HHS Public Access

Author manuscript

Int J Ind Ergon. Author manuscript; available in PMC 2024 March 01.

Published in final edited form as:

Int J Ind Ergon. 2023 March ; 94: . doi:10.1016/j.ergon.2023.103428.

Establishment-level occupational safety analytics: Challenges and opportunities

Anne M. Foreman[a,*], **Jonathan E. Friedel**[a,b], **Timothy D. Ludwig**[c], **Maira E. Ezerins**[d], **Yalçin Açikgöz**[c], **Shawn M. Bergman**[c], **Oliver Wirth**[a]

[a]National Institute for Occupational Safety and Health, USA

[b]Georgia Southern University, USA

[c]Appalachian State University, USA

[d]University of Arkansas, USA

Abstract

In occupational safety and health, big data and analytics show promise for the prediction and prevention of workplace injuries. Advances in computing power and analytical methods have allowed companies to reveal insights from the "big" data that previously would have gone undetected. Despite the promise, occupational safety has lagged behind other industries, such as supply chain management and healthcare, in terms of exploiting the potential of analytics and much of the data collected by organizations goes unanalyzed. The purpose of the present paper is to argue for the broader application of establishment-level safety analytics. This is accomplished by defining the terms, describing previous research, outlining the necessary components required, and describing knowledge gaps and future directions. The knowledge gaps and future directions for research in establishment-level analytics are categorized into readiness for analytics, analytics methods, technology integration, data culture, and impact of analytics.

Keywords

Safety analytics; Big data; Occupational safety; Injuries; Near misses; Data culture

Businesses worldwide are collecting and analyzing ever-increasing amounts of data (Marr, 2018). Vast amounts of customer data, health data, production and process data, and data from other various organizational functions are being collected and analyzed to better understand consumer preferences (Bradlow et al., 2017), medical care (Wang et al., 2018), and supply chain management (Waller and Fawcett, 2013), among others. Advances in computing power and analytical methods have allowed companies to reveal insights from

*Corresponding author. vpc3@cdc.gov (A.M. Foreman).

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, Centers for Disease Control and Prevention.

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.

the “big” data that previously would have gone undetected. For example, Monsanto—an agricultural biotechnology company—analyzed 150 billion soil observations and 10 trillion weather-simulation points to improve crop yields and reduce losses. The insights gained direct “smart” machinery to plant seeds at specific locations and depths to maximize yields (Schumpeter, 2014).

In occupational safety and health (OSH), big data and analytics show promise for the prediction and prevention of workplace injuries. For example, a gold mining company (Goldcorp) hired a professional services company (Deloitte) to examine a five-year period of data that included 2,000 safety incidents and 1.8 million days worked as well as demographics, production data, operations data, and the weather (Stewart, 2013). Analytics techniques identified relations between compensation and injuries, injury rates and age, injury rates and job roles, among others.

Despite the fact that businesses are collecting more data than ever before, it is estimated that 60%–73% of all data collected remains unanalyzed (Gualtieri, 2016), data that is often termed “dark data” (Schembera and Durán, 2020). In OSH, large scale data collection and adoption of sophisticated data analytic approaches has lagged even further behind. Numerous barriers and challenges have hindered the implementation of large-scale analytics in OSH, including a lack of knowledge and experience of OSH professionals in data science, employee privacy concerns, absence of centralized databases or IT expertise, and a lack of knowledge about the potential benefits of their use (Wagner, 2014). The purpose of the present paper is to illuminate these challenges and propose several promising steps towards realizing the full potential of safety analytics.

1. What is safety analytics?

The term “analytics” is ubiquitous in contemporary business settings, and yet the definitions are numerous and inconsistent (Gandomi and Haider, 2015; Van Barneveld, Arnold and Campbell, 2012). Although analytics is often used as a synonym for applications of statistics, data science, or any other type of quantitative approach that guides decision making (Rose, 2016), we prefer the definition of analytics as “the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data” (Cooper, 2012, p. 3). The statistical techniques in analytics can vary from simple linear regression techniques to complex unsupervised algorithms depending on the nature of the data and the problem being addressed. As described in the definition, for these statistical models to be considered analytics they must produce information that allows people the ability to solve or abate the problem at hand. In 2014, a NIOSH Science Blog post drew attention to the potential application of analytics techniques in OSH (i.e., safety analytics) where they could be used to predict and prevent injuries and illness among workers (Wagner, 2014).

Much like analytics, the term “big data” has become a catch-all buzzword. A popular way to conceptualize big data and distinguish it from features of data used in traditional statistical and quantitative methods is in terms of five features: *volume*, *velocity*, *variety, value, and veracity* (Ramadan, 2017). Volume simply refers to the amount of data, and big data volume

historically refers to an amount of data that surpasses the capacity of typical database systems (Dumbill, 2013). However, as the capacity of databases grow with distributed databases and other technological advances (e.g., Hadoop; White, 2015), the volume feature of big data becomes less limiting. The volume of data that is collected, however, allows for the application of statistical models and analysis that cannot reliability be used on the amount of information collected used in traditional statistical analyses. In OSH, volume is the amount of data that can be incorporated into analytics across the organization.

Velocity refers generally to the speed at which the data are generated. For example, many industrial machines now include sensors that can provide second-to-second measurements of stressors (e.g., temperature, force) that can lead to breakdowns. In OSH, injury reports represent a relatively slow velocity data generation because the data tend to be gathered infrequently and intermittently. On the other hand, daily behavior-based observations of hazards and risks (Agraz-Boeneker et al., 2007) and hazard sensor data (Pavón et al., 2018) can generate dozens to thousands of data points every day. This continual collection of data makes it possible to conduct “real time” analyses and update retrospective and prediction models quicker than what was possible with traditional data collection methods.

Variety refers to the diversity of information types, forms, and data structures. Variety implies not only numeric data, but also video, audio, still-images, direct sensor data, and text. An important aspect of data variety is whether the data are structured or unstructured. Structured data are sometimes perceived as more desirable because they are in a format that can be readily analyzed (Baars and Kemper, 2008), such as a database of patient information in a hospital. Unstructured data are not recorded in a format that can be readily analyzed (Katal et al., 2013), such as hand written notes from a doctor in that same hospital, audio files, or images. Although analysis of unstructured data can be more labor intensive and require more expertise, converting unstructured data into structured data can be advantageous because it may provide more detailed and actionable information (Inmon and Nesavich, 2007). In OSH, examples of unstructured data include text from behavioral observation checklists, narrative summaries on incident reports, or change data from equipment sensors, which are usually stored in data lake. Other systems may or may not be structured depending on the system that is used to record the data. For example, safety checklists that are entered via a tablet or computer and automatically uploaded to a database or spreadsheet in which the information is structured into appropriate rows and columns might be amenable to analysis. This data variety means that traditional data storage and management techniques need to be expanded to collect, process, and analyze a more robust sampling of events happening in OSH.

Value and veracity relate to all types of data, not just big data. Value refers to how useful or beneficial the data are. In analytics, the value of data is in terms of providing information that leads to actional solutions or insights. Some data may not be worth the effort required to collect them. Additionally, educated guesses and inferences about the value of data can be made, but it may only be possible to tell if data are valuable after analysis has begun. Veracity refers to how accurate or trustworthy the data are. Veracity has also been defined as the uncertainty of the data (Bellazzi, 2014). Veracity of the data may depend on the source. For example, weather data automatically collected from National Oceanic and Atmospheric

Administration sources may have more veracity than a worker collecting observational data on a shop floor. Veracity is a critically important in big data analytics when the volume of data can easily encompass hundreds of thousands or millions of data points.

It is important to note that the terms “big data” and “analytics” are often used interchangeably, but they are distinct concepts. Analytics can be used with any data set whether or not it can be considered “big data.” Many OSH studies apply analytics methods with data sets that lack volume, velocity, or variety, but the analysis may still reveal valuable insights. For example, a study analyzing 309 case reports of fatal occupational injuries—a small number in terms of big data—over a five year period in the construction industry in Taiwan used association rule mining (a type of analytics) to identify factors contributing to injuries (e. g., weather conditions, worker tenure, etc.; Liao and Perng, 2008). Other examples of analytics that have been used in occupational safety research with traditional data, sometimes affectionally called “small data” (Guilfoyle et al., 2016), include applications of classification and regression trees (Cheng et al., 2012), structural equation modeling (Manzoor et al., 2018), and k-means clustering (Raviv, Fishbain, & Shapira, 2017). Regardless of the size of data sets, analytics can still provide valuable insights to organizations.

Although these studies show the power of using small data with analytics, unlocking the full potential of safety analytics requires that information be gathered from a variety of sources throughout the organization. Advancements in technology, automated processes, and shared databases now make it more possible than ever to bring together different types of information from all organizational areas in order capture a more representative picture of the organization and how functions across the enterprise relate to OSH outcomes. Combining traditional sources of safety data (injury and near miss reports, safety audits, inspection checklists, and behavioral observations) with other sources of organizational data, such as human resources, operational, customer and sales information, and financial data represents a unique opportunity to realize the full potential of safety analytics and enhance the safety of the workplace.

Regardless of the type of data being used in analytics, the properties and needs of the data are the same as those of traditional statistical methods. There is a temptation to assume that just because data is “big”, it is valid and reliable. This assumption is inaccurate as the issues with “dirty” data (i.e., data that is inaccurate, incomplete or inconsistent) in analytics is a continued concern (Gartner, 2017) and is captured in one of the five features of big data, veracity (Ramadan, 2017). Given the speed at which analytics can deliver information and, in some cases automate decision-making, the accuracy, representativeness, and reliability of the data, big and small, flowing into analytic algorithms is more important than ever. In other words, analytic solutions that can reveal undiscovered patterns in organizational data are only as good as the quality of the information being captured.

Analytic strategies are often classified into four categories: descriptive, diagnostic, predictive, and prescriptive (Institute for Operations Research and the Management Sciences, 2014). Descriptive analytics are often used in the initial stages of data analysis and consist of visualizing historical data through the creation of scatter plots, histograms, or

other types of graphs as well as calculating descriptive statistics for the variables of interest (e.g., mean, mode, variance, etc.). In some areas, the term exploratory analysis can be used to refer to descriptive analysis. Diagnostic analytics are used to assess the relationship between predictor and outcome variables to explore why something happened. Diagnostic analytics can often involve traditional statistical techniques such as correlations, linear and logistic regressions, and analysis of variance but can also include other non-traditional techniques such as decision-trees, artificial neural networks, and support vector machines to examine the relationships between predictor variables and safety outcomes. Descriptive and diagnostic analytics are both retrospective and are conducted on predictor and outcome data that has already been collected.

Predictive and prescriptive analytics, on the other hand, focus on forecasting the likelihood of an outcome based on continuously updated predictor data which has been collected. More specifically, predictive analytics consists of using statistical approaches to analyze historical data to evaluate what could happen in the future. This occurs when new predictor data is collected, but the outcome data has not yet been captured. The patterns found in diagnostic models are applied to the new predictor data to forecast the likelihood of an incident. Diagnostic models used in predictive analytics are thoroughly examined using a variety of cross-validation techniques to look for overfitting or selection bias and estimate how the pattern will generalize to future data. Once the outcome measure is captured, the models are reexamined, and the forecasts are updated. Prescriptive analytics involves using analytic strategies to forecast the outcomes of different interventions, courses of action, or scenarios in a proactive effort to change the likelihood of an event that is predicted to occur. Prescriptive analytics automates the predictive modeling and proactively provides suggestions for a course of action that will improve a desired outcome. It should be noted that these categories provide general descriptions of approaches in analytics, but the distinctions between them are sometimes blurry. Additionally, applications of analytics may involve more than one type of analytics.

2. The current scope of analytics in occupational safety

Advances in computing power, combined with growing awareness of the potential in predictive and prescriptive analytics to improve safety, have led to greater interest and investment toward analytics in OSH (Wagner, 2014). The scope of analytics can be as narrow as within one establishment or as broad as across industry sectors. Because applications of analytics in OSH are still in its early stages, the existing examples in the safety literature are quite varied in scope and methods, which usually are dictated by the question(s) of interest. We will describe several examples of safety analytics in the literature that vary in terms of scope and methodology. Ultimately, we will argue that while cross-sector and industry-level analyses can provide insights which improve OSH, the use of safety analytics at the enterprise- or establishment-level is an under-utilized approach that holds a great deal of promise.

It is helpful to categorize the studies described in this section according to the North American Industry Classification System (NAICS) hierarchy. The hierarchy consists of the sector, subsector, industry, enterprise, and establishment. To illustrate the different levels of

the hierarchy, consider a chain of five grocery stores. The sector would be Retail Trade, the subsector would be Food and Beverage Stores, and the Industry would be Grocery Stores. The company that owns all the grocery stores would be considered an enterprise and each individual grocery store in the chain would be considered an establishment. There are advantages and disadvantages of expanding or limiting the scope of analysis across or within industries, including trade-offs between generalizability and specificity (discussed below).

At the cross-sector and sector level of analysis, several studies have used analytic approaches to analyze precursors to injuries (Ghodrati et al., 2018; Goh and Chua, 2013; Matías et al., 2008). When the scope of analysis is very broad, at either the cross-sector or sector level, the analyst may have a large amount of data (large volume) to examine across a limited number of variables. There may be a low variety in the sources data because studies conducted with data collected across a sector are often using data collected from national databases that may only require establishments to include a handful of details when reporting a work-place incident. For example, a study concerning occupational injuries in the construction sector was conducted with data collected by the Taiwanese government (Cheng et al., 2012). The variables that were investigated were accident type (e.g., electric shock, traffic accident), project type, company size, project contract amount, source of injury, worker gender, worker age, unsafe condition (e.g., PPE not provided), unsafe acts (e.g., PPE not used), and safety management systems (e.g., safety training was offered, self-inspections were conducted, etc.). The authors used a classification and regression tree to identify the factors associated with different occupational incidents. One example finding from the study was that falls were more likely in construction projects by small businesses (fewer than 10 people) that cost less than five million Taiwanese dollars (approximately 166,000 US dollars). A limitation of these sorts of analyses is a lack of data representing internal organizational operations making it more difficult to identify specific actionable interventions, the goal of safety analytics.

Other sector or cross-sector level analytics projects include studies of worker compensation injury narratives and other large databases of injury-related narratives (Bertke et al., 2016; Kakhki et al., 2020; Marucci-Wellman et al., 2015; Vallmuur et al., 2016). Typical analyses of these sorts of databases are time consuming because the data are unstructured and thus often must be evaluated by human raters. Text mining is an analytic approach that is used to identify patterns and trends in textual data with machine learning algorithms and natural language processing (Hotho et al., 2005). In an example study, two naïve Bayes algorithms were used to classify 15,000 workers' compensation narratives into Bureau of Labor Statistics two-digit event codes (Marucci-Wellman et al., 2015). The algorithms were 87% accurate, and the accuracy was high across all of the event codes, including those that were relatively uncommon. After classifying the narratives with the algorithms, only 32% of the cases required additional manual coding. Although there is promise in text-mining approaches to analyze injury narratives, sector and cross-sector data sets are usually limited by minimal governmental reporting requirements that often include only immediate events and conditions surrounding an injury. Despite the hard-wired limitations in the types of variables collected, actionable insights have been obtained from such databases (Meyers et al., 2018).

Industry-level analyses also have limitations related to the variety of data available similar to sector and cross-sector levels of analyses. Researchers analyzing data at the industry level often use national compulsory industry reporting databases (e.g., OSHA, national or state workers' compensation databases). For example, a study analyzing injuries in the wood manufacturing industry in a region of Italy relied on the Italian National Workers' Compensation Authority's database, and thus the data were limited to the specifics of the incident and workers' ages. One industry with a robust voluntary incident reporting system is aviation which instituted the Aviation Safety Reporting System in 1975 (Billings et al., 1976). Although there have been calls for industry-level incident reporting in the railroad industry (Federal Railroad Administration, 2003) and voluntary *patient* safety reporting systems are prevalent in the healthcare industry (Herzer et al., 2012), there are known barriers to the implementation of voluntary worker incident reporting systems, including confidentiality and effort burden (Gifford and Anderson, 2010). The lack of robust, standardized reporting among industries is reflected in the dearth of industry-level analytic approaches in the OSH literature.

Enterprise- and establishment-level analytics offer several potential advantages over analyses with a broader scope and obviate some of the previously discussed limitations of not having data from multiple sources within an organization. In contrast to databases of injuries collected by state and national governments like those described above, enterprises and establishments can leverage data from more sources that capture the larger context in which incidents and near misses occur. For example, Lingard et al. (2017) report an analytics project that used data from a large infrastructure construction program within a single enterprise that incorporated data from the enterprise's safety program (e.g., toolbox meetings, prestart meetings, safety observations, hazards reported, etc.) as well as the frequency of injuries. By incorporating numerous leading indicators, the researchers were able to identify a cyclical pattern between leading and lagging indicators in which injuries were often followed by preventative measures such as an increase in toolbox meetings, which then decreased in frequency as the time from the injury increased. In another enterprise-level analysis conducted on data collected by a construction contracting company in Singapore, some variables analyzed were related to characteristics of the construction projects (e.g., project type, ownership, cost, etc.) and other variables were related to internal safety inspections (e.g., falling hazards, scaffold safety, etc.; Poh et al., 2018). With these data, the authors constructed five different types of machine learning models to predict the occurrence of no accidents, minor accidents, and major accidents in historical data. The best-fitting model in their analysis, a type of classification model called a random forest, achieved 78% accuracy in predicting accidents. The findings from this study are demonstrative of the more nuanced relations that can potentially be revealed at the enterprise and establishment level.

Analytics conducted at enterprise and establishment levels may hold the most promise for predicting and preventing workplace injuries given it can incorporate local information from across the entity. As described above, exploiting the data collected by state and national governing bodies provides incomplete insight because of the limited amount of organizational information collected (although there are some research questions that will always be best addressed at the sector or cross-sector level, such as extremely rare events).

In contrast, conducting analytics within an enterprise or establishment greatly increases the types and sources of data available. Table 1 shows variables that could be collected categorized by topic area. The table is not intended to be an exhaustive list. The examples of variables within the table are not industry specific, and some of the relevant variables will differ across industries (for example, wind speed data may be highly relevant on an oil rig but may be irrelevant in furniture manufacturing). The divisions in which different variables are housed may also vary across industries and organizations. The variables differ in terms of their type, either numeric or alphanumeric, and may drastically differ in terms of their structure. For example, if machine inspections are done by hand with pencil and paper, then it may take several steps (e.g., transcribing, formatting, and entering into a database) before they become structured data. With the availability and decreasing cost of tablet computers, the step of transcribing information to a computer is likely becoming less common and is likely a contributing factor to the growing interest in big data and analytics in OSH. The data amenable to analysis within enterprises and establishments is growing in amount and variety along with technological advancements (e.g., tablet computers, wearable sensors), making the current time ripe for more sophisticated approaches to addressing OSH concerns.

The safety metrics listed in Table 1 are categorized as leading or lagging safety indicators and are particularly important in predicting and preventing injuries with analytics. Leading indicators are measures that can potentially help prevent safety incidents and include aspects of safety management systems, such as observations, audits, and toolbox talks, whereas lagging indicators are more traditional safety metrics that measure outcomes, such as injury frequency, injury severity, and lost work days (Lingard et al., 2017; Wurzelbacher and Jin, 2011). Although the empirical evidence for a distinction between the two measures may be lacking (Kongsvik et al., 2011; Lingard et al., 2017), the distinction can serve as a useful heuristic for assessing safety and reducing injury within establishments. Safety climate is a leading indicator that measures the attitudes and perceptions of employees concerning safety in their workplaces (Zohar, 1980) and is typically measured via self-reported questionnaires (Shannon and Norman, 2009).

3. Readiness for establishment-level analytics

Although analytics offers promise as an approach to reducing injuries in occupational settings, many establishments may not be best positioned to conduct them. There are several prerequisites for conducting analytics that produce useful, actionable results. Although there are no published recommendations for assessing analytics readiness for occupational safety, guidance can be gleaned from readiness assessments developed for other applications. For example, to help academic institutions assess their readiness for learning analytics in education, Arnold et al. (2014) described four factors that are integral to analytics readiness: *ability,* having the right people to establish the appropriate IT systems and conduct analytics, *data*, having appropriate, reliable, and accessible sources of data, *culture and process*, having an organizational culture that engages in data-driven decision making, and *governance and infrastructure,* having leadership invested in the success of analytics by allocating resources and personnel to the process. By using these factors as a model, a readiness framework can be developed for safety analytics. Although the Arnold et al. criteria identify and organize several general programmatic and organizational elements needed for conducting analytics

in educational settings, we made a few modifications to make them more applicable for conducting analytics in occupational safety. For example, we created a separate data culture category with features that are better aligned with current conceptualizations of safety culture.

Table 2 lists the key readiness factors modified from Arnold et al. (2014) to establish a framework for analytics readiness in OSH: expertise, IT infrastructure, data, and data culture. Also listed in the table are potential organizational barriers or concerns related to the implementation of each factor and examples of personnel, programmatic, or organizational remedies to address these concerns. The relationships among the key readiness factors are shown in Fig. 1.

Expertise refers to having the skills necessary to conduct all phases of analytics, and a lack of analytics or data science expertise may be an organizational concern. The specifics of analytic phases related to occupational safety data have been well-described in recent literature, from collecting and storing the data to interpreting and applying analytic models (Huang et al., 2018). Executing these processes may require skills in database administration, configuring computer software, and data encryption in addition to expertise in descriptive, diagnostic, predictive, and prescriptive analytics. Successfully deploying an analytics approach to occupational safety data may necessitate collaboration among experts within an establishment, including personnel from the information technology department, human resources department, and safety and health department. If employees lack the required skills, a vendor or contractor could be outsourced to provide the services, appropriate training could be provided to existing employees, or new employees could be hired. In an analysis of analytics job listings, two types of jobs were identified: technology-enabler professionals and business-impacting professionals (De Mauro, Greco, Grimaldi, Ritala and Management, 2018). Technology-enabler professionals are concerned with developing the infrastructure of big data analytics, such as managing server platforms or developing dashboards. Business-impacting professionals are concerned with turning the data into actionable steps for the organization through data analysis and project management. Organizations may need to hire or train existing employees or partner with external experts to fill one or both types of roles.

IT Infrastructure refers to possessing both the hardware and software components necessary for analytics. Analytics often requires a substantial investment in hardware and software infrastructure, and therefore the leaders of an establishment must be willing to divert or prioritize funds to invest in acquiring the necessary tools in order to conduct analytics. This is particularly important for predictive or prescriptive analytics which require predictor data to continually flow, ideally in-real time, into the analytic algorithms and allow for near-time predictions and recommendations. Within an establishment, data are often siloed across divisions or departments (Ransbotham and Kiron, 2017), and combining data across divisions may require creating a data lake (Miloslavskaya and Tolstoy, 2016) or relational data warehouse (Watson, 2014). The details of the approaches to data integration are beyond the scope of the present paper, but numerous books and papers detailing the processes are available in the literature (e.g., Loshin, 2013; Oussous et al., 2015; Sherman, 2014; Zikopoulos and Eaton, 2011). The IT Infrastructure factor overlaps with the Expertise

factor because an establishment needs to have the appropriate personnel to develop, build, and execute the IT infrastructure to conduct analytics and effectively deploy the results to end-users.

Data is an obvious factor required for analytics readiness, but having data is not the only prerequisite, as old adages about the quality of data entering and exiting a data analysis process attest (e.g, “garbage in, garbage out”), and lack of quality data is a potential concern when instituting analytics. Including data quality in an assessment of readiness serves as a prompt for an establishment to reflect on the types, amounts, and accuracy of data that are collected across departments or divisions. Establishments that are best positioned for analytics likely have robust behavioral safety programs that collect large amounts of data on a daily basis, including worker safety observations, hazard analyses, and safety audits (McSween, 2003). As shown in Table 1, these data can be analyzed along with data from other divisions to extract new insights in contrast to analyzing safety data in isolation, and measures of data quality can be undertaken to ensure that the data meet appropriate standards for analytics.

There are many different dimensions of data quality described in the literature (Pipino et al., 2002) but five were identified as particularly relevant for healthcare analytics and are also relevant for OSH analytics. The five dimensions are completeness, correctness, plausibility, currency, and concordance. Completeness is whether the appropriate data exist and there is a lack of missing data within and across forms. Correctness is that data are measuring phenomena that they purport to be measuring, and the data do not contain errors. Plausibility is concerned with the “believability” of the data, as in the data are inaccurate or contain unrealistic values. Currency is concerned with the availability and timeliness of the data, and concordance is that the data are entered in a consistent fashion and remain reliable across time. The data collected can be assessed for these data quality dimensions through a number of methods including comparison to a gold standard, data element agreement, and distribution comparisons, among others (Weiskopf and Weng, 2013). By reflecting on data quality dimensions as they relate to their big and small data, establishments may be prompted to increase the frequency of collection of safety related data, improve the data quality, improve their behavioral safety programs, or streamline the aggregation of data across divisions.

A common concern of data quality that may be particularly important in OSH is range restriction among safety outcome variables. For example, safety performance is often evaluated with frequencies of OSHA recordables or injuries requiring first aid. These lagging indicator variables can have low velocity, particularly among establishments with robust safety programs who are exploring potential analytic approaches or for establishments in industries that have relatively lower frequencies of injury (e.g., some types of manufacturing). Thus, databases consisting of only injuries or other low frequency events may be too sparse for conducting potentially insightful analytics.

Range restriction concerns are not limited to big data analytic approaches and have been discussed in studies using more traditional approaches in OSH. For example, in a study examining the relation between worker safety climate and injuries in a sample of nurses,

the authors acknowledged that range restriction was a limitation when examining relations among variables related to safety climate and frequency of injury (Nixon et al., 2015). Their solution was to also examine turnover intentions, safety workarounds, and hazards.

Additionally, analytics of events where an actual injury occurred may miss important variability in other variables of interest when those injuries were not occurring. In other words, the best analytics would provide actionable insights on the variables predicting injuries as well as the variables predicting safe working conditions. An option available to enterprises or researchers is to aggregate the data across multiple departments or establishments to obtain a greater range of values. For example, in a multi-billion dollar construction infrastructure project in Australia, data across numerous contracting organizations were aggregated to conduct a temporal analysis of the relationships among a number of safety indicators (Lingard et al., 2017). By including a large number of projects, it is likely that they incorporated construction sites that had both relatively high and low frequencies of injuries and obtained a wider range of values.

Strong data governance is another important data subfactor. Data governance involves establishing goals and objectives for analytics, developing data policies and procedures, defining roles and responsibilities, among others (Alhassan et al., 2016). Although many companies already collect and track safety-related data such as safety observations and injury recordables without a centralized system or governance describing what data will be collected and stored, post-hoc attempts to perform safety analytics are likely to be challenging. In larger organizations consisting of multiple divisions, it is likely that there would be significant disparities between the ways divisions manage their safety data. Each division may be collecting different data, naming the same variable in different ways, and storing data in different and potentially unconnected data storage technologies. To avoid such a situation, organizations should impose data governance policies across different units, consistent to the extent possible, which describe which data to collect and how to structure the data being collected (e.g., variable names, types, etc.), as well as establish centralized data storage facilities to enable safety analytics across the enterprise.

Data governance can also address concerns about privacy and security. Employees may have concerns about how the data will be used, and these concerns have grown as technologies such as wearable sensors for tracking workers' physiological state or environmental conditions have been developed and implemented (Schall et al., 2018). Deidentifying data is standard practice in analytics (Sweeney, 2002), but precautions should be taken to prevent potential re-identification (Ohm, 2009). Furthermore, data security ensures that unauthorized access to the data is prevented. Models have been proposed that address security concerns and provide recommendations at all phases of the data lifecycle, from data collection to the extraction of insights from the data (Alshboul et al., 2015; Kanika and Khan, 2018; Xu et al., 2014). Establishments can develop protocols related to both data privacy and security outlining who has access to the data, how the data will (and will not) be used, and how the data will be protected. For example, policies related to data security and contingencies for security breaches should be developed as a prerequisite to analytics readiness.

The *Data Culture* factor of analytics readiness is concerned with norms and practices related to analytics within the establishment, such as data collection and handling. A "data-driven culture" has been defined as "the extent to which organizational members (including top-level executives, middle managers, and lower-level employees) make decisions based on the insights extracted from the data" (Gupta and George, 2016, p. 5). This definition necessitates investment and participation in the process from all employees (Diaz et al., 2018) and that insights obtained from analytics are available to workers who need them (Kiron and Shockley, 2011). As an example, behavioral safety programs typically rely on worker participation to promote safe work practices. But such programs also can serve as a backbone for occupational safety analytics because front-line employees play a key role in collecting relevant and timely data, often from voluntary peer-to-peer safety observations. Programs that encourage workers' direct participation in this way been shown to improve safety climate and culture (DePasquale and Geller, 1999). Similarly, an organization's data culture may be improved by such worker-initiated safety practices. For example, if a worker at a manufacturing plant completes and submits a safety checklist of any observed safe and at-risk behaviors or conditions, and if the worker also sees that information led to improvements in safe work practices that are communicated through safety briefings, toolbox talks, and other mechanisms, they may be more inclined to collect that information with greater fidelity and regularity. In a strong safety-data-driven culture, employees at all levels are actively involved in the analytics process by collaborating safety experts and IT to leverage the collected data to improve workplace safety.

4. Knowledge gaps and future directions

The limited application of analytics in OSH to help guide decision making leaves numerous knowledge gaps and thus many opportunities for research and development. Table 3 lists several topic areas and specific research objectives that we believe merit further attention. Although the list of research objectives is not exhaustive, it reinforces and extends calls for research made by others in the field (Abioye et al., 2021; Huang et al., 2018; Ouyang et al., 2018; Tan et al., 2016). We organized the topic areas around these central lines of inquiry: readiness for analytics, analytics methods, technology integration, data culture, and impact of analytics.

Analytics readiness.

Establishments intent on conducting safety analytics with their organizational data would benefit from an assessment of their current data systems and their "readiness" for meaningful analytics. Readiness assessment tools have been developed within other industries to help organizations identify their readiness for conducting analytics, and if not, what steps they need to take to reach that point (Arsenijević et al., 2018; Nemati and Udiavar, 2013; Venkatraman et al., 2016). For example, the Healthcare-Analytics Pre-Adoption Readiness Assessment (HAPRA) Instrument (Venkatraman et al., 2016) allows healthcare organizations to self-rate their maturity under five factors: digital medical technologies, IT infrastructure, user adoption of technology, quality of available data, and management alignment. The assessment provides scores compared with ideal scores and guidance on ways to improve readiness going forward. In OSH, a similar measurement tool

could be developed by establishments to assess their analytics readiness and reveal insights on how to improve current data systems.

Analytics methods.

Another topic for which there is a great need relates to the specific analytical techniques and processes that can be applied to occupational safety. Specific areas of need include exploring advanced analytics methods and modeling techniques to better describe and predict OSH-related outcomes, scaling analytics for organizations of different sizes, and demonstrating the integration of technology (e.g., worker body sensors) into the analytics process. Many differences in work environments and data types exist across industries, but one of the primary goals of analytics in OSH is injury reduction. Therefore, there may be analytic approaches that are better suited for these tasks, such as time series methods (e.g., autoregressive models).

Research examining different phases of advanced analytics methods (e.g., descriptive, diagnostics, predictive, and prescriptive) and modeling techniques (e.g., classification, clustering, regression) and examining the outcomes from these approaches would help expand the OSH analytics knowledge base. From these results, “rules of thumb” may be developed for handling particular types of data or addressing particular types of research questions, which will help enterprises implement analytics more efficiently. To obtain this information, survey research could be conducted in which enterprises and establishments are asked to report what types of analytics they perform, what specific analytic approaches they employ, and what OSH questions they are addressing with analytics. It would also be useful to know which analytics implementations failed and the reasons for their failures. After further development of the field, systematic reviews could contribute valuable information as well.

Additionally, studies that show implementation of analytics with establishments across a variety of industries will help demonstrate what amounts and types of data are necessary to conduct analytics and obtain useful results. As described in a previous section, most of the analytics OSH research published to date has been conducted on large governmental data sets, and fewer studies have been conducted on data collected within enterprises or establishments. There may be particular challenges to implementing OSH analytics in small and medium-sized enterprises, and indeed challenges have been identified for these groups when implementing business intelligence (a term for more general business analytics; (Scholz et al., 2010). Investigating what challenges and particular benefits there may be to implementing OSH analytics in smaller enterprises would be worthwhile as there are 28 million small and medium-sized enterprises in the United States alone (Office of the United States Trade Representative, 2020).

Last, in the current era of rapid technological progress, advancements have led to the development and implementation of sensors and other devices that can be incorporated into OSH analytics to further improve workplace safety. For example, wearable sensors can monitor the physical health of a worker, including their vital signs (e.g., heart rate, temperature, respiration; Lee et al., 2017), vibration exposure (Pavón et al., 2017), and ergonomics (Nath et al., 2017). Other examples of sensor technology in the workplace

include radio frequency identification (RFID) sensors that have been developed to track PPE use (Musu et al., 2014) and alert workers to their proximity to hazardous machinery (Kim and Kim, 2012; Ruff, 2007), ultrasound technology that has been used to monitor corrosion within pipelines (Cegla and Allin, 2017) and acoustic emission monitoring that has been used to identify cracks and microcracks in civil and industrial engineering (Behnia et al., 2014). The benefits to wearable and environmental sensors to both the employee and the enterprise are extensive as the sensors can collect real-time data on environmental conditions and exposures to potential hazards. The sensors can provide timely feedback to workers to mitigate the hazard and prevent an adverse event. Data collected by sensors could be incorporated into OSH analytics via real-time dashboards to attempt to improve worker health and safety, although privacy concerns associated with such measures need to be addressed (Schall et al., 2018).

Data culture.

Research has established that several key features of an organization's safety culture can promote and effective safety performance. It seems reasonable to extend the concept of culture to an organization's attitudes, believes, and norms around data collection, the individual behaviors involved in collecting data and using information for decisions, data systems, and data-based decision making. A better understanding of an organization's data culture requires an assessment tool— an instrument to evaluate data culture of establishments or enterprises, particularly around their safety data. For example, one goal of the instrument could be to evaluate the front-line employees' and managers' commitment to collecting safety data and trust in the appropriate use of the data by upper management. Another research objective related to data culture is examining how organizational data culture correlates with the application of analytics. A research question related to this objective could be: Do safety programs with robust behavioral observation programs, for example, see more actionable insights from conducting analytics? Investigating this hypothesis could involve disseminating the data culture survey to establishments and enterprises and also asking questions related to their safety management systems, data collection processes, and analytics processes. A hypothesis related to this potential project would be that establishments and enterprises with strong data cultures are conducting more mature stages of analytics.

Impact of analytics.

The application of analytics in OSH is an emerging field, and thus safety researchers and professionals would benefit from seeing many examples of successful and unsuccessful applications. Although there have been calls for applications of analytics to occupational safety, there is a dearth of research demonstrating its effectiveness in reducing injury and illness among workers. Additional published accounts of successful implementation of analytics may help OSH professionals obtain management and worker buy-in and help improve and refine the methodology of analytics. The examination of "dark" OSH data (i.e., information that has been collected but is not used to derive insights or inform decision making) would help increase the understanding of the situations that lead to incidents and help in their prevention. One would be to use text analytics to examine open-ended text information contained in incident or audits. Incorporating data sources that are not currently

traditionally incorporated in OSH analyses (e.g., human resources, operational, customer and sales information) would also increase the validity of OSH prediction models. This type of data aggregation and analysis should be possible given the technology and automation available to most organizations and the push to centralize much of the operational data. More research is also needed to develop standardized protocols to improve generalizability and aggregation of data sets across industries.

5. Practical impact

In 2012, then American Society of Safety Engineers president Terrie Norris stated that "A statistical plateau of worker fatalities is not an achievement but evidence that this nation's effort to protect workers is stalled," and she called for a new paradigm for addressing the plateau (Smith, 2012). In this information era that ushered in advances in computer software and hardware technology, it is now possible for companies to gather a wealth of safety-related information in real time. In the present paper, we extended the concepts and principles of big data and analytics to OSH and highlighted the unique advantages of conducting analytics at the enterprise and establishment levels. Several knowledge gaps remain, and more research and demonstrations of effective analytics in OSH are needed to provide more practical guidance on how to use analytics effectively. Nevertheless, the guidance provided in this paper may help safety professionals and researchers accept the challenges and opportunities that analytics present towards breaching the plateau of safety performance and further reducing work-related injuries and fatalities.

Data availability

No data was used for the research described in the article.

References

Abioye SO, Oyedele LO, Akanbi L, Ajayi A, Delgado JMD, Bilal M, Ahmed A, 2021. Artificial intelligence in the construction industry: a review of present status, opportunities and future challenges. J. Build. Eng 44, 103299.

Agraz-Boeneker R, Groves WQ, Haight JM, 2007. An examination of observations and incidence rates for a behavior based safety program. J. SH&E Res 4 (3), 1–22.

Alhassan I, Sammon D, Daly M, 2016. Data governance activities: an analysis of the literature. J. Decis. Syst 25 (Suppl. 1), 64–75.

Alshboul Y, Nepali R, Wang Y, 2015. Big data lifecycle: threats and security model. In: Paper Presented at the Proceedings of the Twenty-First Americas Conference on Information Systems, Puerto Rico

Arnold KE, Lonn S, Pistilli MD, 2014. An exercise in institutional reflection: the learning analytics readiness instrument (LARI). In: Paper Presented at the Proceedings of the Fourth International Conference on Learning Analytics and Knowledge

Arsenijević D, Stankovski S, Ostojić G, Baranovski I, Oros D, 2018. An overview of IoT readiness assessment methods. In: Paper Presented at the Zbornik Radova 8th International Conference on Information Society and Technology–ICIST

Baars H, Kemper H-G, 2008. Management support with structured and unstructured data—an integrated business intelligence framework. Inf. Syst. Manag 25 (2), 132–148.

Behnia A, Chai HK, Shiotani TJC, Materials B, 2014. Advanced structural health monitoring of concrete structures with the aid of acoustic emission. Construct. Build. Mater 65, 282–302.

Bellazzi R, 2014. Big data and biomedical informatics: a challenging opportunity. Yearbook Med. Inf 9 (1), 8.

Bertke S, Meyers A, Wurzelbacher S, Measure A, Lampl M, Robins D, 2016. Comparison of methods for auto-coding causation of injury narratives. Accid. Anal. Prev 88, 117–123. [PubMed: 26745274]

Billings C, Lauber J, Funkhouser H, Lyman E, Huff E, 1976. NASA Aviation Safety Reporting system(Tech. Report TM-X-3445) Retrieved from Moffett Field, CA.

Bradlow ET, Gangwar M, Kopalle P, Voleti S, 2017. The role of big data and predictive analytics in retailing. J. Retailing 93 (1), 79–95.

Cegla F, Allin J, 2017. Ultrasonics, corrosion and SHM: the story of Permasense Ltd. In: Paper Presented at the 11th International Workshop on Structural Health Monitoring

Cheng C-W, Leu S-S, Cheng Y-M, Wu T-C, Lin C-C, 2012. Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. Accid. Anal. Prev 48, 214–222. [PubMed: 22664684]

Cheng C-W, Yao H-Q, Wu T-C, 2013. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. J. Loss Prev. Process. Ind 26 (6), 1269–1278.

Cooper A, 2012. What Is Analytics? Definition and Essential Characteristics

De Mauro A, Greco M, Grimaldi M, Ritala P, Management, 2018. Human resources for Big Data professions: a systematic classification of job roles and required skill sets. Inf.Process 54 (5), 807–817.

DePasquale JP, Geller ES, 1999. Critical success factors for behavior-based safety: a study of twenty industry-wide applications. J. Saf. Res 30 (4), 237–249.

Diaz A, Rowshankish K, Saleh T, 2018. Why data culture matters. McKinsey Q 3, 37–53.

Dumbill E, 2013. Making sense of big data. Big Data 1 (1), 1–2. [PubMed: 27447028]

Gandomi A, Haider M, 2015. Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manag 35 (2), 137–144. 10.1016/j.ijinfomgt.2014.10.007.

Ghodrati N, Yiu TW, Wilkinson S, Shahbazpour M, 2018. A new approach to predict safety outcomes in the construction industry. Saf. Sci 109, 86–94. 10.1016/j.ssci.2018.05.016.

Gifford ML, Anderson JE, 2010. Barriers and motivating factors in reporting incidents of assault in mental health care. J. Am. Psychiatr. Nurses Assoc 16 (5), 288–298. [PubMed: 21659279]

Goh YM, Chua D, 2013. Neural network analysis of construction safety management systems: a case study in Singapore. Construct. Manag. Econ 31 (5), 460–470.

Gualtieri M, 2016. Hadoop Is Data's Darling for Reason Retrieved from. https://go.forrester.com/blogs/hadoop-is-datas-darling-for-a-reason/.

Guilfoyle S, Bergman SM, Hartwell C, Powers J, 2016. Social Media, Big Data, and Employment Decisions: Mo' Data, Mo' Problems? In: Landers R, Schmidt G (Eds.), Social Media in Employee Selection and Recruitment Springer, Cham. 10.1007/978-3-319-29989-1_7.

Gupta M, George JF, 2016. Toward the development of a big data analytics capability. Inf. Manag 53 (8), 1049–1064.

Herzer KR, Mirrer M, Xie Y, Steppan J, Li M, Jung C, Mark LJ, 2012. Patient safety reporting systems: sustained quality improvement using a multidisciplinary team and "good catch" awards. Joint Commission J. Quality Patient Saf. Surg 38 (8), 339–AP331.

Hotho A, Nürnberger A, Paaß G, 2005. A brief survey of text mining. LDV Forum 20, 19–62.

Huang L, Wu C, Wang B, Ouyang Q, 2018. Big-data-driven safety decision-making: a conceptual framework and its influencing factors. Saf. Sci 109, 46–56.

Inmon WH, Nesavich A, 2007. Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence. Pearson Education

Institute for Operations Research and the Management Sciences, 2014. Certified Analytics Professional (CAP) Examination Study Guide INFORMS, Catonsville, MD.

Jebelli H, Choi B, Lee S, 2019. Application of wearable biosensors to construction sites. II: assessing workers' physical demand. J. Construct. Eng. Manag 145 (12) 10.1061/(ASCE)CO.1943-7862.0001710.

Kakhki FD, Freeman SA, Mosher GA, 2020. Applied machine learning in agro-manufacturing occupational Incidents. Procedia Manuf 48, 24–30.

Kanika A, Khan R, 2018. An improved security threat model for big data life cycle. Asian J. Comput. Sci. Technol 7 (1), 33–39.

Katal A, Wazid M, Goudar R, 2013. Big data: issues, challenges, tools and good practices. In: Paper Presented at the 2013 Sixth International Conference on Contemporary Computing (IC3)

Kim K, Kim M, 2012. RFID-based location-sensing system for safety management. Personal Ubiquitous Comput 16 (3), 235–243.

Kiron D, Shockley R, 2011. Creating business value with analytics. MIT Sloan Manag. Rev 53, 57–63.

Kongsvik T, Johnsen SÅK, Sklet S, 2011. Safety Climate and Hydrocarbon Leaks: an Empirical Contribution to the Leading-Lagging Indicator Discussion. J. Loss Prevent. Process Ind 24, 405–411, 4.

Lee W, Lin K-Y, Seto E, Migliaccio GC, 2017. Wearable sensors for monitoring on-duty and off-duty worker physiological status and activities in construction. Autom. ConStruct 83, 341–353.

Liao C-W, Perng Y-H, 2008. Data mining for occupational injuries in the Taiwan construction industry. Saf. Sci 46 (7), 1091–1102.

Lingard H, Hallowell M, Salas R, Pirzadeh P, 2017. Leading or lagging? Temporal analysis of safety indicators on a large infrastructure construction project. Saf. Sci 91, 206–220.

Loshin D, 2013. Big Data Analytics: from Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph. Elsevier

Manzoor MAB, Hussain S, Ahmad W, Jahanzaib M, 2018. An empirical analysis of a process industry to explore the accident causation factors: a case study of a textile mill in Pakistan. J. Basic Appl. Sci 14, 72–79.

Marr B, 2018. How much data do we create every day? The mind-blowing stats everyone should read Retrieved from. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#2a028eab60ba.

Marucci-Wellman HR, Lehto MR, Corns HL, 2015. A practical tool for public health surveillance: semi-automated coding of short injury narratives from large administrative databases using Naïve Bayes algorithms. Accid. Anal. Prev 84, 165–176. 10.1016/j.aap.2015.06.014. [PubMed: 26412196]

Matías J, Rivas T, Martín J, Taboada J, 2008. A machine learning methodology for the analysis of workplace accidents. Int. J. Comput. Math 85 (3–4), 559–578.

McSween TE, 2003. Values-based Safety Process: Improving Your Safety Culture with Behavior-Based Safety John Wiley & Sons.

Meyers AR, Al-Tarawneh IS, Wurzelbacher SJ, Bushnell PT, Lampl MMP, Bell JL, Wei C, 2018. Applying machine learning to workers' compensation data to identify industry-specific ergonomic and safety prevention priorities: Ohio, 2001 to 2011. J. Occup. Environ. Med 60 (1), 55. [PubMed: 28953071]

Miloslavskaya N, Tolstoy A, 2016. Big data, fast data and data lake concepts. Procedia Comput. Sci 88, 300–305.

Mistikoglu G, Gerek IH, Erdis E, Mumtaz Usmen PE, Cakan H, Kazan EE, 2015. Decision tree analysis of construction fall accidents involving roofers. Expert Syst. Appl 42 (4), 2256–2263. 10.1016/j.eswa.2014.10.009.

Musu C, Popescu V, Giusto D, 2014. Workplace safety monitoring using RFID sensors. In: Paper Presented at the 2014 22nd Telecommunications Forum Telfor (TELFOR)

Nath ND, Akhavian R, Behzadan AH, 2017. Ergonomic analysis of construction worker's body postures using wearable mobile sensors. Appl. Ergon 62, 107–117. [PubMed: 28411721]

Nemati H, Udiavar A, 2013. SCAX: measuring organizational readiness to embrace supply chain analytics. Int. J. Bus. Intell. Res 4 (2), 19–38.

Nixon AE, Lanz JJ, Manapragada A, Bruk-Lee V, Schantz A, Rodriguez JF, 2015. Nurse safety: how is safety climate related to affect and attitude? Work. Stress 29 (4), 401–419.

Office of the United States Trade Representative, 2020. Small and Medium-Sized Enterprises Retrieved from. https://ustr.gov/trade-agreements/free-trade-agreements/trans-pacific-partnership/tpp-chapter-chapter-negotiating-8.

Ohm P, 2009. Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev 57, 1701.

Oussous A, Benjelloun F-Z, Lahcen AA, Belfkih S, 2015. Comparison and classification of nosql databases for big data. In: Paper Presented at the Proceedings of International Conference on Big Data, Cloud and Applications

Ouyang Q, Wu C, Huang L, 2018. Methodologies, principles and prospects of applying big data in safety science research. Saf. Sci 101, 60–71.

Patel D, Jha K, 2015. Neural network model for the prediction of safe work behavior in construction projects. J. Construc. Eng.: Management and Compliance Series 141 (1), 04014066.

Pavón I, Sigcha L, López J, De Arcas G, 2017. Wearable technology usefulness for occupational risk prevention: smartwatches for hand-arm vibration exposure assessment. In: Occupational Safety and Hygiene V. CRC Press, pp. 77–82.

Pavón I, Sigcha L, Arezes P, Costa N, de Arcas G, Lopez-Navarro JJOS, 2018. Wearable technology for occupational risk assessment: potential avenues for applications. In: Occupational Safety and Hygiene VI: Book Chapters from the 6th International Symposium on Occupation Safety and Hygiene (SHO 2018), pp. 447–452.

Pereira E, Hermann U, Han S, AbouRizk S, 2018. Case-based reasoning approach for assessing safety performance using safety-related measures. J. Construct. Eng. Manag 144 (9), 04018088.

Pipino LL, Lee YW, Wang RY, 2002. Data quality assessment. Commun. ACM 45 (4), 211–218.

Poh CQ, Ubeynarayana CU, Goh YM, 2018. Safety leading indicators for construction sites: a machine learning approach. Autom. ConStruct 93, 375–386.

Polyvyanyy A, Pika A, Wynn MT, ter Hofstede AH, 2019. A systematic approach for discovering causal dependencies between observations and incidents in the health and safety domain. Saf. Sci 118, 345–354.

Railroad Administration, Federal, 2003. Voluntary Reporting of Safety Information: the Feasability of Developing Such Programs in the US Railroad Industry and a Proposed Pilot Demonstration Project

Ramadan RA, 2017. Big Data Tools: an Overview. International Journal of Computer Software Engineering, 2017

Ransbotham S, Kiron D, 2017. Analytics as a source of business innovation. MIT Sloan Manag. Rev 58 (3).

Rose R, 2016. Defining analytics: a conceptual framework. OR/MS Today 43 (3).

Ruff TM, 2007. Recommendations for Evaluating and Implementing Proximity Warning Systems Onsurface Mining Equipment

Sanmiquel L, Rossell JM, Vintró C, 2015. Study of Spanish mining accidents using data mining techniques. Saf. Sci 75, 49–55.

Schall MC, Sesek RF, Cavuoto LA, 2018. Barriers to the adoption of wearable sensors in the workplace: a survey of occupational safety and health professionals. Hum. Factors 60 (3), 351–362. 10.1177/0018720817753907. [PubMed: 29320232]

Schembera B, Duran JM, 2020. Dark data as the new challenge for big data science and the introduction of the scientific data officer. Philosophy Technol 33 (1), 93–115.

Scholz P, Schieder C, Kurze C, Gluchowski P, Böhringer M, 2010. Benefits and challenges of business intelligence adoption in small and medium-sized enterprises. ECIS 2010 Proceedings

Schumpeter J, 2014. Digital Disruption on the Farm, 64. The Economist

Shannon HS, Norman GR, 2009. Deriving the factor structure of safety climate scales. Saf. Sci 47 (3), 327–329. 10.1016/j.ssci.2008.06.001.

Sherman R, 2014. Business Intelligence Guidebook: from Data Integration to Analytics. Newnes

Smith S, 2012. ASSE President: Efforts to Protect U.S. Workers Is Stalled Retrieved from. http://ehstoday.com/safety/management/worker-safety-efforts-stalled-0109.

Stewart D, 2013. A Look at Safety Analytics Retrieved from. http://www.canadianminingjournal.com/features/a-look-at-safety-analytics/.

Sweeney L, 2002. k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowledge-Based Syst 10, 557–570, 05.

Tan KH, Ortiz-Gallardo VG, Perrons RK, 2016. Using Big Data to manage safety-related risk in the upstream oil & gas industry: a research agenda. Energy Explor. Exploit 34 (2), 282–289.

Vallmuur K, Marucci-Wellman HR, Taylor JA, Lehto M, Corns HL, Smith GS, 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–i42. [PubMed: 26728004]

Van Barneveld A, Arnold KE, Campbell JP, 2012. Analytics in higher education: establishing a common language. EDUCAUSE learning initiative 1 (1) l–ll.

Venkatraman S, Sundarraj RP, Mukherjee A, 2016. Prototype design of a healthcare-analytics pre-adoption readiness assessment (HAPRA) instrument. In: Paper Presented at the International Conference on Design Science Research in Information System and Technology

Wagner GR, 2014. Can predictive analytics help reduce workplace risk? Retrieved from. https://blogs.cdc.gov/niosh-science-blog/2014/10/02/pa/.

Waller MA, Fawcett SE, 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. J. Bus. Logist 34 (2), 77–84.

Wang Y, Kung L, Byrd TA, Change S, 2018. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol. Forecast 126, 3–13.

Watson HJ, 2014. Tutorial: big data analytics: concepts, technologies, and applications. Commun. Assoc. Inf. Syst 34 (1), 65.

Weiskopf NG, Weng C, 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J. Am. Med. Inf. Assoc 20 (1), 144–151.

White T, 2015. Hadoop: the Definitive Guide. O'Reilly Media

Wurzelbacher S, Jin Y, 2011. A framework for evaluating OSH program effectiveness using leading and trailing metrics. J. Saf. Res 42 (3), 199–207. 10.1016/j.jsr.2011.04.001.

Xu L, Jiang C, Wang J, Yuan J, Ren Y, 2014. Information security in big data: privacy and data mining. IEEE Access 2, 1149–1176.

Zikopoulos P, Eaton C, 2011. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data McGraw-Hill Osborne Media.

Zohar D, 1980. Safety climate in industrial organizations: theoretical and applied implications. J. Appl. Psychol 65 (1), 96–102. [PubMed: 7364709]

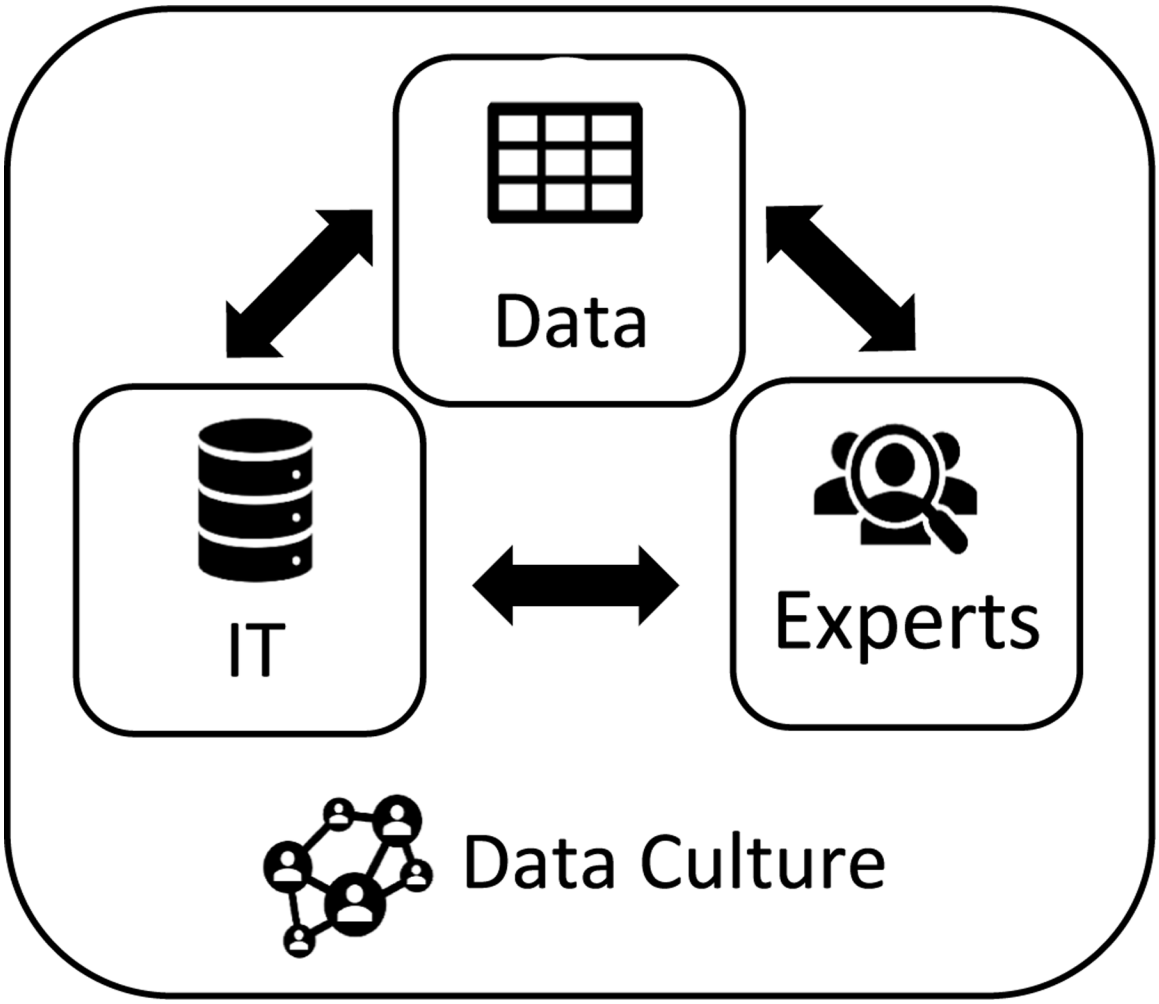

Fig. 1. Diagram showing key readiness factors for conducting analytics in occupational safety and health; IT = information technology.

Table 1

Variety of data examples by category and example studies.

Category	Data Examples	Example Studies
Production		
	Overall Equipment Efficiency	
	Staffing loads	Poh et al. (2018)
	Volume trends	
Human Resources		
	Overtime	
	Absenteeism	
	Job Tenure	Cheng et al. (2013)
	Disciplinary actions	
Safety Metrics		
Leading Indicators		
	Peer observations	Pereira et al. (2018)
	Manager observations	Lingard et al. (2017)
	Identified hazardous conditions	Polyvyanyy et al (2019)
	Wearable fatigue monitoring	Jebelli et al. (2019)
	Safety climate	Patel and Jha (2015)
Lagging Indicators		
	Incident type	Sanmiquel et al. (2015)
	Injury type	Cheng et al. (2012)
	Injury severity	Mistikoglu et al. (2015)
	Temperature	Pereira et al. (2018)
	Wind speed average	Pereira et al. (2018)
Maintenance		
	Maintenance schedules	Goh and Chua (2013)
	Failures (Equipment)	
	Action item backlog	

Table 2

Readiness factors and related needs implementing analytics for occupational safety.

Readiness Factor	Needs	Examples
Expertise	Analytics and data science expertise	The establishment has access to data scientists or statisticians.
	Subject matter expertise	The establishment hires or consults with experts in information technology, data management, statistics, and occupational safety.
IT Infrastructure	Hardware and software	The establishment possesses IT hardware and software suitable for analytics.
	Data storage/accessibility	Data systems across the establishment are merged into accessible databases.
Data	Data quality	Guidance documents are created to assess the accuracy of safety and organizational data.
	Data fidelity	Standard operating procedures are created to detect and remedy missing data on leading indicators.
	Standardized format	Data are reformatted and restructured to be suitable for analytics, and new data types are collected and formatted with analytics in mind.
Data culture	Data governance	Policies and procedures are developed to support the effective and efficient use of information.
	Participation	Employees are committed to collecting data on a regular basis and attend to data quality and completeness.
	Analytics value	Establishment leadership is committed to and understands the value in the analytics process.
	Employee protection	Clear guidelines indicate how employee data will be used in employee decision making and maintaining employee privacy will be prioritized.

Table 3

Suggestions for future research on analytics in occupational safety and health (OSH) by topic area.

Topic Area	Example Research Objectives
Analytics readiness	Identify key factors that are critical for effective analytics in OSH Develop an assessment tool to evaluate organizational readiness for implementation of analytics
Analytics methods	Explore advanced analytical methods and modeling techniques to better describe and predict OSH-related outcomes Develop scalable analytics methods and solutions to suit small and medium-sized enterprises = and large-scale industries Demonstrate the integration of technology (e.g., radio frequency identification, worker body sensors) into the analytics process
Data culture	Identify key elements of effective and supportive data cultures Develop and validate a data culture assessment tool
Impact of analytics	Demonstrate the effectiveness of analytics in reducing injuries and near misses Develop standardized protocols to improve generalizability and aggregation of data sets within and across industries Document and disseminate case examples of successful implementation of analytics in aiding OSH-related decision making and improved safety outcomes.