Machine learning models in health prevention and promotion and labor productivity: A co-word analysis

Sergio Arturo Domínguez Miranda¹, Roman Rodriguez-Aguilar²

¹ Facultad de Ciencias Económicas y Empresariales, Universidad Panamericana, Augusto Rodin 498, 03920, Mexico City, México.
² Facultad de Ciencias Económicas y Empresariales, Universidad Panamericana, Augusto Rodin 498, 03920, Mexico City, México.
Email: 0246533@up.edu.mx
Corresponding author.

ABSTRACT
Objective. This article aims to carry out a co-word study on the application of machine learning models in health prevention and promotion and its effect on labor productivity.
Design/Methodology/Approach. The analysis of the relevant literature on the proposed topic, identified in the last 15 years in Scopus, is considered. Articles, books, book chapters, editorials, conference papers, and reviews of refereed publications were considered. A thematic mapping analysis was performed using factor analysis and strategy diagrams to derive primary research approaches and identify frequent themes and thematic evolution.
Results/Discussion. The results of this study show the selection of 87 relevant publications with an average annual growth rate of 23.25% in related production. The main machine learning algorithms used, the main research approaches, and key authors derived from the analysis of thematic maps were identified.
Conclusions. This study emphasizes the importance of using co-word analysis to understand trends in research on the impact of health prevention and promotion on labor productivity. The potential benefits of using machine learning models to address this issue are highlighted and anticipated to guide future research on improving labor productivity through prevention and health promotion.
Originality/Value. Identifying the relationship between work productivity and health prevention and promotion through machine learning models is relevant, but little has been analyzed in recent literature. The analysis of co-words allows us to establish the reference point of the state of the art in this regard and future trends.
Keywords: co-word analysis; research trends; bibliometrics; machine learning models; health prevention and promotion; labor productivity.
INTRODUCTION

Non-communicable diseases (NCDs) stand as the leading causes of death and a major public health concern globally (Córdova-Villalobos et al., 2008). The rise in NCDs is closely associated with an acute presentation of dyslipidemia, constituting one of the primary risk factors along with smoking, sedentary activities, improper nutrition, and genetic factors that contribute to the potentiation of diseases such as metabolic syndrome (diabetes, hypertension, and obesity), oncological, cardiological, and neurological conditions. Improving social conditions and the adoption of prevention and health promotion strategies such as diet quality, body weight, smoking cessation, and increasing physical activity can significantly alleviate the disease burden of these conditions. The evolution of technology and its implementation has revolutionized our perception of the world, positively impacting the health sector through a range of possible applications with a strong emphasis on remote prevention and diagnosis, aiming to reduce inequities in access to healthcare and the prevalence of non-communicable diseases (NCDs) (Dominguez-Miranda & Rodriguez-Aguilar, 2022).

Considering these elements, emphasis has been placed on labor productivity, various sectors have proposed strategies focused on maintaining the well-being of human resources and avoiding negative impacts on corporate objectives, seeking to implement preventive and health promotion processes. Haldar and Mallik (2010) state that better health improves workforce productivity by reducing disability, frailty, and the number of days lost due to sick leave. Furthermore, it contributes to shaping production with any combination of skills, physical capital, and technological knowledge (Schultz, 1997).

However, analyzing the feasibility and efficiency of NCD prevention initiatives is challenging because the various strategies implemented globally require substantial investments and implementation time can be prolonged, making it difficult to deduce their impact on labor productivity. An increasingly common alternative involves the use of technological resources focused on discovering patterns in data through machine learning, which has revolutionized analysis, modeling, and decision-making. However, the landscape of machine learning models is diverse. Therefore, the objective of this article is to carry out a co-word study on the application of machine learning models in health prevention and promotion, and its effect on labor productivity. The structure of the work is as follows. The first section describes the conceptual framework for NCDs, prevention, and promotion of health, and labor productivity. As well as the application of machine learning models in the analysis of these concepts. The second section addresses the design of the analysis. Subsequently, the analysis of results and conclusions is presented.

THEORETICAL FRAMEWORK

Non-communicable diseases, health prevention and promotion, and labor productivity

Non-communicable diseases (NCDs) account for approximately 70% of all deaths worldwide and are defined as chronic diseases, including heart diseases, strokes, cancer, chronic respiratory diseases, and diabetes (WHO, 2021). NCDs refer to a group of diseases not primarily caused by acute infection, resulting in health consequences that often necessitate long-term treatment and care. It is crucial to note that most of these diseases arise from the interaction of genetic and environmental factors. Predisposition to NCDs becomes evident when individuals are exposed to an unfavorable lifestyle, characterized by increased calorie consumption, simple sugars, fats, and reduced physical activity (Aguilar, 1999).

As the global population ages and the pace of life and work accelerates, the incidence of chronic diseases increases year by year, accompanied by a rise in chronic disease expenditures (Baig et al., 2017). Dietary habits leading to high triglyceride levels and abdominal obesity are key risk factors that escalate the risk of metabolic syndrome, significantly increasing the likelihood of severe heart diseases and strokes and diminishing the quality of life (Finkelstein et al., 2015; Lee et al., 2020).

Pan American Health Organization (OPS, 2022) estimates a projected expenditure of $47 trillion USD from 2010 to 2030 on NCDs.
treatment. This equates to a 48% loss of the annual global GDP, impacting 4% of the annual GDP in medium to low-income countries. Studies reveal that out-of-pocket expenses represent 20% to 30% of income for approximately 60% of cases reviewed (Jaspers et al., 2015). In the United States, cardiovascular diseases result in a productivity loss ranging from $8,539 to $10,175 USD per person per year, while diabetes mellitus incurs an annual cost of $1,962,314 USD with a 49% premature mortality rate. South Korea also faces productivity losses ranging from $171,157 to $537,745 USD per person per year due to strokes. This evidence underscores the need for a health system that monitors individuals with chronic diseases, emphasizing a proactive approach through continuous remote health monitoring for employees (Chaker et al., 2015).

Governments and organizations seek innovative methods for early patient care, acknowledging that late-stage care incurs significantly higher costs. Currently, only regular medical institutions can provide systematic health tests for chronic diseases. However, due to the large population base, existing medical resources fall short of meeting people’s medical health needs, prompting researchers and organizations to explore new intervention routes (Zhang et al., 2017).

The effect of a working-age population with poor health results in absenteeism, presenteeism, poor job performance, and decreased productivity of workers. Research on the impact of health promotion and prevention in the working-age population indicates that employees with low levels of physical activity and sedentary behavior are less productive, exhibit greater presenteeism, have reduced work capacity, and are more prone to getting sick (Rongen et al., 2013). Therefore, the need to delve deeper into the advantages of health promotion and prevention and its effect on work productivity is evident. Although sufficient evidence was not found on the effects of preventive health schemes on work productivity, it can be inferred that the impact may be positive.

At a global level, a concept of great relevance has been defined for understanding and applying different health models. This concept is known as health promotion/prevention. Both prevention and health promotion involve elements to prevent or delay health problems and reduce the burden of disease. Activity and interest in the fields of workplace health promotion and prevention, awareness of healthcare costs and presenteeism, as well as an aging population are contributing to increased demand for health promotion/prevention, particularly for working populations (Chapman, 2005). However, doubts about the economic performance associated with these efforts are a constant reality. Nutbeam and Muscat (2021) argue that the concept of health promotion involves the process that allows individuals to increase control over their health and improve it. It covers all political and social strategies aimed at modifying and improving health conditions, not only for individuals but also for communities, generating a positive impact on public and personal health (Arco-Canoles, 2019). Thus, health promotion in the work environment represents a vital comprehensive social and political process, understood as policies and activities developed to help workers and employers improve and control their health, promoting the productivity and competitiveness of organizations (Sánchez et al., 2018, Suárez et al., 2023, Cano et al., 2023).

As part of the elements of productivity within the health of the working-age population, two important concepts must be clarified: absenteeism and presenteeism. Absenteeism refers to time lost from work due to changes in work status, such as a reduction in routine working hours, temporary cessation of work, job loss, or early retirement (Zhang et al., 2011). Presenteeism, on the other hand, refers to reduced productivity at work due to health problems (Mayne et al., 2004). Health risks associated with absenteeism and presenteeism include lack of physical activity, poor nutrition, high body mass index, high stress, and diabetes (Mansyur, 2021). Tompa et al., (2006) conducted a literature review on economic evaluations of workplace interventions for occupational health and safety, providing examples of published studies that credit all productivity increases to the interventions performed. Koopmanschap et al., (1995) calculated the value of lost production, considering unemployment and the possibility of someone replacing a sick employee. Rozjabek et al., (2020) carried out an empirical study.
that evaluates the relationship between physical components, activation, vitality, and general health with absenteeism, presenteeism, activity performance behavior, and loss of productivity. Showing that the body mass index had a negative effect on commitment to the company and on activity issues.

**Machine learning and health issues**

Machine learning is an application of artificial intelligence that allows a system to learn from a data-centric environment, seeking to learn through explicit programming (El Naqa & Murphy, 2015). It can turn a sample of data into a computer program capable of drawing inferences from new data sets for which it has not been previously trained. This construct is part of level two of artificial intelligence, as shown in Figure 1. In machine learning, algorithms are trained to find patterns and correlations in large data sets and to make the best decisions and predictions with a primary focus in predictive analysis. According to Mahesh (2020), Machine Learning can be understood as a set of algorithms that continuously learn, its function being to extract, process, group, and predict.

There are several ways to categorize different algorithms. Yang and Wu (2021) classified them into three basic categories: supervised learning, unsupervised learning, and reinforcement learning. Fatima and Pasha (2017) presented six methods for diagnostic tasks using ML: supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, evolutionary learning, and deep learning (Fig. 2).

In the healthcare sector, the use of machine learning algorithms to discover patterns and achieve early diagnoses has recently gained a lot of attention. Healthcare, with the exploration and adoption of personalized medicine, is using predictive analytics to tailor health screenings and modify treatments for those most likely to be successful in a specific individual. Some health insurers are also using this type of analysis to identify potential high-need/
high-use beneficiaries to provide more intensive services from anywhere (Wagner, 2014). Through large-scale data analysis from various sources, such as genetics and laboratory data, researchers can extract statistical data with the probabilities of developing certain diseases and allow the development of more effective strategies for the treatment of diseases (T-systems, 2016).

Tools that can accelerate people’s discoveries about diseases, especially non-communicable ones, increasingly focus on things called artificial intelligence. In recent years, the speed of innovation in predictive health applications has reached an all-time peak thanks to the development of new ecosystems. Major medical technology players and high-tech companies are beginning to implement solutions that enable accelerated diagnosis (Biundo et al., 2020). Some authors have carried out analyzes on health indicators derived from various databases. Rangel-Baltazat et al., (2019) used the 2016 National Health Survey database to find correlations between height-weight ratio and cardiovascular problems. Mora Brito and Herrera (2023) used the INEGI database in Mexico, using Random Forest models to understand the construct of obesity in Mexico. Bello-Chavolla et al., used Neural Networks to characterize diabetes in Mexico. On the other hand, Mhasawade and Chunara (2021) developed a machine learning-based model to understand the general risks that may exist in the United States workforce. There is a pressing need to understand health indicators to transform them into information. Given the complexity and difficulty of grouping such complex data, machine learning models can be very useful for characterizing information. However, there is little literature focused on the working population.

There is a growing interest in modeling and health based on automatic learning, but the relationship between the prevention and promotion of health and labor productivity has not been analyzed exhaustively. Because bibliometric study allows for establishing the state of the art, the trends in research, as well as the key concepts addressed, will allow for establishing a frame of reference for future studies.

**ANALYSIS DESIGN**

The Scopus database was used as the information source from 2008 - 2023, where the necessary references for the research were obtained. There were considered articles, books, book chapters, editorials, conference papers, and reviews refereed publications as part of the research, and during the analysis of the results, by filtering the information, a total of 87 documents were obtained. Subsequently, the database obtained was downloaded to carry out a bibliometric study with the use of the R application and the Bibliometrix library. Additional filtering was also carried out, seeking to combine terms into one, derived from their results, as observed in Table 1.

<table>
<thead>
<tr>
<th>•Patients</th>
<th>•Machine-learning</th>
<th>•Healthcare</th>
<th>•Middle-aged</th>
<th>•Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>Machine learning</td>
<td>Diagnosis</td>
<td>Adults</td>
<td>Algorithms</td>
</tr>
<tr>
<td>Humans</td>
<td></td>
<td>Health care delivery</td>
<td></td>
<td>Learning algorithms</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>Clinical practice</td>
<td>Adult</td>
<td>Model</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>Mental health</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients</td>
<td></td>
<td>Patient treatment</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Health care</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.** Consolidation of terms.

As the next step, a co-word analysis was developed, also known as co-occurrence analysis or co-word mapping, which is a quantitative method used in bibliometrics to explore relationships between concepts or terms within documents. This approach involves identifying frequently co-occurring terms and mapping their relationships to uncover underlying themes or patterns (Callon et al., 1991). Using co-word analysis, it can be examined the
thematic landscape of machine learning models in health prevention and promotion and allows the uncovering of the most prominent and interconnected concepts within the literature, providing valuable insights into the underlying structure and trends within the field. Mapping words is useful in co-word analysis as it’s reflected in Table 2.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Description</th>
<th>What researchers can obtain?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-occurrence Matrix</td>
<td>Creates a matrix indicating the number of times two terms co-occur within a specified context.</td>
<td>A quantitative representation of the relationships between terms, forming the basis for further analysis.</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>Identifies underlying factors or dimensions that explain the patterns of word co-occurrence, reducing the dimensionality of the data.</td>
<td>Insights into latent themes or concepts within the dataset, aiding in thematic identification and interpretation.</td>
</tr>
<tr>
<td>Cluster Analysis</td>
<td>Groups terms into clusters based on similarities in co-occurrence patterns, revealing thematic clusters within the data.</td>
<td>Identification of thematic clusters or groups of related terms, facilitating the understanding of thematic structures.</td>
</tr>
<tr>
<td>Strategic Diagrams</td>
<td>Provides visual representations of word relationships, such as MDS plots or cluster dendrograms.</td>
<td>Visual insights into the structure of the co-word network, aiding in the interpretation and communication of findings.</td>
</tr>
<tr>
<td>Network Analysis</td>
<td>Analyzes the structure of the co-word network, including centrality measures, community detection, and network visualization.</td>
<td>Insights into the organization and connectivity of terms within the co-word network, highlighting important nodes and subnetworks.</td>
</tr>
<tr>
<td>Term Mapping</td>
<td>Maps terms onto a conceptual space, often using semantic analysis techniques, to visualize relationships or similarities between terms.</td>
<td>Insights into the conceptual relationships between terms, facilitating the identification of clusters or thematic areas within the data.</td>
</tr>
<tr>
<td>Word Cloud</td>
<td>Visualizes the frequency of terms in the corpus by displaying them in varying font sizes, with more frequent terms appearing larger.</td>
<td>A visual representation of the most prominent terms in the corpus, providing an intuitive overview of the main topics or themes.</td>
</tr>
</tbody>
</table>

Table 2. Techniques used in co-word analysis. Data compiled from Leydesdorff & Rafols (2009), Golder & Macy (2011), and Roberts Jr (2000).

Many researchers have engaged in co-word analysis to identify relevant research fields across diverse areas of inquiry. Some applications of this type of study in the business field have shown their usefulness in visualization, topic detection, information intersection, and knowledge management (Gonzalez-Valiente et al., 2021, Gonzalez-Valiente, 2023), identifying leading thematic areas within the research field related to a sustainable organization (LIS, 2018), or for determining the core concepts in the domain of e-learning enabled workforce development (Cheng & Wang, 2011). In the health sector has been used for understanding the thematic evolution of medical tourism (De la Hoz-Correa et al., 2018), examining the intellectual structure of health literacy area (Baji et al., 2018), drug abuse (Varmazyar et al., 2023), or examining to what extent different co-word methods capture items related to the primary and secondary symptoms associated with major depressive disorder (Kjell et al., 2021).

In this research, it was considered the usage of 3 main elements in the co-word analysis. Term mapping for visually representing the relationships between health prevention and promotion and how is connected to productivity using machine learning models. A word cloud was used for a graphical representation of the main terms along the 87 articles where the size of each word corresponds to its frequency within the dataset. Finally, factorial analysis was used to uncover underlying relationships between the relevant terms selected. The indicators of machine learning, health, and work productivity were studied according to Figure 3. The objective is to carry out a co-word study on the application of machine learning models in health prevention and promotion, and its effect on labor productivity using the 3 techniques selected.
RESULTS

It is observed that since 2008, there is information related to the terms used; from there, the growth of scientific production has had an average annual increase of 23.25%. Figure 4 shows the variety of terms used where we combine the main terms used in the 87 articles found. This mosaic of words aids in understanding the various dimensions of the approach to scientific production found.

In more depth, we see the trend of the relevant topics in Figure 5, where it is observed that the focus on health is a first, although other terms have been of relevance in the research, this is in accordance with the largest number of terms found within the summaries of the documents generated. The number of times the terms appear in the 87 selected articles can be seen in better detail in Table 3.
On the other hand, according to Figure 6, Machine Learning is the primary focus of the 87 identified documents. It is evident that the emphasis has primarily been on health and disease-related topics. Additionally, the production has aimed to differentiate analyses by gender. Furthermore, Deep Learning stands out as the most referenced algorithm in the selected scientific production.

The way in which the themes around the use of machine learning have evolved in various aspects is observed as a great derivation in the use of algorithms to understand various behaviors. This is seen in Figure 7, where the use of these algorithms has increased its production on health and disease issues, as well as the focus on loss of productivity.
In the analysis of relationships, two approaches were pursued. The first involved utilizing all relevant words from both keywords and the titles and abstracts of the documents, as depicted in Figure 8. As an initial observation derived from this analysis, it is evident that the primary relationship lies precisely in the use of Machine Learning. Examining the clustering of sectors reveals primarily three levels of association: one focused on productivity, another on artificial intelligence algorithms, and a third on diseases. The utilization of cohort analysis in adults using artificial intelligence in productivity emerges as a significant finding among the addressed topics. On the other hand, the application of learning systems in diseases is another prominent theme highlighted in the conducted mapping.

In a subsequent analysis, as depicted in Figure 9, a more in-depth exploration of the topics considered for the analysis of the selected terminology is presented. Concerning diseases, the focus has been on cardiovascular diseases, COVID, and mental health issues. The predominant theme in these articles has been the utilization of predictive elements. The use of big data derived from population statistics and the Internet of Things has been the primary data source for the literature analysis. On the other hand, algorithms such as random forest, decision trees, adaptive boosting, Bayesian theorems, KNN, and neural networks, including convolutional ones, are the preferred choices in the analyzed documents.
Ultimately, as part of the analysis construction, upon examining the factor analysis, it is observed that the two main dimensions accumulate 71.92% of the variability according to Figure 10. Natural groupings are apparent, such as in the upper-right quadrant, where models like decision trees, adaptive boosting, and data mining have primarily been employed in health indicators for prediction. In the upper-left quadrant, algorithms like natural language processing and random forest are noticeable in electronic health records. In the lower-left quadrant, it is observed that support vector machines have been used for adult patients in clinical studies. On the other hand, the use of deep learning has been focused on elements that enhance quality of life, such as diseases like COVID in the lower-right quadrant. Lastly, the lower-right quadrant shows that some articles demonstrate the use of KNN for the classification of indicators, considering the selected terms.

The thematic mapping is observed in Figure 11. The relevant topics that have gained prominence include the study of social media, mental health, and health policies. The articles have been driven by artificial intelligence, the use of specialized algorithms in electronic health records, and productivity topics that show relevance for basic but also trending analysis engines.
Finally, the topics observed to be of less use, either due to their emergent nature or a decline in their utilization, include diagnostic imaging and COVID-related themes.

**DISCUSSIONS AND CONCLUSIONS**

There has been a significant increase in scientific production related to the selected terms, with an average annual growth rate of 23.25%. The analysis reveals a predominant focus on health-related topics, particularly in utilizing machine learning and deep learning algorithms. In addition, concerning the impact analyses on productivity, despite notable growth and considerable research relevance, it has not yet manifested a significant impact. The identified studies predominantly focus on addressing specific indicators within companies, particularly specific health indicators, with a pronounced emphasis on mental health. There exists a gap in the comprehensive treatment of non-communicable diseases as a preventive and health promotion model, aiming to minimize the impact on labor productivity.

On the other hand, machine learning emerges as the primary focus of the identified documents, with a significant emphasis on health and disease-related issues. The relationship analysis highlights three primary associations: productivity, artificial intelligence algorithms, and diseases. Notable themes include the application of learning systems in diseases, particularly cardiovascular diseases, COVID-19, and mental health issues, utilizing predictive elements and big data sources. Additionally, the utilization of cohort analysis and factor analysis reveals distinct patterns in algorithm selection for various health indicators and patient groups. Prominent topics driving scientific production include the study of social media, mental health, and health policies, driven by artificial intelligence and specialized algorithms. However, certain topics, such as diagnostic imaging and COVID-related themes, show either emergent nature or declining utilization.

The preferred algorithms in the observed documents include random forest, decision trees, adaptive boosting, Bayesian theorems, KNN, and neural networks, emphasizing their relevance in health-related analyses,
nevertheless, deep learning stands out as the most referenced algorithm, indicating a trend toward employing advanced techniques in understanding various behaviors and addressing health challenges. Through Machine Learning models, intriguing solutions to the problems outlined by several researchers can be found, offering suitable resolution models. Further research is imperative to enhance productivity indicators, and health models could play a crucial role in improving these metrics within companies.

The study carried out demonstrates a substantial increase in scientific production since 2008, particularly in health-related areas and Machine Learning. There’s a trend towards using algorithms to understand behaviors, especially in health and productivity contexts, and Deep Learning is the most prominent algorithm, but also, other algorithms like random forest and neural networks are preferred. Cohort analysis and disease-focused learning systems are significant themes as well. Predictive elements and big data are extensively used in disease research. Factor analysis reveals groupings based on algorithm types and applications.

This research underscores the importance of employing co-word analysis to grasp the nexus between health and productivity in companies. It sheds light on the potential advantages of integrating machine learning models and serves as a compass for future studies to bolster labor productivity and foster employee well-being. Moreover, the analysis highlights a remarkable expansion and diversification in scientific output within health-related domains, propelled by strides in machine learning and deep learning algorithms. The dedicated attention given to predictive analytics, leveraging big data, and thoughtfully selecting algorithms indicates a joint effort to address urgent health challenges and improve healthcare delivery and decision-making processes. These findings anticipate guiding future investigations focused on enhancing labor productivity and benefiting employees.

Conflict of interest

The authors declare that there is no conflict of interest.

Statement of data consent

The data generated during the development of this study has been included in the manuscript.

Contribution statement

Conceptualization, investigation, formal analysis, methodology, writing – original draft: Sergio Arturo Domínguez-Miranda

Conceptualization, formal analysis, methodology, supervision, validation, writing – review & editing: Román Rodríguez-Aguilar.

REFERENCES


Kumar, Y., & Mahajan, M. (2020). Recent advancement of machine learning and deep learning in the field of healthcare system. *Computational Intelligence for Machine Learning and Healthcare Informatics, 1*, 77. https://doi.org/10.15171/9783110648195-005


ORIGINAL ARTICLE

Machine learning models in health prevention...


