Vol. 04, No. 2 (2022) 217-220, doi: 10.24874/PES04.02.011

Proceedings on Engineering Sciences

www.pesjournal.net

SAFETY DATA SCIENCE: BASIC FOUNDATIONS AND CRITICAL FACTORS

Rodrigo F S Gomes[1]

Received 03.02.2022.
Accepted 14.04.2022.
UDC – 004.056.5

Keywords:

Safety Data Science; SDS; Critical factors; Barriers; Safety Management

A B S T R A C T

This paper explores critical factors related to the application of data science in the field of occupational safety and discusses plausive causes to explain why the herewith called safety data science (SDS) is still a distant reality in most organizations. To do that, existing literature was screened and inductive reasoning was used to identify potential barriers for the massive use of SDS. As a result, basic foundations are introduced and three critical factors are discussed: (1) the lack of theoretical foundations in SDS to guide empirical applications; (2) unavailability or poor safety data at the organizational level, and (3) the lack of expertise related to data science from the perspective of safety professionals. The study has important implications. From the theoretical perspective, it offers an initial conceptual baseline for SDS and presents its critical factors. Also, directions are given for safety practitioners to explore SDS at the organizational level.

1. INTRODUCTION

Data science has been largely employed in different fields in the government and private sectors. Applications using data mining, analytics, and machine learning have been used to predict patterns, trends and to decipher causalities. In this context, benefits are already known in several dimensions, notably when it comes to accuracy and precision. Moreover, in a world with vast amounts of available data, the capacity of extracting, processing, and analyzing large data sets is widely known as a source of competitive advantage, not only from a business perspective but also to improve other aspects of society like education and public health. This explains why data has been considered the world's most valuable resource (Ashwin Ram, 2019; Economist, 2017), and why the quote "data is the new oil" has become a common refrain.

The use of data science to improve occupational safety is surely beneficial at all levels including governments, organizations, and society. This is because a healthy and safe work environment not only is desirable from the workers' perspective, but also contributes considerably to labor productivity, promotes economic growth, and relieves pressure on public and private social systems (Heuvel et al., 2017).

In the existing literature, some applications using data science techniques in the field of occupational safety are found, for instance, mining association rules (Agrawal et al., 1993) in accident prevention analysis (Changhai & Shenping, 2019; Mirabadi & Sharifian, 2010; Qiao et al., 2018; Verma et al., 2014), road safety assessment (Rekha Sundari et al.; Yang et al., 2019), and incident cause-and-effect analysis (Cheng et al., 2010; Li et al., 2017).

[1] Corresponding author: Rodrigo F S Gomes
Email: rodrigofrank@edu.unisinos.br

Also, from the practical perspective, some modern data analytics tools capable of extracting, integrating, and analyzing previously inaccessible and siloed data are also found in the marketplace. Yet, the use of such techniques seems to be a distant reality from the practitioners that are tasked to improve safety in their organizations. As a result, disharmony between theory and application is verified, and not much progress in accident prevention has been observed from a global perspective (ILOSTAT, 2021).

This is consistent with the fact that the field of safety science is advancing very slowly (Rae et al., 2020). On the one side, only a few researchers have drawn attention to the application of data science concepts in the field of safety, exploring benefits and challenges in practice. On the other side, organizations are demanding safety data scientists to make significant progress in the field of safety, instead of establishing safety systems based on relatively outdated tools.

This article outlines that safety data science is not defined in the existing literature. Therefore, in the next session a definition is given, and plausive critical factors are discussed to explore why SDS is still a distant reality in most organizations.

2. SAFETY DATA SCIENCE (SDS)

No definition has been found to the meaning of safety data science (SDS) in peer-review literature screened in the main databases to date. To close this gap, a plausive proposal to define SDS should take into consideration recognized definitions of both terms 'safety science' and 'data science'.

On one side, there are many definitions of safety science and to pin down one is not an easy task. Yet, it is very acceptable that safety science can be understood as a safety knowledge-generating process, comprising knowledge about safety-related phenomena, processes, events, etc., as well as conceptual tools which cover the development of concepts, theories, principles, and methods to understand, assess, communicate and manage (in a broad sense) safety (Aven, 2014). On the other side, at a high level, data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data (Provost & Fawcett, 2013).

Based on these consolidated concepts, we offer the following definition of SDS:

Safety data science is a set of principles and techniques that support and guide the extraction, processing, and analysis of safety data, to reveal what are the relevant factors, how they are interconnected, and to draw causal conclusions.

This definition establishes that safety data scientists must be able to deal with issues from a data perspective grounded on principles like data-analytic thinking, which gives the data scientist a framework to systematically extract useful knowledge from data, improving decision-making (Provost & Fawcett, 2013).

In this context, organizations interested in SDS but facing deficiencies in data science resources should still understand basic concepts and the steps to engage employees and consultants. These basic steps shown in Table 1 are the following and more detail can be found in specialized literature found elsewhere (der Aalst, 2016).

Table 1. Steps for implementing Safety Data Science

Step	Description
Data cleaning and structuring	Safety data might be available in a form of different types of variables. For example, the quantity of incidents, injuries, working hours, or accident rate. Also, qualitative data might be available, such as behavior assessments. A safety data scientist must be able to organize data in a form to be extracted and properly processed.
Data extraction	Information technology is highly recommended as a means of collecting a large amount of safety data. Workers need means to report safety events at job sites in real-time, notably in decentralized operations. Also, auditors conducting safety inspections using online checklists into mobile devices increase their productivity and eliminate waste.
Data processing	Safety usually offers large data sets and requires the use of adequate resources for processing. This is related, for instance, to either computing techniques such as data mining algorithms, or IT infrastructure to support the data processing (e.g., databases).
Data analysis	Evaluating safety data science results requires careful consideration. It depends on several aspects like the context in which information will be used, and the technique and method used in the data processing. Important concepts such as correlation and causation must be considered strictly and never neglected. Generalization also requires close attention, making clear the assumptions in which the general pattern is proposed.

These basic foundations consider the necessary steps represent a minimum body of knowledge and practical information to be considered by organizations and safety professionals interested in applying SDS. Also, depending on the availability of resources including specialized knowledge, SDS might be difficult to become realistic. The next session outlines some critical factors concerning the use of SDS.

3. CRITICAL FACTORS

Several aspects might offer incentives for the use of SDS in academic research and enterprise applications. First, the need for improvements in safety performance (e.g. reducing accidents and injuries) by considering that traditional analysis is no longer effective. In longitudinal studies, for instance, results can be compared when SDS is confronted with other types of analysis over time. Second, to identify patterns and co-occurrence of factors. This is useful for an organization to prevent incidents based on the correlation of relevant antecedent factors (e.g. excess of worked hours) and consequence factors, such as lost time accidents. Third, to identify causation. In this case, it is sought to verify what factor is supposed to cause what effect. Besides the aforementioned aspects, some critical factors create barriers for the increase of studies and applications using principles of SDS. These factors were identified by following inductive reasoning and are useful for practitioners and future studies in the field of safety data science (see Table 2).

From the researchers' perspective, the critical factor (1) intend to draw the attention to the need of more attention regarding theoretical and empirical studies published into specialized literature to boost the application of SDS presented in a form of case studies, short communications, accident prevention studies etc. This is important to check the efficacy of SDS when confronted with other thecniques. From the managerial's perspective, the critical factor (2) should be considered whenever SDS is been considered to be applied in an organization. Safety data in a structured form has been critical for many applications in safety, but is crucial for SDS. Managers and practitioners should address this issue as a primary target in order to benefit from SDS. Finally, safety education programs are asked to review its content to cover all these modern techniques rather than only basic statistics. This is absolutely necessary and it represents a constraint for the advance of SDS. All these critical factors presented in Table 2. are encompassed into 3 dimensions, i.e. academia, managerial, and safety education programs, and represent barriers for the implementation and to the extend of SDS in the literature and in-practice

Table 2. Critical factors for implementing safety data science within organizations

Critical factor	Description
(1) Lack of basic foundations in SDS to boost empirical studies	In general, empirical applications require a previous theoretical base. On the one side, by considering that data science is still a body of knowledge too far from the field of safety science, the lack of basic foundations of the so-called safety data science represents a plausive factor that explains the relatively low attention from scholars. On the other side, an emerging conceptual base is fundamental for drawing the attention of researchers and practitioners to the potential benefits in the field of safety science.
(2) Unavailability or poor safety data	Organizations frequently deal with a lack of qualified data for decision-making. In the field of safety management, this represents a fundamental factor. Safety data must be structured and as accurate as possible to avoid misinterpretation of the results in the stage of data analysis. Therefore, the means of workers to report safety events, as well as other sources of data (e.g. investigation reports) should require close management attention since it represents a baseline for SDS.
(3) Lack of expertise related to data analytics from the perspective of safety professionals	Practitioners in the field of occupational health and safety have not been prepared to work under the data science perspective. Expertise concerning principles of data analytics and knowledge about techniques used in data science are not part of the majority of safety programs. As a result, safety professionals are still stuck on the basics of statistics and taking decisions based on simplistic analysis methods.

4. CONCLUSION

Despite the application of data science in the field of occupational health and safety being relatively recent, organizations claim for short-term advances. This is clearly verified by searching "safety data scientist" on the internet. By browsing this term, several job opportunities requiring such expertise are found. This disharmony represents an existing gap between researchers' attention and the organizations' needs.

To close this gap this paper offers a definition for the so-called safety data science (SDS). Also, it discusses critical factors related to the application of data science in the field of safety: (1) the lack of theoretical foundations in SDS to guide empirical applications; (2) unavailability or poor safety data at the organizational level, and (3) the lack of expertise related to data analytics from the perspective of safety professionals. The study has important implications. From the theoretical implications, it introduces an initial conceptual baseline for SDS and presents its critical factors. Also, directions are given for safety practitioners to explore SDS at the organizational level.

Opportunities for future studies are therefore clear. Theory-building studies on SDS are welcome to broaden comprehension about the foundations presented in this paper. Also, empirical applications concerning the basic concepts herewith presented, and reality-based studies with consistent results about the advantages and challenges associated with the implementation of SDS are highly encouraged.

Despite the consistency of this short communication with the understanding that most of what passes for theory in organizational studies consists of approximations (Weick, 1995), further studies are encourage either to confirm or to claim for modifications in the basic foundations herewith offered.

References:

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association in Large Databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207–216.

Ashwin Ram. (2019, November 15). Data Is The New Oil -- And That's A Good Thing. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/?sh=26ebfac27304

Aven, T. (2014). What is safety science? *Safety Science*, *67*(0925), 15–20.

Changhai, H., & Shenping, H. (2019). Factors correlation mining on maritime accidents database using association rule learning algorithm. *Cluster Computing*, *22*, 4551–4559. https://doi.org/10.1007/s10586-018-2089-z

Cheng, C.-W., Lin, C.-C., & Leu, S.-S. (2010). Use of association rules to explore cause-effect relationships in occupational accidents in the Taiwan construction industry. *Safety Science*, *48*(4), 436–444. https://doi.org/10.1016/j.ssci.2009.12.005

der Aalst, W. (2016). Process mining: Data science in action. In *Process Mining: Data Science in Action*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49851-4

Economist, T. (2017). *The world's most valuable resource is no longer oil, but data*. The Economist. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Heuvel, S. van den, Zwaan, L. van der, Dam, L. van, Oude-Hengel, K., Eekhout, I., Emmerik, M. van, Oldenburg, C., Brück, C., Janowski, P., & Wilhelm, C. (2017). Estimating the costs of work-related accidents and ill-health: An analysis of European data sources - European Risk Observatory. In *European Agency for Safety and Health at Work*. European Agency for Safety and Health at Work. https://doi.org/10.2802/566789

ILOSTAT. (2021). *Statistics on safety and health at work*. https://ilostat.ilo.org/topics/safety-and-health-at-work/

Li, H., Li, X., Luo, X., & Siebert, J. (2017). Investigation of the causality patterns of non-helmet use behavior of construction workers. *Automation in Construction*, *80*, 95–103. https://doi.org/10.1016/j.autcon.2017.02.006

Mirabadi, A., & Sharifian, S. (2010). Application of association rules in Iranian Railways (RAI) accident data analysis. *Safety Science*, *48*(10), 1427–1435. https://doi.org/10.1016/j.ssci.2010.06.006

Provost, F., & Fawcett, T. (2013). Data Science and its relationship to big data and data-driven decision making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Qiao, W., Liu, Q., Li, X., Luo, X., & Wan, Y. (2018). Using data mining techniques to analyze the influencing factor of unsafe behaviors in Chinese underground coal mines. *Resources Policy*, *59*, 210–216. https://doi.org/10.1016/j.resourpol.2018.07.003

Rae, A., Provan, D., Aboelssaad, H., & Alexander, R. (2020). A manifesto for Reality-based Safety Science. In *Safety Science* (Vol. 126, p. 104654). Elsevier B.V. https://doi.org/10.1016/j.ssci.2020.104654

Rekha Sundari, M., Prasad Reddy, P. V. G. D., & Srinivas, Y. (2019). Traffic risk-safety restraints-awareness through data mining approaches. *International Journal of Engineering and Advanced Technology*, *8*(5), 2608–2613. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069940599&partnerID=40&md5=2f90c218a191e2dd613ac7a0e2c2363f

Verma, A., Khan, S. D., Maiti, J., & Krishna, O. B. (2014). Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. *Safety Science*, *70*, 89–98. https://doi.org/10.1016/j.ssci.2014.05.007

Weick, K. E. (1995). What Theory is Not, Theorizing Is. *Administrative Science Quarterly*, *40*(3), 385. https://doi.org/10.2307/2393789

Yang, Y., Yuan, Z. Z., Sun, D. Y., & Wen, X. L. (2019). Analysis of the factors influencing highway crash risk in different regional types based on improved Apriori algorithm. *Advances in Transportation Studies*, *49*, 165–178. https://doi.org/10.4399/978882552809113

Rodrigo Frank de Souza Gomes
Universidade do Vale do Rio dos Sinos
Porto Alegre, Brazil
rodrigofrank@edu.unisinos.br