ability to compare crash outcomes in relation to the type of legislation that has been enacted in the different States (e.g. we observed a reduced risk of head injuries in States with universal helmet laws even after adjusting for motorcycle helmet use). These efforts provide an example for how future multi-State projects may be carried out.

DOT HS 812 179
July 2015

U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

www.nhtsa.gov

11107-072815-v3