Accident Analysis and Prevention 101 (2017) 117–123

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Linkage of traffic crash and hospitalization records with limited identifiers for enhanced public health surveillance

Sarah Conderino[a,b], Lawrence Fung[a,*], Slavenka Sedlar[a], Jennifer M. Norton[a]

[a] New York City Department of Health and Mental Hygiene; 125 Worth St Room 620, New York, NY 10013, USA
[b] Centers for Disease Control and Prevention/Council of State and Territorial Epidemiologists Applied Epidemiology Fellowship Program, USA

ARTICLE INFO

Article history:
Received 18 July 2016
Received in revised form 20 January 2017
Accepted 13 February 2017

Keywords:
Probabilistic
Linkage
Hospital
Crash
Traffic

ABSTRACT

Background: Motor vehicle traffic (MVT) crashes kill or seriously injure approximately 4250 people in New York City (NYC) each year. Traditionally, NYC surveillance practices use hospitalization and crash data separately to monitor trends in MVT-related injuries, but key information linking crash circumstances to health outcomes is lost when analyzing these data sources in isolation. Our objective was to match crash reports to hospitalization records to create a traffic injury surveillance dataset that can be used to describe crash circumstances and related injury outcomes. The linkage of the two systems presents a unique challenge since the system tracking crashes and the system tracking hospitalizations and emergency department (ED) visits lack key identifying data such as names and dates of birth.
Methods: NYC Department of Transportation provided electronic records based on reports of motor vehicle crashes submitted to the New York State Department of Motor Vehicles for all crashes occurring in NYC from 2009 to 2013. New York Statewide Planning and Research Cooperative System (SPARCS) ED and hospitalization administrative data from NYC hospitals were used to identify unintentional MVT-related injuries using external cause of injury codes. Since the two systems do not share unique individual identifiers, probabilistic record linkage was conducted using LinkSolv9.0. Sensitivity/specificity calculations and chi-square analyses of linkage rates were conducted to assess linkage results.
Results: From 2009–2013, there were 1,054,344 individuals involved in MVT crashes in NYC and 280,340 ED visits and hospitalizations from MVT-related injuries. There were 145,003 linked pairs, giving a linkage rate of 52% of the total MVT-related hospital records. This linkage had a sensitivity of 74% and a specificity of 93%. Linkage rates were comparable by age, sex, crash role, collision type, hospital county, injury location, hospital type, and hospital status, indicating no apparent biases in the match by these variables.
Conclusions: Performing a probabilistic linkage between MVT crash reports and hospitalization records is possible with a limited set of identifying variables. These linked data will inform traffic safety policies by providing new information on how crash circumstances translate to health outcomes.

1. Introduction

Motor vehicle traffic (MVT) crashes kill or seriously injure approximately 4250 people in New York City (NYC) each year (Transportation NYCDo, 2016). Multiple independent data systems are used for surveillance of traffic crashes and resulting injuries and death. Traffic crashes are tracked through law enforcement and transportation data systems. These data include details on the crash circumstances and a broad categorization of injury severity based on a police officer's assessment when first arriving at the scene, but lack details on specific injury outcomes. Public health surveillance of traffic injuries include vital statistics death certificate data and hospitalization data. Public health data include detailed information about the victims of traffic crashes and subsequent injury outcomes, but lack information about crash circumstances. Our objective was to link information found on motor vehicle crash reports to information found on hospital records in order to create an enhanced traffic injury surveillance dataset that can be used to describe crash circumstances and related injury outcomes. This enhanced surveillance system will be used to inform and support local efforts to eliminate traffic deaths in NYC as part of the Vision Zero initiative (Operations NYCMsOo, 2014).

Methods to link crash report records and hospitalization data have been developed nationally and used at the state-level as part of the United States National Highway Traffic Safety Administration

* Corresponding author at: 125 Worth St Room 620, New York, NY, 10013, USA.
E-mail addresses: sconderino1@health.nyc.gov (S. Conderino), lfung1@health.nyc.gov (L. Fung).

http://dx.doi.org/10.1016/j.aap.2017.02.011
0001-4575/

(NHTSA) Crash Outcome Data Evaluation System (CODES) (Cook et al., 2015). State CODES linked datasets have been used to characterize injury and economic outcomes of crashes related to crash type (Shen and Neyens, 2015; Chitturi et al., 2011), driving behaviors (Shen and Neyens, 2015; Conner and Smith, 2014), and the use of safety equipment (Shen and Neyens, 2015; Han et al., 2016; Olsen et al., 2014). CODES uses probabilistic linkage methodologies to define the probability that a pair of records represents a true match using a combination of weighted variables. In a recent NHTSA study of 8 states CODES datasets, 7 included date of birth and 6 included elements of first and last names in their linkage models, with no state CODES data linkages lacking both dates of birth and names (Cook et al., 2015). To our knowledge, the unique challenge of our project was performing a data linkage between crash reports and hospital records without access to key identifying data such as names and dates of birth in both data systems (Han et al., 2016; Cryer et al., 2001; Watson et al., 2015; Rosman and Knuiman, 1994; Wilson et al., 2012; Lujic et al., 2008) and implementing this linkage at a city-level rather than state-level. Here we describe the linkage methodology using personal identifiers limited to age and sex and high information-content variables, such as date of crash/admission and crash/hospital location, and we assess whether there are any biases in which records were successfully linked.

2. Methods

2.1. Data sources

New York State (NYS) Statewide Planning and Research Cooperative System (SPARCS) emergency department (ED) and in-patient hospitalization administrative records from NYC hospitals were obtained through a data use agreement with the NYC Department of Health and Mental Hygiene (New York State Department of Health Statewide Planning and Research Cooperative System, 2013). Data represent discharge events. ED data represent 'treat and release' visits that do not result in an in-patient hospital stay. Hospitalizations represent in-patient hospitalization stays. A patient who is admitted from the ED of the same healthcare facility is represented as a single in-patient hospitalization record. Only visits classified as unintentional MVT related injuries (ICD-9 CM external cause of injury code (E-Code) of E810–E819) with an admission and discharge date between 2009 and 2013 were included. NYS mandates and enforces the reporting of E-Codes, and during this time period both ED and hospitalization records were over 99% complete for E-Codes (Barrett et al., 2016). There were a total of 286,445 ED/hospitalization MVT discharges during this time period.

Multiple discharge events for a single individual can be due to multiple crash events or due to transfers to different healthcare facilities. Therefore, in order to match the hospital data to the crash reports, it was necessary to remove any ED/hospitalization records for the same individual that were related to the same crash event but retain any records for a person that could be related to a new crash event. Using a unique identification number, a total of 14,982 individuals with more than one ED/hospitalization MVT event were identified. We calculated the number of days between the date of discharge and the date of subsequent admission and found that 36% of the repeat visits occurred within one week of the prior discharge, with 11% occurring on the same day. The occurrence of multiple ED/hospital records dropped off substantially after one week, with only 21% of repeat visits occurring from one week to six months and 43% of repeat visits occurring from over six months to five years after the prior discharge. Thus, ED/hospital admissions for the same individual that occurred within one week from the prior discharge date were considered to be same crash event visits (n = 6105) and were removed from the final dataset used in the linkage (n = 280,340). Other variables (such as disposition, facility, admission source, and E-Code) were examined but were found to be too unreliable or inconsistent to be used to define event visits for the same crash. ED visits comprised 91% of the final hospital dataset.

NYC Department of Transportation (DOT) provided electronic records based on MV-104 reports of motor vehicle crashes submitted to the New York State (NYS) Department of Motor Vehicles (DMV) and NYS DOT for all crashes occurring in NYC from 2009 to 2013. NYS DMV designates an injury severity score (KABCO) based on a standard algorithm using the reporting police officer's judgment of victim, type of injury complaint, and location of injury complaint. Non-reportable cases (having property damage of less than $1000, no serious injuries, and no deaths) within NYS DOT files and cases not within both NYS DMV and NYS DOT files were excluded (Fig. 1). Reports were de-duplicated by identification number. There were 365,973 crash events occurring within NYC from 2009 to 2013 with 1,054,344 total individuals involved in these events. However, not all crashes involve injuries that would result in a visit to the ED/hospital. Based on the police-based KABCO injury severity in the crash reports, about one-third of the approximately one million individuals involved in these crashes had a reported injury, so we would expect that around 350,000 individuals would have visited an ED or hospital for treatment during this time period.

2.2. Data linkage method

Probabilistic record linkage is a preferred methodology for efficiently linking large datasets that do not share common unique identifiers or contain errors/omissions (McGlincy, 2004; Dean et al., 2001; Clark, 2004). Briefly, probabilistic record linkage calculates match probabilities for all possible merged pairs based on agreements and disagreements between match variables. Match variables are weighted based on probabilities of error in the data and probabilities of agreement by chance alone. Blocking variables reduce the possible number of merged pairs by forcing all merged pairs to agree on those variables, and multiple passes can be conducted varying the combinations of blocking variables to evaluate more potential linked pairs, which are termed merged pairs (Dean et al., 2001; McGlincy, 2006; Sayers et al., 2015).

We performed a Bayesian probabilistic record linkage using LinkSolv9.0 (McGlincy, 2006), which implements Marcov Chain Monte Carlo (MCMC) statistics to estimate the model parameters (McGlincy, 2004). This software and methodology was chosen since it is accepted and implemented at the national level and reduces potential biases related to selecting only high probability pairs by conducting multiple imputations that select weighted samples of all potential merged pairs (Cook et al., 2015). Our final linked dataset was the Maximum Likelihood Estimate (MLE), which is the iteration with the maximum likelihood based on match probabilities of linked pairs and the remaining merged pairs that were not linked.

Using a subset of the data from 2009, we assessed the following variables for inclusion in the linkage model using forward stepwise selection: date of crash/hospital admission, hour of crash/hospital admission, county of crash/hospital, patient age, patient sex, crash role, collision type, injury location, injury type, and injury occurrence. Based on Bayesian model outputs and distributions of match probabilities, our final model included the match variables reported in Table 1. Within the hospital data, crash role and collision type were determined using the ICD-9 E-codes. Injury body location was determined by applying the Barell Matrix (a coding matrix that classifies ICD-9-CM injury diagnosis codes by body region and nature of the injury) (Barell et al., 2002) to the admitting diagnostic code,