HHS Public Access

Author manuscript

Int J Occup Saf Ergon. Author manuscript; available in PMC 2024 June 01.

Published in final edited form as:

Int J Occup Saf Ergon. 2024 June ; 30(2): 559–570. doi:10.1080/10803548.2024.2325301.

Establishment-level safety analytics: a scoping review

Anne M. Foreman[a], **Jonathan E. Friedel**[b], **Maira E. Ezerins**[c], **Riggs Matthews**[d], **Royale E. Nicholson**[d], **Logan Wellersdick**[d], **Shawn Bergman**[d], **Yalcin Açıkgöz**[d], **Timothy D. Ludwig**[d], **Oliver Wirth**[a]

[a]Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, WV, USA

[b]Department of Psychology, Georgia Southern University, Statesboro, GA, USA

[c]Department of Management, The Sam M. Walton College of Business, University of Arkansas, Fayetteville, AR, USA

[d]Department of Psychology, Appalachian State University, Boone, NC, USA

Abstract

The use of data analytics has seen widespread application in fields such as medicine and supply chain management, but their application in occupational safety has only recently become more common. The purpose of this scoping review was to summarize studies that employed analytics within establishments to reveal insights about work-related injuries or fatalities. Over 300 articles were reviewed to survey the objectives, scope and methods used in this emerging field. We conclude that the promise of analytics for providing actionable insights to address occupational safety concerns is still in its infancy. Our review shows that most articles were focused on method development and validation, including studies that tested novel methods or compared the utility of multiple methods. Many of the studies cited various challenges in overcoming barriers caused by inadequate or inefficient technical infrastructures and unsupportive data cultures that threaten the accuracy and quality of insights revealed by the analytics.

Keywords

data analytics; occupational safety; injuries; data mining

CONTACT Anne M. Foreman amforeman@cdc.gov; vpc3@cdc.gov Health Effects Laboratory Division, National Institute for Occupational Safety and Health, 1095 Willowdale Road, Morgantown, WV 26505, USA.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the National Institute for Occupational Safety and Health (NIOSH), Centers for Disease Control and Prevention.

Disclosure statement

No potential conflict of interest was reported by the authors.

Supplemental data

Supplemental data for this article can be accessed at http://dx.doi.org/10.1080/10803548.2024.2325301.

1. Introduction

There has been growing interest in the application of data analytics to aid research and practice in occupational safety. This growth is demonstrated by the rapid increase in the number of Scopus results for the search ‘safety analytics’ from 2003 to 2022 (see Figure 1). The popular press coverage on safety analytics has increased in kind. A post on the NIOSH Science Blog titled ‘Can Predictive Analytics Help Reduce Workplace Risk?’ outlines the potential benefits and barriers to the application of analytics to workplace safety, using the logic that ‘if injuries can be predicted accurately, they can be prevented’ [1]. Similarly, a 2021 article in *Industrial Safety & Hygiene News* titled ‘Rethinking Predictive Analysis: Learn How to Stop Workplace Incidents Before They Occur’ [2] described the advantages of analytic approaches to safety in identifying patterns associated with risk. The number of companies applying analytics to their occupational safety data may be unknown but, given the number of consultation companies now providing occupational safety solutions within their data analytic services and considering the substantial increase in the number of published studies related to this topic, we can infer the number is likely increasing.

The growing number of articles in the occupational safety analytics literature (reflected in Figure 1) cover a variety of topic areas including sensor technology, hazard evaluation, personal protective equipment (PPE) identification and evaluating injuries across sectors, within industries and within organizations. Sensor technology studies are concerned with implementing sensors for detecting worker-specific variables like fatigue [3], muscle force [4] or energy expenditure [5], then analyzing the sensor data with analytic techniques. Hazard evaluation studies are concerned with using analytics to identify hazards in the work environment. Hazards in this context are aspects of technology, tasks or activities with potential for harm [6]. Some examples include implementing a video detection algorithm to detect safety equipment [7] and developing machine learning algorithms to predict hearing loss in workers associated with industrial noise [8]. PPE identification studies are concerned with using computer vision technology to identify PPE use among workers. Studies have been conducted targeting hard hat [9] and harness [10] use, among other types of PPE. Finally, analytics studies have evaluated injury data to identify the cause of incidents across multiple levels of analysis such as across national sectors [11], within industries [12] and within organizations [13].

Acknowledging a lack of agreement as regards the definition of analytics in the literature [14,15], we started with a consensus definition: ‘the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data’ (p.3) [16]. For the purposes of this review, we further limited the definition to only include existing data, not simulated data. Although this consensus definition is rather broad and non-specific, its inclusivity permits us to capture a wide range of approaches in the review. While we limited our scope to include only existing data and excluded research in occupational safety using simulated data, a wide range of analytic approaches are captured in the review. There is a great deal of heterogeneity in the existing body of studies as the context for analytics in occupational safety involves a rather recent shift away from traditional approaches, which often consisted of simple

frequency counts of injuries across time, to applied data analyses and more complex analytic approaches like machine learning.

The purpose of the present article is to conduct a review of studies that use analytics within enterprises or establishments to reveal insights about work-related injuries or fatalities. Our review focuses on enterprise-level and establishment-level analytics instead of industry-level analytics, which are usually based on analyses of national databases such as those recorded by the Occupational Safety and Health Administration or the Bureau of Labor Statistics. Injury reporting to industry or regulatory bodies often only includes a small number of variables, such as days away from work or restricted time at work. Reviews have been published on industry-level safety analytics [17], but to date no reviews have focused on enterprise-level or establishment-level analytics. The present review may serve as a useful resource for researchers, safety professionals and others interested in implementing analytics within organizations by summarizing the methods and findings of research that has been conducted at similar levels.

2. Method

We performed a scoping review – a preliminary assessment of the size and scope of available research literature – of selected articles in which occupational safety analytics was conducted at the enterprise or establishment level. A scoping review approach was used instead of a systematic review approach, which seeks to identify and synthesize evidence related to a specific question, because the application of analytics in this area of investigation is a relatively new phenomenon and there is substantial heterogeneity in types of applications of analytics in occupational safety and the methods used. In conducting the scoping review, we followed guidelines recommended by the Preferred Reporting Items for Systematic reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR [18]).

2.1. Eligibility criteria and search details

We searched Scopus and Web of Science for relevant articles. There were two sets of search terms: one set related to analytics and the other set related to occupational safety. The search terms were developed through consultation with experts and based on results of preliminary attempts to search various databases. The same terms were used across databases, but the search strategy was tailored to the particular syntax of each database. The search string was as follows:

> TITLE-ABS-KEY ((‘big data’ OR analytics OR ‘data mining’ OR ‘machine learning’ OR clustering OR ‘decision tree∗’ OR ‘neural network’ OR ‘artificial intelligence’ OR ‘association rule’) AND (‘occupational safety’ OR ‘occupational injur∗’ OR ‘work∗ injur∗’ OR ‘occupational health and safety’ OR ‘industrial safety’ OR ‘construction safety’ OR ‘manufacturing safety’))

The search was limited to articles published after 2006, and only peer-reviewed journal articles and conference publications were considered. Furthermore, articles had to be written in English, involve the use of analytics techniques, have injuries or near misses as an outcome of interest (i.e., the dependent variable) and use data collected at the enterprise or

establishment level. The inclusion and exclusion criteria are listed in Table 1. These studies were excluded to both keep the number of resulting articles to a manageable number and to focus on studies that evaluated injuries or near misses as the outcome variable. Studies may refer to injuries as 'incidents' or 'accidents' as these terms are often used interchangeably [19]. Examples of studies that were excluded included those that were not concerned with occupational safety (e.g., pedestrian safety, patient safety, etc.); dealt with occupational health issues instead of injuries (e.g., pneumoconiosis, cardiovascular disease, etc.); were concerned with assessing risks or hazards; or were literature reviews, commentaries or other types of non-empirical research papers.

Additional searches, beyond the initial Scopus and Web of Science search, were conducted to ensure that as many qualifying studies as possible were identified. The *Safety Science* journal had the largest number of qualifying articles identified from the database searches (described earlier). Therefore, we conducted a manual search of the titles and abstracts of every paper published in *Safety Science* from 2005 to September 2023. Additionally, citation searches were conducted for the studies that met the inclusion criteria. This involved searching the references of a particular study and subsequent citations of that study listed by Google Scholar for inclusion criteria.

2.2. Selection and data charting

At least two co-authors reviewed each citation, and the first author (A.M.F.) reviewed all citations. If two authors disagreed on an article's inclusion, a third author would break the tie. The searches were conducted in September 2023. A data chart was populated for each article selected for inclusion. The data chart contained the following columns: year, title, authors, affiliation, author keywords, journal, study question(s)/objective(s), description of data source(s), data preprocessing, predictor variables, outcome variables, analytics techniques, analytics-derived insights, challenges described and notes. A thematic synthesis was used to identify and summarize patterns in the included articles and identify knowledge gaps. These findings will be described in the following section.

3. Results

A flow chart summarizing the scoping review process is shown in Figure 2. Searches of Scopus and Web of Science resulted in 2845 citations. After the removal of duplicates, 2617 unique citations remained. After reviewing the abstracts for meeting the exclusion criteria, 2292 studies were excluded and 325 studies required further review. The full text of the remaining 325 studies was further examined to assess their eligibility. Because the present review is focused on analytics conducted at the enterprise or establishment level, 71 studies were excluded because their data sources were collected across sectors or industries (e.g., occupational injuries by country) and 225 studies were excluded because their data were collected within a sector or industry. The remaining 29 studies from the initial search satisfied all the inclusion criteria. From the manual search of *Safety Science*, we identified three additional citations. We also identified another 17 studies via the forward and backward citation search method. The final sample included 49 studies.

Figure 3 shows summary characteristics of the included studies. Figure 3(A) displays the years of publication for the 49 studies. The first study that met the inclusion criteria was published in 2008. There was a marked increase in the number of studies in 2017, and the greatest number of studies were published in 2019. The countries in which the studies were conducted are shown in Figure 3(B). Most studies were conducted in India (n = 29; 59% of the included studies). Italy and Iran were the second most frequent with three (6%) studies each. Figure 3(C) displays the sectors in which the studies were conducted. These included manufacturing (n = 36; 73%), construction (n = 7; 15%), utilities (n = 3; 6%), transportation and warehousing (n = 2; 4%), and agriculture, forestry, fishing and hunting (n = 1; 2%). Finally, Figure 3(D) shows the types of manufacturing represented in the scoping review studies. Most of the studies conducted in manufacturing were in steel manufacturing (83%; n = 30/36).

3.1. Study objectives

All studies included in the scoping review attempted to predict injuries or near misses, or, alternatively, to identify common patterns related to injuries or near misses. This is not surprising as one of the inclusion criteria for a study was that it be related to analytics of injuries or near misses. Most of the studies were concerned with exploring data from incident reports, safety inspection reports and other safety data, although there were a few exceptions with more specific research objectives. For example, Marques et al. [20] examined the effects of drug and alcohol testing frequency on the occurrence of injuries. Additionally, Tsang et al. [21] explored the effects of employee characteristics (e.g., age, body mass index [BMI], average heart rate) and environmental/behavioral variables (e.g., ambient temperature, recovery time) on injuries among cold-storage employees.

3.2. Descriptions of data sources

Many studies did not disclose the sources of their data. This is problematic, as it impedes the ability to replicate or extend that research by others in the field. There were some common themes in the studies that did disclose information on the sources and types of data sources. Twenty-five studies that included information on data sources were conducted in steel plants, and a few studies were conducted in construction organizations or refineries. The datasets usually spanned 3–5 years, with the longest duration of time being 16 years [22]. Various people collected the data, including front-line workers [23,24] and managers [25], and data were typically stored in the organization’s online safety management system (SMS) [23,26–31].

Data reported in the studies featured both structured and unstructured data with a mix of data types (e.g., categorical, numerical, textual). Structured data can be thought of as searchable data that are easily processed by a computer and have a clearly defined organizational system [32]. Unstructured data cannot be easily processed with conventional tools and has no clearly defined organizational system (e.g., the text of written incident reports, photographs, etc.) [32]. Commonly used structured data included the frequency of injuries and incidents. Unstructured data included text-based descriptions of incidents and observations as well as occasional photographic observations.

There was a high degree of variability in the types of data collected. Most of the variables included in the studies were reactive or lagging indicators. These are variables such as incident reports, injuries and lost workdays that occur after a safety incident. Figure 4(B) shows the most frequently included incident details. Only 10 (22%) studies included proactive or leading indicators. Leading indicators occur before a safety incident and include preventive variables such as safety audit reports, frequency of toolbox meetings and safety training history. A few studies included human resource management data such as employee age, education and marital status. Figure 4(A) shows the most frequently included employee characteristics. Although some studies in the construction sector included details related to project cost, stage of completion or complexity [33,34], limited studies used data from organizational areas outside safety, such as production or maintenance. For example, Tsang et al. [21] did include production data, but these data were used for analyses separate from those investigating injury.

Some of the studies included upwards of 20 variables in their analyses, and although a majority provided definitions for each variable, 14 (29%) of the studies did not provide any definitions. For example, one study examined several predictor variables, including ‘cause of accident’, but did not describe what causes were captured by that variable [35]. Although some variables may be self-explanatory or not require detailed definitions (e.g., year, day of the week, season), other variables necessitate further explanation (e.g., nature of incident, equipment damage score, etc.). To fully understand and interpret the results of the analyses conducted, the reader needs to know what the variables are measuring so that they may draw informed conclusions from the study results, make comparisons to their own organizational data or compare results across studies.

3.3. Data preprocessing

The preprocessing stage of analytics involves selecting variables to analyze, handling missing data and restructuring text data. This stage typically takes the most amount of time, and there are different methods for approaching the preprocessing steps. Figure 5(A) shows the approaches to variable selection reported in the studies included in this review. Variable selection involves deciding which variables to include in the analysis. For 27% ($n = 13$) of the studies, there was no description of the process for variable selection [1222, 24,25,34–42]. For studies that described the process, the most common approaches were statistical in nature, including using Boruta feature selection [34,44,45], χ^2 [27,28,33] and random forest [46], among others [22,24,47–53]. Fifteen studies stated inclusion or exclusion criteria which varied from including all variables [29,31,54–58], the most common variables [59] or variables that aligned with specific International Organization for Standardization recommendations [21], among others [20,60–64]. Last, five studies stated that their variables were selected based on domain knowledge or consultations with experts [30,65–68].

Accounting for and handling missing data is an important step in preprocessing and analysis. The approaches to handling missing data are shown in Figure 5(B). Most of the studies (59%; $n = 29$) did not include a description of how missing data was handled. Eight studies used a statistical approach to handling missing data, which included imputation [33,50], random forest [27,44–47] and expectation maximization [65]. Seven

studies mentioned that missing data were accounted for but did not provide specific details [24,26,29,39,42,57,61]. Four studies stated that data with missing values were omitted from the analysis [35,41,59,60]. One study consulted with experts to inform the techniques to handle missing data, although did not describe the chosen techniques in their manuscript [30].

Free-text data (e.g., incident narratives) can often provide a great deal of information, but it first must be converted into a format that can be analyzed. That is, unstructured, open-ended text (e.g., explanations on an incident report) must be converted to a structured format that can be read by data analysis software. There are numerous methods to convert free text to structured data. The methods to convert free text reported by the studies in this review are shown in Figure 5(C). The most common approach was latent dirichlet allocation [24,27,44,47,53,56] followed by term frequency-inverse document frequency (TF-IDF) [13,24,38,54,64]. Other studies used structural topic modeling [45,57] and expectation maximization [24,28].

Class imbalances in the data were another concern that was addressed during the data preparation phase in several studies. Class imbalance occurs when one outcome variable occurs disproportionately more often than another, creating bias toward the majority variable [69]. In such datasets, the variable of greatest interest is often the minority variable, such as workplace injuries, compared to more common near-miss events [70]. Data-level techniques are the most popular methods for addressing class imbalance and consist of two approaches, undersampling and oversampling. In oversampling, values from the minority variable are replicated to increase the size of the class, and in undersampling, values from the majority variable are deleted to decrease the size of the class. In the present studies, undersampling was used in one study [50] and oversampling was used in a handful of studies to address class imbalance issues. Oversampling techniques implemented included the synthetic minority oversampling technique (SMOTE) [22,33,34,44], majority weighted minority oversampling technique (MWMOTE) [44], borderline SMOTE (BLSMOTE) [69] and *k*-means SMOTE (KMSMOTE) [44].

The software used to conduct analysis is not strictly part of data preprocessing. However, there are an increasing number of widely available software programs developed for conducting analytics in which different implementations of the algorithms (e.g., non-convergence rates) within the software lead to slightly different analysis outcomes. We have no reason to suspect that any such issue is related to the outcomes within the studies under review, but it is good practice to report the software used for preprocessing and data analysis. The software used in the reviewed studies are shown in Figure 5(D). Most of the studies (78%) did describe which software was used to analyze data. The most common software was R [25,26,31,33,38,43–47,52,57,66], which is an open-source, free software program. The next two most frequently used software programs were SAS [23,30,58,60,61,67] and SPSS [20,35,37,40,59,68].

3.4. Analytics approaches

Approaches to analytics are sometimes divided into classification and regression tasks [71]. In classification methodologies, the output of a statistical model is assigned to a particular

class. In the case of occupational safety, an example of the output may be an injury event or a non-injury event. In regression methodologies, the output of a model is a continuous variable. In occupational safety research, an example of a continuous variable may be the frequency or rate of injuries. For example, Ajayi et al. [33] compared several different machine learning techniques predicting the number of hand injuries in power infrastructure operations workers. For most of the selected studies, the output of the models was assigned to a class (e.g., fatal injury, serious injury, first aid [56]). Most analytic techniques can be used for both classification and regression.

Figure 6 shows the techniques that were used in studies focusing on a single analytics approach or method. The most common analytics techniques among these studies were association rule mining [21,25,30,39,54,59,63,64,72] and Bayesian networks [26,29,36,41,62]. Many of these single-approach studies employed techniques that examine the relations among temporal variables (e.g., antecedent events preceding near misses or injuries) including association rule mining, multiple correspondence analysis, vector autoregression, object role modeling, cause-and-effect diagrams and axiomatic design framework. As an example, association rule mining is a technique that was originally developed to detect patterns of transactions in retail stores to identify the frequency of patterns in data by identifying conditional associations (e.g., if/then or antecedent/consequent relationships). By examining these conditional associations, events or characteristics that are commonly correlated with injury (the ‘then’ or ‘consequent’ portion of the association) can be detected. The resulting rules can vary in the number of items (e.g., a four-item rule could be: young workers [Item 1] with shorter job tenures [Item 2] who work in Department 3 [Item 3] are more likely to have lower limb injuries [Item 4]), although the analysis becomes more complex with additional items. Studies in this scoping review obtained two-item [21,25,39], three-item [20,24,29,38,53,58,62,63,71] all eight studies previously listed, four-item ([21,30,39,40,54,72] and five-item [30,39,54] rules. In general, the researchers were able to identify more specific scenarios in which injuries were likely to happen through association-rule mining. For example, Buddhakulsomsiri et al. [63] found that incidents that resulted in major injuries and high costs were associated with work performed by outside contractors, less experienced workers and workers between 41 and 50 years of age. The studies that used association-rule mining were in oil refining [25], warehousing [21], construction [59], steel manufacturing [39,54,64,67] and car manufacturing [40].

One finding that emerged was that many studies (41%) were comparing the performance of several different analytics approaches. The studies and the techniques compared are presented in Table 2. The objective of those comparisons was to identify techniques that were most accurate in predicting injuries (or near misses, depending on the study). The techniques that were identified as the most accurate in each study are highlighted in Table 2 and included random forest and classification and regression tree techniques. A summary table of the analytics techniques performed in all 49 included studies can be found in the Supplemental data.

3.5. Analytics-derived insights

Most of the studies provided detailed findings from the analytics conducted. Of the studies that reported detailed findings, many of the relations identified were rather complex and were unlikely to be detected by using more traditional, simpler analytical approaches. For example, Ajayi et al. [33] identified complex interactions among project complexity, time of year, worker age and experience, and task type that were more predictive of injury. Sarkar et al. [44] developed 19 specific safety rules that described scenarios under which incidents were more likely to happen in different divisions of a steel manufacturing facility. Lingard et al. [43] identified complicated interactions among safety indicators over time that suggested a cyclical relationship between the behavior of management and the occurrence of injury. Through a χ^2 automatic interaction detector (CHAID) analysis, Marques et al. [20] were able to identify an optimal schedule for drug and alcohol testing of employees that reduced injuries but would not place undue burden on either the organization or its employees. Although the findings of some studies were somewhat unsurprising (e.g., attributes like incident types and primary causes were related to injury risk) [22,28,67], many analyses in these studies made use of free-text, narrative data which often go unanalyzed. Finally, six studies were methodologically focused and only reported information related to model fit (e.g., accuracy, robustness) [28,36,38,41,45,66].

Although a desirable end goal of any analytics effort is to reveal actionable insights, the findings may be less actionable than desired. In fact, they may reveal more fundamental or underlying problems with a company's existing data reporting systems. For example, through an expectation-maximization-based clustering analysis of free-text injury narratives, Verma et al. [67] identified misunderstandings on the part of workers about operational definitions for incident variables. The analysis revealed that workers were categorizing some events related to falling materials as slip/trip/fall events instead of 'struck by' events. This study demonstrates that the analytics process is not unidirectional and can also inform and improve processes related to data collection. For example, it can lead to unexpected outcomes, like the improvement of data collection systems, by identifying unclear or incorrect instructions for incident reporting.

As analytics in occupational safety is a relatively new area of exploration, most of the studies were designed as methodological demonstrations or considered pilot studies from which future studies could be based. Only one study in the review described actions taken based on the analytics that were conducted. Tsang et al. [21] reported that because of the implementation of the analytics software that was developed, actions such as allocating additional recovery time from low-temperature environments or wearing at least three layers of clothing insulation decreased the frequency of employee injuries from 12 injuries per week to five.

3.6. Limitations and future directions identified by the included studies

There were many limitations identified in the included studies. Two categories of limitations were cited most frequently: model choice and data integrity. Whereas all models had some degree of success in predicting safety incidents, some data did not meet the assumptions of the chosen model which undermined validity. In one example, which concerned text

analytics, researchers found that the use of short-hand narratives to describe accidents led to too sparse an amount of contextual data to allow for valid conclusions [36]. To mitigate this shortcoming, Sarkar et al. [28] used more advanced methods such as expectation-maximization algorithms in follow-up studies. The authors also stated that more in-depth analyses using similar statistical techniques may yield richer insights. For another example, Guo et al. [59] suggested using additional multidimensional association rule mining of unsafe behavior to learn more about behavior patterns.

Other common limitations identified by the studies in the review included data collection methods, data integrity and data variety/breadth. Recurring problems with collection methods included biased observers, a lack of subject matter experts and the underreporting of injuries; a frequently proposed solution within the identified studies to these problems is the use of standardized surveys made by subject matter experts. Common data integrity issues included confidentiality concerns and inconsistency of record-keeping across different departments. The final limitation within datasets concerned the breadth of data. In one case, while building a rule mining database, the researchers found that their case study was not sufficiently large or broad enough to allow for generalizable results [59]. Another study found that the sample size limited their ability to extract decision rules from the decision tree they had generated [65].

4. Discussion

For the current scoping review, we identified 49 studies conducted since 2007 that implemented analytics techniques within an enterprise or establishment to improve workplace safety. The identified studies were conducted predominantly in the steel manufacturing industry within India, although a smaller number of studies were conducted in other manufacturing industries, warehousing and construction. More than half (n = 27, 55%) of the included studies were performed by Maiti and colleagues at the Indian Institute of Technology Kharagpur, which accounts for the predominance of studies conducted in an Indian steel manufacturing plant. Their productivity is demonstrative of the depth of occupational safety insights that can be obtained through data analytics. The review uncovered a large variety in the types of analytics that have been implemented and the software used to conduct the analyses. The insights revealed with analytics approaches in many studies were likely more complex than those that can be identified by traditional approaches to injury data analysis.

4.1. Variety of data sources

The reviewed studies primarily analyzed safety-related data. Most of the data consisted of lagging indicators such as the details of incidents and near misses. More recent studies have incorporated leading indicators, including safety audit reports, safety training records and the frequency of toolbox talks. By incorporating these leading indicators into safety analytics, organizations can assess how the effort put toward proactive safety programs is borne out quantitatively in improvements to lagging indicator metrics. Additionally, by conducting the somewhat onerous data preprocessing stages for leading indicators, organizations may be able to identify areas in need of improvement or gaps in their

safety programs. Extending data sources beyond safety-specific data to other divisions of an organization could have the potential to detect new relations among variables. For example, incorporating production data (e.g., production pressure), human resource management data (e.g., overtime) or weather data (e.g., temperature, humidity, precipitation) could provide a more detailed picture of the potential causes of injury.

Providing more information about the variables analyzed is another area that could be improved. More than one-quarter of the studies did not provide definitions for the variables in the analysis, and on a few occasions the names of the variables were shorthand labels used in the data analysis programs, making interpretation by the reader more difficult. Explicitly stating the variable type, whether it was Boolean, numeric, categorical or free text, would also improve the understanding of the data, including if and how the data were converted from one format to another.

4.2. Data preprocessing and analytics approaches

As with the sources of data, more information about the data preprocessing stages could improve future manuscripts. A large proportion of the studies did not include information related to how variables were selected for analysis or how missing data were dealt with. In some cases, variable selection may be straightforward if the variables are limited to those collected as part of incident reporting. In any case, explicitly stating what variables were included and excluded would help the reader better understand the authors' analytics process. Regarding missing data, there are a variety of approaches that can be implemented, including omitting files with missing data from the analysis or addressing them through imputation [73]. More than half of the included studies did not mention missing data, and of those that did only 11 described the specific procedures used. Given that missing data is an inevitability, particularly with data collected by managers and workers, how this is handled is an important part of the data preprocessing details that should be included.

There were a variety of analytics techniques implemented across the included studies, but often the reasons for selecting techniques were not described. Stating a rationale for the chosen methodology would assist others in deciding what approach to take with their own organizational data. Similarly, providing more detailed information about the software and packages utilized would provide further assistance to the research community. For example, providing a list of the R packages selected for data analysis would help others replicate an approach with their own data.

4.3. Findings

The analytics conducted in the included studies resulted in findings that often would not be obtained with approaches to injury data analysis that solely involve descriptive statistics. In many cases, more complex relations indicating the specific circumstances under which types of incidents occurred were detected and described. One potential area of improvement is to more clearly describe detected relationships among variables. Jargon that is specific to an organization may be difficult for the reader to interpret, and thus better-defined terms would improve understanding. This includes ensuring that all acronyms are spelled out and adequately defined within tables, figures and text. Further, studies often failed to

describe their findings in great practical depth. While technique comparisons were common, these models were not compared to the base rate of safety predictions without modeling [45]. Studies also failed to report the practical application of their models; trained models were excellent at predicting another already-gathered dataset, but there was no mention of accident reduction when implemented at a facility [41,65].

Other studies generated conclusions using an inductive approach, with generation of their hypotheses post hoc [33]. While this method allows scientists to be more liberal in their investigations, a lack of a-priori hypothesis generation is a principal component that leads to reproducibility issues and to 'just so' explanations that overgeneralize results [74]. Once again, without follow-up studies confirming the reduction in outcome variables pre and post model integration, it is difficult to assess whether these inductive conclusions are valid.

Studies often attributed blame to safety observers, citing a lack of informative data entry [30,67] or data integrity issues related to a diversity of reporting methods and different storage mediums [63]. While there very well may be data integrity issues associated with method of entry, a productive safety culture dissuades blaming individuals, instead focusing on aligning goals [75].

4.4. Future directions

Future studies can overcome the limitations described in the previous section with additional focus on prescriptive rather than simply predictive outcomes. Although models have shown success in predictive power, follow-up studies demonstrating reduction in real-world injuries would better inform practitioners in the field and demonstrate the utility of analytics. Model comparisons to more traditional analytics tools, such as multiple regression, could offer practitioners better perspective on the comparative advantages of more sophisticated models. Additionally, 37 out of 49 studies focused on steel manufacturing or construction. Expansion of these models to new sectors could both improve safety in those industries and possibly offer new insights into current predictive models.

5. Conclusions

The present scoping review is the first of its kind to review applications of analytics to occupational safety-related concerns at the establishment and enterprise levels. More than 300 articles from databases and journal reviews were reviewed to survey the objectives, scope and methods used in this emerging field. Despite widespread interest and long-term reliance on data analytics in other fields, we conclude that the promise of analytics for providing actionable insights to address occupational safety concerns is still in its infancy. Our review shows that most of the articles were focused on method development and validation, including studies that tested novel methods or compared the utility of multiple methods. Despite these promising efforts, few studies reported actionable insights derived directly from the analytics. Therefore, the espoused goals and promise of analytics for occupational safety have yet to be fully realized. Nevertheless, we are optimistic that increasing use of and reliance on analytics by safety practitioners and researchers will spur rapid progress in this field, and the work described in the studies included in this review has resulted in a relative treasure trove of references for those interested in applying

particular methods to their organizational data. Our review also revealed a final point worth emphasizing, and that is the importance of establishing ‘readiness’ for analytics. Many of the studies cited various challenges in overcoming barriers caused by inadequate or inefficient technical infrastructures and unsupportive data cultures that threaten the accuracy and quality of insights revealed by the analytics. The old adage ‘garbage in, garbage out’ characterizes a common threat to many well-intentioned analytics initiatives within companies. Indeed, many establishments or enterprises are simply not ready for analytics because inadequate measurement systems are in place. An ‘analytics readiness audit’ seems to be a good first step before embarking on further analytics inquiries.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

[1]. Wagner GR. Can predictive analytics help reduce workplace risk?’ [Internet]. NIOSH science blog; 2014. Available from: https://blogs.cdc.gov/niosh-science-blog/2014/10/02/pa/ [cited 2024 March 20].

[2]. Suman A. Rethinking predictive analysis: Learn how to stop workplace incidents before they occur [Internet]. ISHN; 2021 [cited 2024 Mar 20]. Available from: https://www.ishn.com/articles/113203-rethinking-predictive-analysis-learn-how-to-stop-workplace-incidents-before-they-occur

[3]. Maman ZS, Chen Y-J, Baghdadi A, et al. A data analytic framework for physical fatigue management using wearable sensors. Expert Syst Appl. 2020;155:113405. doi:10.1016/j.eswa.2020.113405

[4]. Patel V, Chesmore A, Legner CM, et al. Trends in workplace wearable technologies and connected-worker solutions for next-generation occupational safety, health, and productivity. Adv Intell Syst. 2022;4(1):2100099. doi:10.1002/aisy.202100099

[5]. Jebelli H, Choi B, Lee S. Application of wearable biosensors to construction sites. II: assessing workers’ physical demand. J Constr Eng Manag. 2019;145(12):1–12.

[6]. Manuele FA. Risk assessment & hierarchies of control. Prof Saf. 2005;50(5):33–39.

[7]. Phuc LTH, Jeon H, Truong NTN, et al. Applying the Haar-cascade algorithm for detecting safety equipment in safety management systems for multiple working environments. Electronics (Basel). 2019;8(10):1079. doi:10.3390/electronics8101079

[8]. Zhao Y, Li J, Zhang M, et al. Machine learning models for the hearing impairment prediction in workers exposed to complex industrial noise:apilotstudy. EarHear.2019;40(3):690.doi:10.1097/AUD.0000000000000649

[9]. Shrestha K, Shrestha PP, Bajracharya D, et al. Hard-hat detection for construction safety visualization. J Constr Eng. 2015;2015(1):1–8. doi:10.1155/2015/721380

[10]. Fang W, Ding L, Luo H, et al. Falls from heights: a computer vision-based approach for safety harness detection. Autom Constr. 2018;91:53–61. doi:10.1016/j.autcon.2018.02.018

[11]. Rivas T, Paz M, Martín J, et al. Explaining and predicting workplace accidents using data-mining techniques. Reliab Eng Syst Saf. 2011;96(7):739–747. doi:10.1016/j.ress.2011.03.006

[12]. Mohammadian F, Sadeghi M, Hanifi SM, et al. Modeling important factors on occupational accident severity factor in the construction industry using a combination of artificial neural network and genetic algorithm. Work. 2022;73(1):189–202. doi:10.3233/WOR-205271 [PubMed: 35871380]

[13]. Sarkar S, Ejaz N, Kumar M, et al. Root cause analysis of incidents using text clustering and classification algorithms. In: Singh PK, Panigrahi BK, Suryadevara NK, editors. Proceedings of ICETIT 2019. Cham, Switzerland: Springer; 2019. p. 707–718.

[14]. Longbing B. Data science and analytics: a new era. Int J Data Sci Analytics. 2016;1:1–2. doi:10.1007/s41060-016-0006-1

[15]. Almosallam EA, Ouertani HC. Learning analytics: definitions, applications and related fields. In: Herawan T, Deris MM, Abawajy J, editors. Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). Singapore: Springer; 2014. p. 721–730.

[16]. Cooper A. What is analytics? Definition and essential characteristics. CETIS Analytics Ser. 2012;1(5):1–10.

[17]. Sarkar S, Maiti J. Machine learning in occupational accident analysis: a review using science mapping approach with citation network analysis. Saf Sci. 2020;131:104900. doi:10.1016/j.ssci.2020.104900

[18]. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467–473. doi:10.7326/M18-0850 [PubMed: 30178033]

[19]. Nemmers P. 2023. The differences between incidents vs. accidents in the workplace. National Association of Safety Professionals. 2023 [cited March 20, 2024]. Available from https://naspweb.com/blog/the-differences-between-incidents-and-accidents-in-the-workplace

[20]. Marques PH, Jesus V, Olea SA, et al. The effect of alcohol and drug testing at the workplace on individual's occupational accident risk. Saf Sci. 2014;68:108–120. doi:10.1016/j.ssci.2014.03.007

[21]. Tsang Y, Choy K, Koo P, et al. A fuzzy association rule-based knowledge management system for occupational safety and health programs in cold storage facilities. VINE J Inf Knowl Manage Syst. 2018;2:199–206.

[22]. Hoenigsberger F, Saranti A, Angerschmid A, et al., editors. Machine learning and knowledge extraction to support work safety for smart forest operations. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Cham: Springer; 2022.

[23]. Verma A, Chatterjee S, Sarkar S, et al. Data-driven mapping between proactive and reactive measures of occupational safety performance. In: Maiti J, Ray PK, editors. Industrial safety management. Singapore: Springer; 2018. p. 53–63.

[24]. Singh K, Maiti J, Dhalmahapatra K. Chain of events model for safety management: data analytics approach. Saf Sci. 2019;118:568–582. doi:10.1016/j.ssci.2019.05.044

[25]. Bevilacqua M, Ciarapica FE. Human factor risk management in the process industry: a case study. Reliab Eng Syst Saf. 2018;169:149–159. doi:10.1016/j.ress.2017.08.013

[26]. Verma A, Rajput D, Maiti J, editors. Prioritization of near-miss incidents using text mining and Bayesian network. In: Advances in Computing and Data Sciences: First International Conference, ICACDS 2016, Ghaziabad, India, November 11–12, 2016, Revised Selected Papers 1. Singapore: Springer; 2017. p. 183–191.

[27]. Sarkar S, Vinay S, Raj R, et al. Application of optimized machine learning techniques for prediction of occupational accidents. Comput Oper Res. 2019;106:210–224. doi:10.1016/j.cor.2018.02.021

[28]. Sarkar S, Lodhi V, Maiti J. Text-clustering based deep neural network for prediction of occupational accident risk: a case study. In: 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). Pattaya, Thailand; 15–17 November 2018. p. 1–6.

[29]. Sarkar S, Kumar A, Mohanpuria SK, et al. Application of Bayesian network model in explaining occupational accidents in a steel industry. In: 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). Kolkata, India; 03–05 November 2017, 1–6.

[30]. Verma A, Khan SD, Maiti J, et al. Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. Safety Sci. 2014;70:89–98. doi:10.1016/j.ssci.2014.05.007

[31]. Dhalmahapatra K, Shingade R, Maiti J. An innovative integrated modelling of safety data using multiple correspondence analysis and fuzzy discretization techniques. Safety Sci. 2020;130:104828. doi:10.1016/j.ssci.2020.104828

[32]. Rusu O, Halcu I, Grigoriu O, et al. Converting unstructured and semi-structured data into knowledge. In: 2013 11th RoEduNet International Conference; Sinaia, Romania; 17–19 January 2013. p. 1–4.

[33]. Ajayi A, Oyedele L, Akinade O, et al. Optimised big data analytics for health and safety hazards prediction in power infrastructure operations. Safety Sci. 2020;125:104656. doi:10.1016/j.ssci.2020.104656

[34]. Poh CQ, Ubeynarayana CU, Goh YM. Safety leading indicators for construction sites: a machine learning approach. Autom Constr. 2018;93:375–386. doi:10.1016/j.autcon.2018.03.022

[35]. Shirali GA, Noroozi MV, Malehi AS. Predicting the outcome of occupational accidents by CART and CHAID methods at a steel factory in Iran. J Public Health Res. 2018;7(2):74–80. doi:10.4081/jphr.2018.1361

[36]. Sarkar S, Vinay S, Maiti J. Text mining based safety risk assessment and prediction of occupational accidents in a steel plant. In: Computational Techniquesin Information and Communication Technologies (ICCTICT), 2016 International ConferenceNew Delhi, India; 11–13 March 2016. p. 1–6.

[37]. Bevilacqua M, Ciarapica FE, Giacchetta G. Data mining for occupational injury risk: a case study. Int J Reliab Qual Saf Eng. 2010;17(4):351–380. doi:10.1142/S021853931000386X

[38]. Sarkar S, Pateshwari V, Maiti J. Predictive model for incident occurrences in steel plant in India. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICC-CNT). Delhi, India; 03–05 July 2017. p. 1–5.

[39]. Sarkar S, Lohani A, Maiti J. Genetic algorithm-based association rule mining approach towards rule generation of occupational accidents. In: Mandal J, Dutta P, Mukhopadhyay S, editors. Computational Intelligence, Communications, and Business Analytics. CICBA 2017. Communications in Computer and Information Science, vol 776. Singapore: Springer; 2017. doi:10.1007/978-981-10-6430-2_40

[40]. Khosrowabadi N, Ghousi R, Makui A. Decision support approach on occupational safety using data mining. Int J Ind Eng Prod Res. 2019;30(2):149–164.

[41]. Pekel E, Akşchir ZD, Meto B, et al. A Bayesian network application in occupational health and safety. In: 2018 3rd International Conference on Computer Science and Engineering (UBMK). Sarajevo, Bosnia and Herzegovina; 20–23 September 2018. p. 1–5.

[42]. Muthusamy K, Gunasegaran HR, Natarajan E, et al. Analysis of potential project work accidents: a case study of a construction project in Malaysia. In: 2021 IEEE European Technology and Engineering Management Summit (E-TEMS). Dortmund, Germany; 18–20 March 2021. p. 1–6.

[43]. Lingard H, Hallowell M, Salas R, et al. Leading or lagging? Temporal analysis of safety indicators on a large infrastructure construction project. Safety Sci. 2017;91:206–220. doi:10.1016/j.ssci.2016.08.020

[44]. Sarkar S, Pramanik A, Maiti J, et al. Predicting and analyzing injury severity: a machine learning-based approach using class-imbalanced proactive and reactive data. Safety Sci. 2020;125:104616. doi:10.1016/j.ssci.2020.104616

[45]. Sarkar S, Gaine S, Deshmukh A, et al. A structural topic modelling-based machine learning approach for pattern extraction from accident data. In: Raju KS, Senkerik S, Lanka SP, Rajagopal V, editors. Data engineering and communication technology. Singapore: Springer; 2020. p. 555–564.

[46]. Sarkar S, Patel A, Madaan S, et al. Prediction of occupational accidents using decision tree approach. In: 2016 IEEE Annual India Conference (INDICON). Bangalore, India; 16–18 December 2016. 1–6.

[47]. Sarkar S, Raj R, Vinay S, et al. An optimization-based decision tree approach for predicting slip–trip–fall accidents at work. Safety Sci. 2019;118:57–69. doi:10.1016/j.ssci.2019.05.009

[48]. Polyvyanyy A, Pika A, Wynn MT, et al. A systematic approach for discovering causal dependencies between observations and incidents in the health and safety domain. Safety Sci. 2019;118:345–354. doi:10.1016/j.ssci.2019.04.045

[49]. Sarkar S, Baidya S, Maiti J. Application of rough set theory in accident analysis at work: a case study. In: 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). Kolkata, India; 03–05 November 2017. p. 1–6.

[50]. Oyedele A, Ajayi A, Oyedele LO, et al. Deep learning and boosted trees for injuries prediction in power infrastructure projects. Appl Soft Comput. 2021;110:107587. doi:10.1016/j.asoc.2021.107587

[51]. Ugur O, Arisoy AA, Ganiz MC, et al. Descriptive and prescriptive analysis of construction site incidents using decision tree classification and association rule mining. In: 2021 International Conference on Innovations in Intelligent Systems and Applications (INISTA). Kocaeli, Turkey; 25–27 August 2021. p. 1–6.

[52]. Dhalmahapatra K, Shingade R, Mahajan H, et al. Decision support system for safety improvement: an approach using multiple correspondence analysis, t-SNE algorithm and k-means clustering. Comput Ind Eng. 2019;128:277–289. doi:10.1016/j.cie.2018.12.044

[53]. Sarkar S, Pramanik A, Maiti J. An integrated approach using rough set theory, ANFIS, and Z-number in occupational risk prediction. Eng Appl Artif Intell. 2023;117:105515. doi:10.1016/j.engappai.2022.105515

[54]. Sarkar S, Vinay S, Djeddi C, et al. Text mining-based association rule mining for incident analysis: a case study of a steel plant in India. In: Djeddi C, Kessentini Y, Siddiqi I, Jmaiel M, editors. Mediterranean Conference on Pattern Recognition and Artificial, Intelligence. Hammamet, Tunisia; 20–22 December 2020, 257–273.

[55]. Pramanik A, Sarkar S, Sai Siddharth V, et al. Semi-automated ontology creation and upgradation for rail-road incidents: a case of a steel plant in India. In: João Manuel Tavares RS, Satyajit Chakrabarti, Abhishek Bhattacharya, et al., editors. Emerging technologies in data mining and information security. Singapore: Springer; 2021. p. 285–294.

[56]. Sarkar S, Ejaz N, Promod C, et al. Pattern extraction using proactive and reactive data: A case study of contractors' safety in a steel plant. In: Pradeep Kumar Singh, Bijaya Ketan Panigrahi, Nagender Kumar Suryadevara, et al., editors. Proceedings of ICETIT 2019 Emerging Trends in Information Technology. Cham, Switzerland: Springer. p. 731–742.

[57]. Sarkar S, Ejaz N, Maiti J. Application of hybrid clustering technique for pattern extraction of accident at work: a case study of a steel industry. In: 2018 4th International Conference on Recent Advances in Information Technology (RAIT). Dhanbad, India; 15–17 March 2018. p. 1–6.

[58]. Verma A, Maiti J. Text-document clustering-based cause and effect analysis methodology for steel plant incident data. Int J Inj Contr Saf Promot. 2018;25(4):416–426. doi:10.1080/17457300.2018.1456468 [PubMed: 29629618]

[59]. Guo S, Zhang P, Ding L. Time-statistical laws of workers' unsafe behavior in the construction industry: a case study. Physica A. 2019;515:419–429. doi:10.1016/j.physa.2018.09.091

[60]. Versteeg K, Bigelow P, Dale AM, et al. Utilizing construction safety leading and lagging indicators to measure project safety performance: a case study. Safety Sci. 2019;120:411–421. doi:10.1016/j.ssci.2019.06.035

[61]. Verma A, Maiti J, Boustras G. Analysis of categorical incident data and design for safety interventions using axiomatic design framework. Safety Sci. 2020;123:104557. doi:10.1016/j.ssci.2019.104557

[62]. Ghasemi F, Kalatpour O, Moghimbeigi A, et al. Selecting strategies to reduce high-risk unsafe work behaviors using the safety behavior sampling technique and Bayesian network analysis. J Res Health Sci. 2017;17(1):372.

[63]. Buddhakulsomsiri J, Pannakkong W, Nanthavanij S. Application of association rule algorithm to industrial safety data mining. Int J Ind Syst Eng. 2015;21(4):415–437. doi:10.1504/IJISE.2015.072728

[64]. Verma A, Dhalmahapatra K, Maiti J. Forecasting occupational safety performance and mining text-based association rules for incident occurrences. Safety Sci. 2023;159:106014. doi:10.1016/j.ssci.2022.106014

[65]. Dhalmahapatra K, Singh K, Jain Y, et al. Exploring causes of crane accidents from incident reports using decision tree. In: Satapathy Suresh Chandra, Amit Joshi, editors. Information and communication technology for intelligent systems. Singapore: Springer; 2019. p. 175–183.

[66]. Sarkar S, Vinay S, Pateshwari V, et al. Study of optimized SVM for incident prediction of a steel plant in India. In: 2016 IEEE Annual India Conference (INDICON), 16–18 December 2016, Bangalore, India; 2016.

[67]. Verma A, Maiti J, Gaikwad V. A preliminary analysis of incident investigation reports of an integrated steel plant: some reflection. Int J Inj Contr Saf Promot. 2018;25(2):180–194. doi:10.1080/17457300.2017.1416482 [PubMed: 29280419]

[68]. Bevilacqua M, Ciarapica F, Giacchetta G. Industrial and occupational ergonomics in the petrochemical process industry: a regression trees approach. Accid Anal Prev. 2008;40(4):1468–1479. doi:10.1016/j.aap.2008.03.012 [PubMed: 18606280]

[69]. Sarkar S, Khatedi N, Pramanik A, Maiti J. An ensemble learning-based undersampling technique for handling class-imbalance problem. In: Singh PK, Panigrahi BK, Suryadevara NK, et al., editors. Proceedings of ICETIT 2019. Cham: Springer; 2020. p. 586–-595.

[70]. Cordón I, García S, Fernández A, et al. Imbalance: oversampling algorithms for imbalanced classification in R. Knowl Based Syst. 2018;161:329–341. doi:10.1016/j.knosys.2018.07.035

[71]. Mishra N, Silakari S. Predictive analytics: a survey, trends, applications, oppurtunities & challenges. Int J Comput Sci Inf Technol. 2012;3(3):4434–4438.

[72]. Singh K, Maiti J. Mining frequent patterns with temporal effect: a case of accident path analysis. In: Singh PK, Panigrahi BK, Suryadevara NK, et al., editors. Proceedings of ICETIT 2019. Singapore: Springer; 2019. 596–603.

[73]. Collins LM, Schafer JL, Kam C-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods. 2001;6(4):330. doi:10.1037/1082-989X.6.4.330 [PubMed: 11778676]

[74]. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533:452–454. doi:10.1038/533452a [PubMed: 27225100]

[75]. Milch V, Laumann K. Interorganizational complexity and organizational accident risk: a literature review. Saf Sci. 2016;82:9–17. doi:10.1016/j.ssci.2015.08.010

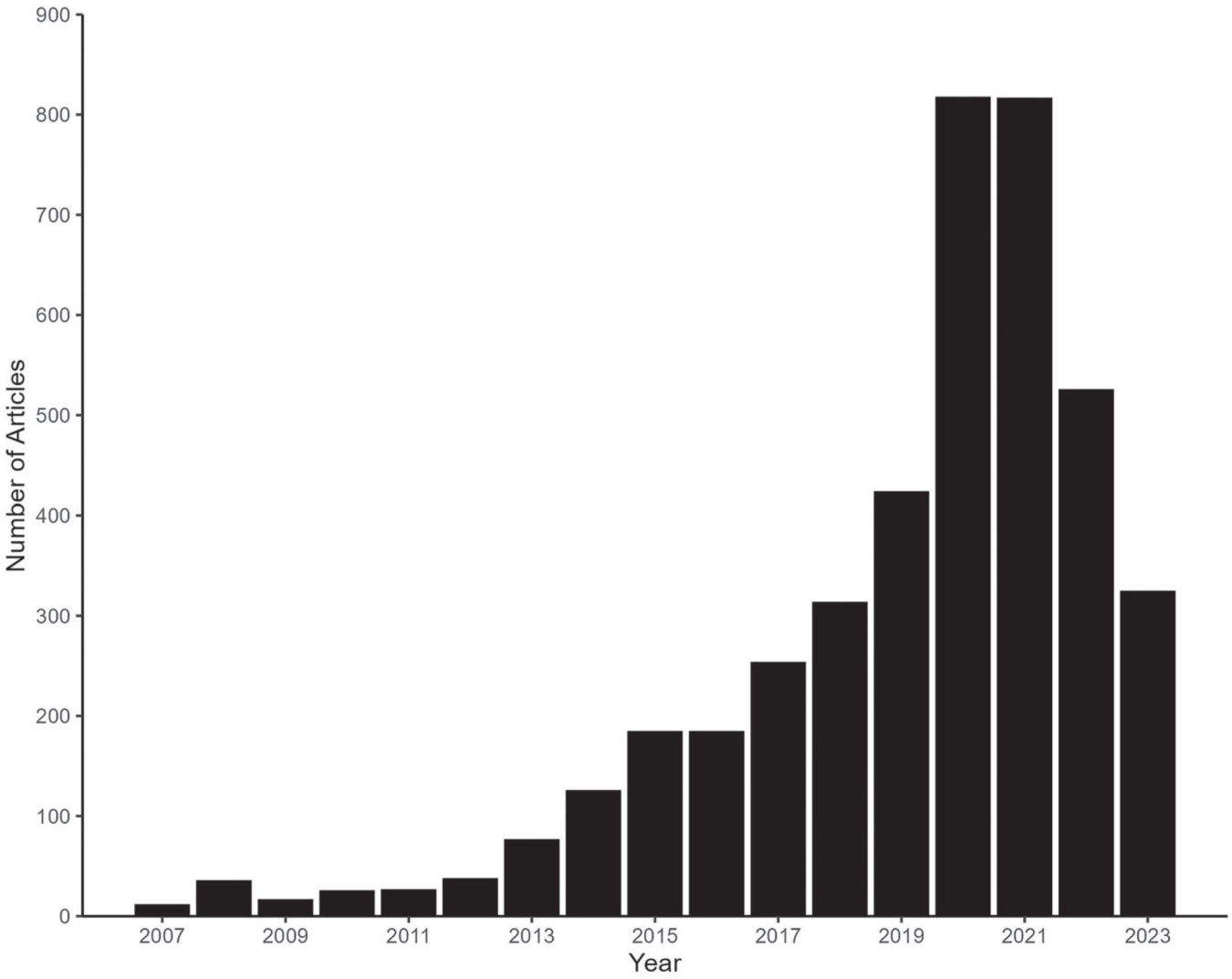

Figure 1.

Number of Web of Science search results from 2007 to 2022 for the search ‘safety analytics’ in the article title, abstract and key words.

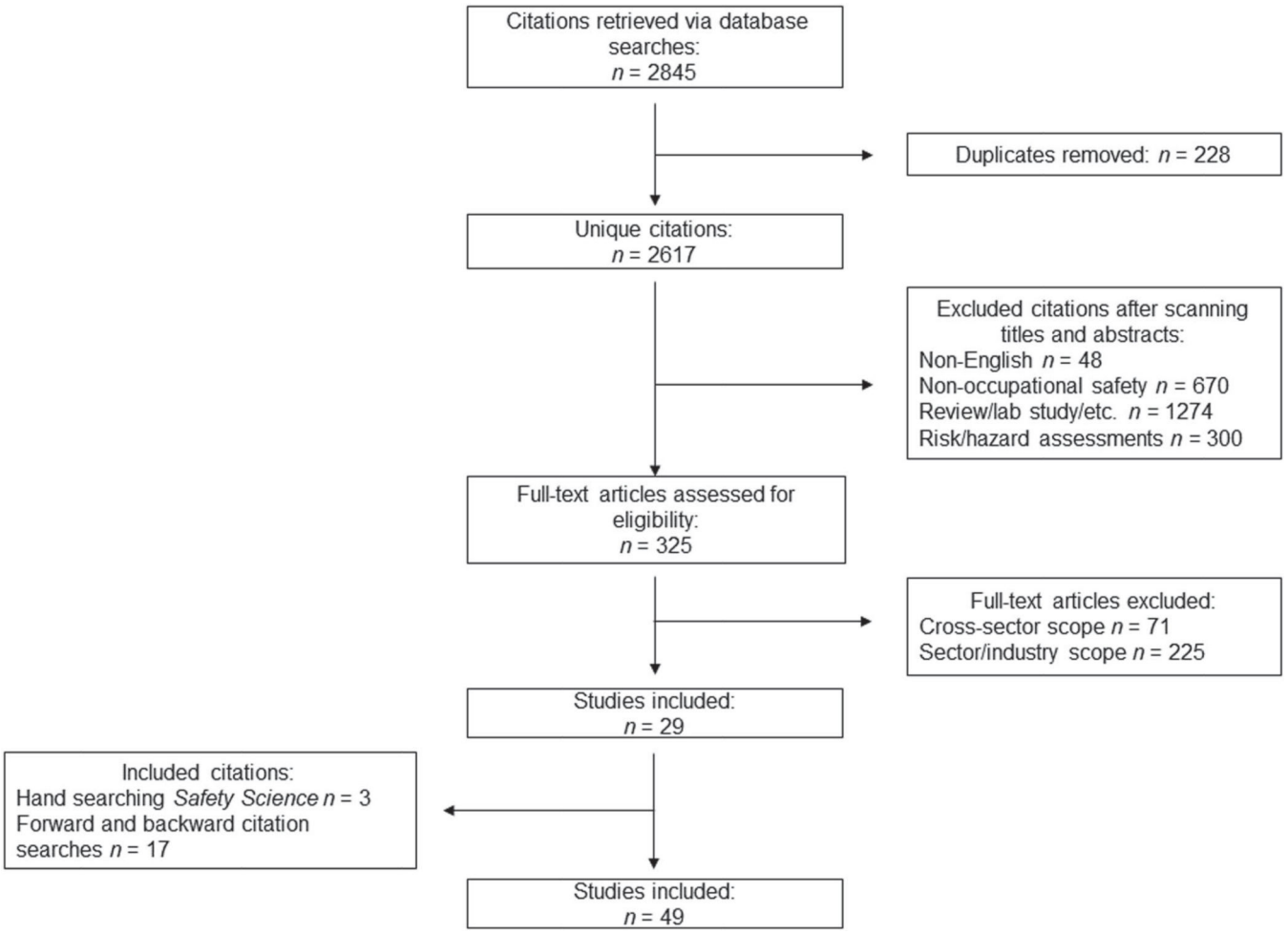

Figure 2.

Flow diagram for the scoping review study selection process. The *n* denotes the number of articles under consideration in each step of the process.

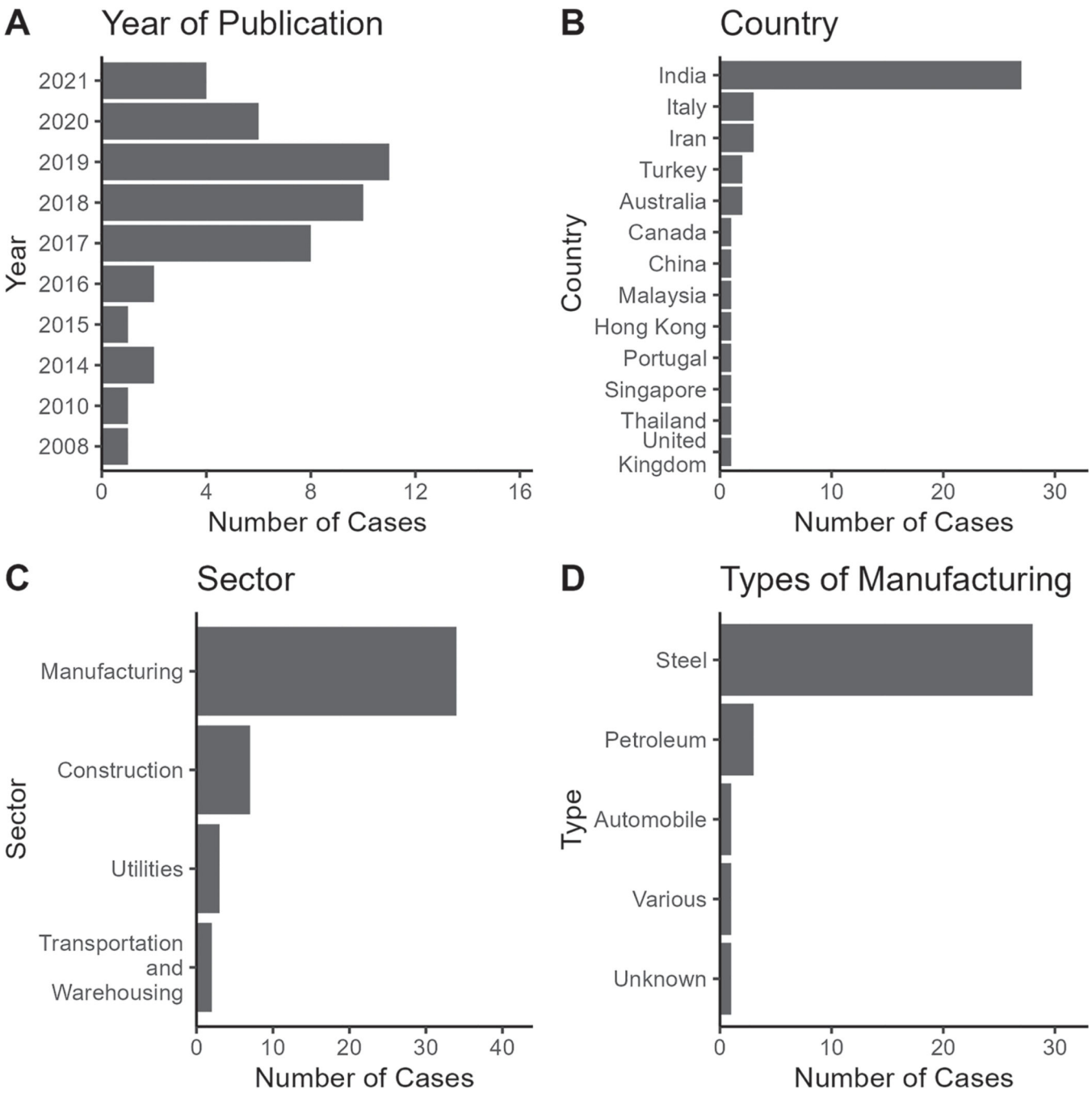

Figure 3.

Frequency counts: (A) year of publication; (B) country where the research was conducted; (C) sector in which the research was conducted; (D) industry in which the research was conducted if the sector was manufacturing.

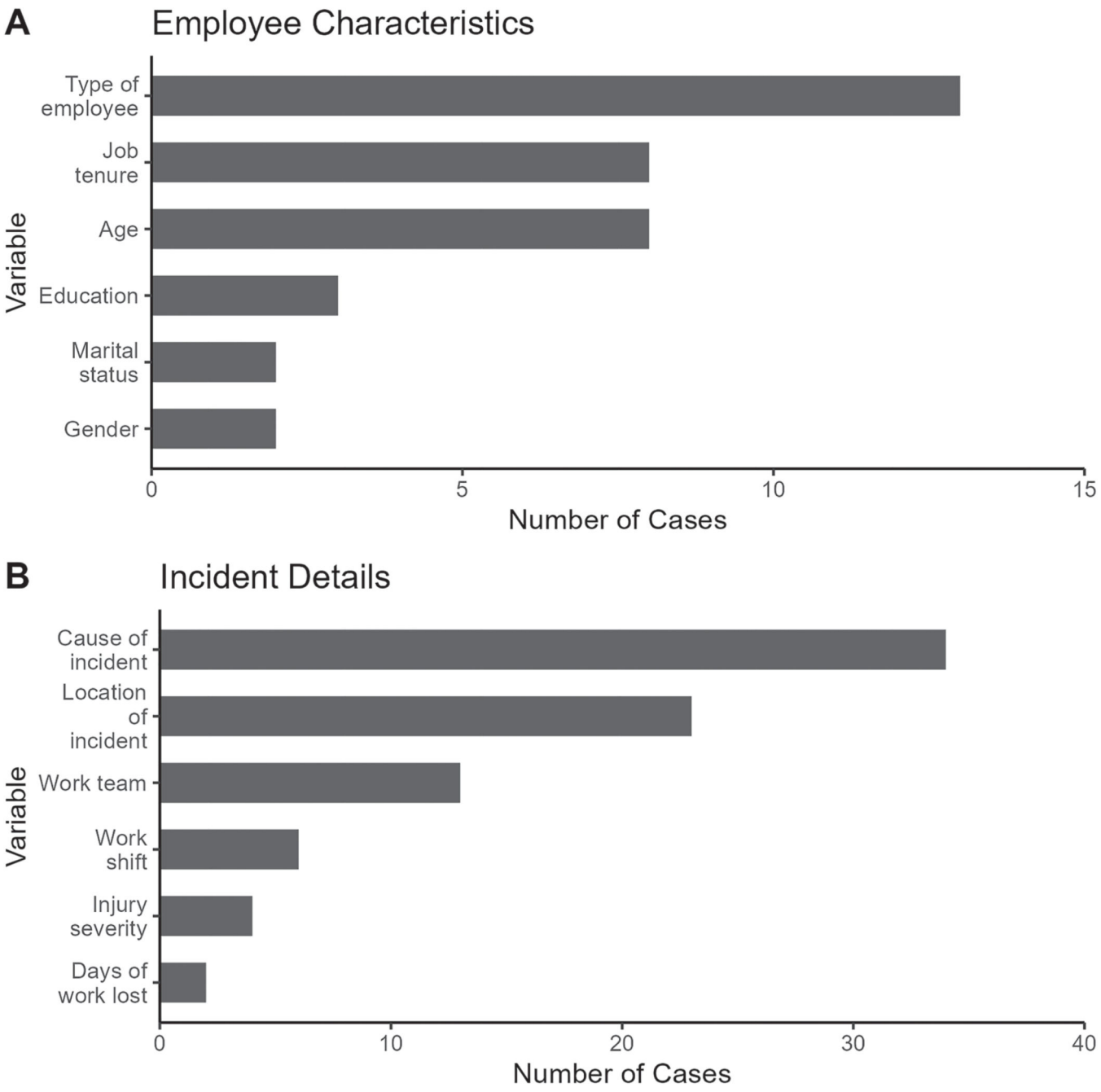

Figure 4.

Most frequently analyzed variables for (A) employee characteristics and (B) incident details in the included studies.

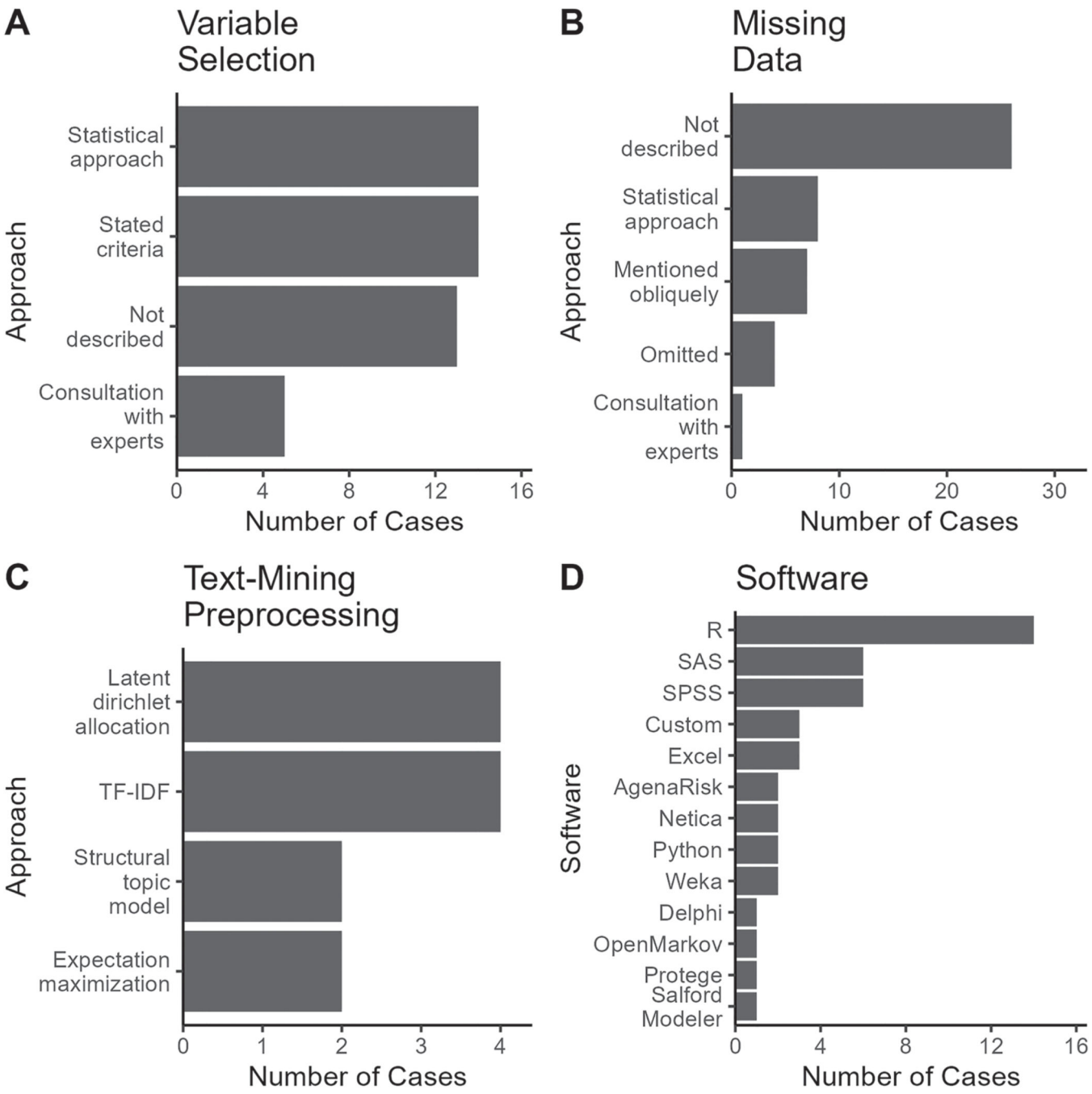

Figure 5.

Summary data for different aspects of data preprocessing: (A) methods employed for variable selection; (B) missing data; (C) text-mining preprocessing; (D) software used for the included studies. Note: TF-IDF = term frequency-inverse document frequency.

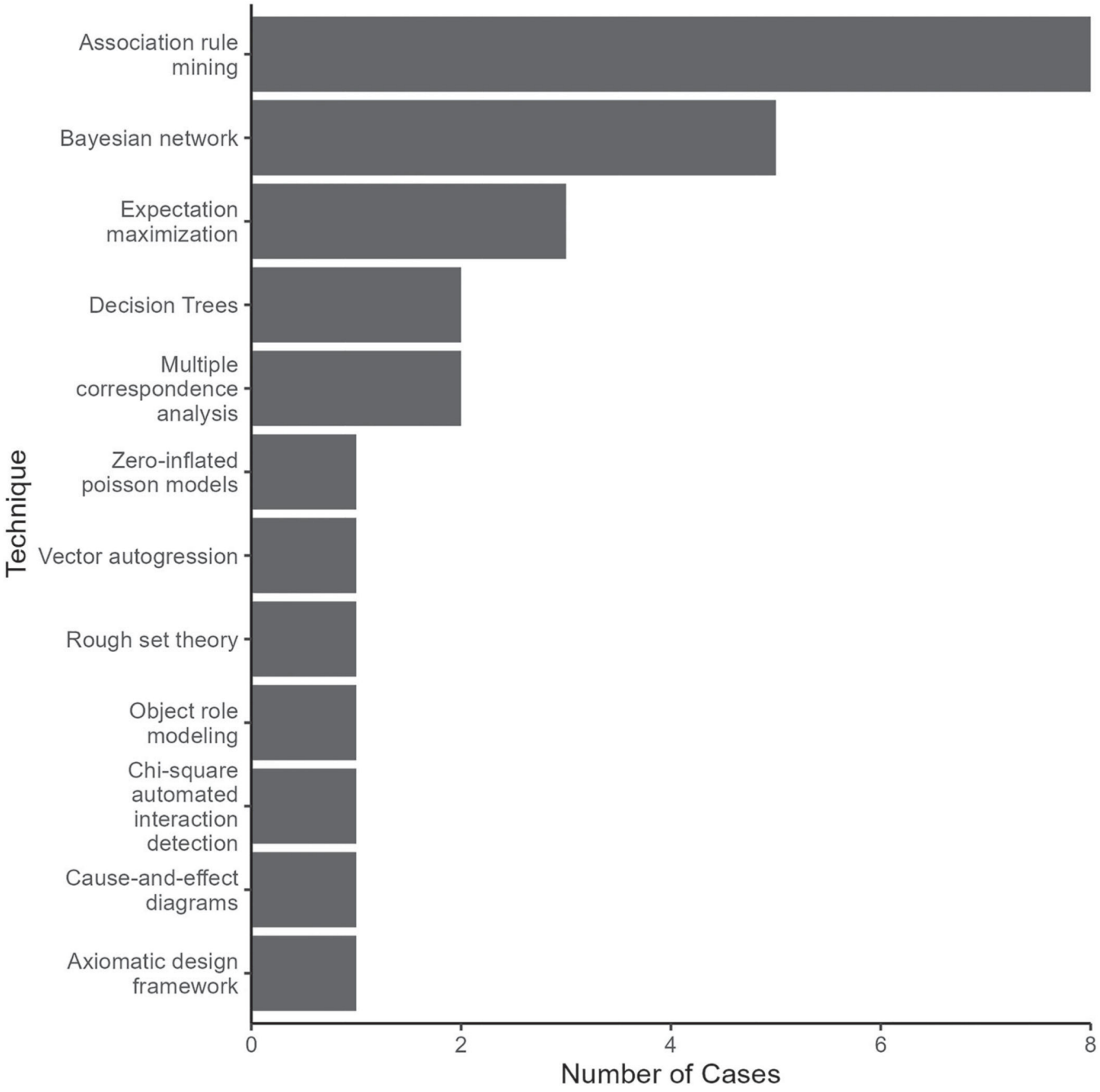

Figure 6.

Techniques employed by studies that used only one technique (in contrast to the studies that compared two or more techniques).

Table 1.

Inclusion and exclusion criteria for studies considered for the scoping review.

Inclusion criteria	Exclusion criteria
Reports, observational studies, experimental studies, case studies	Reviews, meta-analyses,
Available in English	laboratory studies, commentaries
Concerned with occupational safety	Unpublished documents
Involve the use of analytics	Non-English language
Published between 2007 and 2021 Conducted within an establishment or enterprise	Published before 2007
Dependent variable is injuries, fatalities or near misses	Conducted across multiple establishments or enterprises, industries or sectors
	Focused on risk or hazard assessment
	Only self-reported data analyzed

Table 2.

Studies included in the scoping review that compared the performance of different analytics techniques listed by citation and the techniques that were compared.

Citation	Techniques compared
[33]	CART, **GBM**, RF
[68]	**CART,** CHAID, eCHAID, QUEST
[37]	ANN, CART, CHAID, eCHAID, negative binomial regression, **neuro-fuzzy systems**, QUEST
[50]	**Deep neural network,** gradient-boosted machines, extreme gradient boosting, SVM, KNN
[34]	DT, **RF**, logistic regression, KNN, SVM
[55]	**RF**, SVM
[69]	RF, **SVM**
[57]	*k*-Means clustering, **SOM-based *k*-means clustering**, SOM-based hierarchical clustering
[56]	C5.0, **CHAID**
[45]	**RF**, SVM, KNN
[28]	DNN, **expectation-maximization-based DNN**, SVM, RF
[46]	CART-tuning algorithms: **genetic algorithm**, grid-based, pruned grid-based
[38]	SVM, **RF**, **ME**
[53]	ANN, CART, KNN, NB, **RF**, SVM
[47]	CART, C5.0, **RF**
[66]	Parameter optimization of SVM: **grid search**, genetic algorithm, BAT algorithm
[27]	ANN, **SVM**
[35]	**CART**, CHAID
[51]	**DT, ARM**
[53]	**ANFIS**, SVM, ANN, KNN, NB, RF

Note: Techniques in bold were the techniques deemed the best for the study data. ANFIS = adaptive neuro-fuzzy inference system; ANN = artificial neural network; ARM = Association-rule mining; BAT = metaheuristic bio-inspired algorithm; CART = classification and regression tree; CHAID = χ^2 automatic interaction detector; DT = decision tree; eCHAID = exhaustive χ^2 interaction detector; GBM = gradient boosting machine; KNN = *k*-nearest neighbors; ME = Maximum entropy; NB = naïve Bayes; QUEST = quick, unbiased and efficient statistical tree; RF = random forest; SOM = Self-organizing map; SVM = support vector machine.