Accident Analysis and Prevention 207 (2024) 107749

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Estimating occupation-related crashes in light and medium size vehicles in Kentucky: A text mining and data linkage approach

Caitlin A. Northcutt [a,b,*], Nikiforos Stamatiadis [c], Michael A. Fields [d], Reginald Souleyrette [c,d]

[a] Department of Epidemiology & Environmental Health, 111 Washington Avenue, University of Kentucky, Lexington, KY 40506, United States
[b] Kentucky Injury Prevention and Research Center, 2365 Harrodsburg Road, Southcreek Building B, Suite B475, Lexington, KY 40504, United States
[c] Department of Civil Engineering, 161 Raymond Building, Lexington, KY 40506, United States
[d] Kentucky Transportation Center, 176 Raymond Building, Lexington, KY 40506, United States

ARTICLE INFO

Keywords:
Occupational Injuries
Traffic Crashes
Safety
Machine Learning
Data Linkage

ABSTRACT

Occupational motor vehicle (OMV) crashes are a leading cause of occupation-related injury and fatality in the United States. Statewide crash databases provide a good source for identifying crashes involving large commercial vehicles but are less optimal for identifying OMV crashes involving light or medium vehicles. This has led to an underestimation of OMV crash counts across states and an incomplete picture of the magnitude of the problem. The goal of this study was to develop and pilot a systematic process for identifying OMV crashes in light and medium vehicles using both state crash and health-related surveillance databases. A two-fold process was developed that included: 1) a machine learning approach for mining crash narratives and 2) a deterministic data linkage effort with crash state data and workers compensation (WC) claims records and emergency medical service (EMS) data, independently. Overall, the combined process identified 5,302 OMV crashes in light and medium vehicles within one year's worth of crash data. Findings suggest the inclusion of multi-method approaches and multiple data sources can be implemented and used to improve OMV crash surveillance in the United States.

1. Introduction

In 2022, occupational motor vehicle (OMV) crashes continue to be a leading cause of occupational fatality in the U.S., including Kentucky (Bureau of Labor Statistics, 2022, 2023). Reducing occupational injuries and fatalities resulting from OMV crashes is a priority and a strategic goal for the National Institute for Occupational Safety and Health (NIOSH)/Centers for Disease Control and Prevention (CDC) and both federal and state departments of transportation (Hsiao and Stout, 2010). To estimate OMV crash injury burden, many states rely on single-source occupational health indicators (OHIs). OHIs, standardized by the Council of State and Territorial Epidemiologists with guidance from the NIOSH/CDC, measure work-related injury and disease and help guide federal and state priorities for prevention and intervention in conjunction with other guidelines for state-based surveillance (Council of State and Territorial Epidemiologists, 2021). OHIs allow for systematic state comparisons of occupational injury trends and risk factors and are typically derived from easily attainable state-specific surveillance data sources that capture crash injuries and fatalities (Bunn et al., 2023; Council of State and Territorial Epidemiologists, 2021). While reliance on single-source data sources is common when calculating OHIs, it presents limitations in estimating the full magnitude of injury incidence and severity (Mirani et al., 2020; Thomas et al., 2012). Timely surveillance data from multiple methods is key to reducing the burden of occupational health injury and disease.

To date, OMV crash injury surveillance efforts using single-source crash data have primarily focused on heavy vehicles, as most crash record databases lack occupational data fields. Commercial vehicle crashes are easily identified as occupational by vehicle type. Conversely, OMV crashes involving light and medium vehicles are harder to systematically identify as occupational based solely on vehicle type (e.g., a four-door sedan could be used for both leisure and work purposes). Limitations are also present for sole reliance on workers' compensation (WC) claims data. Certain industries and occupations are exempt from or opt out of WC coverage (e.g., self-employed). Additionally, only using a WC claim or WC insurance payer in hospitalization data to identify these crashes likely underrepresents the burden from OMV crash injuries and fatalities (Bush et al., 2021). To combat these surveillance limitations,

* Corresponding author at: 111 Washington Avenue, Lexington, KY 40506, United States.
E-mail address: caitlin.pope@uky.edu (C.A. Northcutt).

https://doi.org/10.1016/j.aap.2024.107749
Received 22 March 2024; Received in revised form 9 August 2024; Accepted 11 August 2024
Available online 17 August 2024

narrative text mining (Boggs et al., 2020; Emu et al., 2022; Gao and Wu, 2013; Kim et al., 2021; Nayak et al., 2010; Rakotonirainy et al., 2015; Scott et al., 2021; Trueblood et al., 2019; Vallmuur et al., 2016) and data linkage (Burch et al., 2014; Burdett et al., 2015; Conner and Smith, 2014; Cook et al., 2015; Curry et al., 2021; Han et al., 2017; Hosseinzadeh et al., 2022; Milani et al., 2015; Shen and Neyens, 2015; Tainter et al., 2020) approaches have gained popularity across occupational health and transportation safety efforts, with few studies specifically focused on improving the broad surveillance of OMV crashes.

Crash record narratives can provide insight into contextual crash details that otherwise cannot be obtained using preexisting tabular surveillance data. Text mining of crash record narratives using machine learning techniques has been successfully used to help identify other crashes that are systematically difficult to track, such as agricultural crashes (Kim et al., 2021; Scott et al., 2021; Trueblood et al., 2019) and secondary crashes (Zhang et al., 2020; Zheng et al., 2015). While machine learning methods vary across studies and crash types, identification of crash type using text mining applications has been established in both crash and hospitalization databases. Utilizing crash narratives from multiple years of crash data from Texas and Louisiana, Scott et al. (2021) applied different data representations and classification algorithms to identify agriculture crashes. Through a systematic process of applying machine learning models trained on a test set of narratives from both states and independently, findings showed that model training on narratives from both states using the bag-of-keywords model was advantageous for correct agriculture crash classification. Differences were seen in performance metrics for the document classification algorithms (support vector classifier vs. multinomial Naïve Bayes classifier) compared to the baseline algorithm that is likely attributed to the length of narratives.

Another example focusing on the identification of secondary crashes is Zhang et al. (2020) who used a four-step process (tokenization, counting, vectorization, and normalization) using the Natural Language Toolkit Python library was applied to a year of crash records from Kentucky to identify secondary crashes. After a comparison across Random Forest, Naïve Bayes, support vector machine method, and logistic regression classification models, it was determined that the unigram logistic regression model had the best overall ability to accurately classify a secondary crash. While accuracy was the highest for the support vector machine method the precision was also high, indicating a large proportion of false negatives. Outside of secondary crashes, other crash record investigations have successfully applied narrative text mining to also assess the impact of crash risk factors (Boggs et al., 2020; Chen, 2010; Gao and Wu, 2013; Nayak et al., 2010; Rakotonirainy et al., 2015). This includes using crash narratives to elucidate key risk factors in both road-curve crashes (Rakotonirainy et al., 2015) and automated vehicle crashes (Boggs et al., 2020).

While utilizing crash narratives is beneficial to contextualizing crashes outside of structured data fields, manual reading of crash narratives is not without limitations. With the substantial number of crashes seen per year in the U.S., limitations of manually coding crash narratives for injury surveillance purposes will include many hours of manual review (Vallmuur et al., 2016). For example, in Taylor et al. (2014) it took three independent coders approximately 25 hours each to complete coding and an additional 25 hours for reconciliation with only 1,000 injury narratives. Machine learning techniques such as text mining provide a solution to these narrative coding limitations but come with additional issues regarding advanced computer programming that may not be easily adopted in practice without training and dedicated resources (Trueblood et al., 2019).

Linkage methods between crash records and health-related databases also provide a more comprehensive picture of crash risk factors and outcomes outside of reliance on single-source databases (Clark, 2004). While not occupation-focused, the Crash Outcome Data Evaluation System (CODES) program has historically supported probabilistic linkage of crash data with hospital databases to improve injury surveillance and to provide a more comprehensive picture of crash risk factors and outcomes (Cook et al., 2015). As of 2015, 30 U.S. states have participated in the CODES program (Cook et al., 2015), producing a variety of crash-related studies including investigations on the financial costs of crashes (Conner and Smith, 2014; Shen and Neyens, 2015), injury severity ratings (Burch et al., 2014; Burdett et al., 2015), identification of at-fault driver status (Sagar et al., 2020), motorcycle helmets and head injury (Singleton, 2016), and seat belt injury reduction effectiveness (Benedetti et al., 2022; Han et al., 2017).

While not focused on OMV crashes, using CODES data from Kentucky, Singleton (2016) found that motorcycle helmets significantly reduced head injuries such as skull fractures, cerebral contusions, and intracranial hemorrhages. Additionally, Sagar et al. (2020) found that using crash data linked to hospital records was more appropriate in defining at-fault driver status compared to sole reliance on crash records. Lastly, Benedetti et al. (2022) found using CODES data across Ohio, Maryland, and Kentucky that seat belts provided significant protection against fatal, serious, and torso injuries, but protection against torso injury attenuated with age. Outside of the CODES project, other linkage efforts in Kentucky using a heuristic, adaptive rule-based approach has also provided greater insight into injury severity by also including linkages to emergency medical services (EMS) data when linking crash and trauma hospitalization records (Hosseinzadeh et al., 2022). Findings from this method of linkage showed that a higher percentage of injury crashes were present in the dataset and bias towards inclusion of certain crash characteristics, such as having an injury severe enough for EMS transport, that would be more likely selected for inclusion than a crash with minimal injuries.

In addition to probabilistic and heuristic linkage, other state surveillance programs and studies have also relied on deterministic linkage (Milani et al., 2015). Deterministic linkage requires a predetermined set of data fields to match across multiple databases, yielding a simpler linkage procedure that is easier to implement in public practice in comparison to probabilistic linkage (Milani et al., 2015). The downside to deterministic linkage is that this linkage methodology can be restrictive based on the inclusion criteria and highly sensitive to data quality issues (e.g., completeness of data fields) and false negatives, issues that are also relevant to all linkage efforts (Clark, 2004). As seen with other types of crash-related injuries, linkage of crash records with health-related databases that include a variety of occupation-related data fields can improve the identification and assessment of OMV crash injuries (Boufous and Williamson, 2009; Thomas et al., 2012).

To better estimate the public health burden of OMV crashes outside of single-source, tabular surveillance data, this study developed and piloted a two-fold surveillance approach. This consisted of 1) a systematic evaluation of crash narrative text mining and 2) deterministically linking a year of Kentucky crash records to WC claims and EMS data. Given the complexity of probabilistic linkage (Clark, 2004) and the resources needed to implement it across states (Milani et al., 2015), a deterministic linkage was piloted. The overall goal of this pilot study was to develop a robust process for identifying OMV crashes in light and medium vehicles that could be extended to additional health-related databases and multiple years of data to pilot the Kentucky Occupational Motor Vehicle Injury Surveillance (OMVIS) database. From this study we aimed to identify OMV crashes in Kentucky in 2019, outside of commercial motor vehicle and special use vehicle crashes, through crash narrative text mining and deterministic linkage.

2. Data and methodology

2.1. Data sources and study population

To develop the methodological approach for identifying OMV crashes, a full year of data was used. This included deidentified data from the Kentucky Collision Reporting and Analysis System for Highways (CRASH) database, the Kentucky WC claims database, and the

Kentucky EMS database. The COVID-19 pandemic resulted in a decrease in overall vehicle miles traveled in 2020, but a higher number of deliveries and an increase in the frequency of the use of services that are typically provided with small and light vehicles (Bureau of Transportation Statistics, 2021). Due to the focus of the study on OMV crashes in light and medium vehicles, data from 2019 was used to avoid potential COVID-19 bias.

2.1.1. Kentucky CRASH database

Kentucky CRASH, obtained from the Kentucky State Police, is a census of reported motor vehicle crashes that occur on publicly travelled roadways and parking lots in Kentucky. For this study, three samples of crashes were created from the 2019 Kentucky CRASH database, including all injury severity types and categorized based on vehicle type. Classifications for vehicle type were determined by the unit type code in the crash record. The first classification included any crash involving a unit with a large commercial vehicle. The second classification included any unit involving a special use vehicle. Finally, the last classification, light and medium vehicle crashes, included all crashes that were not included in either the large commercial vehicle or a special use vehicle classification. A breakdown of the specific unity type codes used to classify crashes is included in Appendix A Table 1. This sampling of the 156,758 unique crashes in 2019 resulted in 13,451 large commercial vehicle crashes, 2,187 special use vehicle crashes, and 141,030 light and medium vehicle crashes.

2.1.2. Kentucky WC claims database

First report of injury (FROI) data from 2019 was obtained from the Kentucky Department of Workers' Claims (KDWC). FROIs from KDWC does not include injuries from self-employed workers and are based on the following acceptance criteria as reported in Bunn et al. (2022):

> Reported worker injuries require at least one day off from work or must result in a disability that extends beyond 60 days;
> The worker is eligible for indemnity and/or lump sum payments when a worker has lost at least seven days of work due to an injury or has a permanent partial disability. This includes payments associated with FROIs or claims (litigated FROIs) defined as paid income benefits to compensate for lost wages, functional impairment, or death and;
> The worker is eligible for lost wage compensation retroactive to the first day of work lost when a worker has lost at least two weeks of work due to an injury.

In 2019, there were 34,026 FROI WC claims of which 1,167 claims had a motor vehicle crash-related cause code and were used for inclusion for deterministic linkage given all the FROI reports are occupation-related (see Fig. 1). In terms of data completeness, the FROI WC claims database had negligible missingness, with only 6 records missing an injury cause code in the 2019 database.

2.1.3. Kentucky EMS database

EMS run data from 2019 were obtained from the Kentucky Board of EMS which follows the National EMS Information System (NEMSIS) data standard v3.4 (National EMS Information System, 2023) and are collected in accordance with Kentucky Administrative Regulations. As seen in Fig. 2, there were 472,554 runs recorded in 2019. To determine occupation-relatedness, EMS run data were initially filtered by the traffic/transportation incident data field listed in complaint reported by dispatch which yield 60,391 records. Next, a filter using three occupation-related fields was applied (work-related, work-related illness or injury, and primary payment). A combination of these three data fields was used as Kentucky Board of EMS does not require completion of the occupation-related fields by EMS reporting agencies. For example, in 2019 79.5% of the records that were traffic/transportation incidents had a null or unanswered value for work-related illness or injury. Therefore, if any of the three occupation-related fields was present, the record was included in the analysis.. This resulted in 2,315 OMV crash related EMS runs that were used for inclusion for deterministic linkage.

2.2. Crash narrative text mining procedure

The sample of crash records used for the narrative text mining procedure consisted only of light and medium vehicles. Using the Natural Language Toolkit Python library, the logistic regression model employed by Zhang et al. (2020) to identify secondary crashes was adapted for OMV crashes. Logistic regression models are a type of supervised machine learning algorithms that are used for discriminant classification (i.e., classifies observations into one of two classes;

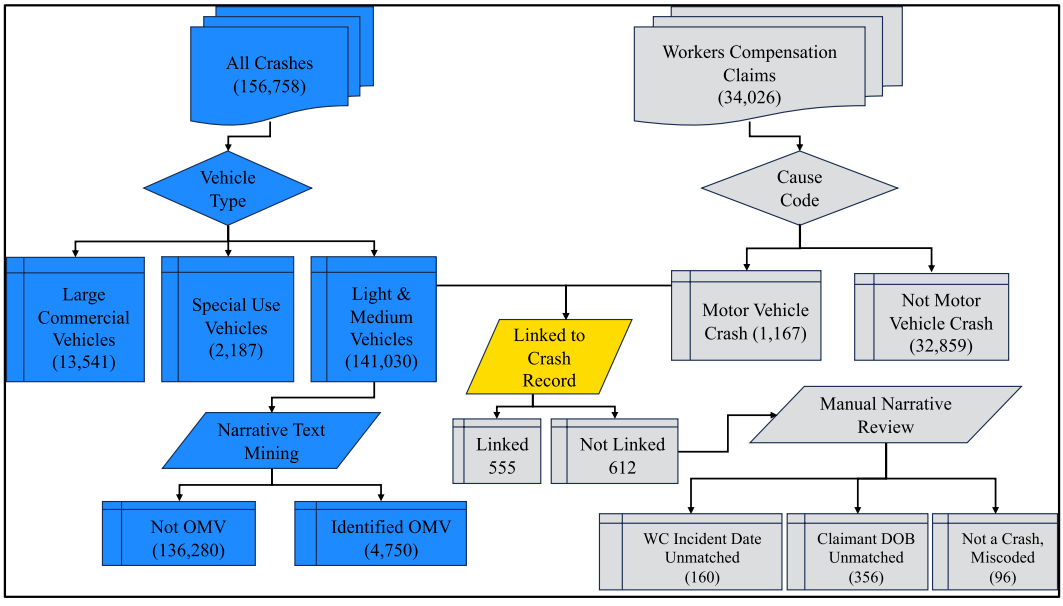

Fig. 1. *Kentucky CRASH and WC Claim linkage to identify OMV crashes. Note.* This figure provides an overview of the deterministic linkage between Kentucky CRASH and Kentucky WC Claim databases for 2019. The linked crash number represented is the total number of linked records prior to deduplication.

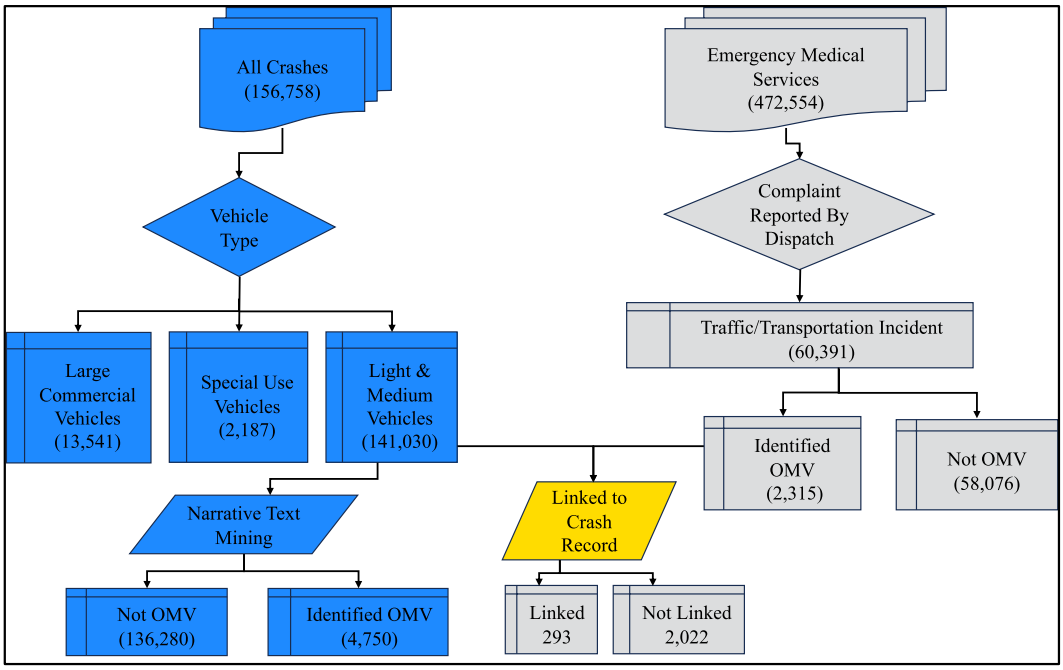

Fig. 2. *Kentucky CRASH and EMS linkage to identify OMV crashes. Note.* This figure provides an overview of the deterministic linkage between Kentucky CRASH and Kentucky EMS databases for 2019. The linked crash number represented is the total number of linked records prior to deduplication.

Jurafsky and Martin, 2023). Based on the information provided, the model is trained to distinguish classes of interest. In this case, the class of interest is OMV crashes. Narrative text mining was conducted at the crash level using a two-step process: a training and testing model was created to develop the logic necessary to classify OMV crashes in the input sample, and a prediction model was deployed on the full set of narratives to classify all crashes.

The initial sample of narratives was selected through a semi-random process. Narrative review was completed across four team members who convened multiple times for discussion on review issues and procedures. Narratives were only reviewed once by the assigned team member. First, the study team reviewed 414 crashes randomly selected from the 141,030 light and medium vehicle crashes, looking for patterns that might identify occupation-related crashes. From the review, a small proportion of randomly selected crashes that could be classified as occupation-related (~4%) was identified. The team then preferentially selected 2,000 records that contained words thought to suggest occupation-relatedness (deliver, work, company, etc.). Records from both selections were carefully reviewed, classified as occupation-related or not, and compiled into the model's input dataset. Based on the narrative review, 407 crashes (16.9%) were classified as occupation-related.

While occupation-related crashes are overrepresented in this sample, it is important (for the predictive power of the regression model) to minimize classification imbalance. Ideally, a binary classification would be close to half-and-half. The more imbalanced the classes, the more likely the model will bias its predictions toward the larger class. In this case, if the proportion of non-OMV records in the sample is drastically higher than that of OMV records, the model would be very unlikely to classify a record in the population as OMV-related.

The initial model used a proportion of crash narratives in the input data as the training set, assessing the text and identified keywords to be used in the classification of unread narratives, and the remaining records as the testing set, to test the model's logic and show the results of the model's agreement with the input dataset. Each record was randomly assigned to either a training set or a testing set. An 80/20 split produced a model with the highest predictive power, and it is also the most used split in literature. A prediction and a probability were generated to determine which class the crash record belonged to. Next, to estimate the model's performance metrics, manual and model classifications were compared to determine if the crash record was classified correctly as OMV-related. Model parameters were assessed through a series of iterative model runs to improve model performance. This involved testing hyperparameter C values as well as unit of analysis.

- The model uses a hyperparameter *C* to inform selection of internal parameters not explicitly set by the user. Higher values of *C* instruct the model to give more weight to the training data, suggesting that it is a strong representation of the population which may lead to overfitting. Lower values of *C* suggest more variation in the data, decreasing the weight of extreme internal parameter values and thus leading to smoothing of the regression model outputs. Several *C* values between 1 and 100 were tested. According to the model outputs, a value of 30 resulted in the most accurate results.
- Assessment of each narrative with the goal of predicting its classification involves using, as a unit of analysis, each individual word in the narrative or contiguous multi-word phrases. To determine which model unit of analysis produced the most accurate predictions, 1-word, 2-word, 1-or-2-word, and 1-or-2-or-3-word phrases were tested. Here, 1-or-2-word phrases were used.

Model performance is characterized by the balance of true positives, true negatives, false positives, and false negatives. The following four indicators consider typical model misclassification error when evaluating the model's predictive ability.

1. Accuracy: The percent of correctly classified records among all records (true negatives and true positives divided by total number of cases).
2. Precision: The percent of correct positive classifications among the model predicted positives (true positives divided by true positives and false positives).
3. Recall: The percent of correct positive classifications among predefined positives (true positives divided by true positives and false negatives).

4. F1 score: The combination of recall and precision (Recall x Precision divided by their sum and multiplied by 2). The F1 score represents the model's tendency to correctly classify OMV crashes while considering the percentage of misclassifications (false positives and false negatives).

Once the model that most accurately predicted each record's classification was determined, those parameter values were used as constants in subsequent model runs. This allowed for the development of a repeatable process to be used to classify additional years of crash data. As noted above, models are evaluated using the four indicators, developed based on the model training, as measures of goodness of fit.

2.3. Deterministic data linkage procedure

Based on how FROI WC and EMS records are stored, data linkage was performed at the person level and aggregated for inclusion into the OMVIS database at the crash level. If a crash was linked to multiple FROI WC or EMS records the crash was considered occupation-related and included once in the database. Linkage was completed in two steps: 1) with Kentucky WC claims data, and 2) with Kentucky EMS data.

Using the FROI cause code field as a filter resulted in 1,167 records that were identified as having a motor vehicle crash-related cause code. Those records were merged via a structured query language table join with the corresponding Kentucky CRASH records. As mentioned earlier, sensitivity analyses were conducted across several iterations of join criteria to determine the set of fields from the two tables that resulted in the most reliable data linkage. The final set of inclusion criteria for the crash and FROI WC data linkage included: 1) the FROI WC county of injury matched the county of crash occurrence in CRASH, 2) the FROI WC claim date of injury matched the CRASH collision date, and 3) the claimant's date of birth in the FROI WC record matched the date of birth for any vehicle occupant in the crash record. For two records to be considered "linked" they must match across all three criteria. If the two records did not match across all three criteria they were considered not linked. For linkage between Kentucky CRASH and EMS records, a similar structured query language table join was used. The final set of inclusion criteria for the crash and EMS data linkage included: 1) the county of the EMS run matched the county of crash occurrence in CRASH, 2) the EMS dispatch date matched the CRASH collision date, and 3) the EMS patient's date of birth matched the date of birth for any vehicle occupant in the crash record. Once the records were joined, using the complaint reported by dispatch and work-related data fields to filter for occupation related crashes resulted in 2,315 records.

A sensitivity analysis was used to determine the combination of data fields that accurately linked OMV crashes while balancing the inflation of false negatives and positives. Requiring more data fields to be identical across databases resulted in more restricted output, or false negatives (incident county, incident date, birthdate, and gender resulted in 27,556 matched records). Requiring fewer data fields to be identical across databases resulted in too many potentially matched records, or false positives (only birthdate and county give 417,646 matched records; and only incident date and county give 4,518,718). A random record review of these join results suggested that most of these were not true matches.

3. Results

3.1. Crash narrative text mining

Model performance metrics capturing accuracy, precision, recall, and the F1 score varied by the model unit of analysis (Table 1). While accuracy and precision was similar across models, the model outputs using 1-or-2-word phrases as the unit of analysis resulted in the highest accuracy and precision. Using 1-word phrases produced the highest recall and F1 score. Across all four metrics, differences arose by the unit of analysis, with similar metrics seen between using 1-or-2-word phrases and 1-or-2-or-3-word phrases. Review of the keyword outputs showed that, of the phrases that correlated with OMV crashes, only one phrase out of 50 used 3 words ("operator of the"). This suggests that a maximum of two words per phrase is ideal over three words and provides subtle advantages in correctly classified records among all records. A list of keywords, in order of their frequency in the population of narratives in the included class, was produced by each model run. The most predictive words found by the model over several iterations are provided in Appendix A Table 2.

Table 1
Crash Narrative Text Mining Model Metrics of Predictive Power.

Unit of Analysis	Accuracy	Precision	Recall	F1 Score
1-word	0.892	0.626	**0.621**	**0.623**
2-word	0.877	0.618	0.379	0.470
1-or-2-word	**0.895**	**0.663**	0.540	0.596
1-or-2-or-3-word	0.890	0.656	0.492	0.562

Note. Bolded values represent the highest unit of analysis score in that category of predictive power.

Based on the accuracy and precision metrics, the 1-or-2-word phrase model was used for all subsequent model runs. Of the 141,030 crashes, the crash narrative text mining process classified 4,750 OMV crashes. This translates to 3.0% crashes in 2019 which involved light and medium vehicles and was occupation related.

Crash characteristics are provided for the identified crashes by method in Table 2. On average, 2 units were involved in OMV crashes that were identified with both the crash narrative text mining model and linkage methodology efforts. Of the crashes identified with narrative text mining, 51.5% had at least 1 unit that was classified as a passenger

Table 2
Crash Characteristics by Identification Methods.

Crash Characteristic	CRASH Narrative Text Mining	CRASH-WC Linkage	CRASH-EMS Linkage
Vehicle Type			
Passenger car	51.5	41.9	35.2
Van, SUV, or Light truck	38.0	46.9	27.8
Time of Day			
Daytime hours (6:00 am – 5:59 pm)	75.6	83.6	84.9
Nighttime hours (6:00 pm – 5:59 am)	24.4	16.4	15.1
Day of week			
Sunday	12.9	5.4	4.9
Monday	9.1	15.5	11.0
Tuesday	14.1	17.3	18.9
Wednesday	14.9	17.0	16.3
Thursday	16.2	19.1	15.5
Friday	15.1	17.7	21.2
Saturday	17.8	7.9	12.1
Driver age			
15–18 years old	8.8	7.6	4.8
55 + years old	42.7	51.4	39.1
Driver gender (male)	93.2	77.0	90.0
Injury Severity			
K (fatal crash)	0.2	0.1	3.3
A (incapacitating injuries)	0.5	0.1	10.3
B (non-incapacitating injuries)	1.6	1.9	35.8
C (possible injuries)	1.7	2.0	35.8
O (no reported injuries)	96.0	95.7	14.8
Rurality			
Rural	31.4	23.6	35.6
Urban	68.6	76.4	64.4

Note. All percentages are reported at the crash level. Vehicle type is the percentage of crashes with at least 1 unit with the listed vehicle type. Driver age and gender is the percentage of crashes with at least 1 unit with a driver of that age category or of the male gender. Injury severity is the worst reported injury for any occupant of any unit involved in 1 crash.

car and 38.0% of the crashes had at least 1 unit that was classified as a van, sports utility vehicle (SUV), or light truck. Most OMV crashes happened during daytime hours (6:00 am – 5:59 pm; 75.6%), in urban areas (68.6%), and were evenly distributed across the day of the week with the most crashes reported on a Saturday (17.8%) and the least on a Monday (9.1%). The majority of OMV crashes had at least 1 driver that was male (93.2%). When assessing driver age, 42.7% of OMV crashes had at least 1 driver that was aged 55 and older and only 8.8% of crashes had at least 1 driver aged 15–18. Regarding injury severity of the crash, almost all the crashes identified with the crash narrative text mining model (96.0%) were crashes with no reported injuries, followed by 1.7% with possible injuries, 1.6% with non-incapacitating injuries, 0.5% with capacitating injuries, and 0.2% reported as fatal crashes.

3.2. Deterministic Kentucky CRASH and WC claims data linkage

There were 621 links from the Kentucky CRASH and WC Claims databases (i.e., matched on all three criteria; see Fig. 1). After consolidating records to consider events with more than one person involved in the same crash, the final linkage between FROI WC claims and crash records resulted in 555 unique linked crashes and 612 unlinked crashes. When categorized by the vehicle type, of the 555 matched records 160 records involved a heavy vehicle, 57 involved a special use vehicle, and 338 involved a light and medium vehicle. There were 6 crashes among the 338 light and medium vehicle crashes already identified from the crash narrative text mining process. This resulted in the identification of an additional 332 crashes from the linkage process that could be added in the OMVIS database.

Of the 332 crashes identified through the CRASH – WC linkage, 41.9% had at least 1 unit that was classified as a passenger car and 46.9% of the crashes had at least 1 unit that was classified as a van, SUV, or light truck. As expected, most OMV crashes happened during daytime hours (6:00 am – 5:59 pm; 84.9%) and in urban areas (76.4%). The crash frequency was evenly distributed across weekdays. Of the CRASH-WC linked crashes, 77.0% had at least 1 driver that was male and 51.4% had at least 1 driver that was aged 55 and older. Like the crashes identified through the narrative text mining, almost all the crashes identified with the CRASH – WC linkage were crashes with no reported injuries (95.7%), followed by 2.0% with possible injuries, 1.9% with non-incapacitating injuries, 0.1% with capacitating injuries, and 0.1% reported as fatal crashes.

3.3. Deterministic Kentucky CRASH and EMS data linkage

There were 293 links from the Kentucky CRASH and EMS databases (see Fig. 2). When categorized by vehicle type, the data showed that 15 records involved a heavy vehicle, 48 records involved a special use vehicle, and 230 involved only light and medium vehicles. Only 10 crashes among the 230 light and medium vehicle crashes already identified from the crash narrative text mining process, suggesting that many crash narratives do not contain sufficient detail to identify an occupation-related crash and that information from additional data sources is critical to this type of analysis. The 293 linked crashes resulted in 220 newly identified crash records that could be added to the OMVIS database and 2,022 unlinked crashes.

As a validation of the deterministic linkage methodology, two additional data fields not used in linkage from the CRASH and EMS records were compared. These included comparing the time of collision to the EMS notification time and, when present, the geographic coordinates from both datasets. The time comparison included subtracting one time from the other to return a time differential in minutes. Given the CRASH and EMS run occurred on the same day, this was an additional check to see if they were the same event. The results (Table 3) show that 87% of the time comparisons are within 15 min and 96% are within 30 min. The second effort involved the latitude/longitude from each linked record pair. All crash records contain this information, but for EMS records this information is optional. Of all the EMS records in 2019, only 7% of the runs had usable latitude/longitude pairs. Approximately half of crashes identified as OMV crashes (46%) had usable latitude/longitude pairs. All available coordinate data from the linked EMS records were collected and compared to their linked crash record using a formula that converts two latitude/longitude pairs to a distance in feet. The validation effort showed that 85% of the pairs were within 2,000 feet, a distance that suggested they were the same event with high confidence.

Table 3
Time and Distance Validation of Crash to EMS Linkage.

Distance (ft) Between Crash and EMS Points	Time (min) Between Crash and EMS Dispatch			
	Under 15	15 to 30	Over 30	Total
Under 200	434	42	17	493
200 to 2000	369	29	15	413
Over 2000	125	28	13	166
Total	928	99	45	1,072

Of the 220 OMV crashes identified in the CRASH – EMS linkage, 35.2% had at least 1 unit that was classified as a passenger car and 27.8% of the crashes had at least 1 unit that was classified as a van, SUV, or light truck. Again, most OMV crashes happened during daytime hours (6:00 am – 5:59 pm; 83.6%) and in urban areas (64.4%). Of the CRASH-WC linked crashes, 90.0% had at least 1 driver that was male and 39.1% had at least 1 driver that was aged 55 and older. Regarding injury severity of the crash, the results were opposite of the crashes identified through crash narrative text mining and CRASH – WC linkage. Almost all the OMV crashes identified with the crash narrative text mining model were injurious, with 3.3% reported as a fatal crash, 10.3% with incapacitating injuries, 35.8% with non-incapacitating injuries, 35.8% with possible injuries, and 14.8% as no reported injuries.

The total number of light and medium OMV crashes from 2019 identified through a combination of crash narrative text mining and deterministic linkage totaled to 5,302. This can be broken down by the 4,750 OMV classified crashes from the crash narrative text mining, the 332 OMV crashes identified via CRASH – FROI WC claim linkage, and the 220 OMV crashes identified via CRASH – EMS linkage.

4. Discussion

Given the burden of OMV crashes in the U.S., improvement of surveillance to aid in the development and prevention of occupation-related injuries and fatalities is a high priority (Hsiao and Stout, 2010). In Kentucky and many U.S. states, the inability to accurately track OMV crashes involving light and medium vehicles (e.g., vans, passenger cars) is due to a reliance on single-source surveillance methods and a lack of occupation-related data fields in crash databases (Mirani et al., 2020). Recently, studies have begun to use machine learning to identify agriculture-related crashes (Kim et al., 2021; Scott et al., 2021; Trueblood et al., 2019) and linkage methodology (Boufous and Williamson, 2009; Thomas et al., 2012) to identify OMV crashes, but a systematic application of both approaches OMV crashes, broadly, is lacking. To help fill this gap, the Kentucky OMVIS system was proposed to develop a robust combination of machine learning and linkage methods to improve identification of OMV crashes in light and medium vehicles. Initial pilot testing of the two-step process of crash narrative machine learning and deterministic linkage included creating a filtering process that incorporates special use vehicles, a vehicle type not currently captured in the single-source Kentucky CRASH database derived OHI. This process identified 5,302 OMV crashes in light and medium vehicles in 2019, but not without limitations that will be discussed.

When assessing the overall performance of the crash narrative text

mining, the model identified 4,750 OMV crashes in light and medium vehicles. This exceeds the current method of only reporting fatal and nonfatal heavy and special use vehicle crashes from the Kentucky CRASH database. In addition, the model metrics with the best accuracy performance included 1-or-2-word phrases compared to 1-word only, 2-word only, and 1-or-2-or 3-word phrases. Using 1-or-2-word phrases correctly classified OMV crashes from non-OMV crashes approximately 89.5% of the time. While there was room for improvement after assessing the balance between false positives and negatives, the 1-or-2-word phrase model still correctly classified OMV crashes greater than half of the time (59.6%). In comparison to Zhang et al. (2020), the same logistic regression model was more effective for identification of secondary crashes from unstructured crash narratives than OMV crashes when taking precision and recall into consideration. In Zhang et al. (2020) it was determined that a logistic regression model outperformed Random Forest, Naïve Bayes, and the Support Vector Machine models in the identification of secondary crashes from the Kentucky CRASH database. While the reasons behind the differences in model performance across studies are unknown, one possible explanation is the level of contextual details required for accurate identification of OMV crashes may be less prevalent in unstructured crash narratives. For example, the most predictive words of OMV crash classification from the current model included words such as *deliver; trailer;* and *employee*. These words reflect the nature of work and are not always known from the reporting officer completing the crash reports, especially in the absence of occupation-related data fields. While this may also be an issue with secondary crashes, Zhang et al. (2020) found that words reflecting the causal association between primary and secondary crashes were the most important for identification. Given the stronger model performance metrics for secondary crashes, it could be interpreted that relevant keywords for secondary crashes may be more prevalent in the crash narrative and overall crash database.

In addition to a lack of included occupation-related keywords in crash narratives, a higher frequency of non-OMV crashes in comparison to OMV crashes (i.e., classification imbalance) could also be a contributing source of misclassification when employing a binary classification prediction modeling approach, such as the current methodology, on uncertain records. As noted in Kim et al. (2021) in the application of narrative text mining models to identify agriculture-related crashes, imbalanced datasets can be misleading in that performance metrics will have high accuracy due to classification to assigning cases more frequently to the larger class. If OMV crashes are indeed less prevalent compared to secondary crashes in the crash database, a larger and more balanced training dataset across multiple years of crash data may be required to improve the model F1 metric. Future research should consider the benefit-cost analysis of manually coding multiple years of crash narratives in efforts to provide a larger training dataset to improve model precision and recall performance metrics. Additionally, practice standards around performance metrics for machine learning applications will help with implementation and adoption.

Deterministic data linkage between Kentucky CRASH and health-related databases was piloted in addition to the narrative text mining. Linkage across databases as noted in other areas of occupational health, removes reliance on single-source surveillance methods which increases the possibility of underestimating events that may not be fully represented within the respective data source (Mirani et al., 2020). In the current study, the feasibility of a deterministic linkage approach was tested between two health-related databases (Kentucky WC Claims and Kentucky EMS) with the Kentucky CRASH database. Deterministic linkage was chosen over probabilistic linkage (see Cook et al., 2015) and heuristic linkage (see Hosseinzadeh et al., 2022) methods as a first step to identifying OMV crashes through linked datasets. This involved determining what occupation-related indicators were available in each dataset for filtering along with inclusion linkage criteria for cases. This filtering process varied across the linked datasets.

While internal OMV crash codes are assigned to FROI WC claims, there is not a 100% overlap with crash records. In fact, only 48% of the FROI WC OMV crash claims from 2019 overlapped with crash records from 2019 across the three data fields used for matching. Observational reasons from descriptive review of the FROI WC claim narratives included OMV crashes where a crash report was not generated or the narrative revealed the event was not crash related in nature (i.e., data error). Furthermore, in comparison to the crash narrative text mining approach tested there was negligible overlap (1%) with FROI WC claim and crash record matched OMV crashes, with 60% of the matched records being newly identified OMV crashes.

Similar linkage effort limitations were seen with EMS records. While EMS records are generated like crash records (i.e., paramedics are dispatched to crashes in the same fashion as law enforcement) in comparison to WC claims, minimal overlap still existed. In states such as Kentucky where occupation-relatedness is not a mandatory field, large amounts of missingness exist. Ultimately, occupation-relatedness becomes reliant on payer type (e.g., WC payer) which incorporate the same fallacies as relying on single-source surveillance. In fact, Bush et al. (2021) found that relying only on WC payer type with emergency department injury visits was less efficient, with 36% more occupation-related visits being accounted for when also incorporating occupation-related International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) codes. Additionally, while not the focus of the current study, the relationship between crash injury severity and linkage cannot be ignored. Prior linkage efforts with probabilistic (Burch et al., 2014; Thomas et al., 2012) and heuristic (Hosseinzadeh et al., 2022) linkage methods have noted that linked databases that relies on EMS and hospitalization records will inevitably be positively biased to the inclusion of linked crashes that have higher severity. For example, EMS is less likely to be dispatched to a crash that is property damage only compared to a crash where additional medical treatment is needed. This could, in part, explain the| low percentage of overlap across both pilot linkage efforts with the FROI WC claims and EMS databases.

Qualitative analysis of crash characteristics revealed similarities and differences across the identification methods. Across all three subsets of identified OMV crashes, the majority of crashes involved 2 units, occurred in urban areas, during daytime hours, and mostly on weekdays or on a Saturday. When assessing the driver, a larger proportion of crashes had at least 1 driver that was male and aged 55 and older. These findings are in line with recent BLS statistics which have shown an increasing trend of older workers (55 +) being represented in occupational injury and fatal statistics (Bureau of Labor Statistics, 2023; Smith and Pegula, 2020) and tracking with the aging population (Anderson et al., 2012). For injury severity, almost all the identified crashes from text narrative mining and the CRASH – WC linkage were crashes with no reported injuries. This was in a clear juxtaposition to identified crashes from the CRASH – EMS linkage which were more likely result in an injury, ranging from possible injuries to fatalities. One possible explanation for the difference could be due to more severe crashes needing immediate medical evaluation and or transport, versus a crash that did not result in an immediate injury. While it was not anticipated that the CRASH – WC linkage would be overly dispersed with crashes with no reported injuries, this could be partially reflecting delayed injury care (e.g., seeking medical care post-crash). Discrepancy was also seen for the distribution of unit types involved in the crash. Compared to the text narrative mining and the CRASH – EMS linkage, the CRASH – WC linkage had more crashes that involved a one van, SUV, or light truck. A possible explanation could be that employers who are required to report to WC in the state of Kentucky are more likely to utilize these types of vehicles compared to employers who do not. More investigation into the associations derived from the linkage as well as how these associations vary and change with probabilistic methods is greatly needed.

4.1. Limitations

This study like others has limitations that should be considered when

interpreting the results. The first includes data quality of administrative data sources. Administrative databases are vulnerable to reporting inconsistencies, data errors, and missingness that can impact narrative text mining and data linkage applications. This includes commonly seen issues in our data such as occupation-related data fields not being consistently populated in the EMS database and the heterogeneity in the length and detail of unstructured crash narratives. Second is the amount of data that was used as a training set for the crash narrative text mining. A larger training dataset, while reliant on additional manual coding of narratives, comprised of multiple years of data could strength the model's precision and recall performance and allow for more in-depth comparison across the unit of analysis. Next, the low percentage of linked crashes with FROI WC claims and EMS databases was concerning. While it is still unclear what the true prevalence of OMV crashes is in Kentucky and the U.S., it is likely that with deterministic linkage methods we are missing potential linkages due to strict inclusion criteria. More complex linkage methodologies, such as probabilistic linkage, would likely be inclusive of more cases that are due to data errors such as transposition (e.g., one number off on a birthdate or incident date). Lastly, complete validation to obtain ground truth on the piloted deterministic links between administrative datasets was not feasible. Commonly seen with administrative datasets, personal identifiers (e.g., name, driver license, etc.) are not provided in all datasets and data use agreements restrict linkage that can result in direct identification. A time- and distance-based validation was conducted on the CRASH-EMS links where usable latitude/longitude pairs were available. While this data is missing for a large proportion of the EMS records (7% of all EMS runs in 2019 and 46% of identified OMV crashes), the links between the two datasets were within distance and time intervals that suggested high confidence in the deterministic linkage between the two datasets. Additional validation efforts are needed for future probabilistic linkage, specifically for datasets like WC that do not have an included spatial or time record for the event. Efforts to systematically validate linkage techniquesare of great importance given this limitation is not unique to Kentucky administrative data.

5. Conclusions

The current project aimed to develop and test a two-step approach to identifying OMV crashes within a year of crash reports in Kentucky. Using crash narrative text mining techniques on crash narratives and deterministic data linkage with FROI WC claims data it was determined that OMV crashes could be accurately identified from non-OMV crashes. This two-step approach resulted in 5,302 OMV crashes identified within one year's worth of data over. While not without limitations, including a substantial reduction in model performance after accounting for false positive and negatives and data errors, the current study is among the first to use a combination of machine learning and data linkage methodology to systematically improve OMV crash identification broadly across all areas of transportation.

Future research would benefit from the incorporation of multiple years of crash data for crash narrative text mining. In addition to narrative text mining of crashes, other health-related databases such as EMS also have narratives available for review in the patient care reports. With proper deidentification of sensitive patient care information, these narratives could also be mined for occupation-related detail and used to improve inclusion of cases. Future research on the feasibility of probabilistic linkage is also needed. This pilot study focused on deterministic linkage as the first step in feasibility of probabilistic linkage between crash and health-related databases. Given the reported limitations of implementing probabilistic linkage into state-wide surveillance efforts (Milani et al., 2015) a benefit-cost analysis between the gains of deterministic versus probabilistic linkage should be further considered and tested. In addition to more flexible linkage methods, the use of injury severity in linkage efforts should also be considered. Given more minor crashes may not result in EMS transport or immediate medical care utilization, understanding how crash linkage rates vary across injury severity types could be useful.

In addition to research, practice recommendations to data stakeholders can also be considered. First, the lack of an occupation-related data field in state crash databases is an oversight that, if included, would have immediate impact on the field of OMV crash surveillance. Working with departments of transportation and state police agencies as they redesign crash record collection to include a simple occupation-related indicator is essential. Second, while occupation-related indicators exist in most EMS databases, they are not required data fields across states. This leads to a large proportion of null or not assigned values in these indicators. More communication with NEMSIS at both the federal and state levels to stress the importance of including these variables as required is needed. Practice recommendations are not independent of research efforts in this space as it is the burden of occupational injury scientists to use innovative methods, such as combining machine learning and data linkage efforts, to educate and work with stakeholders to improve data capturing methods and quality.

CRediT authorship contribution statement

Caitlin A. Northcutt: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Nikiforos Stamatiadis:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Michael A. Fields:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Reginald Souleyrette:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

Funding: The authors have no conflicts of interest, actual or perceived, to declare. This publication was supported by Grant or Cooperative Agreement number National Institute for Occupational Safety and Health (NIOSH) U60-award (5U60OH008483-19 PI: Northcutt Pope) titled "Kentucky Occupational Safety and Health Surveillance Program'', funded by the NIOSH/Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aap.2024.107749.

References

Anderson, L.A., Goodman, R.A., Holtzman, D., Posner, S.F., Northridge, M.E., 2012. Aging in the United States: opportunities and challenges for public health. Am. J. Public Health 102 (3), 393–395. https://doi.org/10.2105/AJPH.2011.300617.

Benedetti, M.H., Humphries, K.D., Codden, R., Sagar, S., Kufera, J.A., Cook, L.J., Norris, J., Stamatiadis, N., Vesselinov, R., Zhu, M., 2022. Age-based variability in the

association between restraint use and injury type and severity in multi-occupant crashes. Ann. Epidemiol. 76, 114–120.e112. https://doi.org/10.1016/j.annepidem.2022.10.003.
Boggs, A.M., Wali, B., Khattak, A.J., 2020. Exploratory analysis of automated vehicle crashes in California: a text analytics & hierarchical Bayesian heterogeneity-based approach. Accid. Anal. Prev. 135, 105354 https://doi.org/10.1016/j.aap.2019.105354.
Boufous, S., Williamson, A., 2009. Factors affecting the severity of work related traffic crashes in drivers receiving a worker's compensation claim. Accid. Anal. Prev. 41 (3), 467–473. https://doi.org/10.1016/j.aap.2009.01.015.
Bunn, T., Honaker, R., Maloney, P., 2023. Kentucky Occupational Health Indicators Report. K. I. P. a. R Center.
Bunn, T.L., Liford, M., Turner, M., Bush, A., 2022. Driver injuries in heavy vs. light and medium truck local crashes, 2010–2019. J. Saf. Res. 83, 26–34. https://doi.org/10.1016/j.jsr.2022.08.001.
Burch, C., Cook, L., Dischinger, P., 2014. A comparison of KABCO and AIS injury severity metrics using CODES linked data. Traffic Inj. Prev. 15 (6), 627–630. https://doi.org/10.1080/15389588.2013.854348.
Burdett, B., Li, Z.X., Bill, A.R., Noyce, D.A., 2015. Accuracy of injury severity ratings on police crash reports. Transp. Res. Record 2516, 58–67. https://doi.org/10.3141/2516-09.
Bureau of Labor Statistics. (2022). *National Census of Fatal Occupational Injuries in 2021*.
Bureau of Labor Statistics. (2023). *2021 Kentucky Fatal Occupational Injuries (CFOI) Data* https://www.bls.gov/iif/state-data.htm.
Bureau of Transportation Statistics. (2021). *Daily Vehicle Travel During the COVID-19 Public Health Emergency* https://www.bts.gov/covid-19/daily-vehicle-travel.
Bush, A.M., Bunn, T.L., Liford, M., 2021. Identification of work-related injury emergency department visits using International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) codes. Inj. Prev. 27 (S1), i3–i8. https://doi.org/10.1136/injuryprev-2019-043507.
Chen, S.H., 2010. Mining patterns and factors contributing to crash severity on road curves. Queensland University of Technology].
Clark, D.E., 2004. Practical introduction to record linkage for injury research. Inj. Prev. 10 (3), 186–191. https://doi.org/10.1136/ip.2003.004580.
Conner, K.A., Smith, G.A., 2014. The impact of aggressive driving-related injuries in Ohio, 2004–2009. J. Saf. Res. 51, 23–31. https://doi.org/10.1016/j.jsr.2014.08.003.
Cook, L.J., Thomas, A., Olson, C., Funai, T., Simmons, T., 2015. Crash Outcome Data Evaluation System (CODES): An Examination of Methodologies and Multi-State Traffic Safety Applications. National Highway Traffic Safety Administration, Washigton, D.C. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812179.
Council of State and Territorial Epidemiologists. (2021). *Occupational Health Indicators: A Guide for Tracking Occupational Health Conditions and Their Determinants* Council of State and Territorial Epidemiologists n Collaboration with the Centers for Disease Control and Prevention National Institute for Occupational Safety and Health. https://www.cste.org/group/OHIndicators.
Curry, A.E., Pfeiffer, M.R., Metzger, K.B., Carey, M.E., Cook, L.J., 2021. Development of the integrated New Jersey safety and health outcomes (NJ-SHO) data warehouse: catalysing advancements in injury prevention research. Inj. Prev. 27 (5), 472. https://doi.org/10.1136/injuryprev-2020-044101.
Emu, M., Kamal, F.B., Choudhury, S., Rahman, Q.A., 2022. Fatality prediction for motor vehicle collisions: mining big data using deep learning and ensemble methods. IEEE Open Journal of Intelligent Transportation Systems 3, 199–209. https://doi.org/10.1109/OJITS.2022.3160404.
Gao, L., Wu, H., 2013. *Verb-Based Text Mining of Road Crash Report* Transportation Research Board 92nd Annual Meeting, Washington, DC. https://trid.trb.org/view/1241434.
Han, G.M., Newmyer, A., Qu, M., 2017. Seatbelt use to save money: Impact on hospital costs of occupants who are involved in motor vehicle crashes. Int. Emerg. Nurs. 31, 2–8. https://doi.org/10.1016/j.ienj.2016.04.004.
Hosseinzadeh, A., Karimpour, A., Kluger, R., Orthober, R., 2022. Data linkage for crash outcome assessment: Linking police-reported crashes, emergency response data, and trauma registry records. J. Saf. Res. 81, 21–35. https://doi.org/10.1016/j.jsr.2022.01.003.
Hsiao, H., Stout, N., 2010. Occupational injury prevention research in NIOSH. Saf. Health Work 1 (2), 107–111. https://doi.org/10.5491/shaw.2010.1.2.107.
Jurafsky, D., Martin, J. H. (2023). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed.) https://web.stanford.edu/~jurafsky/slp3/.
Kim, J., Trueblood, A.B., Kum, H.C., Shipp, E.M., 2021. Crash narrative classification: Identifying agricultural crashes using machine learning with curated keywords. Traffic Inj. Prev. 22 (1), 74–78. https://doi.org/10.1080/15389588.2020.1836365.
Milani, J., Kindelberger, J., Bergen, G., Novicki, E.J., Burch, C., Ho, S.M., West, B.A., 2015. Assessment of characteristics of state data linkage systems (DOT HS 812 180).
Mirani, N., Ayatollahi, H., Khorasani-Zavareh, D., 2020. Injury surveillance information system: a review of the system requirements. Chin. J. Traumatol. 23 (3), 168–175. https://doi.org/10.1016/j.cjtee.2020.04.001.
National EMS Information System (NEMSIS). (2023). NEMSIS Version 3.4.0.200910CP Data Dictionary. In https://nemsis.org/technical-resources/version-3/version-3-data-dictionaries/.
Nayak, R., Piyatrapoomi, N., Weligamage, J., 2010. Application of text mining in analysing road crashes for road asset management. Engineering Asset Lifecycle Management, London.
Rakotonirainy, A., Chen, S., Scott-Parker, B., Loke, S.W., Krishnaswamy, S., 2015. A novel approach to assessing road-curve crash severity. Journal of Transportation Safety & Security 7 (4), 358–375. https://doi.org/10.1080/19439962.2014.959585.
Sagar, S., Stamatiadis, N., Wright, S., Green, E., 2020. Use of codes data to improve estimates of at-fault risk for elderly drivers. Accident Analysis & Prevention 144, 105637. https://doi.org/10.1016/j.aap.2020.105637.
Scott, E., Hirabayashi, L., Levenstein, A., Krupa, N., Jenkins, P., 2021. The development of a machine learning algorithm to identify occupational injuries in agriculture using pre-hospital care reports. Health Inf. Sci. Syst. 9 (1), 31. https://doi.org/10.1007/s13755-021-00161-9.
Shen, S., Neyens, D.M., 2015. The effects of age, gender, and crash types on drivers' injury-related health care costs. Accid. Anal. Prev. 77, 82–90. https://doi.org/10.1016/j.aap.2015.01.014.
Singleton, M., 2016. Differential protective effects of motorcycle helmets against head injury. Traffic Inj. Prev. 18 (4), 387–392. https://doi.org/10.1080/15389588.2016.1211271.
Smith, S.M., Pegula, S.M., 2020. Fatal occupational injuries to older workers (Monthly Labor Review, Issue.
Tainter, F., Fitzpatrick, C., Gazillo, J., Riessman, R., Knodler Jr., M., 2020. Using a novel data linkage approach to investigate potential reductions in motor vehicle crash severity - an evaluation of strategic highway safety plan emphasis areas. J. Saf. Res. 74, 9–15. https://doi.org/10.1016/j.jsr.2020.04.012.
Taylor, J.A., Lacovara, A.V., Smith, G.S., Pandian, R., Lehto, M., 2014. Near-miss narratives from the fire service: a Bayesian analysis. Accid. Anal. Prev. 62, 119–129. https://doi.org/10.1016/j.aap.2013.09.012.
Thomas, A.M., Thygerson, S.M., Merrill, R.M., Cook, L.J., 2012. Identifying work-related motor vehicle crashes in multiple databases. Traffic Inj. Prev. 13 (4), 348–354. https://doi.org/10.1080/15389588.2012.658480.
Trueblood, A.B., Pant, A., Kim, J., Kum, H.C., Perez, M., Das, S., Shipp, E.M., 2019. A semi-automated tool for identifying agricultural roadway crashes in crash narratives. Traffic Inj. Prev. 20 (4), 413–418. https://doi.org/10.1080/15389588.2019.1599873.
Vallmuur, K., Marucci-Wellman, H.R., Taylor, J.A., Lehto, M., Corns, H.L., Smith, G.S., 2016. Harnessing information from injury narratives in the 'big data' era: understanding and applying machine learning for injury surveillance. Inj. Prev. 22 (Suppl 1), i34 https://doi.org/10.1136/injuryprev-2015-041813.
Zhang, X., Green, E., Chen, M., Souleyrette, R.R., 2020. Identifying secondary crashes using text mining techniques. Journal of Transportation Safety & Security 12 (10), 1338–1358. https://doi.org/10.1080/19439962.2019.1597795.
Zheng, D., Chitturi, M.V., Bill, A.R., Noyce, D.A., 2015. Analyses of multiyear statewide secondary crash data and automatic crash report reviewing. Transp. Res. Rec. 2514 (1), 117–128. https://doi.org/10.3141/2514-13.