RBSO Revista Brasileira de Saúde Ocupacional
ISSN: 2317-6369 (online)

Review article
Dossier RBSO 50 Years

Big data on Occupational Health: how far are we?

Big data em Saúde do Trabalhador: o quão distantes estamos?

Thales Pardini Fagundes[a]
https://orcid.org/0000-0002-3302-9913

Roberta Moreira Wichmann[b]
https://orcid.org/0000-0002-4951-9327

Maria Eduarda de Lima e Silva[b]
https://orcid.org/0000-0003-3100-8817

Tiago Almeida de Oliveira[c]
https://orcid.org/0000-0003-4147-7721

[a] Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto, Hospital das Clínicas. Ribeirão Preto, SP, Brazil.
[b] Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa, Programa de Pós-graduação em Economia. Brasília, DF, Brazil.
[c] Universidade Estadual da Paraíba, Departamento de Estatística, Programa de Pós-graduação em Desenvolvimento Regional. Campina Grande, PB, Brazil.

Contact:
Thales Pardini Fagundes
E-mail:
tpfagundes@hcrp.usp.br

Abstract

Objective: to identify strategies and challenges in the use of big data and Artificial Intelligence (AI) in Occupational Health, as well as practices and obstacles to their implementation. **Methods:** scoping review using terms related to occupational health, big data, and AI in four databases (Medline, Embase, BVS, and SciELO) considering articles in Portuguese, Spanish, and English published up to 2022. Studies using large databases and AI for occupational health-related analyses were included. Article selection was performed independently by two researchers, and the conflicts were resolved by consensus. **Results:** of the 505 articles identified, 16 were selected. The low number may be associated with the scarcity of data that address worker's health systemically, considering demographic, technological, socioeconomic, and environmental factors. The selected studies showed that big data and AI have a good potential to support occupational health by identifying health indicators and enabling accurate predictions. Implementation faces challenges such as data storage and ethical issues. **Conclusion:** big data and AI can be useful tools for analyzing the complex interactions of variables to improve the identification of health determinants and record data on work environments and individuals exposed to them.

Keywords: Big Data; Occupational Diseases; Artificial Intelligence; Machine Learning; Algorithms; Occupational Health.

Resumo

Objetivo: identificar estratégias e desafios no uso de *big data* e inteligência artificial (IA) em Saúde Ocupacional, assim como práticas e obstáculos na sua implementação. **Métodos:** revisão de escopo utilizando termos relacionados à saúde ocupacional, *big data* e IA em quatro bases de dados (Medline, Embase, BVS e SciELO), considerando artigos em português, espanhol e inglês publicados até 2022. Foram incluídos estudos com uso de grandes bases de dados e IA para análises relacionadas à saúde ocupacional. A seleção dos artigos foi feita independentemente por dois pesquisadores, com conflitos resolvidos por consenso. **Resultados:** de 505 artigos identificados, 16 foram selecionados. O baixo número pode estar associado à escassez de dados que tratam da saúde do trabalhador de maneira sistêmica, considerando fatores demográficos, tecnológicos, socioeconômicos e ambientais. Os estudos selecionados mostraram que o *big data* e a IA têm bom potencial para subsidiar a saúde ocupacional ao identificar indicadores de saúde e possibilitar previsões precisas. A implementação enfrenta desafios, como armazenamento de dados e questões éticas. **Conclusão:** *big data* e IA podem ser ferramentas úteis para analisar interações complexas de variáveis visando aprimorar a identificação de determinantes de saúde e dados de registros sobre ambientes de trabalho e indivíduos a eles expostos.

Palavras-chave: *Big Data*; Doenças Ocupacionais; Inteligência Artificial; Aprendizado de Máquina; Algoritmos; Saúde do Trabalhador.

How to cite (Vancouver):
Fagundes TP, Wichmann RM, Silva MEL, Oliveira TA. Big data on Occupational Health: how far are we? Rev. Bras. Saúde Ocup. [Internet]. 2024;49:edcinq11. Available from: https://doi.org/10.1590/2317-6369/40522en2024v49edcinq11

Introduction

Economically active individuals are exposed to a variety of physical, chemical, biological, and social stimuli during their occupational activities[1]. There is a growing recognition of the interaction between these agents and the emergence of diseases that cause individual, social, and economic losses[2]. Changes in work organization have already been associated with adverse effects on workers' health and may be related to layoffs, outsourcing, role clarity, and predictability[3].

Given these changes, various occupational exposures may contribute to the emergence and worsening of workers' suffering. Knowledge about the association between these exposures and work-related diseases can benefit the decision-making from governments to doctors who attend to workers daily in their practice. In Brazil, in 2007, an initiative was implemented to use epidemiological criteria to establish associations between occupational exposure and illness, the Technical Epidemiological Nexus of Social Security (NTEP), during the evaluation of eligibility for disease-related benefits[4,5]. The data supporting such associations in Brazil are found in a database called *Sistema Único de Benefícios* (SUB), which contains records of benefits provided by the Brazilian social security system[5].

Machine learning (ML) models are used in a wide range of human knowledge areas and have strongly impacted the health sciences, mainly in oncology, radiology, nuclear medicine, therapeutic targets, emergency medicine, nephrology, and ophthalmology[6]. Through ML models, it is even possible to predict the risk of biological accidents among health professionals who provide primary and emergency care[7]. One definition of ML is that it is a subdivision of Artificial Intelligence (AI) that allows a system to learn and improve automatically from experiences[7]. ML algorithms are capable of learning by trial and error and enhance their performance over time[8].

Different learning methods vary according to the type of input and output data and the problem to be analyzed[9]. Initially, they can be categorized as supervised, unsupervised, semi-supervised, and reinforcement learning algorithms[7]. Supervised learning is used when the learning function maps inputs to defined outputs. One of the most common tasks is label classification. Unsupervised learning is applied when data are not labeled. In this case, the algorithm attempts to identify underlying patterns in the data and group them according to their characteristics. An example is the clustering algorithm. Semi-supervised learning is a combination of the previous methods and combines labeled and unlabeled data. Initially, the unlabeled data are classified by the clustering model, and the labeled data are used to classify the unlabeled data. The reinforcement algorithm involves a system of rewards and penalties, where the results of actions are inferred from interactions with the environment[10].

Some authors have characterized five components of big data, which are often referred to as the "five Vs": volume, variety, velocity, veracity, and value[11,12]. Other parameters to define big data consider a diversity of resources and providers, such as social media, mobile devices, databases, and records from different sources. This requires contemporary methods of analysis and storage to deal with the exponential growth of data, considering heterogeneity and differences in data architecture, using unstructured or semi-structured data[13]. Big data analysis allows for new sources and the creation of hypotheses that would otherwise not be possible from the combination of data from diverse and non-traditional sources, the generation of links between different types of information, which contribute to the elaboration of more efficient preventive health and safety public policies, detecting causes of diseases, optimizing health service offerings, costs, and improving the quality of health care[11,13].

However, it should be noted that since big data analysis depends on data contained in electronic records and mobile devices of different natures, socially vulnerable groups may be excluded or underrepresented in these datasets due to barriers to accessing such data, associated with age, gender, race, and access to health services[8]. At this point, the application of this technique can reinforce inequalities and make them prone to non-coverage of actions and policies oriented toward protection by environmental factors and safety strategies in the workplace. Given these caveats, Stieb[8] highlights that big data has been recognized as a powerful tool for exploring the determinants of occupational and environmental health. The combination of a wide variety of data integrates occupational, genetic, behavioral, technological, political, and economic factors related to work, which interact with each other throughout an individual's life and can affect occupational health. Including diverse and intricate factors in the analysis models can improve the data closer to the true distribution of the problem[14].

The objective of this study is to identify strategies and challenges in the use of big data and artificial intelligence in occupational health, as well as practices and obstacles in the implementation of large databases and ML tools to improve the analysis and prevention of occupational diseases.

Methods

Search strategy: A scoping review was conducted using terms related to occupational health, big data, and AI to identify experience reports on the use of big data to improve information on work-related diseases. The search terms were applied to four databases: Medical Literature Analysis and Retrieval System On-line – Medline (via PubMed), Embase, BVS, and Scientific Electronic Library On-line – SciELO, covering all available years up to December 2022. The complete search strategy is described in **Table 1**. In addition, manual searches were conducted in the references cited in the returned articles to obtain additional information sources.

Table 1 Search strategy

MEDLINE	("Occupational Health"[MeSH] OR "Health, Occupational"[MeSH] OR "Hygiene, Industrial"[MeSH] OR "Employee Health"[MeSH] OR "Occupational Medicine"[MeSH] OR "Occupational Health Services"[MeSH]) AND ("Artificial Intelligence"[MeSH] OR "Machine Intelligence"[MeSH] OR "Machine Learning"[MeSH] OR "Deep Learning"[MeSH] OR "Big Data"[MeSH] OR "Information Systems"[MeSH] OR "Databases, Factual"[MeSH] OR "Health Information Systems"[MeSH]) Total articles: 75 Selected articles: 15
EMBASE	PART 1 ('occupational health'/exp OR 'occupational health') AND ('artificial intelligence'/exp OR 'artificial intelligence' OR 'machine learning'/exp OR 'machine learning' OR 'big data'/exp OR 'big data') AND 'data base' Total: 45 Selected articles: 6 PART 2 ('occupational health'/exp OR 'occupational health') AND ('artificial intelligence'/exp OR 'artificial intelligence' OR 'machine learning'/exp OR 'machine learning' OR 'big data'/exp OR 'big data' OR 'data base'/exp OR 'data base') AND (2010:py OR 2011:py OR 2012:py OR 2013:py OR 2014:py OR 2015:py OR 2016:py OR 2017:py OR 2018:py OR 2019:py OR 2020:py OR 2021:py OR 2022:py) AND ('cohort analysis'/de OR 'human experiment'/de OR 'longitudinal study'/de OR 'observational study'/de) AND 'article'/it Total articles: 409 Selected articles:15
BVS	(mh:("Inteligência Artificial")) AND (mh:("Saúde Ocupacional")) Total articles: 32 Selected articles: 20
SciELO	("saúde ocupacional" OR "saúde do trabalhador") AND ("inteligência artificial" OR "aprendizado de máquina" OR "machine learning" OR "big data") AND ("base de dados" OR "banco de dados" OR "database" Total articles: 20 Selected articles: 1

Inclusion criteria: Studies describing experiences with large databases containing information on workers' health; studies reporting experiences in obtaining information on occupational health and related factors from available information systems; and studies that use AI algorithms to improve the process of retrieving and analyzing occupational health data.

Article selection: The selection of articles was carried out independently in two stages by two researchers. Initially, the articles were screened based on their titles and abstracts. Subsequently, the selected articles were read in full for final inclusion. Conflicts in selection were resolved by consensus among the researchers.

Language restrictions and types of studies: Only articles written in Portuguese, Spanish, and English were selected. Original articles, reviews, and preprints that met the inclusion criteria were included.

Tools and software used: Reference management was performed using EndNote software.

Protocol registration: There was no protocol registration for this review.

Data extraction and synthesis: A data extraction spreadsheet created in Google Sheets was used and completed independently by two researchers. The methods employed in the synthesis of the results included qualitative analysis and categorization of extracted data according to the main identified themes.

Results

A total of 505 articles were identified through database searches: 75 from Medline, 409 from Embase, 20 from BVS, and 1 from SciELO. After screening the titles and abstracts, 43 studies were selected for full-text reading. Of these, 26 were excluded due to unavailability for full reading or not meeting the inclusion criteria described in the previous section. This resulted in 17 full-text articles that were assessed for eligibility. After full-text reading, 1 article was excluded, leaving 16 studies included in the review (**Figure 1**). The selected articles were published between 2003 and 2022 (**Figure 2**). There has been an increase in the number of publications in recent years. Between 2003 and 2014, six articles were published. Between 2015 and 2022, 10 papers were published, indicating growing interest in the application of big data and ML in occupational health research.

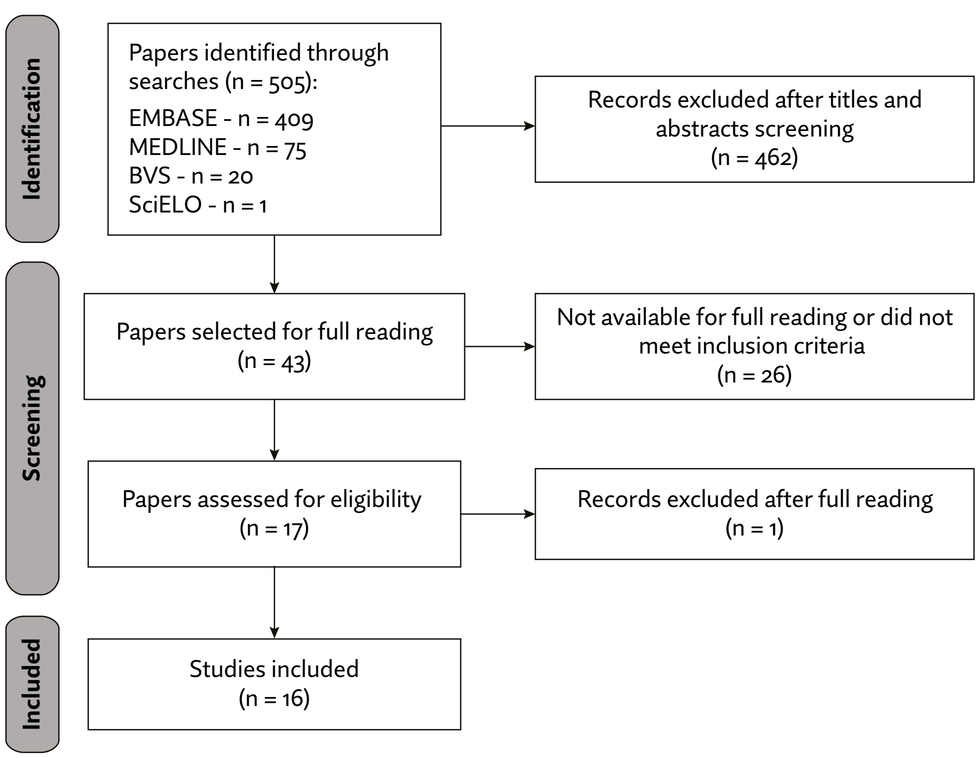

Figure 1 PRISMA Flowchart of literature search

https://doi.org/10.1590/2317-6369/40522en2024v49edcinq11

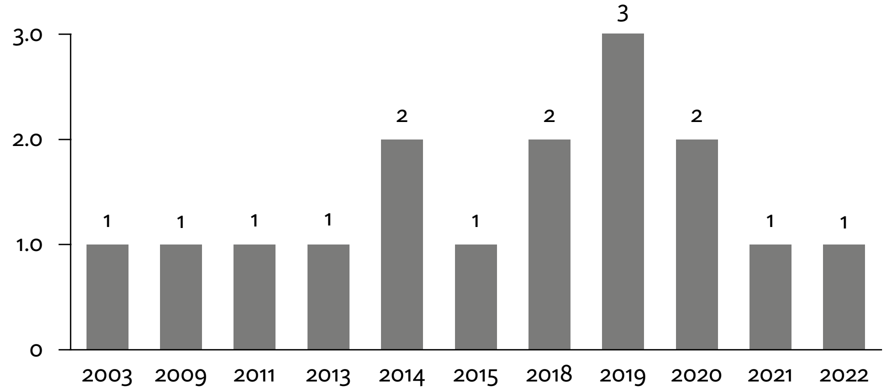

Figure 2 Number of studies selected by year of publication
Source: Articles selected by this study in Medline, Embase, BVS and SciELO databases.

Table 2 lists and describes the objectives of the articles selected by the study that use big data and ML techniques applied to occupational health.

Table 2 Selected articles

Title	Identification	Country	Study type	Is Big Data?	Summary of the study's contribution to the area of occupational health
REDECA: A Novel Framework to Review Artificial Intelligence and Its Applications in Occupational Safety and Health	Pishgar et al. 2021[6]	EUA	Review	No	The REDECA methodology is used to review articles that apply artificial intelligence to assess and prevent accident risks in the work environment in companies from five sectors (oil and gas, mining, transportation, construction, and agriculture). As AI applications across industries continue to increase, further exploration of the benefits and challenges of AI applications in occupational safety and health is needed to optimally protect worker health, safety and well-being.
Predictors of working beyond retirement in older workers with and without a chronic disease - results from data linkage of Dutch questionnaire and registry data	Wind et al. 2018[13]	Netherlands	Original	No	Aims to identify the relevant factors that explain the supply of work among individuals after retirement. These were divided into two groups, those with chronic diseases and those without chronic diseases. Factors related to health (better perception of physical health and height) and work (level of job qualification and workload) are significant to explain the supply of work beyond retirement for both groups. The greater difficulty in forecasting among workers with chronic diseases may be associated with the greater degree of heterogeneity of the group compared to their peers without chronic diseases.

Continue

 https://doi.org/10.1590/2317-6369/40522en2024v49edcinq11

Table 2 Selected articles. Continuation

Determinants of metabolic syndrome in obese workers: gender differences in perceived job-related stress and in psychological characteristics identified using artificial neural networks	Vigna et al. 2019[14]	Italy	Original	No	The study investigated the associations among gender, psychosocial variables, job-related stress, and the presence of metabolic syndrome (MS) in a cohort of obese Caucasian workers. Data analyses were performed with an artificial neural network algorithm called Auto Semantic Connectivity Map (AutoCM), using all available variables, age, BMI, waist circumference, fasting glucose, blood pressure, triglycerides, and HDL cholesterol were collected to define MS. In addition, were evaluated eating behaviors, depressive symptoms, and work-related stress. Was found a complex gender-related association between MS, psychosocial risk factors, and occupational determinants.
Prediction of pneumoconiosis by serum and urinary biomarkers in workers exposed to asbestos-contaminated minerals	Yang 2019[16]	Taiwan	Original	No	The objective of this study was to assess the diagnostic accuracy of serum and urinary biomarkers for pneumoconiosis in workers exposed to asbestos-contaminated minerals. The study used machine learning algorithms to identify markers relevant to the diagnosis of pneumoconiosis, a lung disease.
A comparison of confounder selection and adjustment methods for estimating causal effects using large healthcare databases	Benasseur et al. 2018[17]	Canada	Original	No	Study using simulations to compare the performance of different machine learning methods to reduce bias by unmeasured and measured variables in large healthcare databases. Were compared Bayesian adjustment for confounding (BAC), generalized Bayesian causal effect estimation (GBCEE), Group Lasso and Doubly robust estimation, high-dimensional propensity score (hdPS), and scalable collaborative targeted maximum likelihood algorithms. The technique helps to reduce bias in LHDs, but the results are still inconclusive.

Continue

Table 2 Selected articles. Continuation

Differential occupational risks to healthcare workers from SARS-CoV-2 observed during a prospective observational study	Eyre et al. 2020[18]	United Kingdom	Original	No	The study evaluated risk differentials for sars-cov-2 between symptomatic and asymptomatic staff of a teaching hospital. The results indicated a higher rate of positive results among professionals who work in an area with direct contact with patients, who were more likely to be contaminated. The lowest rate was observed among intensive care unit staff. Positive results were more likely in Black and Asian staff, independent of role or working location, and porters and cleaners.
Availability and accuracy of occupation in cancer registry data among Florida firefighters	McClure et al. 2019[19]	EUA	Original	No	The objective of this study is to determine the frequency and predictors of missing and inaccurate occupation data for a cohort of career firefighters in a state cancer registry. Among the registries of cancer patients, the results of the logistic regression model present a greater probability of attribution of the occupation of firefighters among men, with a more recent diagnosis, and among younger ones.
Empirical estimation of the grades of hearing impairment among industrial workers based on new artificial neural networks and classical regression methods	Farhadian et al. 2015[20]	Iran	Original	No	This study aimed to analyze the potential of artificial neural networks and logistic regression techniques for the estimation of hearing impairment among industrial workers. Five features including the age of workers, work experience, noise exposure level, smoking status, and using status of HPD were determined as the final variables to develop the prediction model. The result confirmed that neural networks could be a suitable tool for analysis of a phenomenon such as hearing loss in which required data for different variables are being collected while the mechanisms of interaction effects are complex and not fully understood.

Continue

 https://doi.org/10.1590/2317-6369/40522en2024v49edcinq11

Table 2 Selected articles. Continuation

Non-cancer morbidity among Estonian Chernobyl cleanup workers: A register-based cohort study	Rahu et al. 2014[21]	Estonia	Original	No	To examine non-cancer morbidity in the Estonian Chernobyl cleanup workers cohort compared with the population sample with special attention to radiation-related diseases and mental health disorders. Elevated morbidity in the exposed cohort was found for diseases of the nervous system, digestive system, musculoskeletal system, ischemic heart disease and for external causes. The most salient excess risk was observed for thyroid diseases, intentional self-harm and selected alcohol-related diagnoses. No obvious excess morbidity consistent with biological effects of radiation was seen in the exposed cohort, except for benign thyroid diseases.
Solbase: a databank of solutions for occupational hazards and risks	Swuste et al. 2003[22]	Spain, Italy, Ireland, Germany, the UK and The Netherlands	Original	Yes	Solbase is a database that gathers information from the work environment of different companies and processes in different sectors to map solutions for occupational accident risks. It is a jointly built base between different European countries. It is concluded that the base is a good search tool for work environment solutions, so that the information extracted can be transferred between countries, companies and branches, attributing scalability to the solutions found.
Research on workplace health promotion in the Nordic countries: a literature review, 1986-2008	Torp et al. 2011[23]	Nordic countries: Norway, Sweden, Denmark, Finland and Iceland.	Review	No	This literature review aimed to identify studies on workplace health promotion in the Nordic countries, to describe when, where and how the studies were performed and to further analyze the use of settings approaches and empowerment processes. The results identify three levels of intervention: 1) individual focused on avoiding stress with relaxation and meditation strategies; 2) role definition and peer support, and 3) interventions aimed at physical, social and organizational aspects, potential sources of stress.
Intelligent Robotics Incorporating Machine Learning Algorithms for Improving Functional Capacity Evaluation and Occupational Rehabilitation	Fong et al. 2020[27]	Canada	Review	No	This paper reviews efforts to develop robotic functional capacity evaluations (FCE) solutions that incorporate machine learning algorithms. The Rehabilitation System works by simulating more skillful maneuvers and functional tasks. Machine learning-based approaches combine the benefits of robotic systems with the expertise and experience of human therapists.

Continue

Table 2 Selected articles. Continuation

Identification and classification of high-risk groups for Coal Workers' Pneumoconiosis using an artificial neural network based on occupational histories: a retrospective cohort study	Liu et al. 2009[28]	China	Original	No	The study was performed using longitudinal retrospective data with artificial neural network to predict the probability for coal workers' pneumoconiosis (CWP) and to categorize the levels of risk. The duration of dust exposure and occupational category were the two most important factors for CWP. Coal miners at different levels of risk for CWP could be classified by the three-layer neural network analysis based on occupational history.
Ethical Considerations of Using Machine Learning for Decision Support in Occupational Health: An Example Involving Periodic Workers' Health Assessments	Dijkstra et al. 2020[30]	Netherlands	Review	Yes	The aim of this study was to explore ethical considerations and potential consequences of using ML based decision support tools (DSTs) in the context of occupational health. Were deliberated about biomedical ethical principles: respect for autonomy, beneficence, non-maleficence and justice. To minimize undesirable adverse effects, the quality of the data (completeness and reliability, including information from different actors) must be observed to reduce possible biases in the results, adhere to social responsibility when designing the models, and observe epidemiological issues around its strengths and weaknesses.
From worker health to citizen health: moving upstream	Sepulveda 2013[31]	EUA	Review	Yes	The study discusses the use of longitudinal data related to occupational health to analyze and identify opportunities to support prevention strategies, increasing the return on investments in health and safety. The value of big data lies in the generalization of results by attributing scalability to interventions. The interventions can achieve better outcomes from the multisource data.
Quality of the data in the information system for work accidents under exposure to biological materials in Brazil, 2010 to 2015	Gomes & Caldas 2017[32]	Brazil	Original	Yes	The study aims to analyze accessibility, opportunity, and completeness as criteria for the quality of the information provided by System of Information for Notifiable Conditions (Sistema de Informação sobre Agravos de Notificação – SINAN) on Work Accidents under Exposure to Biological Materials (Acidentes de Trabalho com Exposição a Material Biológico – ATEMB) from 2010 to 2015. Information is accessible and timely. Despite the accessibility of the database and the relevance of its variables, SINAN-ATEMP exhibits problems in its quality that indicate an indisputable need to improve the completeness of information.

Source: Articles selected by this study in Medline, Embase, BVS and SciELO databases.

Big Data and Occupational Health

Only one article[13] reported experience with big data to generate information on occupational health. This study investigated whether demographic, socioeconomic, job characteristics, and social and health factors can be used to predict the possibility of working after retirement in workers with and without chronic diseases. They found that feeling full of life, being physically active, being taller, and experiencing less physical workload at the time of the study accurately predicted working after retirement. The study was developed with a population of 1,125 workers selected based on inclusion and exclusion criteria from a large Dutch dataset called STREAM (Study on Transitions in Employment, Ability, and Motivation),[15] containing information from 15,118 workers. Multivariate logistic regression was applied to investigate the outcome[13].

Artificial Intelligence and Occupational Health

Two articles used ML and/or deep learning techniques to predict diseases in workers. One of these studies employed an algorithm called the Auto Semantic Connectivity Map (AutoCM) to predict metabolic syndrome in 210 obese workers using demographic and anthropometric features, such as age, body mass index (BMI), waist circumference, and fasting glucose, as well as work-related variables, for example, work-related stress[14]. Information about the metrics used to evaluate the performance of the ML algorithm was not provided in this article. Another study investigated the diagnostic accuracy of serum and urinary biomarkers for pneumoconiosis in workers exposed to asbestos[16]. In this case-control study, ML algorithms were used to link clinical variables to biomarkers, serological variables, and urinary variables and generate more accurate models for the diagnosis of pneumoconiosis. The Area Under the Receiver Operating Characteristic (Auroc) of these algorithms ranged from 0.7 to 1.0[16].

Machine Learning Algorithms

We identified a total of seven AI algorithms used in two studies[14,16], namely: decision tree, extreme gradient boosting, random forests, support vector machines, generalized linear models, artificial neural networks, and the Auto Semantic Connectivity Map (AutoCM).

Data

Other studies have made important contributions to our understanding of occupational health and the application of big data and AI. A study from Canada used simulations to compare the performance of different ML methods to reduce bias caused by unmeasured and measured variables in large healthcare databases[17]. This study compared Bayesian adjustment for confounding, generalized Bayesian causal effect estimation, Group Lasso and Doubly robust estimation, high-dimensional propensity scores, and scalable collaborative targeted maximum likelihood algorithms. The proposed technique helps reduce bias in large healthcare databases; however, the results are still inconclusive[17].

Another study, from the United Kingdom, A study conducted in the United Kingdom evaluated the occupational risks of SARS-CoV-2 infection among staff at a teaching hospital. The researchers compared the rates of positive COVID-19 tests among different groups of workers. The results indicated that professionals working in areas with direct patient contact were more likely to test positive for the virus. The intensive care unit staff had the lowest rate of positive tests. The study also revealed that Black and Asian staff members, as well as porters and cleaning personnel, were more likely to test positive, regardless of their role or workplace location[18].

A study from the United States focused on the accuracy and availability of occupation data from a cancer registry of Florida firefighters. The study determined the frequency and predictors of missing and inaccurate occupation data for a cohort of career firefighters[19].

In Iran, a study aimed to analyze the potential of artificial neural networks and logistic regression techniques to estimate hearing impairment in industrial workers. The study confirmed that neural networks could be a suitable tool for analyzing the complex interactions of variables related to hearing loss[20].

In Estonia, a register-based cohort study examined non-cancer morbidity among Chernobyl cleanup workers compared with the general population. The study found elevated morbidity for diseases of the nervous, digestive, and musculoskeletal systems, and ischemic heart disease[21].

A study conducted in multiple European countries developed a database called *Solbase*, which gathers information from various work environments to map solutions for occupational accident risks. This database is a valuable tool for sharing and scaling solutions across different sectors[22].

A literature review of Nordic countries identified studies on workplace health promotion that described interventions focused on individual stress reduction, role definition, peer support, and physical, social, and organizational aspects[23].

Discussion

In our literature search, we did not find any reviews that directly explored the use of big data and ML to address occupational diseases. As a topic with limited evidence, the scoping review was adopted as a strategy to describe the current state of knowledge about workers' health, based on more accurate prediction models through the application of big data and AI[6]. This approach serves as a valid basis for formulating study strategies aimed at identifying health determinants and guiding possible improvements in the dataset of records on workplace characteristics and individuals. Lack of standardization and incomplete data are among the main barriers to the application of big data in occupational health[10,24].

Previous reviews have described the use of AI to predict occupational accidents[25,26]. Pishgar et al.[6] conducted a narrative review exploring how big data could address and predict workplace risks, concluding that this technology could be useful for discovering new relationships between diseases and exposures, revealing how people react to potential hazards, and investigating how workplace practices and laws impact workers' health. However, their review specifically focused on the application of AI in occupational safety and health in five industries, whereas our scoping review aims to provide a broader view of the use of big data and ML in the context of occupational diseases.

Our scoping review evidences examples of how AI can enhance the investigation of occupational situations, which often involve complex interactions between environmental, mental, and genetic variables. Through our literature search, we identified studies by Fong et al.[27] and Liu et al.[28] that used big data and ML to predict occupational health outcomes. Specifically, Yang[16] used ML algorithms to identify biomarkers relevant to the diagnosis of pneumoconiosis, while Farhadian et al.[20] analyzed the potential of artificial neural networks and logistic regression techniques to estimate hearing loss among industrial workers. McClure et al.[19] investigated the availability and accuracy of occupational data in cancer registries, highlighting the importance of reliable data for occupational health studies.

Big data analysis allows for a data distribution that more closely reflects the reality of the investigated phenomenon by including multiple potential predictors of worker characteristics, health, socioeconomic, and social factors[22], facilitating the identification of complex relationships and the underlying non-linear differences between the response variable and predictors. Due to the characteristics of the technique, Wind et al.[13] used big data to mitigate the difficulties associated with predicting work after retirement, especially among individuals with chronic diseases. The greater difficulty is associated with the influence of subjective issues, such as social support and work-associated autonomy, and greater heterogeneity among individuals with chronic diseases compared to those without chronic diseases. The characteristics of decision factors and individuals make the relationship between the determinants of labor supply even more intricate.

Vigna et al.[14] applied ML techniques to take a systemic approach to assess the determinants of metabolic syndrome, considering work-related factors and psychopathological variables, such as binge eating and depressive symptoms. The adopted empirical strategy aims to overcome the limitations of traditional statistical models, which have linear properties that arise in the presence of complex and non-linear relationships, characteristic of biological systems. The AutoCM method allows finding trends and associations between variables and reconstructing partial transmission channels in a complex dataset. Yang[16] employed different ML algorithms, decision trees, and neural networks to build a prediction model aimed at evaluating the accuracy of pneumoconiosis diagnosis in mining workers exposed to asbestos. The model included demographic variables, such as age and sex, and serum and urinary biomarkers. The results indicate that the data extracted from the ML models generated predictive models with high accuracy.

The inclusion in this review of only a few articles that use big data in occupational health may be associated with the scarcity of data that address worker health in a systemic manner, considering demographic, technological, socioeconomic, and environmental factors. The multiplicity of variables results from the diversity and nature of adverse events and risk factors for workers' health. Workers' health status can be affected by ergonomic problems, chemical intoxication and biological contamination, noise exposure, work accidents, interpersonal and organizational relationships, among other factors, and exposure agents related to working conditions and the environment. At this point, the lack of records on work-related health problems is the major issue. Facchini et al.[29] attributed the scarcity of information to the existing conflicts of interest, since in most countries, including Brazil, the recording of these data is carried out by the company itself, which may choose to underreport or not register to avoid sanctions by supervision and control institutions. Santana and Nobre[24] further emphasized that a lack of quality in occupational health information systems hinders the development of effective prevention strategies.

Facchini et al.[29] argue that the limitation of data availability may result from inadequate collection and non-standardization of variables, leading to inconsistency, lack of harmonization, and poor articulation between different information systems. Furthermore, there are problems such as underreporting of information and low reliability of important variables[29]. The lack of data may also be associated with the absence of guidance for the professionals responsible for records, who may not understand the significance of the data for their daily activities and may lead to errors in data entry[29]. On the other hand, the scarcity of studies on big data in occupational health may be associated with limited knowledge about available data, which is sometimes dispersed across different data repositories[29].

The application of big data to health science may face technical and ethical barriers when dealing with decentralization and data sharing across different repositories. Dijkstra et al.[30] discuss the ethical considerations and potential consequences of using machine-learning-based decision support tools in occupational health, emphasizing the importance of data quality, adherence to social responsibility when designing models, and observance of epidemiological issues. Another problem is that the partitioning of records can reduce the volume of information, resulting in insufficient data for building big data. Issues related to the absence of periodic and systematic collection schedules and poor quality records on workers' health hinder the construction and maintenance of large datasets, limiting the application of predictive models with ML[23].

Sepulveda[31] emphasized the importance of using longitudinal data related to occupational health to analyze and identify opportunities to support prevention strategies, thus increasing the return on investments in health and safety. The value of big data lies in the generalization of results, attributing scalability to interventions, and allowing them to achieve better outcomes from multi-source data. Gomes and Caldas[32] investigated the quality of data in the Brazilian information system for work accidents involving exposure to biological materials. Despite the accessibility of the database and the relevance of its variables, the researchers found that the system presented problems with data quality, indicating a clear need to improve the completeness of information. The study highlights the importance of ensuring data quality in occupational health information systems to enable effective analysis and decision-making.

These data quality and availability challenges may explain the scarcity of studies applying ML to occupational health data analysis. Moreover, many of the existing studies have limitations in their data structures. For example, the analysis of metabolic syndrome conducted by Vigna et al.[14], which is not necessarily a work-related disease, uses cross-sectional data. Yang's[16] study on asbestosis, a work-related disease, is retrospective. While these approaches are valuable, they limit the ability to establish causal relationships and track changes over time. Sociodemographic, clinical, and laboratory variables were also included in the model for disease prediction in the study by Yang et al.[16]. This approach aligns with the proposition of systemic evaluation based on internal and external factors that interact as determinants of worker health.

Conclusions

The application of big data and ML in occupational health has the potential to improve our understanding of the complex interactions between work-related factors and health outcomes. However, there are important challenges to be addressed. These include the lack of standardization and quality of occupational health data, ethical considerations in the use of sensitive health information, and the need for more comprehensive and longitudinal datasets. Future research should focus on developing strategies to overcome these barriers and harness the potential of big data and AI to advance occupational health and safety. Collaborative efforts between researchers, industry players, and policymakers are essential to ensure responsible and effective use of these technologies for the benefit of workers worldwide.

References

1. World Health Organization. Risks in ambient work environment. Geneve: World Health Organization; 2024 [cited 2024 June 3]. Available from: https://www.who.int/tools/occupational-hazards-in-health-sector/risks-in-ambient-work-environment
2. Rushton L. The global burden of occupational disease. Curr Environ Health Rep. 2017;4(3):340-8. https://doi.org/10.1007/s40572-017-0151-2
3. Fløvik L, Knardahl S, Christensen JO. The effect of organizational changes on the psychosocial work environment: changes in psychological and social working conditions following organizational changes. Front Psychol. 2019;10:2845. https://doi.org/10.3389/fpsyg.2019.02845
4. Brasil. Câmera dos Deputados. Decreto nº 3.048, de 6 de maio de 1999. Aprova o Regulamento da Previdência Social, e dá outras providências. Diário Oficial União, 1999 May 12.
5. Groto AD, Perlin CM, Andrade SM, Salamanca MAB. Uso da inteligência artificial para predição de acidentes de trabalho com materiais biológicos em profissionais da saúde. Res Soc Dev. 2021;10(12):e93101219743-e. https://doi.org/10.33448/rsd-v10i12.19743
6. Pishgar M, Issa SF, Sietsema M, Pratap P, Darabi H. REDECA: a novel framework to review artificial intelligence and its applications in occupational safety and health. Int J Environ Res Public Health. 2021 Jun;18(13):6705. https://doi.org/10.3390/ijerph18136705
7. Taiwo Oladipupo A. Types of machine learning algorithms. In: Yagang Z, editor. New advances in machine learning. Rijeka: IntechOpen; 2010. p. 19-48.
8. Stieb DM, Boot CR, Turner MC. Promise and pitfalls in the application of big data to occupational and environmental health. BMC Public Health. 2017:372. https://doi.org/10.1186/s12889-017-4286-8
9. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform. 2018 Apr;112:59-67. https://doi.org/10.1016/j.ijmedinf.2018.01.007
10. Nguyen TL. A framework for five big v's of big data and organizational culture in firms. 2018 IEEE International Conference on Big Data (Big Data); 2018 Dec 10-13; Seattle, USA. IEEE; 2018. p. 5411-3.
11. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: a systematic review. Int J Med Inform. 2018 Jun;114:57-65. https://doi.org/10.1016/j.ijmedinf.2018.03.013
12. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digital Med. 2020;3(1):1-7. https://doi.org/10.1038/s41746-020-00323-1

13. Wind A, Scharn M, Geuskens GA, Beek AJ, Boot CR. Predictors of working beyond retirement in older workers with and without a chronic disease-results from data linkage of Dutch questionnaire and registry data. BMC Public Health. 2018 Feb;18:265. https://doi.org/10.1186/s12889-018-5151-0

14. Vigna L, Brunani A, Brugnera A, Grossi E, Compare A, Tirelli AS, et al. Determinants of metabolic syndrome in obese workers: gender differences in perceived job-related stress and in psychological characteristics identified using artificial neural networks. Eat Weight Disord. 2019 Feb;24(1):73-81. https://doi.org/10.1007/s40519-018-0536-8

15. Ybema JF, Geuskens G, Van den Heuvel S, Wind A. Study on transitions in employment, ability and motivation (STREAM): the design of a four-year longitudinal cohort study among 15,118 persons aged 45 to 64 years. Hoofddorp: TNO; 2011.

16. Yang H-Y. Prediction of pneumoconiosis by serum and urinary biomarkers in workers exposed to asbestos-contaminated minerals. PLoS One. 2019;14(4):e0214808. https://doi.org/10.1371/journal.pone.0214808

17. Benasseur I, Talbot D, Durand M, Holbrook A, Matteau A, Potter BJ, et al. A comparison of confounder selection and adjustment methods for estimating causal effects using large healthcare databases. Pharmacoepidemiol Drug Saf. 2022 Apr;31(4):424-33. https://doi.org/10.1002/pds.5403

18. Eyre DW, Lumley SF, O'Donnell D, Campbell M, Sims E, Lawson E, et al. Differential occupational risks to healthcare workers from SARS-CoV-2 observed during a prospective observational study. Elife. 2020 Aug;9:e60675. https://doi.org/10.7554/eLife.60675.

19. McClure LA, Koru-Sengul T, Hernandez MN, Mackinnon JA, Schaefer Solle N, Caban-Martinez AJ, et al. Availability and accuracy of occupation in cancer registry data among Florida firefighters. PloS One. 2019;14(4):e0215867. https://doi.org/10.1371/journal.pone.0215867

20. Farhadian M, Aliabadi M, Darvishi E. Empirical estimation of the grades of hearing impairment among industrial workers based on new artificial neural networks and classical regression methods. Indian J Occup Environ Med. 2015 May-Aug;19(2):84-9. https://doi.org/10.4103/0019-5278.165337

21. Rahu K, Bromet EJ, Hakulinen T, Auvinen A, Uusküla A, Rahu M. Non-cancer morbidity among Estonian Chernobyl cleanup workers: a register-based cohort study. BMJ open. 2014;4(5):e004516. https://doi.org/10.1136/bmjopen-2013-004516

22. Swuste P, Hale A, Pantry S. Solbase: a databank of solutions for occupational hazards and risks. Ann Occup Hyg. 2003 Oct;47(7):541-7. https://doi.org/10.1093/annhyg/meg056

23. Torp S, Eklund L, Thorpenberg S. Research on workplace health promotion in the Nordic countries: a literature review, 1986–2008. Glob Health Promot 2011;18(3):15-22. https://doi.org/10.1177/1757975911412401

24. Santana VS, Nobre L, editors. Sistemas de informação em saúde do trabalhador; 3a Conferência Nacional de Saúde do Trabalhador; 2005. Brasília, DF: Ministério da Saúde, Ministério do Trabalho e Emprego, Ministério da Previdência e Assistência Social; 2005. (Série D. Reuniões e Conferências).

25. Khairuddin MZF, Hasikin K, Abd Razak NA, Lai KW, Osman MZ, Aslan MF, et al. Predicting occupational injury causal factors using text-based analytics: a systematic review. Front Public Health. 2022;10:984099. https://doi.org/10.3389/fpubh.2022.984099

26. Barker TT. Finding pluto: an analytics-based approach to safety data ecosystems. Saf Health Work. 2021;12(1):1-9. https://doi.org/10.1016/j.shaw.2020.09.010

27. Fong J, Ocampo R, Gross DP, Tavakoli M. Intelligent robotics incorporating machine learning algorithms for improving functional capacity evaluation and occupational rehabilitation. J Occup Rehabil. 2020 Sep;30(3):362-70. https://doi.org/10.1007/s10926-020-09888-w

28. Liu H, Tang Z, Yang Y, Weng D, Sun G, Duan Z, et al. Identification and classification of high risk groups for Coal Workers' Pneumoconiosis using an artificial neural network based on occupational histories: a retrospective cohort study. BMC Public Health. 2009;9:366. https://doi.org/10.1186/1471-2458-9-366

29. Facchini LA, daCosta Nobre LC, Faria NMX, Fassa AG, Thumé E, Tomasi E, et al. Occupational Health Information System: challenges and perspectives in the Brazilian Unified Health System (SUS). Cienc Saude Coletiva. 2005 Dec;10(4):857-67. https://doi.org/10.1590/S1413-81232005000400010

30. Six Dijkstra MW, Siebrand E, Dorrestijn S, Salomons EL, Reneman MF, Oosterveld FG, et al. Ethical considerations of using machine learning for decision support in occupational health: an example involving periodic workers' health assessments. J Occup Rehabil. 2020;30(4):343-53. https://doi.org/10.1007/s10926-020-09895-x

31. Sepulveda M-J. From worker health to citizen health: moving upstream. J Occup Environ Med. 2013;55(12 suppl):S52-7. https://doi.org/10.1097/JOM.0000000000000033

32. Gomes SCS, Caldas AdJM. Quality of the data in the information system for work accidents under exposure to biological materials in Brazil, 2010 to 2015. Rev Bras Med Trab. 2017;15(3):200-28. https://doi.org/10.5327/Z1679443520170036

Authors' contributions: Fagundes TP, Wichmann RM, Silva MEL, Oliveira TA contributed to the conception of the study and the writing of the manuscript. All authors participated in the critical review, and approved the final version, and they assume full responsibility for all aspects of the study.

Data availability: The dataset supporting the results of this study is available at SciELO Data, at https://doi.org/10.48331/scielodata.7FNOXU.

Funding: The authors declare that this study did not receive funding.

Competing interests: The authors declare no conflicts of interest.

Presentation at a scientific event: The authors declare that this study has not been presented at any scientific event.

Received: November 8, 2022
Reviewed: February 1, 2023
Approved: April 28, 2023

Editor-in-Chief:
José Marçal Jackson Filho