International Journal of Mental Health Nursing

International Journal of Mental Health Nursing (2023) *32*, 966–978 doi: 10.1111/inm.13114

Review Article

Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: An integrative review

Oliver Higgins, [1,2] ***Brooke L. Short,*** [3,4] ***Stephan K. Chalup*** [5] ***and Rhonda L. Wilson*** [1,6]

[1]School of Nursing and Midwifery, University of Newcastle, Newcastle, [2]Central Coast Local Health District, Gosford, [3]School of Medicine and Public Health (Medicine), University of Newcastle, Newcastle, [4]St Vincent's Hospital, Sydney, [5]School of Information and Physical Sciences (Computer Science and Software Engineering), University of Newcastle, Newcastle, New South Wales, Australia, and [6]Massey University, Wellington, New Zealand

ABSTRACT: *An integrative review investigating the incorporation of artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health care settings was undertaken of published literature between 2016 and 2021 across six databases. Four studies met the research question and the inclusion criteria. The primary theme identified was* trust and confidence*. To date, there is limited research regarding the use of AI-based decision support systems in mental health. Our review found that significant barriers exist regarding its incorporation into practice primarily arising from uncertainty related to clinician's trust and confidence, end-user acceptance and system transparency. More research is needed to understand the role of AI in assisting treatment and identifying missed care. Researchers and developers must focus on establishing trust and confidence with clinical staff before true clinical impact can be determined. Finally, further research is required to understand the attitudes and beliefs surrounding the use of AI and related impacts for the wellbeing of the end-users of care. This review highlights the necessity of involving clinicians in all stages of research, development and implementation of artificial intelligence in care delivery. Earning the trust and confidence of clinicians should be foremost in consideration in implementation of any AI-based decision support system. Clinicians should be motivated to actively embrace the opportunity to contribute to the development and implementation of new health technologies and digital tools that assist all health care professionals to identify missed care, before it occurs as a matter of importance for public safety and ethical implementation. AI-basesd decision support tools in mental health settings show most promise as trust and confidence of clinicians is achieved.*

KEY WORDS: *artificial intelligence, machine learning, mental health, nursing.*

Correspondence: Oliver Higgins, C/O Rhonda Wilson, 77a Holden St, Gosford, NSW 2250 Australia. Email: oliver.higgins@newcastle.edu.au

Declaration of conflict of interest: Prof. Rhonda Wilson is an Editorial Board Member of International Journal of Mental Health Nursing.

Authorship statement: OH – concept development, project design, data collection, data analysis, manuscript preparation. BS – data analysis, contribution to manuscript, supervision of project. SC – data analysis, contribution to manuscript, supervision of project. RW – concept development, project design, data collection, data analysis, manuscript contribution, supervision of project.

Oliver Higgins, RN, BN, BTech (CompSt).
Brooke L. Short, MBBS (Hon), MMed, BSc(Biochem), BMedSc(Path), FRANZCP.
Stephan K. Chalup, PhD, Dipl.-Math.
Rhonda L. Wilson, RN, BNSC, MNurs(Hons), PhD.

Accepted December 28 2022.

INTRODUCTION

In Australian public hospitals, emergency presentations have increased for the 5 years prior to 2019–20, with patients staying longer in Emergency Department (ED) and fewer visits completed within 4 h (Australian Institute of Health and Welfare 2021). The escalating demand for health care resources is known to lead to clinician burnout, emotional exhaustion, a lack of motivation, and a sense of frustration (Mudallal *et al*. 2017). Burnout among healthcare workers, particularly nursing, leads to reduced productivity and inefficiency (Mudallal *et al*. 2017). Additionally, as the demand increases, unfinished or missed care has become a significant issue that affects acute care hospitals globally (Jones *et al*. 2015). Nurses resort to utilizing prioritization strategies that result in patients' educational, emotional, and psychological needs going unmet (Jones *et al*. 2015). Resource burden and burnout point to wider systemic problems with under-resourcing and staff shortages (Jones *et al*. 2015). This integrative review aims to investigate the evidence for the incorporation of Artificial Intelligence (AI) and Machine Learning (ML) based Decision Support Systems (DSS) in the mental health care setting as one possible partial solution in response to these issues. This literature review is conducted using the Whittemore and Knafl (2005) integrative review framework, which was selected because the subject under examination requires information from the disciplines of health to be integrated with computing science research to be reviewed and synthesized.

BACKGROUND

Interventions and solutions to assist the mental health workforce are necessary to support the burden and service gap experienced by health care workers and those who seek comfort face. An example of such an innovation can be found in the rapidly emerging field of AI and ML, which has demonstrated that clinicians possess critical information already documented in their notes, and these can assist with issues such as the prediction of future suicidal behaviours (van Mens *et al*. 2020). AI/ML powered Decision Support Systems (DSS) represent one possible solution to the problematic intellectual load currently placed on clinicians (Walsh *et al*. 2019), providing a way of supporting the clinician with the tools they require to accelerate decision making (Collins *et al*. 2016). While many AI implementations perform exceedingly well, the nature and complexity of their learning algorithms can obscure the rationale or reasons for a recommendation, the resulting AI becomes what is known as a *black-box* AI (Rai 2019). Leading to reluctance or anxiety for some clinicians where they cannot understand the process or logic within the system nor the recommendation it devised (Brown *et al*. 2020). As such, clinicians are less likely to engage or trust *black-box* recommendations, because they require an understanding of the data and features used to make predictions, as they would in clinical documentation (Brown *et al*. 2020). Clinicians assert that imposed black-box recommendations will require a very sensitive recommendation, which may lead to clinicians neglecting to investigate the recommendation altogether if overused (Brown *et al*. 2020).

THE STUDY

Aims

The research questions guiding this integrative review are:

1. Is there evidence to support the use of artificial intelligence or machine learning based decision support systems in the delivery of mental health care?
2. What barriers exist for mental health end-users (clinicians and patients) in the adoption of artificial intelligence or machine learning based decision support systems?

Design

Based on Whittemore and Knafl (2005) framework, this integrative literature review will follow the stages of problem identification, systematic literature search, data retrieval, article evaluation, and data analysis and presentation. This framework was selected in this instance as the review requires information arising from two disciplines that are not traditionally combined to be reviewed and synthesized while providing a rigorous, structured approach to the data analysis stage and allowing for the inclusion of all research designs.

The research process will be guided by the method of the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA- P) statement (2020) checklist.

Search strategy

A literature search of databases was conducted 10th of October 2022, and 4th January 2022. The search was

developed in Medline then adapted as necessary to Scopus, Web of Science and Google Scholar, IEEE Xplore and CINAHL with Full Text (EBSCOhost). Technology innovation in the computer science fields moves at an incredibly fast pace, not only the adoption but the raw power to make this change happen (Lin 2022). Up until 2012 Moore's Law dictated the advancement of most technologies, (computing power doubling every 2 years), but the advancement of AI technologies has changed everything, AI is now outpacing Moore's law by doubling every 3–4 months (Perrault *et al*. 2019). With health AI focused research sees an exponential growth that occurs in 2017 (van de Sande *et al*. 2022). Therefore to allow for the rapid speed of technological adoption and innovation, articles (including conference papers) from 2016 to 2021 as the publication year were chosen, as the term 'Machine Learning" was introduced as a MeSH term in 2016 (National Center for Biotechnology Information 2016). Decision support was chosen rather than the MeSH term "Decision Support Systems, Clinical" because it allowed a consistent search to be applied outside the medical databases and incorporate the computer science field.

Search terms and keywords used to locate relevant literature for the study include the following: ("machine learning" OR "artificial intelligence") AND ("mental health" OR psychiatry) AND ("decision Support"). The following parameters were applied for all databases and searches: articles published between 2016 and 2021, written in English, and primary research. Inclusion for the review required that articles be focused on AI or ML based DSS for use in mental health within the study, written in English, primary research, published post 2016 and be accessible online. This data set was used to determine technological advancement in AI/ML based decision support systems in mental health. All publications prior to 2016, non-primary research, non-English, and not focusing on the use of AI or ML based DSS for mental health or psychiatry were excluded. Table 1 includes the search strategy used in Medline.

TABLE 1 *Medline search strategy*

Search #	Keywords
#1	mental health.mp. or Mental Health/
#2	psychiatry.mp. or psychiatry/
#3	#1 OR #2
#4	machine learning.mp. or machine learning/
#5	artificial intelligence.mp. or artificial intelligence/
#6	#4 OR #5
#7	decision support.mp.
#8	#3 AND #6 AND #7

Search outcomes

A total 46 articles were identified for full-text review after search limits were applied. We excluded a further 42 articles due to not meeting the required inclusion criteria, were not the primary research, did not relate to decision support in mental health using machine learning or artificial intelligence, or were model only (Refer to Fig. 1 and Table 2). Four articles were deemed to eligible and included in the review. The articles chosen included one randomized control trial, one within-subject factorial experiment and two interventional clinical trials, including interviews, feedback and questionnaire forms and observational methods. These studies were conducted in Germany (1), the United States (1) and Canada (2).

Quality appraisal

Table 3 summarizes the critical appraisal skills programme (CASP) principles used to read and analyse papers that met the inclusion criteria. With limited publications on the topic being researched, the examination was completed with the four articles selected for inclusion in the critical review, taking note of their strengths and weaknesses. Overall, the articles reviewed were of a quality standard. All articles varied in the intervention that their research was based upon, but all delivered a decision support system in various settings.

Data abstraction and synthesis

The Whittemore and Knafl (2005) approach will be applied when undertaking synthesis and analysis, following the process of data reduction, data display, data comparison, and conclusion verification. The extent of data reduction includes grouping information from each article according to study demographics, methods, sample population, key findings, and limitations. In addition, the type of AI/ML used, the issue or health problem the DSS targeted, the clinicians/providers the system targeted, and clinician and patient experiences and attitudes towards the technology were also extracted. The articles were grouped based on these themes for data comparison and conclusion verification. Table 2 contains the resulting information arranged display of the data, identifying the themes and relationships.

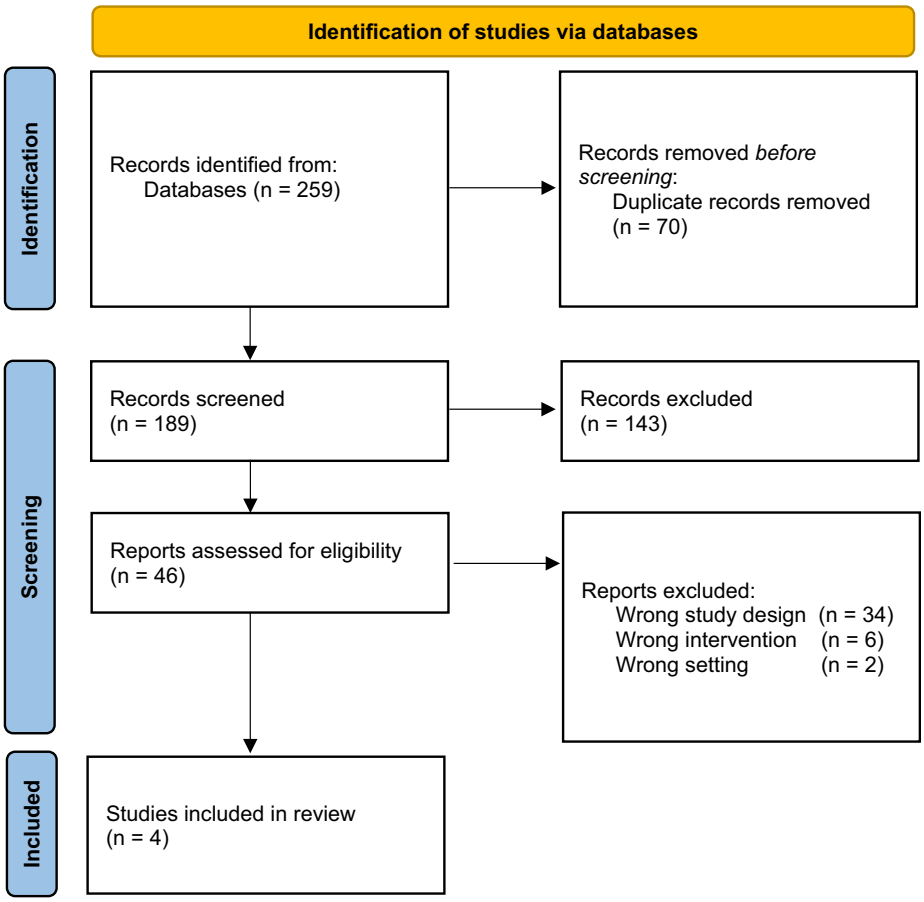

FIG. 1 Prisma.

Ethical considerations

Ethical clearance is not required as this integrative review is based on published literature.

RESULTS

Modern health care is becoming increasingly reliant on technology to deliver more care at the highest quality possible. However, the heart of health care still resides with clinicians delivering the hands-on care. Overall, the literature highlights that trust and confidence act as barriers towards achieving the acceptance of DSS within clinical care settings. Popescu *et al.* (2021), undertook an interventional clinical trial with medical (n = 7) and patients (n = 17) to access the feasibility of AI/ML DSS for individualized Major Depressive Disorder (MDD) treatment remission prediction, reported positive trust components in their research. While Benrimoh *et al.* (2020), conducted a simulation study (n = 20) to evaluate the usability of AI/ML DSS for depression treatment selection, found that 60% of the clinicians involved in the research reported that they were confident and trusted that the AI/ML could help them to assist with treatment selection. The clinicians' conveyed that the tool helped to increase patient trust and understanding of treatment, indicative of an openness for use of the tool in actual clinical practice, notably, Popescu *et al.* (2021) reported that 71% of their small sample size of clinicians reported trust of the system. Jacobs *et al.* (2021), investigated the way in which correct and incorrect AI/ML recommendations influence the selection accuracy of primary care physicians (n = 220), revealing that clinicians with previous experience and familiarity with AI/ML were less likely to engage with AI/ML-based recommendations. This is interesting because it infers that previous experience with ML could impact their engagement, confidence and trust in AI/ML-based DSS when compared to clinicians with lower AI/ML awareness. The expectation that prior experience being beneficial in the use of an intervention yet instead presents a barrier. Jacobs

TABLE 2 *Characteristics of AI/ML based DSS use in mental health sites*

Author	Study aim	Design	Results	Limitations	Recommendations
Benrimoh *et al.* (2020) (Canada)	Use of a simulation in evaluating the usability of AI/ML DSS for depression treatment selection and its impact on physician-patient interaction Deep Learning/Neural Network (AIFred)	CT Simulation $n = 20$ Psychiatry and family medicine (Medical). Clinicians involved in the iterative design cycle of tool Self-report questionnaires (PHQ-9), scenario observations, interviews and standardized patient feedback	Clinicians indicated confidence to use the DSS tool in clinical practice, trusting in the system's predictions to assist with treatment selection Tool helped increase patient confidence and trust in treatment	Simulation only, small sample size	Conduct a longitudinal study of the DSS in a clinical setting RCT study aimed at assessing tool effectiveness and safety Clinicians should be involved throughout the development of DSS tools
Jacobs *et al.* (2021) (USA)	Investigate how correct and inaccurate AI/ML DSS recommendations for antidepressant medication selections influence clinicians' selection accuracy Random Forest	Primary Care Physicians (Medical) $n = 220$ Web-based case study experiment using patient vignettes ANOVA for within-subjects analyses and a two-sided student's t test for post hoc independent pairwise comparisons	Prior experience with AI/ML models influence clinicians' trust and confidence in using AI/ML DSS recommendations Clinicians benefit from meaningful explanations about recommendations to enhance the levels of trust and confidence in AI/ML DSS The performance of an algorithm is insufficient to predict positive clinical outcomes	Did not include medication combinations or nonpharmacological treatments Study was deployed as an online behavioural experiment using hypothetical patient scenarios All psychopharmacology experts involved in this study came from the northeast United States	Strategies to increase trust and confidence in clinicians need to be developed as part of implementing AI/ML based DSS Future research to consider how clinicians' experience with AI/ML may influence their future confidence and trust in AI/ML systems DSS should communicate confidence levels and supporting explanations for the recommendation
Lutz *et al.* (2021) (Germany)	RCT to investigate the effects of a AI/ML DSS and feedback system Nearest Neighbour Classification	RCT $n = 538$. Psychologists & Clinicians delivering CBT Use of AI/ML DSS Treatment provision and selection. Clinicians rated scales (4) completed after each session. Analysis conducted using R version 4.0.1	Clinicians felt confident with the tool, finding DSS recommendations useful in the therapeutic relationship The evaluation showed a positive effect for patients when therapists followed the recommended treatment Some clinicians deviated from DSS recommendations however 'why' this occurred was not explored	The cross-therapist design may not be ideal for this type of study since it is difficult to blind effectively Patient blinding may also cause problems. TAU group may have expected therapists to communicate questionnaire answers DSS did not provide a binary output of recommendations	The clinicians should be informed of recommendations as binary outputs with supporting evidence to inform the underlying decision mechanism Research into why the therapist did/did not follow treatment recommendations

(*Continued*)

TABLE 2 (Continued)

Author	Study aim	Design	Results	Limitations	Recommendations
Popescu *et al.* (2021) (Canada)	Feasibility of an AI/ML DSS for individualized MDD treatment remission prediction Deep Learning/Neural Network (AIFred)	Interventional CT in primary care setting. Medical (n = 7), Patients (n = 17) Patients had access to their own version of the tool that operated in parallel with the clinicians Data was collected through self-report questionnaires (PHQ-9 and GAD-7) and semi structured interviews. Analysis conducted using G*Power package, version 3.1.9.7	Clinicians stated trust and confidence in the DSS Introduction of the tool did not increase appointment length Patients and physicians reported that the tool was easy to use, and both felt the patient-clinician relationship improved significantly	Study design changed due to COVID-19 resulting in a small sample size Unable to evaluate the effectiveness of the DSS on depression scores	DSS is feasible for use and ready for effectiveness studies to determine if the DSS is effective in improving depression outcomes

Abbreviations: AI, Artificial Intelligence; ANOVA, Analysis of variance; CBT, Cognitive Behavioural Therapy; CT, Clinical Trial; DSS=Decision Support System; GAD-7, General Anxiety Disorder-7; MDD, Major Depressive Disorder; ML, Machine Learning; PHQ-9, Patient Health Questionnaire-9; RCT, Randomized Control Trial; TAU, Treatment As Usual.

TABLE 3 *CASP*

Author & Year	Did the study address the aim?	Were the participants recruited acceptably?	Was the outcome accurately measured to minimize bias?	Have relevant outcome factors been identified?	Are the results precise?	Are the results applicable to the local population?	Do the results fit with other known evidence?	Do the results have implications for modern practice?
Benrimoh *et al.* (2020)	Y	Y	CT	Y	Y	Y	Y	Y
Jacobs *et al.* (2021)	Y	Y	CT	Y	Y	Y	Y	Y
Lutz *et al.* (2021)	Y	Y	Y	Y	Y	Y	Y	Y
Popescu *et al.* (2021)	Y	Y	CT	Y	Y	Y	Y	Y

et al. (2021) state that although clinicians' acceptance of the technology and the algorithm's performance are two crucial factors for adoption, they are not enough to predict positive performance outcomes.

Clinician acceptance is a key factor for the uptake of innovative technological change (Wade *et al*. 2014). In addition, while research shows that overall sentiment towards AI DSS is positive when compared to decisions made with the assistance of AI, AI guided decision making is not congruent with the way that clinicians typically make diagnostic decisions (Hah & Goldin 2021). Brown *et al*. (2020)'s research indicates that clinicians need to understand not just what the model has recommended but, more importantly, the logical path the model undertook to arrive at the outcome it did. Clinicians' feedback further indicates the requirement to explain which diagnostic features lead to a patient receiving a particular recommendation and an expectation of knowing how those features will influence treatment (Brown *et al*. 2020). AI systems can demonstrate trustworthiness to clinicians by displaying predictability, procedural transparency, algorithmic transparency, and robustness (Bhatt & Shams 2021). Clinician participation in the development of the system can improve trust, however clinicians are often not present at the developmental stages (Schwartz *et al*. 2021).

Clinicians' trust and confidence

The literature identified that communicating recommendations or predictions to the clinician as a significant barrier. Lutz *et al*. (2021), undertook a Randomized Control Trial (RCT) (n = 538) to investigate the effects of a DSS and feedback system, did not provide their recommendations in a binary yes/no output but rather through the presentation of boxplots with different nearest neighbour predictions. This interface resulted in difficulty for some clinicians understanding which elements of the client's presentation contributed to the formulation of treatment recommendations and this made conclusions challenging to communicate. Jacobs *et al*. (2021) simulated and manipulated the ML recommendations to assess how clinicians respond to algorithmic errors. When paired with incorrect recommendations, interacting with feature-based explanations correlated with lower accuracy scores, suggesting that participants struggled to calibrate their clinical practice against the ML performance. However, Benrimoh *et al*. (2020) found that the tool in their research was acceptable to the clinicians that participated, their feedback indicated a significant degree of trust and confidence in the system's predictions to assist with treatment selection. They expressed a willingness to use the tool in actual clinical practice, stating that they felt it helped to increase patient understanding and trust in treatment plan. Similarly, the clinicians involved in the research conducted by Popescu *et al*. (2021) reported high trust and usability ratings in how the system reported treatment recommendations. The intervention accommodated recommendations that could be modified or rearranged as the clinicians felt were clinically indicated with no automated clinical decisions process in place, the final clinical decision logic was always in clinicians' hands (Popescu *et al*. 2021).

The literature also identified a mistrust from clinicians regarding the validity of AI/ML and its applicability to their practice. Jacobs *et al*. (2021) reported that using AI/ML recommendations to prompt treatment selection did not improve treatment accuracy among 220 antidepressant prescribing clinicians in their evaluation compared with psychopharmacology experts. However, when clinicians engaged with incorrect recommendations, their treatment selection accuracy scores were significantly lower than those engaged with correct recommendations or questions without AI/ML recommendations. The results suggest that incorrect AI/ML recommendations may adversely impact clinician treatment selection, challenging the idea that clinicians that use AI/ML tools will perform better than clinicians or AI/ML algorithms separately. It is also essential to consider the impact of incorrect recommendations by AI/ML systems, as clinicians' prior experience with AI/ML models can influence their trust and confidence in the system's recommendations in treatment selection decisions (Jacobs *et al*. 2021). The inability of the system to communicate the underlying mechanism or process used to provide the recommendation can contribute to clinician mistrust of the system (Jacobs *et al*. 2021). Notably, Lutz *et al*. (2021) highlight that some clinicians may be overwhelmed with the concept of an AI/ML derived decision support tool. Their misinterpretations or overreactions may produce adverse treatment outcomes, with several clinicians reported applying strategies other than the system recommendation.

Designing in development with clinicians

Benrimoh *et al*. (2020) note that clinicians were involved in the DSS the ongoing iterative design and

development process. The authors believed that this led to the design of a tool that ensured clinician autonomy and allowed the clinicians to select any treatment or action they deemed appropriate, contributing to their overall positive levels of trust and confidence. Additionally, identifying algorithmic errors or biases should not be the sole responsibility of clinicians, but if such tools are used in the real world, dealing with imperfect algorithms will be necessary and reinforces the importance of the clinician in the iterative process of the tool's design, testing, and refinement (Jacobs *et al*. 2021).

The clinical interface

Interaction and design decisions, such as the format of the explanation, can significantly affect clinicians' behaviour. However, while there may be a desire to develop clinic facing, visually simple technologies, Jacobs *et al*. (2021) results would indicate that less information is not always better, suggesting that communicating a recommendation's confidence level could influence the use of the recommendation itself. Likewise, Lutz *et al*. (2021) propose that presenting a binary output with supporting evidence to inform the clinician on the underlying decision mechanism would benefit the clinician compared to the nearest neighbour box plots they presented in their intervention.

Patient attitudes and acceptance

The literature provides limited insight into patient acceptance of AI/ML DSS, with only Popescu *et al*. (2021) including patient feedback in their data collection, the research stating 62% of patients reported that they trusted the DSS and that 62% felt that their appointment time did not change and good overall usability of the DSS by 92% of patients. Significantly, 46% of patients felt that the patient-clinician relationship improved upon exit from the program, whereas 54% felt it did not change. Popescu *et al*. (2021) indicated that this positive therapeutic relationship was possibly due to the intervention being directly tied to clinical care, medical officer engagement in use of the intervention and the direct involvement of the patients in the shared decision-making process. Benrimoh *et al*. (2020) noted that some clinicians in their study tended to turn the laptop towards the patient, "inviting them in" to the session, patients reporting that they felt engaged in decision-making.

The literature provides evidence of the role that trust and confidence contribute to clinicians' acceptance and use of AI/ML based DSS. While limited in scope, the literature demonstrates positive benefits for patients and the relationship with clinicians. Clinicians require systems that offer transparent and interpretable results with clearly communicated treatment options and retain the autonomy to select the recommendation that they feel is clinically appropriate. Involving the clinician design and development process contributes to a positive outcome and ensures critical components such as ensuring that the clinicians making the final decision are not overlooked. Therefore, to utilize the opportunity that AI/ ML based DSS offers, all elements must be evaluated to ensure that quality interventions can be delivered safely and consistently.

DISCUSSION

Missed care is a significant problem that affects hospitals worldwide; the cost of this missed care results in multiple negative outcomes for patients, nurses, and organizations (Jones *et al*. 2015). Missed, unfinished, or care left undone results when time is scarce, and is the outcome of rationing that occurs due to clinical priority setting for healthcare staff under high demand (Jones *et al*. 2015). Nurses need to be equipped to recognize situations and evaluate the available evidence that will help guide decisions that, as a result of rationing of time, may result in care being missed (Jones *et al*. 2015). One solution to this problem is the emerging field of AI and ML, which may provide new and innovative interventions to help prevent incidences of missed care, reduce the resource burden, and give time back to clinicians. This literature review will address to this issue by establishing the scope of scientific work undertaken and highlighting the challenges and issues associated with the use and implementation. The literature provides evidence to support the use of artificial intelligence or machine learning based decision support systems in the delivery of mental health care, although it is still in an early phase of maturity (Benrimoh *et al*. 2020; Lutz *et al*. 2021; Popescu *et al*. 2021). While AI/ML systems may perform incredibly well in benchmarking or testing, the results do not necessarily translate to better practice or patient outcomes (Jacobs *et al*. 2021), the resulting clinical tool can only be effective when clinicians are confident in its application.

Our literature review has revealed that implementation of any tool in health can only be achieved when clinicians become confident in its use and trusting of its clinical capabilities. However, the nature of the trust relationship within the context of AI is more complex

than placing trust in a person or system. The human tendency to anthropomorphise AI can lead to an emotional connection (Ryan 2020), a relationship that requires trust to be imparted, the trustee being held accountable for their actions, is something AI cannot be (Ryan 2020). Trust and confidence are earnt through AI being considered reliable rather than trustworthy, placing the burden of responsibility upon the developers and researchers (Ryan 2020). Implementation plans must build confidence through evidence that ensures the intervention can meet the clinical need before belief and trust in the system can be established. For mental health clinicians, the application of trust and the resulting confidence is paramount when working in an environment that can carry a significant level of risk, such as the assessment of suicidality. Mental health is a complicated specialty that includes complex presentations with the potential for extreme adverse outcomes such as suicide. Suicide is the result of many complicated variables and relationships that contribute to a person's mental state. The prediction of such a complicated indicator requires the trust of clinicians yet still requires considerable human clinical judgement to interpret the system recommendations. While this literature review's sample size is small, it has demonstrated that trust and confidence are essential for any DSS to implement successfully. The use AI/ML will not automatically result in better care, regardless of how well it may perform, as a poorly designed and implemented system has the potential to erode pivotal clinician trust even further.

Trust in technology

The potential to trust and have confidence in the model is dependent on its implementation into routine clinical care. Prior to clinical use, a tool must demonstrate that it is accurate, outperforms, or complement clinical judgement (Bentley *et al*. 2021). Therefore, to achieve an optimal level of trust, a system must demonstrate fairness, transparency, and robustness (Asan *et al*. 2020) and validate the system's capabilities as per any other clinical tool to achieve an optimal level of confidence and trust. Clinicians should be presented with treatment recommendations with validity and confidence of prediction, with clinicians having the final decision for the appropriate course of treatment. The opportunity to engage with this emerging technology is dependent on the clinician's acceptance and trust in the technology, especially in risk-averse or sensitive domains such as suicide assessment and mental health. Many have expressed scepticism to varying degrees, citing outright distrust of such a tool, the concept of the use of a DSS is anxiety-provoking for some (Bentley *et al*. 2021). The trust barriers with clinicians must be overcome to achieve the desired outcome of any DSS recommendations, Jacobs *et al*. (2021)'s conclusions indicate that those with previous experience with AI/ML are less likely to engage, especially if they have had a poor prior user experience, revealing the delicate balance for the use of DSS by clinicians.

Clinician involvement

The clinician's role is not limited to that of the end-user; their involvement should start early in the design and development cycle and continue through to practice. The design in Benrimoh *et al*. (2020) DSS included clinician involvement in the ongoing iterative development process, noting that the authors felt it contributed to the overall positive findings, the clinicians designing a system that ensured autonomy and allowed them to select a treatment or action beyond the DSS recommendations that they deemed appropriate. Traditionally, expert clinical involvement is most common when predictive DSS specifications are made or when a system is evaluated, omitting the critical iterative development cycle (Schwartz *et al*. 2021) where they could be available to select model features and verify clinical correctness (Schwartz *et al*. 2021). Clinician involvement through all aspects of development will accelerate the transformation of algorithms into practical clinical application (Verghese *et al*. 2018) and accommodate for appropriate risk mitigation (Seneviratne *et al*. 2020). Ideally, clinicians must be active participants in this emerging technology, participating in all components, from project specification to evaluation, and provide iterative clinical involvement throughout the developmental cycle.

Intelligible, interpretable, and transparent machine learning

The way that the DSS communicated its predictions or recommendations to clinicians was identified as a significant barrier. Systems lacking clear communication of the underlying mechanisms and confidence in recommendations can contribute to difficulty interpreting the results and recommendations (Lutz *et al*. 2021). Clinicians indicated a mistrust of AI/ML based DSS, the closed, *black-box* nature of its design provided a level of ambiguity that is unacceptable for many

clinicians (Jacobs *et al*. 2021). Thus, a clinical algorithm will lose its utility if its underlying clinical features are hidden from the clinicians or if the clinicians do not perceive the clinical features as intuitively meaningful (Brown *et al*. 2020).

A significant step forward in intelligibility is the advent of InterpretML (Nori *et al*. 2019), an open-source package developed by Microsoft that allows for interpretable models. Unlike closed *black-box* models, InterpretML implements a *glass-box* methodology that interprets what the 'machine' has 'learnt' from the data (Nori *et al*. 2019), offering deep insights and data anomaly discovery (Chang *et al*. 2020). The InterpretML approach represents an important advance in the trade-off between model accuracy and interpretability for applications such as healthcare, where verification and debugging are equally important as accuracy (Caruana *et al*. 2015). If the interpretation of the model reveals data heterogeneity, it is possible to edit the model itself, reducing deployment risk and, where possible, ensuring system bias can be accounted for (Caruana *et al*. 2015). Therefore, any AI/ML system for use in health should be intelligible, interpretable, transparent, and clinically validated. It should provide a clear explanation for each prediction it makes and communicates which clinical features contributed to the recommendation (Nori *et al*. 2019). The use of tools such as IntepretML, which utilize *glass-box* design, should be encouraged and incorporated into clinical interfaces while adhering to the World Health Organization's (WHO) *Ethics and Governance of Artificial Intelligence in Healthcare* (World Health Organization 2021).

Clinical interface and the implementation gap

The interaction and design decisions of a DSS can significantly affect clinicians' behaviour and acceptance into practice, the clinical interface used to communicate recommendations is just as important as the recommendation itself. However, attempting to make a visually plain interface or simplifying treatment options can result in a clinician's reluctance to trust the DSS recommendations (Jacobs *et al*. 2021). For a DSS to be accepted and trusted, it requires clear actionability, with outputs directly linked to an intervention by the clinician (Seneviratne *et al*. 2020). Clinicians require supporting evidence on the clinical features incorporated in the recommendation (Lutz *et al*. 2021), communicating clearly and concisely and allowing clinicians to understand what features and processes the system undertook to make its recommendation (Brown *et al*. 2020). Systems should also communicate their confidence level in the prediction recommendations (Lutz *et al*. 2021) and present several treatment options rather than one recommendation (Benrimoh *et al*. 2020). Finally, a system must have straightforward technical integration into the healthcare system and requires evaluation on how well a model integrates within an existing clinical workflow outside of the development setting (Reddy *et al*. 2021). For integration to be successful, it must include clear actions, incorporating how clinical features contributed to the outcome, communicating multiple recommendations, and providing prescribed recommendation confidence. While it is recommended that systems provide transparent and interpretable results, there is limited discussion regarding the underlying knowledge of the system is required by clinical staff. It should also be noted that accuracy of the systems in the literature discussed is limited and warrants further research in the communication of this metric, (Brown *et al*. 2020) highlighting the importance of communicating the underlying mechanisms of the prediction.

Lived experience attitudes and acceptance

The literature does provide evidence that using AI/ML DSS has a positive effect on the clinician/patient relationship. Through the use of the tool, patients felt engaged in the decision-making process of their care when the clinicians turned the laptop screen towards them Benrimoh *et al*. (2020). Patients reported that they felt the use of the system led to a better relationship with the clinician and contributed to a better health outcome through the use of the tool (Popescu *et al*. 2021). However, patients' attitudes and acceptance of AI/ML in the provision of care, especially mental health, is poorly understood. There is a significant potential for harm when using AI/ML (Farthing *et al*. 2021) and the technology industry's mantra of "move fast and breaks things" (Taplin 2017) has limited application in health. Future innovations should consider potential iatrogenic harms to ensure safety (Farthing *et al*. 2021), especially with vulnerable or already disadvantaged populations (World Health Organization 2021). Of particular note, there is very limited involvement overall of the inclusion of lived experience service users (Gooding *et al*. 2022), that is, the very people who are supposed to gain the greatest benefit from these innovations. It is recommended that future research be framed to understand better how the use of AI/ML in the delivery of care will affect those with a lived experience of mental illness, especially psychosis (Higgins *et al*. 2022).

Limitations

There are several limitations to this study, first, the emerging use of AI and ML based DSS in mental health care is relatively recent, resulting in a small number of studies, and it did not include literature composed of model development and testing. Second, DSS that involve complex diagnostic processes such as Magnetic Resonance Imaging (MRI), Computerized Tomography (CT), Imaging or Electroencephalogram (EEG) have been excluded due to their complexity. Third, the review includes studies written in English, with studies in other languages excluded from this analysis. Fourth, the primary focus of care is mental health and does not include other health disciplines that may have a larger diagnostic dataset to draw upon. Finally, we note that this research is limited by the small number of papers and small sample sizes, restricting the evidence base; however, the findings provide an opportunity for further discussion and reaffirm that this research topic is still in its infancy and requires further verification.

Implications

Even from a global context, the literature regarding the use of AI/ML based DSS in mental health is limited. Based on the results of this literature review, it suggests that this innovation has a beneficial effect; however, it also produces its own set of unique problems. Continued use of systems that are not transparent or clinically developed and evaluated will continue to erode the trust of healthcare workers. The notion of negative associations with machine learning remains a decisive factor in incorporating such methods into practice, as a poorly implemented system could result in significant distrust. Further research into the role of the clinician developer could improve the trust and acceptance of AI/ML for clinical practice within the global context. Consideration must also be given to the clinical and professional accountability of AI/ML based supported decision making. The accountability for AI systems in the mental health context must be appropriately distributed, with suitable solutions in place (Gooding *et al*. 2022). The attribution of accountability and responsibility is vital for not only those affected using the tools but also for public trust in AI solutions. There must be adequate opportunity to integrate the objectives, outcomes and trade-offs that come with the use of AI, however clinician accountability is yet to be clearly defined (Gooding *et al*. 2022). Both clinicians and patients, that is, end-users, should be involved in the development and design of new health technologies. The very people who stand to benefit the most from new technologies in mental health care are often missed in the research design and reporting of innovation in the literature (Gooding *et al*. 2022), their omission compromises user-centred design and denies them authentic contribution in partnership with clinicians to promote development and creation of potentially powerful shared decision making platforms. Finally, as more AI/ML based interventions are introduced into healthcare settings, often presented as a panacea for many problems, using an algorithm that performs well will not necessarily equate to better care.

Ethical implications

As AI and data driven technologies may appear to offer the promise of solutions to resource burden and burnout, there is a clear potential for harm to occur (Gooding *et al*. 2022) and this must be considered carefully during design and commissioning of technological solutions. The introduction of AI DSS as a tool to address resource burden and burnout may be perceived to undermine the wider systemic problems with underresourcing and workforce shortages, devaluing the tool in the process. AI is an emerging technology with considerable push momentum for application across many fields. Despite this, the current market, especially in mental health, is still speculative in its potential (Gooding *et al*. 2022). Overall, much of the literature remains in pilot or exploratory stages and is most frequently located in the computer science field (Gooding & Kariotis 2021). Governments that invest in emerging technologies stand to gain political advance from appearance as technologically innovative (Gooding *et al*. 2022), though many of the claims made about AI technologies in mental health are yet to be proven effective (Gooding *et al*. 2022). Algorithmic accountability, lack of lived experience involvement and the potential for techno-solutionism, overmedicalisation and discrimination are serious risks (Gooding & Kariotis 2021) that needs to be addressed before acceptance of the tools can be sought.

CONCLUSION

There is limited research that pertains to the use of AI/ML based DSS in mental health. This integrative review identified that mental health clinicians and those with a lived experience of mental health could benefit from the incorporation of AI/ML based DSS.

However, significant barriers exist regarding its incorporation into practice, such as clinicians' trust, confidence, and system transparency. In order to participate effectively in this new technology, clinicians must be involved throughout all design and development phases, from specification to evaluation. Integration needs to include clear actions, incorporating the clinical measures that contributed to the outcome, communicating multiple recommendations, and providing confidence in the prescribed recommendations. Treatment recommendations should be presented to clinicians with validity and confidence of prediction, with clinicians making the final decision regarding the appropriate course of treatment. Models should utilize *glass-box* design, such as InterpretML and be incorporated into digital patient record interfaces. Finally, more research is required to understand how AI/ML in the delivery of care affects those who have lived experience of mental illness.

RELEVANCE TO CLINICAL PRACTICE

This research could have implications for the delivery of mental health care in a global context, the incorporation of AI/ML in mental health care could help nurses and other mental health clinicians provide timely and relevant care. Mental health clinicians should have meaningful involvement in developing and implementing new health technologies at all stages and embrace the opportunity to develop the right tools that can assist all health care workers in identifying missed care before it occurs.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of Central Coast Local Health District. Open access publishing facilitated by The University of Newcastle, as part of the Wiley - The University of Newcastle agreement via the Council of Australian University Librarians.

FUNDING INFORMATION

Partial financial support was received from NSW Ministry of Health as part of the Towards Zero Suicides initiative.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

Asan, O., Bayrak, A. E. & Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J Med Internet Res*, *22* (6), e15154. https://doi.org/10.2196/15154 [Accessed date 20th December 2021].

Australian Institute of Health and Welfare. (2021). Emergency department care. Available from: URL: https://www.aihw.gov.au/reports-data/myhospitals/sectors/emergency-department-care [Accessed 20th December 2021].

Benrimoh, D., Tanguay-Sela, M., Perlman, K. *et al*. (2020). Using a simulation centre to evaluate preliminary acceptability and impact of an artificial intelligence-powered clinical decision support system for depression treatment on the physician-patient interaction. *BJPsych Open*, *7* (1), e22. https://doi.org/10.1192/bjo.2020.127

Bentley, K., Zuromski, K., Fortgang, R. *et al*. (2021). Implementing machine learning algorithms for suicide risk prediction in clinical practice: A focus group study. https://doi.org/10.31234/osf.io/6m5qd

Bhatt, U. & Shams, Z. (2021). Trust in artificial intelligence: Clinicians are essential. In: A. B. Bhatt (Ed). *Healthcare Information Technology for Cardiovascular Medicine: Telemedicine & Digital Health*. (pp. 127–141). Springer, Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-81030-6_10

Brown, L. A., Benhamou, K., May, A. M., Mu, W. & Berk, R. (2020). Machine learning algorithms in suicide prevention: Clinician interpretations as barriers to implementation. *Journal Clinical Psychiatry*, *81* (3), 19m12970. https://doi.org/10.4088/JCP.19m12970

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. pp. 1721–1730.

Chang, C.-H., Tan, S., Lengerich, B. J., Goldenberg, A. & Caruana, R. (2020). How interpretable and Trustworthy are GAMs? *arXiv e-prints*, *abs/2006.06466*, arXiv-2006.

Collins, H., Calvo, S., Greenberg, K., Forman Neall, L. & Morrison, S. (2016). Information needs in the precision medicine era: How genetics home reference can help. *Interactive Journal of Medical Research*, *5* (2), e13. https://doi.org/10.2196/ijmr.5199

Farthing, S., Howell, J., Lecchi, K., Paleologos, Z., Saintilan, P. & Santow, E. (2021). *Human Rights and Technology*. Sydney, Australia: A. H. R. Commission. Available from: URL: https://tech.humanrights.gov.au/downloads [Accessed 17th January 2022].

Gooding, P. & Kariotis, T. (2021). Ethics and law in research on algorithmic and data-driven technology in mental health care: Scoping review. *JMIR Mental Health*, *8* (6), e24668. https://doi.org/10.2196/24668 [Accessed 17th January 2022].

Gooding, P., Brown, L. X. Z., Myrick, K. *et al*. (2022). Digital futures in mind: Reflecting on technological experiments

in mental health and crisis support. Available from: URL: https://melbourne.figshare.com/articles/report/Digital_Futures_in_Mind_Reflecting_on_Technological_Experiments_in_Mental_Health_and_Crisis_Support/21113899

Hah, H. & Goldin, D. S. (2021). How clinicians perceive artificial intelligence-assisted technologies in diagnostic decision making: Mixed methods approach. *Journal of Medical Internet Research*, *23* (12), e33540. https://doi.org/10.2196/33540

Higgins, O., Chalup, S. K., Short, B. L. & Wilson, R. L. (2022). Interpretations of innovation: The role of technology in explanation seeking related to psychosis. *Perspectives in Psychiatric Care*. https://onlinelibrary.wiley.com/doi/10.1111/inm.13121

Jacobs, M., Pradier, M. F., McCoy, T. H., Jr., Perlis, R. H., Doshi-Velez, F. & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: The example of the antidepressant selection. *Translational Psychiatry*, *11* (1), 108. https://doi.org/10.1038/s41398-021-01224-x

Jones, T. L., Hamilton, P. & Murry, N. (2015). Unfinished nursing care, missed care, and implicitly rationed care: State of the science review. *International Journal of Nursing Studies*, *52* (6), 1121–1137. https://doi.org/10.1016/j.ijnurstu.2015.02.012

Lin, S. (2022). A Clinician's guide to artificial intelligence (AI): Why and how primary care should lead the health care AI revolution. *The Journal of the American Board of Family Medicine*, *35* (1), 175–184. https://doi.org/10.3122/jabfm.2022.01.210226

Lutz, W., Deisenhofer, A. K., Rubel, J. *et al*. (2021). Prospective evaluation of a clinical decision support system in psychological therapy. *Journal of Consulting and Clinical Psychology*, *90*, 90–106. https://doi.org/10.1037/ccp0000642

van Mens, K., Elzinga, E., Nielen, M. *et al*. (2020). Applying machine learning on health record data from general practitioners to predict suicidality. *Internet Interventions*, *21*, 100337. https://doi.org/10.1016/j.invent.2020.100337

Mudallal, R. H., Othman, W. M. & Al Hassan, N. F. (2017). Nurses' Burnout: The Influence of leader empowering behaviors, work conditions, and demographic traits. *Inquiry*, *54*, 46958017724944. https://doi.org/10.1177/0046958017724944

National Center for Biotechnology Information. (2016). Machine learning. [Cited 12 February 2022]. Available from: URL: https://www.ncbi.nlm.nih.gov/mesh/2010029

Nori, H., Jenkins, S., Koch, P. & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Perrault, R., Shoham, Y., Brynjolfsson, E. *et al*. (2019). *The AI Index 2019 Annual Report*. Stanford, CA: AI Index Steering Committee, Human-Centered AI Institute, Stanford University.

Popescu, C., Golden, G., Benrimoh, D. *et al*. (2021). Evaluating the clinical feasibility of an artificial intelligence-powered, web-based clinical decision support system for the treatment of depression in adults: Longitudinal feasibility study. *JMIR Formative Research*, *5* (10), e31862. https://doi.org/10.2196/31862

Rai, A. (2019). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48* (1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Reddy, S., Rogers, W., Makinen, V. P. *et al*. (2021). Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform*, *28* (1), e100444. https://doi.org/10.1136/bmjhci-2021-100444

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, *26* (5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

van de Sande, D., Van Genderen, M. E., Smit, J. M. *et al*. (2022). Developing, implementing and governing artificial intelligence in medicine: A step-by-step approach to prevent an artificial intelligence winter. *BMJ Health & Care Informatics*, *29* (1), e100495. https://doi.org/10.1136/bmjhci-2021-100495

Schwartz, J. M., Moy, A. J., Rossetti, S. C., Elhadad, N. & Cato, K. D. (2021). Clinician involvement in research on machine learning-based predictive clinical decision support for the hospital setting: A scoping review. *Journal of the American Medical Informatics Association*, *28* (3), 653–663. https://doi.org/10.1093/jamia/ocaa296

Seneviratne, M. G., Shah, N. H. & Chu, L. (2020). Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations*, *6* (2), 45–47. https://doi.org/10.1136/bmjinnov-2019-000359

Taplin, J. (2017). *Move Fast and Break Things: How Facebook, Google, and Amazon Have Cornered Culture and What it Means for all of us*. United Kingdom: Pan Macmillan.

Verghese, A., Shah, N. H. & Harrington, R. A. (2018). What this computer needs is a physician: Humanism and artificial intelligence. *Journal of the American Medical Association*, *319* (1), 19–20. https://doi.org/10.1001/jama.2017.19198

Wade, V. A., Eliott, J. A. & Hiller, J. E. (2014). Clinician acceptance is the key factor for sustainable telehealth services. *Qualitative Health Research*, *24* (5), 682–694. https://doi.org/10.1177/1049732314528809

Walsh, S., de Jong, E. E. C., van Timmeren, J. E. *et al*. (2019). Decision support Systems in Oncology. *JCO Clinical Cancer Informatics*, *3* (3), 1–9. https://doi.org/10.1200/CCI.18.00001

Whittemore, R. & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing*, *52* (5), 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. Available from: URL: https://apps.who.int/iris/rest/bitstreams/1352854/retrieve [Accessed 17th January 2022].