Safety Science 201 (2026) 107243

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/safety

Applications of artificial intelligence in the detection and prevention of assaults against healthcare Staff: A systematic review☆

Antonio J. Moreno-Moreno [a], Juan Jesús García-Iglesias [b,*], Antonio Castaño-Seiquer [c], David Ribas-Pérez [c], Javier Fagundo-Rivera [b], Azahara Ruger-Navarrete [d], Regina Allande-Cussó [e], Juan Gómez-Salgado [b,f,*]

[a] Department of Information Technology, Higher Technical School of Engineering (ETSI), University of Huelva, Huelva 21007, Spain
[b] Department of Sociology, Social Work and Public Health, Faculty of Labour Sciences, University of Huelva, Huelva 21007, Spain
[c] Department of Stomatology, Faculty of Odontology, University of Sevilla. Sevilla 41009, Spain
[d] Department of Nursing, Faculty of Health Sciences of Ceuta, University of Granada, Ceuta 51001, Spain
[e] Department of Nursing, Faculty of Nursing, Physiotherapy and Podiatry. University of Seville, Sevilla 41009, Spain
[f] Safety and Health Postgraduate Programme, Universidad Espíritu Santo, Guayaquil 092301, Ecuador

ARTICLE INFO

Keywords:
Artificial intelligence
Natural language processing
Deep learning
Workplace violence
Health personnel
Violence prevention
Risk assessment
Clinical decision support systems

ABSTRACT

Background: The rising incidence of physical, verbal, and psychological assaults on healthcare workers has become a critical occupational and public health concern. Such incidents negatively impact staff well-being, contribute to burnout, and compromise the overall quality of patient care.
Objectives: To explore how artificial intelligence (AI) techniques, including machine learning, deep learning, and natural language processing, can be applied in healthcare settings to detect, prevent, or mitigate workplace assaults against healthcare staff. Additionally, it sought to identify the most effective technical and ethical strategies for implementing AI-based interventions.
Methods: A systematic review was performed following the PRISMA 2020 guidelines. Searches were conducted between July and August 2025 in electronic databases (PubMed, Scopus, Web of Science, CINAHL, IEEE Xplorer, and Google Scholar) and AI-assisted bibliographic tools (Perplexity, SciSpace, and Elicit). Studies were screened and assessed for methodological quality using the Joanna Briggs Institute critical appraisal tools.
Results: Seventeen studies met the inclusion criteria. Evidence indicates that AI can play a substantial role in enhancing staff safety. Approaches include smart sensors for real-time detection of aggressive behaviors, deep learning models analyzing clinical notes to predict violent incidents, advanced natural language processing systems (BERT, GPT, RAG-ECE) for identifying threatening language, and interpretable machine learning algorithms (LightGBM-SHAP-ALE) for pinpointing key risk factors.
Conclusions: Predicting and preventing violence in healthcare is a complex challenge requiring multidisciplinary solutions. AI offers promising tools for early detection, accurate prediction, and effective intervention. However, implementation must consider technical limitations, algorithmic bias, privacy concerns, and the necessity of human oversight.

1. Introduction

In recent decades, assaults on healthcare staff in the form of physical, verbal, or psychological violence have increased significantly, becoming a major public health and occupational challenge (Ministerio de Sanidad, 2025; Cebrino and Portero de la Cruz, 2020; Vicente et al. 2020). This phenomenon has led institutions such as the World Health Organization (WHO) and the International Labour Organization (ILO) to

☆ This article is part of a special issue entitled: 'Emergent Research Trends in Healthcare' published in Safety Science.

* Corresponding authors at: Department of Sociology, Social Work and Public Health, Faculty of Labour Sciences, University of Huelva, Avda. Tres de Marzo, S/N, 21007 Huelva, Spain.

E-mail addresses: antonio.moreno@dti.uhu.es (A.J. Moreno-Moreno), juanjesus.garcia@dstso.uhu.es (J.J. García-Iglesias), acastano@us.es (A. Castaño-Seiquer), dribas@us.es (D. Ribas-Pérez), javier.fagundo@dstso.uhu.es (J. Fagundo-Rivera), azahara.ruger@ugr.es (A. Ruger-Navarrete), rallande@us.es (R. Allande-Cussó), salgado@uhu.es (J. Gómez-Salgado).

https://doi.org/10.1016/j.ssci.2026.107243
Received 16 October 2025; Received in revised form 1 April 2026; Accepted 14 April 2026
Available online 22 April 2026

declare violence against healthcare workers an international emergency. According to recent data, the healthcare sector has become one of the most challenging work environments, with healthcare workers facing a risk of occupational violence up to 16 times higher than that faced by workers in other sectors (Kavanagh et al. 2024).

Violence in healthcare settings takes many forms, ranging from physical assaults to a wide range of other harmful behaviours. It can be defined as any act of physical, verbal, or psychological aggression, manifested through any channel and with varying degrees of intensity, directed against any healthcare professional in the course of their duties (Boyle and Wallis, 2016). This phenomenon presents challenges for researchers due to the lack of a precise definition that encompasses its various possible manifestations. Still, there is a widespread consensus that these assaults violate fundamental rights, impacting individuals and communities (Ceballos-Vásquez et al. 2024; Vallejo-Andrada et al. 2025).

To comprehensively address the multidimensional nature of workplace violence against healthcare staff, this review integrates the actor-based Cal/OSHA typology (California Occupational Safety and Health Administration, 1995) with specific behavioural manifestations. In adapting this classification to the health context, violent incidents are first classified according to the perpetrator's relationship to the healthcare setting: Type I involves external violence, perpetrated by individuals with no legitimate connection to the healthcare institution, primarily for the purpose of committing a crime during the incident; Type II (the most frequent) involves user–professional violence, directed towards healthcare professionals by patients, family members, or visitors during the provision of healthcare services; Type III (internal or institutional violence): conflicts between employees, or hierarchical abuse; and Type IV refers to personal violence that extends to the workplace due to personal relationships of healthcare personnel outside of work.

Regardless of the perpetrator type, the actual acts of violence are operationalised into five standardised behavioural categories: (a) physical violence, involving intentional force such as hitting, pushing, grabbing, or using objects; (b) verbal violence, including insults, shouting, and humiliation; (c) threats, defined as expressions of intent to cause harm; (d) psychological harassment, encompassing bullying, intimidation, and repeated hostile behaviour; and (e) sexual harassment, referring to unwanted sexual attention or conduct. This spectrum ranges from physical assaults, such as hitting and pushing, and the use of objects as weapons, to verbal violence, such as insults, threats, and humiliation, as well as racial violence and sexual harassment (Varol Arısoy et al., 2025).

By cross-referencing who commits the violence (Cal/OSHA Types I–IV) with what specific behaviours are enacted (categories a–e), this dual-axis classification provides a precise and consistent framework for evaluating how artificial intelligence interventions detect and mitigate these varied threats.

The available evidence reveals a global pattern of high prevalence. Between 8% and 38% of healthcare workers experience physical violence during their careers, with one in five experiencing Type II violence annually (Li et al. 2020; World Health Organization WHO, 2025). In Europe, prevalence rates range from 48% to 65% (De Raeve et al. 2023; Milczarek 2010), while Spain recorded 17,070 cases in 2024 (Ministerio de Sanidad, 2025), the United States reported two nurses assaulted every hour over a three-month period in 2022 (Kavanagh et al. 2024), and similarly high rates have been observed in Latin American countries (Travetto et al. 2015; Mautong et al. 2023; Vargas et al. 2022). Underreporting remains a critical concern, with studies documenting rates of unreported incidents as high as 67.6% in Andalusia (Mayorca Yancán et al., 2013) and 73% in Spain and Italy (Martínez and Rodríguez, 2024; Veronesi et al. 2023), attributed to normalisation of violence, complex reporting processes, and perceptions of inefficacy (Minha and Sahiran, 2024).

The most common profile of healthcare professionals who are assaulted is women with a mean age of 37, predominantly nursing professionals working in emergency services or primary care (Ministerio de Sanidad, 2025). Nurses with less than three years' experience face disproportionately high rates of violent incidents (Luo et al. 2025). Emergency units, intensive care units, and psychiatric services are particularly vulnerable given the combination of long waiting times, overcrowding, the severity of the cases treated, and the emotional state of patients and their families (Lee et al. 2023; Johnson et al. 2025).

Behavioural and psychological symptoms in patients with dementia (BPSD) are an important consideration for risk management. Aggressive behaviours, such as hitting, grabbing, and shouting not only endanger the safety of staff, but also affect the quality of life of residents in care facilities (Hatano & Han 2025; Inoue and Moshnyaga, 2024).

The consequences are severe and far-reaching. At an individual level, documented reports include high rates of burnout (37.6%), depersonalisation (59.6%), low personal fulfilment (44%), anxiety, depression, and post-traumatic stress disorder, as well as a significant reduction in quality of life (Mayorca Yancán et al., 2013; Vallejo-Andrada et al., 2025). From an institutional perspective, violence can lead to increased absenteeism, poorer work performance, a deteriorating organisational climate, higher staff turnover, and increased costs related to incident management. Ultimately, it compromises patient safety and the quality of care (Ceballos-Vásquez et al. 2024; Silva-Bajaña & Quinde-Alvear 2024).

Understanding the multifaceted nature of violence against healthcare workers requires a comprehensive conceptual framework. This review is anchored in the ILO/ICN/WHO/PSI Framework Guidelines for Addressing Workplace Violence in the Health Sector (International Labour Office (ILO), International Council of Nurses (ICN), World Health Organization (WHO), Public Services International (PSI), 2002), which provides a multi-level perspective for analysing and addressing occupational violence. Consistent with the socio-ecological model, this framework recognises that workplace violence results from the interaction of individual factors (e.g., patient mental health status, substance use), interpersonal factors (e.g., communication failures, unmet expectations), organisational factors (e.g., staffing levels, waiting times, institutional culture), and societal factors (e.g., normalisation of violence, access to weapons). Artificial intelligence technologies can potentially intervene at each of these levels: predictive models can identify individual risk factors, natural language processing (NLP) systems can detect interpersonal warning signs in communication, and organisational analytics can identify systemic conditions that facilitate violence. This multi-level framing guides the interpretation of findings throughout this review.

Due to the complexity and scale of this issue, artificial intelligence-based technologies are emerging as a promising and potentially transformative solution for early identification of risk situations, prevention of incidents, and efficient management of adverse events within the healthcare environment (Meskó et al. 2018; Chen and Decary, 2020; Alowais et al. 2023; Bekbolatova et al. 2024).

Recent advances in NLP, machine learning (ML), and deep learning (DL) provide a unique opportunity to address this challenge from a variety of perspectives (Chen and Decary, 2020).

A range of AI approaches have been explored in this context. NLP-based systems have demonstrated high accuracy in identifying aggressive language (Varol Arısoy et al., 2025; Dobbins et al., 2024), while computer vision and convolutional neural networks can analyse video content in real time to detect threats (Ahmad et al. 2024). Machine learning models, including Random Forest and Support Vector Machines, have shown promising results in predicting violent behaviour in clinical settings (Menger et al. 2018; 2019; Cheng et al. 2023). However, the integration of these technologies raises significant ethical and technical challenges, including data privacy, algorithmic bias, and the need for human oversight (Gerke et al. 2020; Karimian et al. 2022; Burema et al. 2023). Additionally, model performance depends critically on training data quality and contextual adaptation (Byon et al., 2023;

Waltzman et al., 2024; Varol Arısoy et al., 2025).

Although previous reviews have examined workplace violence in healthcare (De Raeve et al. 2023; Li et al. 2020; Vicente et al. 2020) and the application of AI in health more broadly (Alowais et al. 2023; Bekbolatova et al. 2024), no systematic review has specifically integrated both fields to evaluate AI technologies applied to the detection, prediction, and prevention of assaults against healthcare staff, incorporating a formal methodological quality assessment. This gap is particularly relevant given the rapid proliferation of AI-based tools in healthcare settings and the urgent need for evidence-based guidance on their effectiveness and limitations.

In this context, the objective of this study was to determine how different branches of artificial intelligence (AI), such as machine learning, deep learning, and natural language processing can be applied in healthcare settings to detect, prevent, or mitigate assaults, whether physical, verbal, or psychological, against healthcare staff in their workplace, and which technical and ethical approaches are most effective for this purpose.

2. Methods

2.1. Study design

A systematic review was conducted following the PRISMA 2020 statement guidelines (Preferred Reporting Items for Systematic reviews and Meta-Analyses) (Page et al. 2021). The study protocol has been made publicly available on OSF, a recognized open science platform, and can be accessed via https://doi.org/10.17605/OSF.IO/CM5X3. This research was registered in PROSPERO with code CRD420261358099.

A systematic review approach was chosen, rather than a scoping review, because the objective of this study was not merely to map the breadth of available literature but to critically evaluate the operational capacity, methodological quality, and performance metrics of AI-based interventions applied to the detection and prevention of workplace violence against healthcare staff. This evaluative purpose requires the structured quality assessment (using JBI tools) and analytical rigour characteristic of a systematic review methodology.

2.2. Databases and search strategy

The search was conducted in the Pubmed, Scopus, Web of Science, CINAHL, IEEE Xplorer, and Google Scholar electronic databases using the keywords generated by the research question, following the PICO strategy (Table 1).

To improve the comprehensiveness and sensitivity when capturing published studies relevant to the study topic, a hybrid search strategy combining controlled descriptors and free-text terms was implemented. MeSH (Medical Subject Headings) descriptors were identified using the MeSH Database tool in PubMed, and were supplemented with synonymous terms and terminological variations to maximise the retrieval of relevant literature. These terms were then combined using the Boolean operators 'AND' (to connect the PICO components) and 'OR' (to group synonyms within each component), according to the following logic: (Population) AND (Intervention) AND (Problem/Context) AND (Results).

Population: 'Health Personnel'.

Intervention: 'Artificial Intelligence', 'Machine Learning', 'Deep Learning'.

Problem/Context: 'Workplace Violence', 'Violence', 'Aggression'.

SUBHEADING: 'Prevention and Control'.

For databases without MeSH terms, the search strategy was adapted by using their specific thesauri and equivalent search fields. Truncation (*) was applied to capture morphological variations of terms, and quotation marks were used to retrieve exact phrases when necessary.

Additionally, to complement traditional searches and minimise publication bias, artificial intelligence tools specialising in systematic reviews were used:

- *Perplexity Open Research Explorer (https://www.perplexity.ai)* to identify relevant grey literature and preprints through semantic search.
- *Elicit* (https://elicit.com) to detect related works through citation analysis.
- *SciSpace (*https://scispace.com*)* to perform in-depth conceptual searches and map existing evidence using natural language processing.

These AI tools were used in a complementary manner, not as substitutes, always validating the results obtained compared to traditional systematic searches to ensure the reproducibility and methodological rigour required by the PRISMA guidelines. Table 2 summarises the search strategy used, carried out between July and August 2025, for each of the databases and tools mentioned above during the search process.

Table 1
PICO format: keywords.

Population	Healthcare staff providing direct and indirect care at any level: – Medical staff (physicians, nurses, assistants) – Emergency and urgent care staff – Mental health and psychiatric staff – Primary care staff – Healthcare support staff (orderlies, front-desk administrative personnel interacting with patients and their families, etc.)
Intervention	Application of AI technologies to detect, prevent, or mitigate assaults.
Comparator	Primary comparators: Secondary comparators: – Conventional security measures without AI components – Standard protocols for the prevention of workplace violence – Comparison between different AI approaches – Pre-post implementation analysis of AI systems – Studies without control groups with temporal analysis
Outcomes	Primary outcomes (quantifiable): – Incidence rate of aggression (physical/verbal/psychological) – Time to first aggression post-implementation – Number of lost workdays due to aggression-related leaves – Diagnostic accuracy of predictive systems (sensitivity, specificity, PPV, NPV) Secondary outcomes: Implementation outcomes: – Staff perception of safety (validated scales) – Job satisfaction and burnout – Response time to incidents – Cost-effectiveness of interventions – Adherence to safety protocols – Healthcare quality indicators – Barriers and facilitators – Ethical and privacy issues – Acceptability by staff and patients
Research question	*How can different branches of artificial intelligence, such as machine learning, deep learning, and natural language processing, be applied in healthcare settings to detect, prevent, or mitigate physical, verbal, or psychological assaults against healthcare staff, and what technical and ethical approaches have proven most effective in achieving this goal?*

3. Selection criteria

The following criteria were applied in selecting the articles:

Inclusion criteria:

- Original articles published in English and Spanish.
- Type: original articles, *meta*-analyses and systematic reviews, brief/short communications, and case reports.
- Articles assessing the use of AI to detect, prevent, or combat assaults against healthcare staff of any professional category.

Table 2
Search strategy.

Database	Search strategy	Results
Pubmed	(("Health Personnel"[Mesh] OR "healthcare worker*"[tiab] OR "medical staff"[tiab] OR "health professional*"[tiab] OR "nurs*"[tiab] OR "physician*"[tiab] OR "doctor*"[tiab]) AND ("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Deep Learning"[Mesh] OR "neural network*"[tiab] OR "predictive model*"[tiab] OR "AI"[tiab] OR "computer vision"[tiab] OR "natural language processing"[tiab]) AND ("Workplace Violence"[Mesh] OR "Violence"[Mesh] OR "Aggression"[Mesh] OR "violen*"[tiab] OR "assault*"[tiab] OR "aggress*"[tiab] OR "workplace safety"[tiab] OR "occupational safety"[tiab]) AND ("Prevention and Control"[Subheading] OR "prevent*"[tiab] OR "mitigat*"[tiab] OR "reduc*"[tiab] OR "intervention*"[tiab]))	70
Scopus	TITLE-ABS-KEY((("health* personnel" OR "health* worker*" OR "medical staff" OR "health* professional*" OR "nurs*" OR "physician*" OR "doctor*") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "predictive model*" OR "AI" OR "computer vision" OR "natural language processing") AND ("workplace violence" OR "violen*" OR "assault*" OR "aggress*" OR "workplace safety" OR "occupational safety") AND ("prevent*" OR "mitigat*" OR "reduc*" OR "intervention*")))	149
Web Of Science (Sin incluir Grants Index)	TS=((("health* personnel" OR "health* worker*" OR "medical staff" OR "health* professional*" OR "nurs*" OR "physician*" OR "doctor*") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "predictive model*" OR "AI" OR "computer vision" OR "natural language processing") AND ("workplace violence" OR "violen*" OR "assault*" OR "aggress*" OR "workplace safety" OR "occupational safety") AND ("prevent*" OR "mitigat*" OR "reduc*" OR "intervention*")))	232
IEEE Xplore	((("Document Title":"health personnel" OR "Abstract":"health personnel" OR "Document Title":"health worker" OR "Abstract":"health worker" OR "Document Title":"medical staff" OR "Abstract":"medical staff" OR "Document Title":"healthcare provider" OR "Abstract":"healthcare provider" OR "Document Title":"nurse" OR "Abstract":"nurse" OR "Document Title":"nurses" OR "Abstract":"nurses" OR "Document Title":"physician" OR "Abstract":"physician" OR "Document Title":"physicians" OR "Abstract":"physicians" OR "Document Title":"doctor" OR "Abstract":"doctor" OR "Document Title":"healthcare professional" OR "Abstract":"healthcare professional")) AND ("Index Terms":"artificial intelligence" OR "Index Terms":"machine learning" OR "Index Terms":"deep learning" OR "Index Terms":"neural networks" OR "Document Title":"predictive model" OR "Abstract":"predictive model" OR "Document Title":"computer vision" OR "Abstract":"computer vision" OR "Document Title":"natural language processing" OR "Abstract":"natural language processing" OR "Document Title":"AI system" OR "Abstract":"AI system") AND ("Document Title":"workplace violence" OR "Abstract":"workplace violence" OR "Document Title":"violent" OR "Abstract":"violent" OR "Document Title":"violence" OR "Abstract":"violence" OR "Document Title":"assault" OR "Abstract":"assault" OR "Document Title":"aggression" OR "Abstract":"aggression" OR "Document Title":"aggressive" OR "Abstract":"aggressive" OR "Document Title":"workplace safety" OR "Abstract":"workplace safety" OR "Document Title":"occupational safety" OR "Abstract":"occupational safety" OR "Document Title":"security" OR "Abstract":"security") AND ("Document Title":"prevent" OR "Abstract":"prevent" OR "Document Title":"prevention" OR "Abstract":"prevention" OR "Document Title":"mitigate" OR "Abstract":"mitigate" OR "Document Title":"reduce" OR "Abstract":"reduce" OR "Document Title":"reduction" OR "Abstract":"reduction" OR "Document Title":"intervention" OR "Abstract":"intervention" OR "Document Title":"detect" OR "Abstract":"detect" OR "Document Title":"detection" OR "Abstract":"detection"))	48
CINAHL	((MH "Health Personnel+") OR (MH "Medical Staff") OR (MH "Nursing Staff") OR TI ("health* personnel" OR "health* worker*" OR "medical staff" OR "nurs*" OR "physician*" OR "doctor*" OR "healthcare provider*" OR "healthcare professional*") OR AB ("health* personnel" OR "health* worker*" OR "medical staff" OR "nurs*" OR "physician*" OR "doctor*" OR "healthcare provider*" OR "healthcare professional*")) AND ((MH "Artificial Intelligence+") OR (MH	41

(continued on next page)

Table 2 (*continued*)

Database	Search strategy	Results
	"Machine Learning") OR (MH "Deep Learning") OR (MH "Neural Networks, Computer") OR TI ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "predictive model*" OR "AI" OR "computer vision" OR "natural language processing") OR AB ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "predictive model*" OR "AI" OR "computer vision" OR "natural language processing")) AND ((MH "Workplace Violence") OR (MH "Violence") OR (MH "Assault") OR (MH "Aggression") OR (MH "Patient Abuse") OR TI ("workplace violence" OR "violen*" OR "assault*" OR "aggress*" OR "workplace safety" OR "occupational safety" OR "patient abuse" OR "verbal abuse") OR AB ("workplace violence" OR "violen*" OR "assault*" OR "aggress*" OR "workplace safety" OR "occupational safety" OR "patient abuse" OR "verbal abuse")) AND ((MH "Preventive Health Care") OR (MH "Safety") OR TI ("prevent*" OR "mitigat*" OR "reduc*" OR "intervention*" OR "safety measure*" OR "risk assessment") OR AB ("prevent*" OR "mitigat*" OR "reduc*" OR "intervention*" OR "safety measure*" OR "risk assessment"))	
Google Scholar	("artificial intelligence" OR "machine learning" OR "deep learning" OR "natural language processing" OR "computer vision" OR "neural networks" OR "predictive analytics") AND ("workplace violence" OR "workplace aggression" OR "patient violence" OR "occupational violence" OR "healthcare violence" OR "medical violence" OR "nurse violence" OR "physician violence") AND ("healthcare workers" OR "healthcare professionals" OR "medical staff" OR "nursing staff" OR "hospital staff" OR "health personnel" OR "medical personnel") AND ("prevention" OR "predict*" OR "detect*" OR "mitig*" OR "intervent*" OR "combat" OR "reduce" OR "manage*")	979
Perplexity Open Research Explorer	<PROMPT> "Find any paper or review about how artificial intelligence, in any of its areas (machine	19
SciSpace Deep Review	learning, deep learning, natural language processing, etc.), can help combat or prevent	252
Elicit	attacks (both physical and otherwise) against healthcare personnel in their workplace"	485
Search date: between July and August 2025	Total	2.275

The decision to include systematic reviews, *meta*-analyses, and narrative reviews alongside primary studies was motivated by the emerging nature of this field, where synthesised evidence provides essential context. To manage the risk of overlap, the primary studies cited in Shang et al. (2025) were cross-referenced against our independently included primary studies: no overlap was identified, as the 68 studies in their *meta*-analysis focused on general violence risk assessment tools, whereas our primary studies specifically address AI applications for detection and prevention of assaults against healthcare staff. The two narrative reviews (Kavanagh et al. 2024; Moreno-Moreno et al. 2025) contribute conceptual and contextual insights without quantitative data that could duplicate primary study results.

Exclusion criteria:

- Studies of low scientific and technical quality after applying the quality assessment tool.
- Articles not addressing the research question and not related to the objective of the review: not related to the healthcare sector, not involving AI, or not addressing the detection, prevention, or analysis of assaults on healthcare staff:
  - Studies on violence that do NOT involve healthcare staff as victims.
  - Violence exclusively between patients.
  - Domestic or sex-based violence without a healthcare context.
  - Traditional security systems without an AI component.
  - Studies on AI applied to other healthcare aspects unrelated to the safety of healthcare staff.
  - Conventional video surveillance technologies without intelligent processing.
  - Studies on cybersecurity or data security exclusively.
- Type: opinion articles, editorials, and letters to the editor/publisher.

4. Data collection and extraction

Two researchers independently carried out searches in the traditional databases and supplemented this process by using artificial intelligence-based tools (Perplexity Open Research Explorer; Elicit; SciSpace).

Following the compilation of all sources, duplicates were eliminated using the automated bibliographic manager Zotero, supplemented by Elicit's similarity detection function. The researchers independently selected articles for inclusion after reading the title and abstract, a process assisted by AI tools for the automated extraction and synthesis of relevant information according to the established PICO criteria.

Then, both authors reviewed the full text of potentially eligible studies, using SciSpace to generate structured abstracts and evidence matrices that facilitated assessment against the inclusion/exclusion criteria. The final decision to include or exclude studies was made by consensus, using comparative syntheses generated by Elicit to systematically document the reasons for exclusion. Discrepancies were resolved by a third author, who had access to both the manual assessments and the analyses generated by the AI tools, thus ensuring transparency and reproducibility in the selection process.

5. Methodological quality assessment

Two reviewers independently determined the methodological quality of the selected studies using the critical appraisal tools for studies from the Joanna Briggs Institute (JBI) at the University of Adelaide (Aromatis and Munn, 2024). These tools allow the methodological quality of a study to be assessed and to determine the extent to which a study has excluded or minimised the possibility of bias in its design, conduct, and/or analysis. The appropriate versions were used according to the type of studies selected (Tables 3, 4, 5, 6, and 7).

6. Results

6.1. Study selection and search results

The initial search strategies identified a total of 2,275 references, which were subjected to successive screening in accordance with the topic of this review. A total of 17 studies were finally selected (Fig. 1), 14 of which were quantitative and 3 qualitative.

Seventeen studies were identified, conducted in the United States (n

Table 3
Summary of methodological quality (JBI).

Type of Study (JBI)	No.	Cut-off point
Analytical cross-sectional studies[1]	**12**	6 out of 8
Diagnostic test accuracy studies,[2,3]	**2**	7 out of 9
Systematic reviews and Research syntheses[4]	**1**	9 out of 11
Textual evidence: Expert Opinion[5]	**2**	4 out of 5

[1]Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, Currie M, Qureshi R, Mattis P, Lisy K, Mu P-F. Chapter 7: Systematic reviews of etiology and risk. In: Aromataris E, Munn Z (Editors). JBI Manual for Evidence Synthesis. JBI, 2020. Available from https://synthesismanual.jbi.global.
[2]How to cite: Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.
[3]Campbell JM, Klugar M, Ding S, Carmody DP, Hakonsen SJ, Jadotte YT, White S, Munn Z. Diagnostic test accuracy: methods for systematic review and meta-analysis. Int J Evid Based Healthc. 2015;13(3):154–62.
[4]Aromataris E, Fernandez R, Godfrey C, Holly C, Kahlil H, Tungpunkom P. Summarizing systematic reviews: methodological development, conduct and reporting of an Umbrella review approach. Int J Evid Based Healthc. 2015;13(3):132–40.
[5]McArthur A, Klugarova J, Yan H, Florescu S. Innovations in the systematic review of text and opinion. Int J Evid Based Healthc. 2015;13(3):188–195.

Table 4
Scores for cross-sectional analytical studies.

Study	C1	C2	C3	C4	C5	C6	C7	C8	Score
Varol Arısoy et al., 2025	✓	✓	✓	✓	*	*	✓	✓	**6/8**
Becker et al., 2025	✓	✓	✓	✓	*	*	✓	✓	**6/8**
Byon et al., 2023	✓	✓	✓	✓	*	*	✓	✓	**6/8**
Cheng et al., 2023	✓	✓	✓	✓	✓	✓	✓	✓	**8/8**
Dobbins et al., 2024	✓	✓	✓	✓	✓	*	✓	✓	**7/8**
Hu et al., 2023	✓	✓	✓	✓	✓	✓	✓	✓	**8/8**
Lee et al., 2023	✓	✓	✓	✓	✓	*	✓	✓	**7/8**
Luo et al., 2025	✓	✓	✓	✓	✓	✓	✓	✓	**8/8**
Menger et al., 2018	✓	✓	✓	✓	✓	✓	✓	✓	**8/8**
Menger et al., 2019	✓	✓	✓	✓	✓	✓	✓	✓	**8/8**
Waltzman et al., 2024	✓	✓	✓	✓	*	*	✓	✓	**6/8**
Yi et al., 2023	✓	✓	✓	✓	✓	✓	✓	✓	**8/8**

Assessment Criteria (C1-C8):
- **C1:** Were the inclusion criteria clearly defined?
- **C2:** Were the subjects and the environment described in detail?
- **C3:** Was exposure measured in a valid and reliable way?
- **C4:** Were standard objective criteria used for measurement?
- **C5:** Were confounding factors identified?
- **C6:** Were strategies to manage confounding factors established?
- **C7:** Were the results measured in a valid and reliable way?
- **C8:** Was an appropriate statistical analysis used?

Caption:
✓ – Yes (meets criteria) * – No (does not meet) **NA** – Not applicable.

= 4), China (n = 5, including 2 studies conducted in Taiwan), Japan (n = 1), South Korea (n = 1), the Netherlands (n = 2), Turkey (n = 1), and global reviews (n = 3). In most of these studies, the healthcare staff involved were mainly nurses (five studies focused specifically on nurses), although physicians (one specific study), general healthcare staff (seven studies), and psychiatric staff (four studies) were also involved, depending on the care setting (hospital, home care, or emergencies).

The types of violence analysed, classified according to the dual-axis framework defined in the Introduction (Cal/OSHA actor-based typology cross-referenced with behavioural categories a–e), included physical violence (twelve studies), verbal violence (eleven studies), threats (eight studies), psychological and sexual harassment (four studies), bullying (three studies), escalation of conflicts, inappropriate interactions, and medical disputes as a trigger for violence (one study each). All included studies predominantly addressed Type II violence (user–professional), except for Ma et al. (2020), which focused on Type III (institutional bullying among nursing staff). It should be noted that some studies, such as Yi et al. (2023), addressed medical disputes as indirect triggers for violence rather than direct forms of aggression; these were included given their relevance to the overall violence prevention framework.

The samples and methods used varied widely, ranging from large corpora of clinical notes or EMRs (n = 7) to hospital reports (n = 2). Techniques included advanced machine learning models such as SVM, Random Forest, LightGBM, and decision trees (n = 8); deep learning approaches including CNNs, LSTMs, and transformers such as BERT and Clinical-Longformer (n = 9); rule-based natural language processing (n = 5); and hybrid or interpretable methods using tools like SHAP and ALE (n = 4).

6.2. AI performance

The performance metrics reported across the studies are summarised in Table 9. Thirteen studies reported metrics above 70% accuracy or AUC, five directly compared AI with traditional methods, and two implemented real-time detection systems.

Among predictive models, several studies showed that AI models, particularly those based on deep learning and clinical notes, can achieve high performance in predicting violence in healthcare settings. Cheng et al. (2023) achieved AUC values above 0.80 in psychiatric settings using machine learning on clinical events. Lee et al. (2023) reported an AUC of 0.90 using Random Forest in emergency departments, with patient dissatisfaction as the main predictor. Luo et al. (2025) reported an AUC of 0.761 using LightGBM-SHAP-ALE for identifying predictors of violence against junior nurses, with medical complaints, psychological demands, and professional identity as main predictors. Menger et al. (2019) evaluated their model across two hospital sites, achieving AUC values of 0.797 (site 1) and 0.764 (site 2). Hu et al. (2023) achieved an AUC of 0.684 by combining structured and unstructured data from Chinese-language EMRs, with a violence rate of 19.7% among psychiatric inpatients. Yi et al. (2023) applied text mining and machine learning (SVM, Random Forest) to predict medical disputes, achieving an AUC of 0.684. Menger et al. (2018) demonstrated that deep learning showed small but consistent improvement over classical machine learning methods.

Among NLP-based approaches, Dobbins et al. (2024) achieved an F1 score of 0.75 using Clinical-Longformer for document classification, surpassing the psychiatry team's F1 of 0.50. Varol Arısoy et al. (2025) achieved 97.97% accuracy with the RAG-ECE model for detecting violent expressions, with an F1-score of 97.67%. In the context of home care, Byon et al. (2023) identified a rate of four incidents per 10,000 home visits by applying rule-based NLP and SVM, achieving an F-measure of 0.79 when analysing 600,000 clinical notes. Becker et al. (2025) improved identification and classification of violence reports in hospital settings using NLP and machine learning. Waltzman et al. (2024) demonstrated that NLP doubled the detection of violent events compared to traditional reporting methods, although specificity remained moderate (~47%).

Two studies developed real-time detection systems. Inoue and Moshnyaga (2024) developed an innovative system using wrist acceleration sensors to detect aggressive behaviour in patients with dementia, achieving 100% accuracy in differentiating aggressive from non-aggressive behaviour, and 80% recall with 78.7% accuracy for specific types of aggression. Ma et al. (2020) developed a mobile application using CNN to detect bullying against nurses, achieving 99% accuracy for mild-moderate scenarios, with an AUC of 0.99.

6.3. Reviews and syntheses

Three included studies were reviews rather than primary research. In

Table 5
Scores for diagnostic accuracy studies.

Study	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Score
Inoue and Moshnyaga, 2024	*	✓	✓	✓	✓	✓	*	✓	NA	✓	7/9
Ma et al., 2020	✓	✓	✓	*	✓	✓	*	NA	✓	✓	7/9

Assessment Criteria (C1-C10):
- **C1: Was a consecutive or random sample of patients included?**
- **C2: Was a case-control design avoided?**
- **C3: Did the study avoid inappropriate exclusions?**
- **C4: Were the index test results interpreted without knowing the reference standard?**
- **C5: If a threshold was used, was it pre-specified?**
- **C6: Is it likely that the reference standard will correctly classify the condition?**
- **C7: Were the standard results interpreted without knowing the index test?**
- **C8: Was there an appropriate interval between the index test and the standard test?**
- **C9: Did all patients receive the same standard of care?**
- **C10: Were all patients included in the analysis?**

Table 6
Scores for systematic reviews.

Study	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	Score
Shang et al., 2025	✓	✓	✓	✓	✓	*	✓	✓	✓	✓	✓	10/11

Assessment Criteria (C1-C11):
- **C1:** Is the review question clearly defined?
- **C2:** Were the inclusion criteria appropriate?
- **C3:** Was the search strategy appropriate?
- **C4:** Were the sources and search resources adequate?
- **C5:** Were the criteria for evaluating studies appropriate?
- **C6:** Was the critical evaluation performed by two or more reviewers?
- **C7:** Were there methods to minimise errors in data extraction?
- **C8:** Were the methods for combining studies appropriate?
- **C9:** Was the likelihood of publication bias assessed?
- **C10:** Are the recommendations backed up by data?
- **C11:** Are the guidelines for further research appropriate?

Table 7
Scores for textual evidence studies.

Study	C1	C2	C3	C4	C5	C6	Score
Kavanagh et al., 2024	✓	✓	✓	✓	✓	NA	5/5
Moreno-Moreno et al., 2025	✓	✓	✓	✓	✓	NA	5/5

Assessment Criteria (C1-C6):
- **C1:** Is the source of the opinion clearly identified?
- **C2:** Is the source prestigious in the field of specialisation?
- **C3:** Are the interests of the relevant population the central focus?
- **C4:** Is the position the result of a logical analytical process?
- **C5:** Is there any reference to existing literature?
- **C6:** Can any inconsistency with the literature be logically defended?

their narrative review, Kavanagh et al. (2024) emphasise that AI can improve the early detection of violence through predictive analysis, but also highlight the challenges associated with its implementation. Similarly, Moreno-Moreno et al. (2025) highlight the significant potential of AI in prevention through early detection, emphasising the value of its integration with organisational and educational strategies. In their systematic review and *meta*-analysis of 68 studies, Shang et al. (2025) confirmed that AI tools demonstrate greater predictive accuracy than traditional risk assessment methods, providing compelling evidence for their implementation.

6.4. Methodological quality assessment

The included studies were assessed using the JBI critical appraisal tool, where they obtained medium–high scores. Table 8 shows the characteristics of each of the studies included in this review. These were classified by authors and year of publication, country, design and objective, healthcare staff involved, type of violence, sample, participants or dataset used, techniques or instruments, and main results. In addition, the results of the JBI critical appraisal tool (quality of studies) were added.

7. Discussion

A growing body of literature has focused on leveraging artificial intelligence (AI) technologies to predict, detect, and prevent physical and non-physical assaults against healthcare staff. The strengths of this research include its innovative use of clinical notes and real-time surveillance data, the development of predictive models that outperform human experts, and the integration of AI with wearable devices and cyber-physical systems. However, there are still limitations in terms of model generalisation, ethical considerations, data privacy, and balancing sensitivity and specificity in violence prediction. Additionally, challenges related to implementation in diverse healthcare settings and the need for explainable AI remain significant. This synthesis provides a critical evaluation of these issues to inform future research and practical applications.

7.1. Predictive models

As shown in Results (Table 9), predictive models achieved performance metrics ranging from AUC 0.684 to 0.90 across different healthcare contexts. The use of systematically stored electronic health records (EHRs) facilitates the implementation of scalable applications that can be transferred to different institutional environments (Menger et al. 2019). Moreover, the adoption of federated (collaborative) learning approaches enhances model performance and facilitates the secure and efficient management of data privacy across institutions (Borger et al. 2022).

However, despite the encouraging results, predictive models often lack generalisability across target populations and healthcare settings, and their effectiveness can decline significantly when implemented in different contexts. The most direct evidence comes from Menger et al.

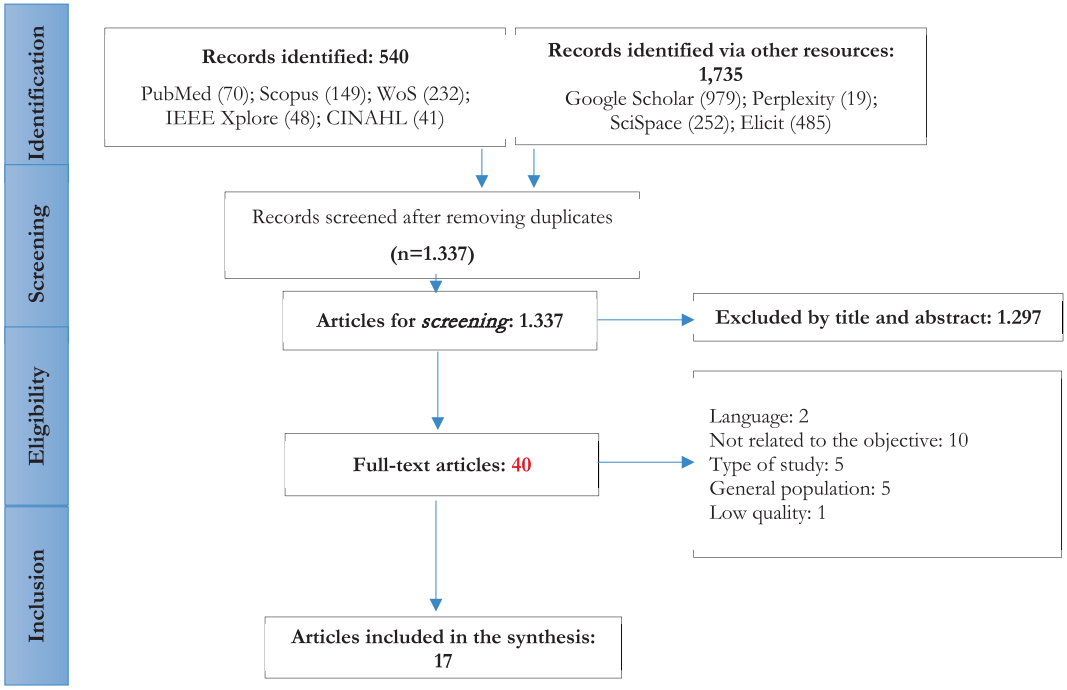

Fig. 1. Search results (Flowchart – PRISMA)(Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis Campbell Systematic Reviews, 18, e1230. https://doi.org/10.1002/cl2.1230).

(2019), who evaluated their model across two hospital sites, reporting AUC values of 0.797 and 0.764, respectively, demonstrating modest but declining performance across settings. Similarly, the diversity of reported performance metrics across studies—ranging from AUC 0.684 (Hu et al. 2023) to 0.90 (Lee et al. 2023)—reflects the influence of different populations, care contexts, and data characteristics on model performance. Notably, 15 of the 17 included studies were validated on data from a single centre, which significantly limits the generalisability of their findings. Multi-centre validation studies, federated learning approaches (Borger et al. 2022), and transfer learning techniques are needed to improve model portability across healthcare environments.

7.2. Natural language processing (NLP)

Analysis of the evidence shows that applying NLP techniques to clinical and nursing notes can effectively detect certain types of non-physical violence, such as verbal abuse and harassment, even when incidents have not been reported. This approach identifies unreported incidents that formal reporting systems might not capture (under-reporting) (Waltzman et al. 2024; Byon et al. 2023). Advanced NLP approaches, including Large Language Models (LLMs) and Named Entity Recognition (NER), can process extensive textual data to detect patterns associated with violence within healthcare settings (Varol Arısoy et al., 2025; Waltzman et al., 2024). In the context of home care, Byon et al. (2023) identified a rate of four incidents per 10,000 home visits by applying rule-based NLP and support vector machines (SVM), achieving an F-measure of 0.79 when analysing 600,000 clinical notes. Becker et al. (2025), meanwhile, made significant improvements to the identification and classification of reports of violent incidents and communication failures in hospital settings by using NLP and machine learning. Additionally, integrating these techniques into wearable devices can enhance the detection of harassment by combining the analysis of physical and emotional states with linguistic analysis (Ghosh 2024).

Although NLP models used in healthcare are highly sensitive, meaning they detect most real cases, they still exhibit moderate specificity (around 47%). In practice, it is often the case that they mislabel cases that are not actually positive, generating false positives. This low level of specificity represents a challenge in healthcare contexts where accuracy is required to prevent unnecessary alerts (Waltzman et al. 2024). In clinical environments already under significant pressure, a false positive rate exceeding 50% could generate alert fatigue, divert attention from genuine threats, and potentially undermine staff trust in these systems. To mitigate this risk, adjustable alert thresholds should be implemented, allowing institutions to calibrate the sensitivity–specificity trade-off according to their specific context. Multi-tiered alert systems (low/medium/high risk) could reduce unnecessary interruptions, and integration with existing clinical workflows, rather than standalone alert systems, is essential to minimise disruption. Despite this limitation, NLP systems still doubled the detection of violent events compared to traditional reporting methods (Waltzman et al. 2024), suggesting that even with moderate specificity, the net benefit may be positive when weighed against the known problem of severe underreporting.

Meanwhile, the use of narrative clinical data, such as handwritten or digitised medical notes and records, introduces noise and variability into analytical processes. This is because medical documentation practices are not always consistent or standardised, and may differ between professionals, hospitals, or regions. Therefore, automatic processing of this data becomes more of a challenge (Menger et al. 2019). Furthermore, large language models (LLMs) require significant computational resources and large amounts of data to function effectively, thereby limiting their implementation in environments with limited resources (Varol Arısoy et al., 2025).

7.3. Real-time detection systems and portable devices

As detailed in the Results, two studies demonstrated real-time detection capabilities with very high accuracy. This edge computing approach significantly reduces data transmission and improves response times. Ma et al. (2020) developed a mobile application that uses Convolutional Neural Network (CNN) to detect bullying against nurses. The application achieved 99% accuracy for mild-moderate scenarios, with

Table 8
Characteristics of the studies included in the systematic review.

Study	Context	Study objective	Type of study	Healthcare staff involved	Type of violence addressed	Sample/Dataset	Methods	Main findings	Quality of studies
Varol Arısoy et al., 2025	Turkey	To detect expressions of violence against healthcare workers using natural language processing techniques.	Experimental study	Physicians	Type II: verbal violence	Violent expressions using textual data	BERT, CNN, LSTM, RAG-ECE, XAI Algorithms	97.97% accuracy rate and F1-score at 97.67% with the RAG-ECE model.	6/8
Becker et al., 2025	United States	To define general categories of workplace violence and communication failures	Retrospective study	Healthcare staff in general	Type II: physical violence, verbal violence, threats	Hospital event reports	NLP, machine learning for classification	Improvement in the identification and classification of violence events and communication failures.	6/8
Byon et al., 2023	United States (Home care)	To identify Type II workplace violence from clinical notes using natural language processing	Retrospective study	Home healthcare nurses	Type II: physical violence, verbal violence, threats, psychological harassment	Nearly 600,000 clinical visit notes, 236 identified clinical notes	NLP based on rules, SVM, vocabulary list analysis	4 incidents per 10,000 home visits. *F*-measure = 0.79.	6/8
Cheng et al., 2023	China	To provide a reference for accurately predicting and preventing of the occurrence of aggressive behaviours	Prospective study	Healthcare staff – psychiatry	Type II: verbal violence, physical violence	Patients with schizophrenia	Machine learning, analysis of clinical events	Risk prediction of aggressive behaviours from patients with schizophrenia.	8/8
Dobbins et al., 2024	United States (University Hospital)	To predict violent events and treats within healthcare settings using clinical notes	Retrospective study	Healthcare staff in general	Type II: physical violence, verbal violence, threats, psychological harassment	560 admissions (280 cases, 280 controls)	Deep learning models (Clinical-Longformer), NLP, SHAP	F1 score of 0.75 for document classification model, surpassing the psychiatry team (0.5 F1).	7/8
Hu et al., 2023	China (Taiwan) (Psychiatry)	To estimate the violence rate for psychiatric inpatients and establish a predictive model	Retrospective study	Healthcare staff – psychiatry	Type II: physical violence	2064 admission records	Text mining of Chinese-language EMRs, machine learning	19.7% violence rate. The model achieved an AUC of 0.684 by combining structured and unstructured data.	8/8
Inoue and Moshnyaga, 2024	Japan (elderly care home)	Real-time detection of patient's aggressive behaviours on a smart sensor	Experimental study	Healthcare staff in general	Type II: physical violence	Patients with dementia in nursing homes	Wrist acceleration sensors, Decision tree, Long Short-Term Memory (LSTM; future)	Distinguishes aggressive from non-aggressive behaviours with 100% accuracy. Detection of specific types with 80% recall and 78.7% accuracy.	7/9
Kavanagh et al., 2024	Global review	To explore the potential of AI to prevent workplace violence by accessing almost instantly a patient's medical history	Narrative review	Nursing staff	Types I–IV: physical violence, verbal violence, psychological harassment	Scientific literature	Literature analysis on AI applications	AI can improve early detection and prevention of violence through predictive analytics.	5/5
Lee et al., 2023	South Korea (Emergency Department)	Predicting workplace violence using electronic health records	Retrospective study	Nursing staff – emergencies	Type II: verbal violence, physical violence	1215 cases of violence, 6044 non-violence	Random Forest, logistic regression, ED factor analysis	Random Forest achieved the best performance (AUC 0.90) with patient dissatisfaction as the main predictor.	7/8
Luo et al., 2025	China (90 tertiary hospitals)	To identify critical predictors of violence against junior nurses	Cross-sectional study	Nursing staff – primary care	Type II: physical violence, verbal violence, threats, sexual harassment	5663 junior nurses	LightGBM, SHAP, ALE (Accumulated Local Effect)	AUC of 0.761. Medical complaints, psychological demands, and professional identity as main predictors.	8/8
Ma et al., 2020	China (Taiwan) (Chi-Mei Hospital)	To develop an app for automatic detection and classification of nurse bullying-levels	Development study	Nursing staff	Type III: psychological harassment (bullying)	960 nurses	CNN, Rasch CAT, mobile app development	99% accuracy for mild-moderate scenario, AUC of 0.99.	7/9
Menger et al., 2018	Netherlands (University	To compare Deep Learning and classical machine	Retrospective study	Healthcare staff – Psychiatry	Type II: verbal violence, physical violence	2521 admissions of 1,796 patients	Deep learning (RNN, Word Embeddings) vs SVM, bag-of-words	Deep learning showed small but consistent improvement over classical methods.	8/8

(*continued on next page*)

Table 8 (*continued*)

Study	Context	Study objective	Type of study	Healthcare staff involved	Type of violence addressed	Sample/Dataset	Methods	Main findings	Quality of studies
	Medical Center Utrecht)	learning techniques to predict violence							
Menger et al., 2019	Netherlands (2 hospitals)	To assess the risk of violence routinely registered clinical notes	Retrospective study	Healthcare staff – Psychiatry	Type II: verbal violence, physical violence	3189 admissions on site 1; 3253 on site 2	Machine learning in EHR notes, cross-validation	AUC of 0.797 (site 1) and 0.764 (site 2). Modest generalisation over different populations.	8/8
Moreno-Moreno et al., 2025	Global review	To review the use of AI to prevent aggressions against health professionals	Narrative review	Nursing staff	Types I–IV: physical violence, verbal violence, psychological harassment	Scientific literature	Literature review on AI applications	AI shows significant potential in prevention through early detection and predictive analytics.	5/5
Shang et al., 2025	Systematic review	To assess the accuracy and applicability of predictive models for workplace violence risk	Systematic review and *meta*-analysis	Healthcare staff in general	Type II: physical violence, verbal violence, threats	68 studies with 25,980 participants	Meta-analysis of risk assessment tools	AI tools show better predictive performance than traditional methods.	10/11
Waltzman et al., 2024	United States (Paediatric hospital)	Surveillance of health care-associated violence using Natural Language Processing	Observational study	Nursing staff	Type II: physical violence, verbal violence, escalation, inappropriate family interactions	70,981 notes from the tertiary care centre for model building and internal validation	NLP for text analysis, traditional reporting methods	NLP doubled the detection of violent events compared to traditional methods.	6/8
Yi et al., 2023	China (Regional Hospital)	To predict medical disputes as a trigger for violence	Retrospective study	Physicians	Type II: medical disputes (indirect trigger for violence)	2064 psychiatric admission records	Text mining techniques, machine learning (SVM, Random Forest)	Violence rate of 19.7%. The model achieved an AUC of 0.684 using clinical notes.	8/8

Dual-axis violence classification key: Cal/OSHA actor type: Type I = external/criminal; Type II = user–professional (patient/family/visitor); Type III = internal/institutional (employee–employee); Type IV = personal relationships extending to workplace.

Behavioural categories: physical violence (a); verbal violence (b); threats (c); psychological harassment (d); sexual harassment (e).

Alphabetical list of acronyms used in Table 8.

Acronym	Description
AUC	Area Under the Curve, metrics for evaluating classification models
BERT	Bidirectional Encoder Representations from Transformers, language model based on Transformers
CNN	Convolutional Neural Network, used in image and text processing
ED	Emergency Department
EMR	Electronic Medical Record
LSTM	Long Short-Term Memory, type of recurrent neural network for time series
LightGBM	Light Gradient Boosting Machine, boosting algorithm based on decision trees
NLP	Natural Language Processing
RAG-ECE	Retrieval-Augmented Generation with Explanation Confidence Estimation, model for generating responses with explanations
RNN	Recurrent Neural Network
SHAP	Shapley Additive Explanations, technique for explaining machine learning models
SVM	Support Vector Machine, classification model
XAI	Explainable Artificial Intelligence

Table 9
Summary of AI performance metrics across included studies.

Study	AI approach	Main technique(s)	Performancemetric	Value	Comparison withtraditional methods
Predictive Models					
Cheng et al., 2023	Predictive model	Machine learning, clinical event analysis	AUC / Accuracy	> 0.80	—
Dobbins et al., 2024	Predictive model / NLP	Clinical-Longformer, SHAP	F1-score	0.75	Yes — surpassed psychiatry team (F1 = 0.50)
Hu et al., 2023	Predictive model	Text mining, ML on Chinese EMRs	AUC	0.684	—
Lee et al., 2023	Predictive model	Random Forest, logistic regression	AUC	0.90	Yes — compared RF vs logistic regression
Luo et al., 2025	Predictive model (interpretable)	LightGBM, SHAP, ALE	AUC	0.761	—
Menger et al., 2018	Predictive model	Deep learning (RNN) vs SVM	AUC	> 0.80	Yes — DL vs classical ML
Menger et al., 2019	Predictive model	ML on EHR notes, cross-validation	AUC	0.797 (site 1)0.764 (site 2)	Yes — cross-site validation
Yi et al., 2023	Predictive model	Text mining, SVM, Random Forest	AUC	0.684	—
Natural Language Processing (NLP)					
Varol Arısoy et al., 2025	NLP	BERT, CNN, LSTM, RAG-ECE, XAI	AccuracyF1-score	97.97%97.67%	—
Becker et al., 2025	NLP / ML	NLP, ML classification	Qualitative improvement	Improved classification	Yes — vs manual review
Byon et al., 2023	NLP	Rule-based NLP, SVM	F-measure	0.79	—
Waltzman et al., 2024	NLP	NLP text analysis	SensitivitySpecificity	High (~90%)Moderate (~47%)	Yes — doubled detection vs traditional reporting
Real-Time Detection Systems					
Inoue and Moshnyaga, 2024	Real-time detection	Wrist acceleration sensors, Decision tree	AccuracyRecall	100% (aggr. vs non-aggr.) 80% (specific types)	—
Ma et al., 2020	Real-time detection / NLP	CNN, Rasch CAT, mobile app	AccuracyAUC	99%0.99	—
Reviews and Syntheses					
Kavanagh et al., 2024	—	Review – Literature analysis	—	—	Conceptual comparison
Moreno-Moreno et al., 2025	—	Review – Literature review	—	—	Conceptual comparison
Shang et al., 2025	—	Review – Meta-analysis of 68 studies	Pooled metrics	AI > traditional	Yes — *meta*-analytic comparison

Abbreviations.
AUC = Area Under the Curve; F1 = F1-score (harmonic mean of precision and recall); DL = Deep Learning; ML = Machine Learning; NLP = Natural Language Processing; RF = Random Forest; SVM = Support Vector Machine; — = Not applicable/not reported.

an area under the curve (AUC) of 0.99. This approach demonstrates the potential of mobile solutions for detecting and preventing workplace violence.

8. Workflow optimisation and staff training

As reported in the Results (Reviews and syntheses subsection), the three included reviews improve the early detection of violence through predictive analysis, but also highlight the challenges associated with its implementation. Similarly, Moreno-Moreno et al. (2025) highlight the significant potential of AI in prevention through early detection, emphasising the value of its integration with organisational and educational strategies. In their systematic review and *meta*-analysis of 68 studies, Shang et al. (2025) confirmed that AI tools demonstrate greater predictive accuracy than traditional risk assessment methods, providing compelling evidence for their implementation.

8.1. Ethical and privacy considerations

The implementation of AI systems to prevent aggressions introduces critical ethical challenges that must be carefully addressed:

- Privacy and data protection: AI-based surveillance systems process sensitive biometric and behavioural data, prompting thoughtful consideration about privacy and confidentiality. Collaborative (or federated) learning is emerging as a potential solution to this issue, as it allows models to be trained without sensitive data being shared between institutions (Borger et al. 2022).
- Algorithmic bias and equity: Poorly designed algorithms can intensify existing biases and have a disproportionate impact on vulnerable groups. Therefore, it is essential to ensure that training datasets are representative and that models are regularly evaluated in order to detect and correct any discriminatory biases (Burema et al. 2023).
- Transparency and explainability: The opaque ('black box') nature of many models limits their interpretability ('unexplainable' models), which complicates their integration into clinical practice and can generate mistrust among healthcare professionals (Menger et al. 2019). To address this issue, Dobbins et al. (2024) employed SHAP (SHapley Additive exPlanations) techniques to enhance the interpretability of their models. They acknowledged the need to balance the transparency of these models with the privacy of sensitive clinical data. Luo et al. (2025), meanwhile, applied ALE (Accumulated Local Effect) techniques to better understand the model's decisions, although they cautioned against possible biases related to demographic and professional factors.
- Informed consent: The use of continuous monitoring systems raises questions about consent from patients and workers, particularly in cases involving patients with limited cognitive capacity or in emergency situations. Inoue and Moshnyaga (2024) highlight the need to monitor patients with dementia who may require alternative approaches to obtain consent. Ma et al. (2020) addressed this issue by involving nursing staff in the design process.

8.2. Real-world impact assessment

A critical observation emerging from this review is that none of the included studies directly measured a reduction in the incidence of violent incidents following the implementation of AI systems. All studies focused on technical performance metrics (AUC, F1-score, accuracy, sensitivity, specificity) or detection capabilities, rather than downstream clinical or safety outcomes such as actual reduction in assaults, staff injury rates, or improvements in perceived workplace safety. This represents perhaps the most significant gap in the current evidence base, as the ultimate validation of these tools must be measured by their capacity to improve healthcare staff safety and reduce the incidence of assaults, not solely by their algorithmic performance. Prospective intervention studies with pre-post or randomised designs that measure actual violence incidence as the primary outcome should be a priority for future research in this field.

8.3. Future lines and perspectives

As revealed by the research analysed in this systematic review, the field is evolving rapidly and offers multiple promising lines of investigation for future developments. Analysis of the literature suggests that significant advances will emerge from the convergence of various technological and methodological approaches, coupled with increased focus on ethical and environmental considerations.

One of the most promising areas of research is the integration of multiple heterogeneous data sources. Evidence suggests that combining electronic health record (EHR) information with real-time sensor data, contextual environmental information, and advanced verbal and non-verbal communication analysis may significantly improve the predictive accuracy of early warning systems for risk situations. The study by Hu et al. (2023) demonstrated that combining structured and unstructured data in predictive models substantially improves performance. Meanwhile, Dobbins et al. (2024) suggested that integrating clinical notes with surveillance data could further increase the accuracy and sensitivity of these systems. This multimodal approach marks a shift toward systems based on multiple data sources, enabling more complex architectures that can capture the multidimensional nature of situations preceding assaults.

Adaptive and personalised models are increasingly recognised as essential for the effective deployment of these technologies. Continuously adaptive systems that account for specific local contexts and patient-group characteristics can improve performance and mitigate the false-positive rates that currently hinder their clinical adoption. Menger et al. (2019) stressed the importance of models that respond to the specific context of each healthcare setting, and Cheng et al. (2023) empirically demonstrated the need to consider specific diagnostic characteristics when developing predictive models. Significant progress in transfer learning and *meta*-learning techniques is enabling pre-trained models to be rapidly adapted to new environments with limited local data. This is particularly relevant for institutions with limited resources or populations that are not typical.

Implementing these lines of research will inevitably require deeper and more structured interdisciplinary collaboration. The effective integration of artificial intelligence systems into complex healthcare environments requires the active participation of not only healthcare providers and technology experts, but also ethics specialists, patient representatives, healthcare policymakers, and regulators. This collaboration must be established in the early stages of system design to ensure that the resulting solutions are technically robust, clinically relevant, ethically acceptable, and legally compliant with the current legal framework.

Finally, the environmental sustainability of these technologies is becoming an increasingly important issue. The debate surrounding the energy consumption and environmental impact of artificial intelligence data centres is becoming more widely discussed in scientific literature and international political discourse. According to the International Energy Agency, estimates suggest that global data centre electricity demand will more than double by 2030, prompting consideration of the long-term sustainability of increasingly complex and computationally intensive AI systems. Future research will therefore need to maintain a balance between pursuing greater accuracy and predictive power and developing more energy-efficient architectures. This will involve exploring approaches such as edge computing, lighter models, and optimised inference strategies.

9. Limitations

This systematic review has several limitations that must be acknowledged when interpreting the results and drawing conclusions about the current state of knowledge in this field.

A key limitation lies in the marked methodological heterogeneity observed across the included studies. The diversity in the research designs used, ranging from retrospective to experimental and cross-sectional studies, makes quantitative synthesis and direct comparison of findings particularly challenging. This heterogeneity also applies to the studied populations, which include contexts as diverse as psychiatric units, emergency services, and home care, each with specific characteristics and risk dynamics. Similarly, the types of technologies tested vary considerably in nature and complexity, and the outcome metrics used to assess their performance are equally diverse, ranging from area under the ROC curve to F1-score and simple accuracy measures. The complexity of the methodological approach is further compounded by the heterogeneity of the subject matter itself. The studies reviewed focus on different forms of violence, ranging from direct physical aggression to medical disputes representing triggers for conflict or situations of workplace bullying among healthcare professionals. While this multiplicity of approaches reflects the richness of the field, it currently limits the ability to draw definitive conclusions.

Another important limitation that must be acknowledged is publication bias. There is a tendency to publish positive or promising results more frequently than negative ones, meaning that failed implementations may be underrepresented in the available scientific literature. This is a common phenomenon in many fields of research and can lead to the actual effectiveness of artificial intelligence technologies in this context being overestimated. It can also make it challenging to identify factors associated with implementation failures.

The evolving nature of the available evidence presents a significant opportunity. As the application of artificial intelligence in healthcare is a new and rapidly emerging field, many of the analysed technologies are still in the experimental or pilot phase. This means that evidence on their long-term effectiveness and ability to be scaled up to broader contexts is limited. Most studies report results from short-term implementations or controlled environments, which limits the ability to assess their viability and sustainability in the context of ongoing clinical use under real-world conditions.

The geographical context of the available evidence raises questions about generalisability. Most studies were conducted in high-income countries, particularly the United States and several Western European countries, as well as in East Asian countries such as China (including Taiwan), South Korea, Taiwan, and Japan. These countries all have advanced technological infrastructures and substantial resources for research and innovation. This geographical clustering greatly restricts the ability to generalise the findings to settings with limited resources, where technological infrastructure may be inadequate, access to digitised data may be restricted, and the cultural and organisational dynamics of violence towards healthcare personnel may differ.

Finally, the real impact of these technologies has not been fully assessed. While few studies currently provide robust data on the effective reduction of violent incidents following the implementation of artificial intelligence systems, the focus is often on technical evaluation metrics such as sensitivity, specificity, and predictive accuracy. The gap

between technical performance and actual clinical and organisational impact is critically limiting, given that the ultimate validity of these tools must be measured by their capacity to improve the safety of healthcare staff and reduce the incidence of assaults, rather than by their algorithmic sophistication or statistical accuracy.

10. Conclusions

This systematic review analyses studies that demonstrate the real capacity of artificial intelligence, in its various technological manifestations (including machine learning, deep learning, natural language processing, and advanced generative models), to detect, predict, and, in certain contexts, prevent physical, verbal, and psychological aggression against healthcare professionals. These systems generally exhibit moderate to high predictive performance, with metrics in several studies reaching optimal values, particularly when deep learning architectures and advanced natural language processing techniques are employed. As the evidence shows, there is a wide variety of applications, ranging from the automated analysis of clinical notes and hospital reports to the use of smart sensors and wearable devices for real-time monitoring. In many cases, these applications surpass the accuracy of traditional methods and the predictive capacity of human experts, and involve the development of interpretable models. However, significant remaining challenges include model generalisation, data representativeness, privacy protection, algorithmic bias, and fundamental ethical considerations. This is compounded by the fact that many studies are based on data from a single centre and lack external validation, which limits the extrapolation of their results to other contexts.

These findings highlight the importance of developing adaptive models that respond to specific local contexts. This approach would foster greater interdisciplinary collaboration, integrating healthcare professionals, technology experts, ethics specialists, and patient representatives. It would also involve developing specific strategies to naturally integrate these systems into everyday workflows while maintaining staff efficiency by avoiding additional burdens. Ideally, these technologies should be implemented by integrating multiple heterogeneous data sources, developing explainable models to generate user trust, and paying careful attention to the environmental sustainability and energy impact of the implemented systems. In conclusion, for artificial intelligence to fulfil its potential as an effective tool in tackling workplace violence in the healthcare sector, the technical, methodological, and ethical limitations must be addressed. This requires the establishment of robust regulatory and ethical frameworks alongside rigorous external validation of the developed models. It is also essential to guarantee that these innovations benefit all professionals and patients in a safe, transparent, equitable, and responsible manner.

CRediT authorship contribution statement

Antonio J. Moreno-Moreno: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Juan Jesús García-Iglesias:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Antonio Castaño-Seiquer:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **David Ribas-Pérez:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Javier Fagundo-Rivera:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation. **Azahara Ruger-Navarrete:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Regina Allande-Cussó:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Juan Gómez-Salgado:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Funding

This study has not received funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

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

References

Ahmad, I., Sahoo, S.R., Ratnakar Dash, 2024. Anomaly Detection in Video Surveillance for Unusual Behavior Identification. In: 2024 2nd World Conference on Communication & Computing (WCONF). https://doi.org/10.1109/WCONF61366.2024.10692100.

Alowais, S.A., Alghamdi, S.S., Alsuhebany, N.d., et al., 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med. Educ. 23 (1), 689. https://doi.org/10.1186/s12909-023-04698-z.

Aromatis, E. and Munn, Z., 2024. Joanna Briggs Institute. Manual for Evidence Synthesis. Adelaide: Australia. Accessed 1 October 2025. https://jbi.global/critical-appraisal-tools.

Becker, M., Hwang, S., Schriver, E., Douma, C., Duffy, C., Atkins, J., McShane, C., Lubken, J., Hanish, A., McGreevey, J.D., Regli, S.H., 2025. Automatically Identifying Event Reports of Workplace Violence and Communication Failures using Large Language Models. AMIA Summits on Translational Science Proceedings 2025, 74.

Bekbolatova, M., Mayer, J., Ong, C.W., Toma, M., 2024. Transformative potential of AI in healthcare: definitions, applications, and navigating the ethical landscape and public perspectives. Healthcare (Switzerland) 12 (2). https://doi.org/10.3390/healthcare12020125.

Borger, T., Mosteiro, P., Kaya, H., et al., 2022. Federated learning for violence incident prediction in a simulated cross-institutional psychiatric setting. Expert Syst. Appl. 199, 116720. https://doi.org/10.1016/j.eswa.2022.116720.

Boyle, M.J., Wallis, J., 2016. Working towards a definition for workplace violence actions in the health sector. Safety in Health 2 (1), 4. https://doi.org/10.1186/s40886-016-0015-8.

Burema, D., Debowski-Weimann, N., von Janowski, A., Grabowski, J., Maftei, M., Jacobs, M., Van Der Smagt, P., Benbouzid, D., 2023. A sector-based approach to AI ethics: Understanding ethical issues of AI-related incidents within their sectoral context. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, pp. 705–714. https://doi.org/10.1145/3600211.3604680.

Byon, H.D., Harris, C., Crandall, M., Song, J., Maxim Topaz, 2023. Identifying type II workplace violence from clinical notes using natural language processing. Workplace Health Saf. 71 (10), 484–490. https://doi.org/10.1177/21650799231176078.

California Occupational Safety and Health Administration (Cal/OSHA). 1995. «Cal/OSHA Guidelines for Workplace Security». San Francisco, CA: State of California Department of Industrial Relations, Division of Occupational Safety and Health.

Ceballos-Vásquez, P., Espinoza-Arrué, F., Solorza-Aburto, J., et al., 2024. Violencia psicológica y física percibida por trabajadores de unidades de emergencias privadas de Chile. Revista Ciencias De La Salud 22 (2), 1–16. https://doi.org/10.12804/revistas.urosario.edu.co/revsalud/a.12528.

Cebrino, J., Portero de la Cruz, S., 2020. A worldwide bibliometric analysis of published literature on workplace violence in healthcare personnel. PLoS One 15 (11), e0242781. https://doi.org/10.1371/journal.pone.0242781.

Chen, M., Decary, M., 2020. Artificial intelligence in healthcare: an essential guide for health leaders. Healthc. Manage. Forum 33 (1), 10–18. https://doi.org/10.1177/0840470419873123.

Cheng, N., Guo, M., Yan, F., et al., 2023. Application of machine learning in predicting aggressive behaviors from hospitalized patients with schizophrenia. Front. Psych. 14. https://doi.org/10.3389/fpsyt.2023.1016586.

De Raeve, P., Xyrichis, A., Bolzonella, F., Bergs, J., Davidson, P.M., 2023. Workplace violence against nurses: challenges and solutions for Europe. Policy Polit. Nurs. Pract. 24 (4), 255–264. https://doi.org/10.1177/15271544231182586.

Dobbins, N.J., Chipkin, J., Byrne, T., et al., 2024. Deep learning models can predict violence and threats against healthcare providers using clinical notes. Npj Mental Health Research 3 (1), 61. https://doi.org/10.1038/s44184-024-00105-7.

Gerke, S., Babic, B., Evgeniou, T., Glenn Cohen, I., 2020. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. npj Digital Med. 3 (1), 53. https://doi.org/10.1038/s41746-020-0262-2.

Ghosh, D., 2024. AI-Enhanced Wearable Devices Integrating Emotion Recognition for Personal Security and Natural Language Processing for Harassment Detection. In: Wearable Devices, Surveillance Systems, and AI for Women's Wellbeing. IGI Global Scientific Publishing, pp. 81–90. https://doi.org/10.4018/979-8-3693-3406-5.ch005.

Hatano, M., Han, H., 2025. Development and Implementation of an AI/IoT-Based Dementia Care Service: Challenges and Interventions. Studies in Health Technology and Informatics 329, 283–287.

Hu, Y.-H., Hung, J.-H., Li-Yu, Hu., Huang, S.-Y., Cheng-Che Shen, 2023. An analysis of chinese nursing electronic medical records to predict violence in psychiatric inpatients using text mining and machine learning techniques. PLoS One 18 (6), e0286347.

Inoue, Y., Moshnyaga, V.G., 2024. A Real-Time Detection of Patient's Aggressive Behaviors on a Smart Sensor. In: 2024 IEEE 67th International Midwest Symposium on Circuits and Systems (MWSCAS). IEEE, pp. 911–914. https://doi.org/10.1109/MWSCAS60917.2024.10658928.

International Labour Office (ILO), International Council of Nurses (ICN), World Health Organization (WHO), Public Services International (PSI). 2002. Framework Guidelines for Addressing Workplace Violence in the Health Sector. Geneva: ILO/ICN/WHO/PSI.

Johnson, G.U., Towell-Barnard, A., Ewens, B., 2025. Predicting violence in the intensive care unit: can artificial intelligence protect those who care? Intensive and Critical Care Nursing 89 (August), 104054. https://doi.org/10.1016/j.iccn.2025.104054.

Karimian, G., Petelos, E., Evers, S.M.A.A., 2022. The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review. AI Ethics 2 (4), 539–551. https://doi.org/10.1007/s43681-021-00131-7.

Kavanagh, K.T., Pontus, C., Cormier, L.E., 2024. Healthcare Violence and the potential promises and harms of artificial intelligence. J. Patient Saf. 20 (5), 307–313. https://doi.org/10.1097/PTS.0000000000001245.

Lee, H., Yun, H., Choi, M., Kim, H., 2023. Predicting workplace violence in the emergency department based on electronic health record data. J. Emergency Nurs. 49 (3), 415–424. https://doi.org/10.1016/j.jen.2023.01.010.

Li, Y.-L., Li, R.-Q., Qiu, D., Xiao, S.Y., 2020. Prevalence of workplace physical violence against health care professionals by patients and visitors: a systematic review and meta-analysis. Int. J. Environ. Res. Public Health 17 (1), 299. https://doi.org/10.3390/ijerph17010299.

Luo, L., Yuze, Wu., Li, S., Li, F., Wang, X., Wei, X., 2025. Identifying the most critical predictors of workplace violence experienced by junior nurses: an interpretable machine learning perspective. J. Nurs. Manag. 2025 (1), 5578698. https://doi.org/10.1155/jonm/5578698.

Ma, S.-C., Chou, W., Chien, T.-W., et al., 2020. An app for detecting bullying of nurses using convolutional neural networks and web-based computerized adaptive testing: development and usability study. JMIR Mhealth Uhealth 8 (5), e16747. https://doi.org/10.2196/16747.

Martínez, G.M., Rodríguez, M., 2024. Agresiones hacia el personal sanitario en España: violencias invisibilizadas, silencios inaceptables. SEMERGEN 50 (8), 102319. https://doi.org/10.1016/j.semerg.2024.102319.

Mautong, H., Camacho-Leon, G.P., Khalid, A., Huespe, I.A., Bhatt, G., Banga, A., Morera Gonzalez, L., Bhattacharjee, P., Hamza, H., Pattnaik, H. and Ramirez, S., 2023. Workplace violence against healthcare workers in Latin America: a subgroup analysis from 12 countries of the Global ViSHWaS Study. https://doi.org/10.2139/ssrn.4567893.

Mayorca Yancán, I.A., Lucena García, S., Cortés Martínez, M.E., Lucerna Méndez, M.Á., 2013. Violencia contra trabajadores sanitarios en un hospital de Andalucía:¿ Por qué hay agresiones no registradas? Medicina y seguridad del trabajo 59 (231), 235–258. https://doi.org/10.4321/S0465-546X2013000200005.

Menger, V., Scheepers, F., Spruit, M., 2018. Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. Appl. Sci. 8 (6), 981. https://doi.org/10.3390/app8060981.

Menger, V., Spruit, M., Van Est, R., Nap, E., Scheepers, F., 2019. Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. JAMA Netw. Open 2 (7), e196709. https://doi.org/10.1001/jamanetworkopen.2019.6709.

Meskó, B., Hetényi, G., Győrffy, Z., 2018. Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Serv. Res. 18 (1), 545. https://doi.org/10.1186/s12913-018-3359-4.

Milczarek, M. 2010. Workplace Violence and Harassment: A European Picture. European risk observatory report / European Agency for Safety and Health at Work. Publ. Office of the European Union. https://books.google.es/books?id=hYOLtwAACAAJ.

Minha, H.S., Sahiran, M.N., 2024. Debunking the obstacles to reporting workplace violence among the healthcare workers in Melaka, Malaysia: a Qualitative Study. Malays. J. Med. Health Sci. 20 (2), 142–150. https://doi.org/10.47836/mjmhs.20.2.19.

Ministerio de Sanidad. Gobierno de España. 2025. «Informe de Agresiones a profesionales del Sistema Nacional de Salud 2024». Ministerio de Sanidad. Gobierno de España. https://www.sanidad.gob.es/areas/profesionesSanitarias/agresiones/docs/InformeAgresiones2024_1.pdf.

Moreno-Moreno, A.J., García-Iglesias, J.J., Gómez-Salgado, J., 2025. Use of artificial intelligence to prevent aggressions against health professionals. Gaceta Médica De México 16986. https://doi.org/10.24875/GMM.25000035.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., et al., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, n71. https://doi.org/10.1136/bmj.n71.

Shang, J., Xue, K., Yang, C., Shi, H., Pan, L., Zeng, Y. 2025. Prediction Models for Health Care Workers' Exposure to Type II Workplace Violence: A Systematic Review and Meta-Analysis». Journal of Nursing Care Quality, advance online publication, 14 July. https://doi.org/10.1097/NCQ.0000000000000897.

Silva-Bajaña, J.J., Quinde-Alvear, Á.G, 2024. Riesgos psicosociales y su impacto en el clima laboral del personal Sanitario, en el área de emergencia del distrito de salud 08d05. Mqrinvestigar 8 (4), 4581–4612. https://doi.org/10.56048/MQR20225.8.4.2024.4581-4612.

Travetto, C.M., Daciuk, N., Fernández, S., et al., 2015. Agresiones hacia profesionales en el ámbito de la salud. Revista Panamericana De Salud Pública 38, 307–315.

Vallejo-Andrada, A., Sánchez-Serrano, J.L.S., Barrera-Algarín, E., Caravaca-Sánchez, F., 2025. Burnout, ansiedad, satisfacción y violencia en el trabajo en trabajadores/as sociales de los servicios sociales municipales de Sevilla (España). Alternativas. Cuadernos De Trabajo Social 32 (1), 81–105. https://doi.org/10.14198/ALTERN.25953.

Vargas, L., Vélez-Grau, C., Camacho, D., Richmond, T.S., Meisel, Z.F., 2022. The permeating effects of violence on health services and health in Mexico. J. Interpers. Violence 37 (13–14), NP10883-911. https://doi.org/10.1177/0886260521990832.

Varol Arısoy, M., Yalçınkaya, M.A., Gürfidan, R., Arısoy, A., 2025. Detection of Expressions of Violence Targeting Health Workers with Natural Language Processing Techniques. Applied Sciences 15 (4), 1715. https://doi.org/10.3390/app15041715.

Veronesi, G., Ferrario, M.M., Giusti, E.M., et al., 2023. Systematic violence monitoring to reduce underreporting and to better inform workplace violence prevention among health care workers: before-and-after prospective study. JMIR Public Health Surveill. 9, e47377. https://doi.org/10.2196/47377.

Vicente, S., Isabel, M., Rodrigo, M.T.F., Dordá, P.J.S., Pardo, F.U., 2020. Agresiones a profesionales del sector sanitario en España, revisión sistemática. Revista Española De Salud Pública 93, e201910097.

Waltzman, M., Ozonoff, A.l., Fournier, K.A., et al., 2024. Surveillance of health care-associated violence using natural language processing. Pediatrics 154 (2), e2023063059. https://doi.org/10.1542/peds.2023-063059.

World Health Organization (WHO). 2025. «Preventing violence against health workers». Accessed 28 August 2025. https://www.who.int/activities/preventing-violence-against-health-workers.

Yi, M., Cao, Y., Wang, L., Gu, Y., Zheng, X., Wang, J., Chen, W., Wei, L., Zhou, Y., Shi, C., Cao, Y., 2023. Prediction of medical disputes between health care workers and patients in terms of hospital legal construction using machine learning techniques: externally validated cross-sectional study. Journal of Medical Internet Research 25, e46854. https://doi.org/10.2196/46854.