Original Research

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–13

DOI: 10.1177/20552076241294182
journals.sagepub.com/home/dhj

Sage

Large language models outperform general practitioners in identifying complex cases of childhood anxiety

Inbar Levkovich[1], Eyal Rabin[2], Michal Brann[3] and Zohar Elyoseph[4]

Abstract

Objective: Anxiety is prevalent in childhood but often remains undiagnosed due to its physical manifestations and significant comorbidity. Despite the availability of effective treatments, including medication and psychotherapy, research indicates that physicians struggle to identify childhood anxiety, particularly in complex and challenging cases. This study aims to explore the potential effectiveness of artificial intelligence (AI) language models in diagnosing childhood anxiety compared to general practitioners (GPs).

Methods: During February 2024, we evaluated the ability of several large language models (LLMs; ChatGPT-3.5 and ChatGPT-4, Claude.AI, Gemini) to identify cases childhood anxiety disorder, compared with reports of GPs.

Results: AI tools exhibited significantly higher rates of identifying anxiety than GPs. Each AI tool accurately identified anxiety in at least one case: Claude.AI and Gemini identified at least four cases, ChatGPT-3 identified three cases, and ChatGPT-4 identified one or two cases. Additionally, 40% of GPs preferred to manage the cases within their practice, often with the help of a practice nurse, whereas AI tools generally recommended referral to specialized mental or somatic health services.

Conclusion: Preliminary findings indicate that LLMs, specifically Claude.AI and Gemini, exhibit notable diagnostic capabilities in identifying child anxiety, demonstrating a comparative advantage over GPs.

Keywords

Large language models, general practitioners, childhood anxiety, artificial intelligence, vignettes

Submission date: 1 July 2024; Acceptance date: 9 October 2024

Introduction

Among the rapid advancements in the domain of artificial intelligence (AI) is the emergence of large language models (LLMs), such as Bard by Google, Claude.AI 2 by Anthropic, and ChatGPT versions 3.5 and 4 by OpenAI. LLMs have demonstrated significant potential within the realm of mental healthcare.[1–4] These technological innovations have the potential to transform the mental healthcare landscape by expediting research processes, augmenting clinical practice by providing valuable assistance to healthcare professionals, and extending support mechanisms to patients.[5,6] Nevertheless, the efficacy and practicality of directly employing LLMs to enhance mental health outcomes, particularly in the case of anxiety disorders (ADs), warrant further investigation. This study evaluates the ability of advanced AI models, specifically LLMs, to diagnose childhood anxiety in complex cases or those with comorbidities, compared to general practitioners (GPs) and mental health

[1]The Faculty of Education, Tel Hai College, Upper Galilee, Israel
[2]Department of Psychology and Education, The Open University of Israel, Ra'anana, Israel
[3]Department of Psychology and Educational Counseling, Max Stern Yezreel Valley College, Afula, Israel
[4]Faculty of Education, University of Haifa, Haifa, Israel

Corresponding author:
Inbar Levkovich, The Faculty of Education, Tel Hai College, Upper Galilee, Israel.
Email: levkovinb@telhai.ac.il

experts. It is noteworthy as one of the first studies to focus on the mental health of children, particularly in the area of childhood anxiety, using AI technology, addressing the significant gap in research in this field.

Undiagnosed ADs have a profound impact on human development and well-being.[7,8] With prevalence rates as high as 25%. ADs represent the most widespread mental health challenge across the lifespan.[9] Indeed, only about 10% of children with ADs, including those exhibiting subthreshold severity, are expected to be free of any mental health issues in adulthood.[10,11] The effectiveness of treatment in mitigating the risks and adversities associated with ADs is well-established.[12,13] Given the pivotal role and accessibility of GPs and their ongoing relationships with families, they are uniquely positioned to identify ADs, which are often characterized by early onset, a chronic or episodic course, physical manifestations, and comorbidities.[14,15]

The challenge of detection is compounded in cases of early onset, lower severity, and less overt manifestations.[16] This difficulty is inherent to the nature of anxiety itself, which is characterized by concealment of core symptoms, a gradual and fluctuating course, and a wide array of accompanying symptoms that are not typically associated with anxiety.[17] These symptoms, which range from temper tantrums and a need for control to social withdrawal, interpersonal difficulties, concentration problems, and somatic complaints, may not be immediately recognized as interconnected or indicative of an underlying AD.[18,19] Detection is further complicated by the overlap of these symptoms with other mental health disorders.[20] The constrained timeframe of GP consultations necessitates swift and accurate interpretation of presented problems, making the initial diagnostic impression critical for effective identification of anxiety in children.[21] Despite the widespread prevalence of ADs, GPs already seem to overlook anxiety in their early diagnostic opinion.[22,23]

This study examined the following research questions:

RQ1: How do various AI tools (ChatGPT-3, ChatGPT-4, Claude.AI, and Gemini) compare to human professionals (MHPs and GPs) in accurately recognizing anxiety, as measured by their performance in total recognition, first identification, and second identification questions?

RQ2: How often do various AI tools (ChatGPT-3, ChatGPT-4, Claude.AI, and Gemini) identify anxiety compared to human professionals (MHPs and GPs)?

RQ3: What is the ideal placement for a child diagnosed with anxiety disorder according to the various AI tools (ChatGPT-3, ChatGPT-4, Claude.AI, and Gemini) compared to human professionals (MHPs and GPs)?

RQ4: How do treatment recommendations for mental and behavioral disorders differ between AI tools and GPs?

Method

AI procedure

The findings were collected during February 2024. The transfer of the vignettes took one month and was conducted using the interfaces of the language models ChatGPT, Claude, and Gemini to identify childhood ADs. These evaluations were compared with the results reported by Dutch mental health professionals (MHPs) and GPs as described by Aydin et al.[23]

Input source: vignettes

To investigate the extent to which LLMs are sensitive to ADs in children, we utilized a series of clinical vignettes developed by Aydin et al.[23] These vignettes were designed to depict the varied symptom presentations commonly seen in pediatric ADs. The original researchers constructed the vignettes to portray a probable underlying AD, while also including symptoms that overlap with other common mental health issues. The vignette development process employed by Aydin et al.[23] began with a review of clinical handbooks and questionnaires to identify relevant symptoms and characteristics. The researchers also analyzed actual referral letters written by GPs for children who were later diagnosed with ADs. These letters provided natural language descriptions of presenting complaints and enabled the researchers to map the correspondence between GPs' stated reasons for referral and the eventual diagnoses. According to Aydin's[23] research, the sample of GPs had varying levels of experience, with 54.1% having over 20 years of practice, indicating a predominantly highly experienced group.

The researchers categorized the extracted descriptions into five domains of symptoms that commonly co-occur with anxiety: somatic complaints, difficult behaviors, depressed mood, developmental problems, and school attendance issues. They then developed vignettes representing each domain. Each vignette depicted a 10- to 12-year-old child with anxiety symptoms as well as attributes suggesting other potential problems. Contextual details were included to enhance realism. The original authors undertook an iterative process of drafting, review, and refinement, in consultation with MHPs and GPs. The vignettes were adjusted to achieve a consistent length of 165 to 172 words each. Audio recordings of the vignettes were created, with the text also included as subtitles, to replicate the verbal nature of clinical encounters while ensuring consistent presentation to participants. This multistep process resulted in a final set of five vignettes, each centered around one key problem area while incorporating a greater number of anxiety cues than other mental health symptoms. The vignettes aimed at authentically capturing the ambiguity of real-world clinical presentations in general practice. The vignettes are described in Appendix 1.

Measures. The LLMs were asked to respond to a series of questions designed to assess their interpretation of the presenting problems and their decision-making with respect to patient referral and management. The survey items were divided into two main sections:

Section A: Vignette-specific questions (A1–A3) posed after each vignette:

A1: First complaint group: *What is the main complaint? (Where do you think this description fits in? With which symptom?)*

1. *Typical development (Option one if it is probably an example of typical development)*
2. *Behavioral problems (Option two for difficult behavior: examples include aggressive behavior or antisocial behavior)*
3. *Complaints regarding establishing contact (If problems likely indicate an autism spectrum disorder, you can choose option three)*
4. *Mood problems (Mood problems and problems that could be related to depressive disorders)*
5. *Somatic complaints (For physical symptoms choose option five, also if a problem might be psychosomatic in nature)*
6. *Eating problems (Option six for eating problems and probable eating disorders)*
7. *Anxiety-related complaints (Option seven for problems related to anxiety and anxiety disorders)*
8. *Complaints regarding attention and activity (Option eight for attention-related complaints that might indicate attention deficit hyperactivity disorder or attention deficit disorder)*
9. *Complaints related to experience of a traumatic event (Option nine for problems related to experience of a traumatic event). What profile would you ascribe this vignette?*

A2: Second complaint group *(If you would like to add a second problem to the main complaint groups you can select it here.)*

1. *No second complaint group (Please choose option 10 if you do not see another complaint.)*

A3: *Where should this child ideally be placed? (Where can this child and the family get the most adequate professional support?)*

1. *Nurse practitioner*
2. *Local youth teams*
3. *Generalized mental healthcare*
4. *Specialized mental healthcare*
5. *Somatic healthcare/hospital*

Section B: General referral tendency questions (B1–B8) presented after all five vignettes have been evaluated:

Where do you think that children with this type of complaint can best be helped? (The eight mental health groups are shown. For each group, please indicate how you would tend to refer children when you suspect these complaints.)

Answer questions B1–B8 by indicating the number of the option you would choose for each question among the options provided.

B1–B8. *Where do you think children with (B1) Behavioral problems/(B2) Complaints regarding establishing contact/(B3) Mood problems/(B4) Somatic complaints/(B5) Eating problems/(B6) Anxiety-related complaints/(B7) Attention and activity/(B8) Complaints related to experiencing a traumatic event can best be helped?*

1. *Watchful waiting*
2. *Nurse practitioner*
3. *Local youth teams*
4. *General mental healthcare*
5. *Specialized mental healthcare.*

Testing the large language models

Each of the five vignettes and the three questions (A1–A3) were input ten times into each of the four AI language models: Claude, ChatGPT 3.5, ChatGPT 4, and Gemini. Each of the ten measurements was conducted in a new tab to avoid the influence of previous information. Questions B1–B8 were asked ten times together in the same tab, with each of the ten measurements conducted in a new tab. Table 1 describes the number of iterations for the different vignettes and questions.

Statistical analysis

The statistical methodology used in this study was designed to provide a rigorous assessment of the effectiveness of various AI tools and human professionals in recognizing anxiety. We calculated three types of recognition rates for anxiety: (1) First Recognition Rate: The percentage of cases where anxiety was identified as the primary concern. (2) Second Recognition Rate: The percentage of cases where anxiety was identified as a secondary concern. (3) Total Recognition Rate: The percentage of cases where anxiety was identified in either the first or second instance. This rate represents the overall ability to recognize anxiety, regardless of whether it was identified as the primary or secondary concern. The first RQ was examined using Chi-squared test of independence to assess whether the differences in total recognition, first identification, and second identification rates across these entities were statistically significant. Due to the multiple comparisons involved in comparing each pair of entities,

Table 1. Number of iterations for the different vignettes and questions.

LLM	V1 iterations	V2 iterations	V3 iterations	V4 iterations	V5 iterations	Total vignette iterations	Question B iterations	Total iterations
Claude	10	10	10	10	10	50	10	**60**
ChatGPT-3.5	10	10	10	10	10	50	10	**60**
ChatGPT-4	10	10	10	10	10	50	10	**60**
Gemini	10	10	10	10	10	50	10	**60**
Total	**40**	**40**	**40**	**40**	**40**	**200**	**40**	**240**

Bonferroni correction method was used. The second RQ was explored using a Chi-squared test of independence to determine whether the likelihood of recognizing anxiety varied among the different evaluators, taking the multiple vignettes into account as repeated measures. The third RQ was analyzed descriptively by comparing the percentage of referrals to different healthcare services by each entity, highlighting differences in their approach to managing anxiety. The fourth RQ was analyzed descriptively by comparing the treatment recommendations for mental and behavioral disorders across AI tools and GPs. This approach quantified the differences in recommendation patterns, highlighting the potential for integrating AI into clinical decision-making.

Results

Comparison of effectiveness of AI tools (ChatGPT-3, ChatGPT-4, Claude.AI, and Gemini) and human professionals (MHPs and GPs) in recognizing anxiety

To evaluate the differences in performance across these entities, Chi-squared tests of independence were conducted to assess the statistical significance of the observed frequencies in each category. Given the multiple comparisons across entities, post-hoc pairwise Chi-squared tests with Bonferroni correction were employed to identify which specific pairs of entities exhibited significantly different recognition rates.

Chi-squared tests of independence revealed significant differences in total recognition rates across the entities ($\chi^2(5, N = 600) = 34.38$, $p < .001$). Similarly, significant differences were found for the first identification ($\chi^2(5, N = 600) = 37.41$, $p < .001$) and second identification ($\chi^2(5, N = 600) = 25.31$, $p < .001$) questions. These results suggest that not all entities performed equally well in recognizing anxiety.

Post-hoc pairwise comparisons with Bonferroni correction revealed significant differences between ChatGPT-4 and both Claude.AI and Gemini for first identification and between ChatGPT-3 and GPs for second identification.

Table 2. Comparison of anxiety recognition rates across different entities.

Entity	Total recognition	First identification	Second identification
ChatGPT-3	54.0%	8.0%	46.0%
ChatGPT-4	38.0%	4.0%	34.0%
Claude.AI	78.0%	34.0%	46.0%
Gemini	78.0%	34.0%	44.0%
MHPs	40.0%	16.4%	24.5%
GPs	14.8%	7.9%	7.8%

Note: The Total Recognition rate represents cases where anxiety was identified in either the first or second instance. It is not necessarily the sum of first and second identification rates.

Claude.AI and Gemini also exhibited significantly better recognition rates compared to GPs on both identification questions. These findings suggest that specific AI tools (Claude.AI and Gemini) outperformed others as well as human professionals in recognizing anxiety. Table 2 and Figure 1 show the anxiety recognition rates across the different entities.

Comparison of number of times anxiety was recognized over five vignettes by various AI entities and medical professionals

A Chi-squared test of independence was conducted to explore the relationship between type of evaluator (ChatGPT-3, ChatGPT-4, Claude.AI, Gemini, MHPs, and GPs) and frequency of anxiety recognition across the five vignettes. The analysis revealed a significant effect of type of evaluator on recognition of anxiety ($\chi^2(25) = 931.66$, $p < .001$), suggesting that the probability of recognizing anxiety

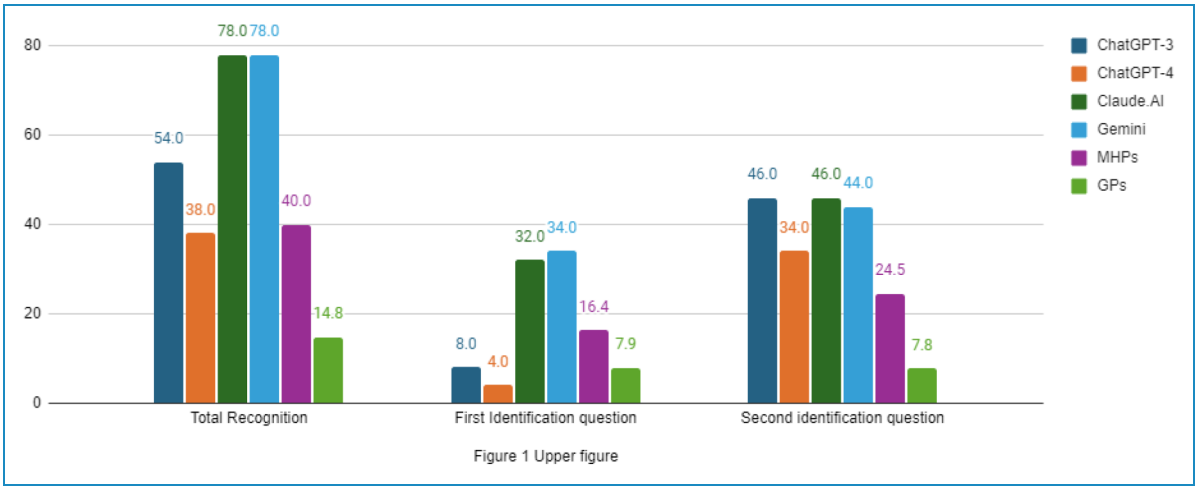

Figure 1. Comparison of anxiety recognition rates across different entities.

Table 3. Percentage of times anxiety was identified by different AI tools and medical professionals, over five vignettes.

AI tool/medical professional	0 vignette (%)	1 vignette (%)	2 vignettes (%)	3 vignettes (%)	4 vignettes (%)	5 vignettes (%)
ChatGPT-3	0.0	10.0	10.0	80.0	0	0.0
ChatGPT-4	0.0	30.0	50.0	20.0	0	0.0
Claude.AI	0.0	0.0	0.0	20.0	70	10.0
Gemini	0.0	10.0	0.0	10.0	50	30.0
MHPs	0.0	27.3	54.5	9.1	0	9.1
GPs	44.1	41.9	11.8	1.7	0	0.4

Table 4. Comparison of referral recommendations among different LLMs and GPs.

LLM	Nurse practitioner	Local youth teams	Somatic health primary MHC	Specialized MHC	MHC
ChatGPT-3	0.0	4.0	34.0	42.0	20.0
ChatGPT-4	0.0	0.0	14.0	84.0	2.0
Claude AI	2.0	22.0	36.0	40.0	0.0
Gemini	0.0	2.0	58.0	38.0	2.0
GPs	39.9	23.5	21.1	13.3	2.1

varied among the different evaluators. Due to the low number of observations, a post-hoc analysis was not conducted.

As can be seen in Table 3 and Figure 2, all the AI tools recognized at least one vignette as anxiety, similar to the MHPs and contrary to the GPs. Claude.AI and Gemini identified at least four of the vignettes as anxiety. The comparison between ChatGPT-3 and ChatGPT-4 shows that ChatGPT-3 identified three vignettes as anxiety in 80% of the cases, whereas ChatGPT-4 identified only 1 or 2 cases of anxiety in 80% of the cases.

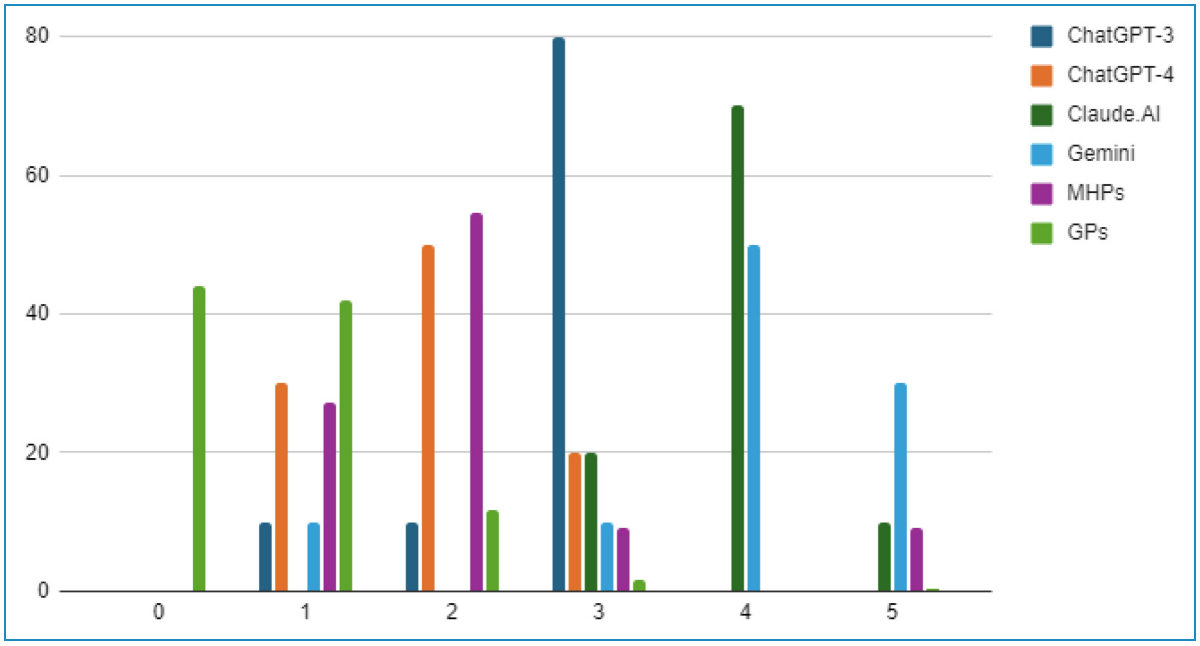

Figure 2. Percentage of times anxiety was selected by different AI tools and medical professionals, over five vignettes.

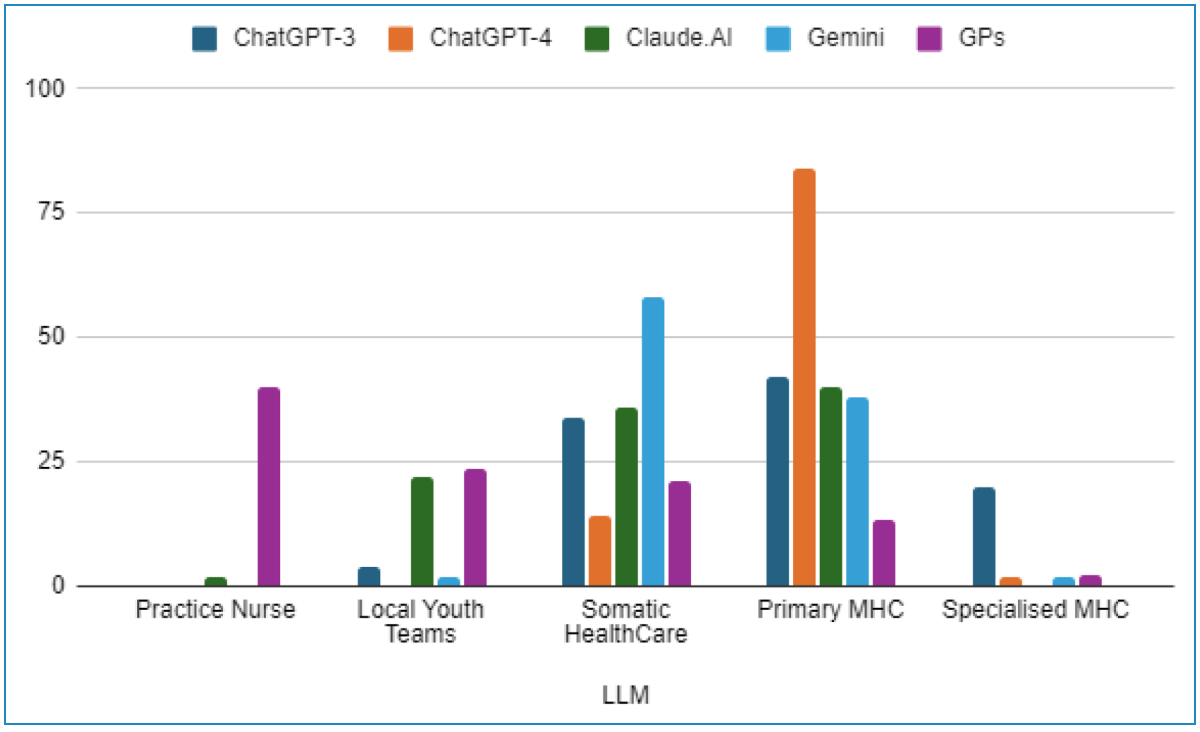

Figure 3. Comparison of referral recommendations among different LLMs and GPs.

Ideal referral for child

The comparison between the four AI tools and the GPs regarding where they would refer a child with a profile similar to the one in the vignette elicited interesting observations (Figure 3). The majority of the GPs responded that they would recommend treating the child in general practice (nurse practitioner = 40%). The AI tools, in contrast, did not recommend treating the child in general practice, except for Claude.AI in one case. Instead, the AI tools recommended referring the child to primary MHC or somatic healthcare.

Claude.AI. recommended referring the child to local youth teams more frequently than the other AI tools, whereas ChatGPT-3 recommended referring the child to specialized MHCs more frequently than the other AI tools (Table 4).

Referral preferences for different types of disorders

The last stage of this research compared the recommendations of the four AI tools—ChatGPT-3, ChatGPT-4, Claude AI, Gemini—and of the GPs for eight distinct mental and behavioral disorders: anxiety, trauma, mood disorders, physical symptoms with psychological basis, eating problems, autism spectrum disorders, attention deficit hyperactivity disorder (ADHD), and difficult behavior. The treatment options ranged from less intensive approaches such as watchful waiting to more specialized interventions such as specialized mental healthcare. Table 5 compares the percentage of treatment recommendations for mental and behavioral disorders of AI tools and GPs.

Anxiety. For ADs, GPs exhibited a balanced approach, recommending a variety of treatment levels, with a notable preference for general mental healthcare (53.6%) and nurse practitioner interventions (25.9%). In contrast, AI tools tended toward more extreme recommendations, with ChatGPT-4 and Claude AI favoring general mental healthcare and ChatGPT-3 and Gemini recommending specialized mental healthcare in all responses.

Trauma. In the case of trauma, the AI tools again showed a propensity for more intensive treatments, with ChatGPT-4, Claude AI, and Gemini predominantly recommending specialized mental healthcare. The GPs demonstrated a more distributed approach, albeit with a tendency toward recommending specialized mental healthcare (37.3%) and general mental healthcare (42.9%).

Mood disorders. For mood disorders the AI responses differed, with ChatGPT-4 and Claude AI significantly recommending general mental healthcare, diverging from ChatGPT-3's sole recommendation of specialized mental healthcare. GPs favored general mental healthcare (52.3%) but also considered nurse practitioner involvement (27.7%).

Physical symptoms. In addressing physical symptoms with psychological underpinnings, AI tools and GPs both recommended a range of treatments. GPs exhibited a notable preference for specialized mental healthcare (25.5%) and nurse practitioner interventions (33.5%). ChatGPT-4 uniquely recommended a balance of general mental healthcare and nurse practitioner interventions, while Gemini leaned toward less intensive options.

Eating problems. Eating problems elicited strong recommendations for specialized mental healthcare from all AI tools, with Claude AI equally endorsing general mental healthcare. In contrast, GPs predominantly recommended specialized mental healthcare (61.2%) but also considered less intensive options to a lesser extent.

Autism spectrum disorders. For autism spectrum disorders, the AI tools favored more intensive interventions, with ChatGPT-4 and Gemini primarily recommending specialized mental healthcare. GPs also preferred specialized mental healthcare (23%) but exhibited significant consideration of nurse practitioner (24.9%) and general mental healthcare (24.9%) interventions.

Attention deficit hyperactivity disorder. In the treatment of ADHD, ChatGPT-4 and Gemini showed a strong inclination toward specialized mental healthcare, whereas Claude AI recommended general mental healthcare. GPs displayed a balanced approach, with no clear preference among the treatment options, suggesting a case-by-case evaluation.

Difficult behavior. For difficult behavior, AI responses varied, with ChatGPT-4 and Gemini recommending specialized mental healthcare, while Claude AI favored general mental healthcare. GPs preferred general mental healthcare (44.4%) and nurse practitioner intervention (32.9%), indicating a preference for a step-up approach based on severity.

In summary, the AI tools tended to recommend more intensive treatment options across most disorders, particularly specialized mental healthcare. GPs, in contrast, displayed a more nuanced approach, considered a wider range of treatments and showed a tendency to prefer general mental healthcare and nurse practitioner interventions, potentially reflecting a more holistic and staged approach to treatment. This comparison highlights the potential differences in treatment recommendation patterns between AI tools and human practitioners, underscoring the importance of integrating AI recommendations with clinical judgment in healthcare decision-making.

Discussion

This study sought to compare AI tools and human professionals in recognizing anxiety among children. Our first research question asked how various AI tools compare to human professionals in accurately recognizing anxiety. Our findings directly answer this question, revealing that Claude.AI and Gemini demonstrated better detection rates than GPs in both identification instances. The research findings reinforce the existing body of literature by highlighting that unnoticed ADs are prevalent among children.[9] Even

Table 5. Comparative analysis of percentage of treatment recommendations for disorders by AI tools and GPs.

		ChatGPT-3	ChatGPT-4	Claude AI	Gemini	GPs
Anxiety	Watchful waiting					2.2
	Nurse practitioner					25.9
	Local youth teams					8.0
	General mental healthcare		70	100	100	53.6
	Specialized mental healthcare	100	30			10.3
Trauma	Watchful waiting					1.4
	Nurse practitioner					14.3
	Local youth teams					4.1
	General mental healthcare				40	42.9
	Specialized mental healthcare	100	100	100	60	37.3
Mood	Watchful waiting					0.9
	Nurse practitioner					27.7
	Local youth teams					6.8
	General mental healthcare		80	100	100	52.3
	Specialized mental healthcare	100	20			12.3
Physical	Watchful waiting	10			30	23.6
	Nurse practitioner	10	40	100	70	33.5
	Local youth teams	20				8.0
	General mental healthcare	60	60			9.4
	Specialized mental healthcare					25.5
Eating problems	Watchful waiting					4.1
	Nurse practitioner					9.1
	Local youth teams					9.6
	General mental healthcare	10		50	10	16.0
	Specialized mental healthcare	90	100	50	90	61.2
Autism spectrum	Watchful waiting					6.1
	Nurse practitioner					24.9
	Local youth teams		80		90	24.9

(continued)

Table 5. Continued.

		ChatGPT-3	ChatGPT-4	Claude AI	Gemini	GPs
	General mental healthcare	90		100	10	23.0
	Specialized mental healthcare	10	20			21.1
Attention hyperactivity	Watchful waiting					1.4
	Nurse practitioner					14.0
	Local youth teams			100		11.7
	General mental healthcare	90	40		90	40.2
	Specialized mental healthcare	10	60		10	32.7
Difficult behavior	Watchful waiting					10.3
	Nurse practitioner					32.9
	Local youth teams		10	80	10	44.4
	General mental healthcare		10	20	90	10.2
	Specialized mental healthcare	100	80			2.3

Note: 1. The numbers in the table represent the percentage of responses for suggested treatment. 2. Empty cells signify that this option was not suggested as a response.

though children typically visit their pediatricians more than twice a year, over two-thirds of anxiety cases remain undiagnosed.[10,11] GPs may fail to recognize ADs, suggesting that unfamiliarity with early symptom presentation may be a key factor. GPs adequately recognized anxiety in vignettes that explicitly mentioned 'fears,' but struggled with less overt presentations. Some GPs focused heavily on school and home functioning while overlooking other relevant domains, such as social relationships.[23] This underscores the potential use of AI to assist healthcare professionals as a supportive tool in detecting ADs. Investigations into the application of AI within the domain of mental health care (MHC) have yielded promising outcomes, with research examining its effectiveness across various dimensions, such as assessment, ongoing observation, and provision of therapeutic measures. Tools employing natural language processing have attracted significant interest due to their proficiency in mimicking human-like discourse and facilitating interactive dialogues. These instruments, often referred to as AI tools, offer potential for administering empirically supported mental health interventions.[4] Such interventions have shown initial success in diminishing symptomatology and enhancing overall mental well-being, with some investigations indicating considerable user satisfaction.[24] For instance, relative to the provision of assistance by human personnel, applications of AI-based tools have demonstrated efficacy in managing patient mental health.[25,26] Furthermore, a recent comprehensive review of these applications revealed that mental health-oriented chatbots are characterized by their user-friendliness, appealing design, prompt responsiveness, reliability, and overall user satisfaction.[27]

The second research question inquired about the frequency of anxiety identification by AI tools compared to human professionals. Our study revealed that each AI instrument successfully identified at least one case study as indicative of anxiety, mirroring MHPs and contrasting with GPs' assessments. Claude.AI and Gemini identified a minimum of four vignettes as instances of anxiety. A comparison of the performances of ChatGPT-3 and ChatGPT-4 revealed that ChatGPT-3 recognized anxiety in three vignettes, whereas ChatGPT-4 recognized anxiety in only one or two vignettes. Even though no studies were found that directly compare Claude.AI and Gemini with medical teams, the current study also highlights the potential use of AI in mental health. The results of this study are in contrast to previous research that examined suicide, depression, and schizophrenia, in which superior capabilities were attributed to the ChatGPT-4 language model over other AI languages.[2,3,5,28] For example, a comparative analysis of different LLMs, such as ChatGPT-3.5, ChatGPT-4, Claude, and Bard, that included mental health

professionals (GPs, psychiatrists, clinical psychologists, and mental health nurses) revealed variations in AI assessments regarding recovery outcomes. Specifically, ChatGPT-3.5 tended to exhibit a consistently pessimistic viewpoint, whereas the assessments of ChatGPT-4, Claude, and Bard were more in line with the perspectives of mental health professionals and the general populace.[3]

Our third research question explored the ideal placement recommendations for children with anxiety according to AI tools versus human professionals. The findings revealed, that on the question of the optimal referral for the child, a significant proportion of GPs indicated a preference to retain the child within the general practice setting, with 40% opting for nurse practitioner. In contrast, the AI tools for the most part did not advocate retaining the child in general practice, with the exception of a solitary instance involving Claude.AI. Instead, the AI tools displayed a tendency to recommend referring the child to primary MHC or somatic health care services. Aydin et al.[23] observed a notable discrepancy in GPs' responses. While a majority of GPs indicated they would consider referring patients when they suspected ADs, in the presented case vignettes for the most part they opted for management within primary healthcare settings rather than referral to MHC. Conversely, AI systems showed a preference for referring cases to mental health services. This discrepancy may suggest that AI is capable of identifying a broader spectrum of symptoms without the influence of factors such as consultation pressure that a GP might experience.[29] Additionally, it may imply that doctors, while occasionally suspecting ADs, only refer to the most severe cases for specialized treatment, possibly due to constraints within their practice environment.[30]

The final research question asked how treatment recommendations for mental and behavioral disorders differ between AI tools and GPs. The current study compared treatment recommendations for eight different mental and behavioral disorders, drawing upon the insights from four AI tools—ChatGPT-3, ChatGPT-4, Claude AI, Gemini—and the responses of GPs. The treatment options under consideration ranged from less intensive approaches such as watchful waiting to more specialized interventions such as specialized mental healthcare. The AI tools exhibited a notable inclination toward recommending more intensive treatments, particularly specialized mental healthcare, across the majority of the disorders. This tendency was most pronounced in the responses for anxiety, trauma, and eating problems, where the AIs almost uniformly suggested the highest level of care. In contrast, GPs demonstrated a more balanced and diversified approach in their treatment recommendations. For most disorders, they considered a range of treatment options, with a tendency to prefer general mental healthcare and nurse practitioner interventions. This suggests a more graduated approach to care, potentially taking into account the individual patient's circumstances and the severity of the disorder. In the field of medicine, preliminary studies have found a good predictive capacity for adapting drug treatment to AI.[31,32] Research in mental health presents conflicting evidence about the predictive capabilities of AI compared to professionals. A review study highlights a shortfall in the diversity of treatment options provided by AI relative to therapists, undermining its reliability and utility for clinicians.[33] In another study, ChatGPT-3.5 and ChatGPT-4 mainly advised psychotherapy for mild depression (95% and 97.5%), unlike primary care physicians who rarely recommended psychotherapy (4.3%). In severe cases, ChatGPT favored psychotherapy, while physicians preferred combined treatments.[2]

The current study found the most significant divergence between AI tools and GPs in the case of trauma and eating problems, with AIs predominantly recommending specialized mental healthcare, whereas GPs exhibited a broader distribution of recommendations. The literature review reveals multiple ways in which AI could enhance MHC, including support in self-management, diagnostic precision, and treatment monitoring.[34] Nevertheless, a recurring concern expressed by family doctors is the impact of AI on therapeutic relationships. Substituting AI for human contact could undermine the therapeutic alliance due to diminished empathy and misinterpretations.[35] Given the low remission rates and the complexity of these disorders, it remains uncertain whether AI can fully grasp and address them. Consequently, skilled clinicians who understand the individual's complexity and background and the nature of the disorder are still needed.

In addition, the category of physical symptoms was unique in that both AIs and GPs suggested a variety of treatment options, indicating no clear consensus regarding the most appropriate level of care. Adolescents frequently mask mental health issues with physical symptoms in primary care, yet over 50% remain unscreened for such concerns. Reviews of 11 studies suggest that pre-consultation electronic screening is effective in fostering open discussions and greater youth disclosure.[36]

In this study, the AI tools' preference for more intensive treatments raises questions about their risk-version or potential for overestimating the need for specialized care. This preference may reflect the algorithms' prioritization of caution in the absence of nuanced clinical judgment. AI should augment rather than replace human decision-makers. Decision-making transcends algorithmic analysis by incorporating distinctly human traits such as creativity, intuition, and ethical judgments, highlighting the need for an integrated approach that combines AI with human insight.[37]

The deployment of AI in the identification of anxiety disorders holds considerable promise, yet it is accompanied by an array of challenges. The precision of AI predictions is contingent upon the caliber and demographic inclusiveness of the datasets utilized for algorithm training. Data that are

biased or lack adequate representation can lead to erroneous forecasts or exacerbate existing healthcare disparities. Moreover, AI algorithms frequently function as opaque entities, obscuring the logic underpinning their predictions. This opacity can hinder the development of trust and acceptance among end-users. Moreover, the application of AI in the diagnosis of anxiety disorders raises several ethical concerns. Paramount among these is the issue of safeguarding data privacy and security, given the particularly sensitive nature of mental health information. It is imperative that users receive clear and thorough information regarding how their data will be used and protected. Additionally, AI should not supplant the clinical acumen of healthcare professionals in diagnosing anxiety disorders. Rather, it should serve as an auxiliary tool that enhances the capability of practitioners to make informed clinical judgments.

Clinical implications

The findings of our research contribute to the discourse on overcoming obstacles in the provision of mental health services, in light of the inadequacy of current resources and intervention strategies to meet the extant and burgeoning demands. A report by the World Health Organization indicates that in excess of 400 million individuals worldwide are afflicted by mental health conditions. While psychotherapeutic interventions and social support mechanisms have demonstrated efficacy, there exists a significant barrier in terms of accessibility to such therapeutic and counseling services for numerous at-risk populations. For example, the psychiatrist-to-population ratio in most nations is less than 1 per 100,000 people, highlighting a critical shortfall in professional workforce and a dearth of available face-to-face treatment modalities.

The contrast between AI and GP recommendations underscores the potential for a synergistic approach in which AI can offer a preliminary assessment that GPs can refine, balancing the efficiency and breadth of AI with the nuanced, patient-centered approach of human practitioners. The varied treatment recommendations by GPs highlight the importance of human judgment in healthcare, which considers the patient's broader context, including factors that AI may not yet adequately account for.

Limitations

The current study highlights AI's potential in identifying anxiety disorders in children, yet certain limitations must be considered. The reliance on standardized vignettes may not fully encompass the complexity inherent in real-life assessments of anxiety disorders, indicating a need for further research that includes more detailed and personalized cases. Additionally, future studies must extend their scope to include a broader spectrum of factors, such as socioeconomic status, cultural backgrounds, and individuals' previous mental health records, to achieve a more comprehensive evaluation. The cross-sectional nature of this research limits the ability to track the evolution of diagnostic precision over time. The inherent opacity of AI algorithms poses a challenge to the interpretation of their decisions. Developing innovative techniques that can demystify and illustrate the logic behind AI judgments would enhance transparency and insight. Moreover, the experimental setting of this study may not adequately reflect the intricate realities of implementing AI in actual clinical environments. Future research should investigate the practical aspects and ramifications of utilizing such AI models in genuine MHC settings, taking into account the interaction between human clinicians and AI technologies and the practicality of their integration.

To enhance the diagnostic accuracy of AI models in identifying childhood anxiety disorders, it is essential to incorporate a comprehensive range of information. This should include detailed symptom profiles encompassing both physical and psychological symptoms, as well as developmental and contextual data, such as the child's age, family history, and environmental influences. Longitudinal data on symptom progression, comorbidity information, and assessments of functional impact across various domains are also critical. Additionally, standardized scores from validated anxiety screening tools, treatment history, and cultural and socioeconomic factors should be integrated. Reports from schools and teachers, along with parental observations of home behavior and family dynamics, further contribute to a holistic understanding of each case. By incorporating this diverse array of data points, AI models can develop a more nuanced and accurate diagnostic capability. However, it is crucial to emphasize that, while AI can efficiently process and analyze complex information, the interpretation and application of AI-generated insights should always be guided by clinical expertise. This approach ensures ethical and patient-centered care, aligning AI's capabilities with the nuanced understanding provided by healthcare professionals.

Conclusion

This study explored the efficacy of AI tools compared to human professionals in identifying anxiety among children. The findings indicate a significant difference in recognition capabilities, with AI tools such as Claude.AI and Gemini demonstrating superior performance over GPs in the initial identification phases. This discrepancy highlights the potential of AI to enhance diagnostic processes in mental health. However, it is essential to consider AI as a supportive adjunct to healthcare professionals, rather than as a standalone diagnostic method. AI has the potential to complement traditional approaches by providing an additional layer of analysis, thereby improving the accuracy

and efficiency of anxiety detection. Nonetheless, the integration of AI-generated insights should be balanced with the clinical judgment, expertise, and nuanced understanding that healthcare professionals bring to patient care.

The study also examined treatment recommendations, revealing differences between AI tools and GPs. AI tools tended to recommend more intensive treatments and specialized mental healthcare, especially for conditions like anxiety, trauma, and eating disorders. In contrast, GPs often preferred less intensive options, such as general mental healthcare or the involvement of a nurse practitioner. This divergence underscores the need for a collaborative approach that leverages both AI and human expertise. While AI can offer valuable insights and help identify cases that may otherwise be overlooked, healthcare professionals play a critical role in contextualizing these insights within the broader framework of the patient's life, preferences, and available resources.

Integrating AI into mental health diagnostics and treatment planning holds significant promise, but it should be implemented within a holistic framework that emphasizes the irreplaceable role of healthcare professionals. Future research and clinical practice should focus on developing methodologies that effectively combine AI capabilities with human expertise to optimize patient outcomes in MHC.

Abbreviations

AI	artificial intelligence
LLMs	large language models
ADs	anxiety disorders
GPs	general practitioners
MHPs	mental health professionals

Availability of data and materials: The authors have the research data, which is available upon request.

Contributorship: IL contributed to the research design, wrote the study protocol, organized the study, coordinated the data collection, drafted the initial manuscript, and approved the final submitted manuscript. ZE contributed to the study design and organization, participated in the data collection, reviewed the manuscript, and approved the final submitted manuscript. MB contributed to the data collection. ER carried out the analysis. All authors read and approved the final version of the manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Disclosure instructions: During the preparation of this work the authors used SPSS, Google sheets and ChatGPT to analyze and visualize the data. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Ethical approval: The study was approved by the Ethics Committee of Oranim College (Authorization No. 1852024) and was conducted in accordance with the Declaration of Helsinki. No patients were involved, and there was no need for informed consent from all participants.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD: Inbar Levkovich https://orcid.org/0000-0003-1582-3889

Patient and public involvement: No patients were involved.

Supplemental material: Supplemental material for this article is available online.

References

1. Blease C, Locher C, Leon-Carlyle M, et al. Artificial intelligence and the future of psychiatry: qualitative findings from a global physician survey. *Digit Health* 2020; 6: 205520 7620968355.
2. Levkovich I and Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. *Fam Med Community Health* 2023; 11: e002391.
3. Elyoseph Z, Levkovich I and Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. *Fam Med Community Health* 2024; 12: e002583.
4. Boucher EM, Harake NR, Ward HE, et al. Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Rev Med Devices* 2021; 18: 37–49.
5. Levkovich I and Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. *JMIR Ment Health* 2023; 10: e51232.
6. Hadar-Shoval D, Elyoseph Z and Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. *Front Psychiatry* 2023; 14: 1234397.
7. Bosman RC, Ten Have M, de Graaf R, et al. Prevalence and course of subthreshold anxiety disorder in the general population: a three-year follow-up study. *J Affect Disord* 2019; 247: 105–113.
8. Copeland WE, Angold A, Shanahan L, et al. Longitudinal patterns of anxiety from childhood to adulthood: the great smoky mountains study. *J Am Acad Child Adolesc Psychiatry* 2014; 53: 21–33.
9. Haller H, Cramer H, Lauche R, et al. The prevalence and burden of subthreshold generalized anxiety disorder: a systematic review. *BMC Psychiatry* 2014; 14: 1–13.
10. Baxter AJ, Scott KM, Vos T, et al. Global prevalence of anxiety disorders: a systematic review and meta-regression. *Psychol Med* 2013; 43: 897–910.
11. Pape K, Bjørngaard JH, Holmen TL, et al. The welfare burden of adolescent anxiety and depression: a prospective study of 7500 young Norwegians and their families: the HUNT study. *BMJ Open* 2012; 2: e001942.

12. Bennett K, Manassis K, Duda S, et al. Preventing child and adolescent anxiety disorders: overview of systematic reviews. *Depress Anxiety*. 2015; 32: 909–918.
13. Moreno-Peral P, Conejo-Ceron S, Rubio-Valera M, et al. Effectiveness of psychological and/or educational interventions in the prevention of anxiety: a systematic review, meta-analysis, and meta-regression. *JAMA Psychiatry* 2017; 74: 1021–1029.
14. Cybulski L, Ashcroft DM, Carr MJ, et al. Management of anxiety disorders among children and adolescents in UK primary care: a cohort study. *J Affect Disord* 2022; 313: 270–277.
15. O'Brien D, Harvey K and Creswell C. Barriers to and facilitators of the identification, management and referral of childhood anxiety disorders in primary care: a survey of general practitioners in England. *BMJ Open* 2019; 9: e023876.
16. Koning NR, Büchner FL, Verbiest ME, et al. Factors associated with the identification of child mental health problems in primary care—a systematic review. *Eur J Gen Pract* 2019; 25: 116–127.
17. Kendall PC, Swan AJ, Carper MM, et al. Anxiety disorders among children and adolescents. In: Butcher JN and Kendall PC (eds). *APA handbook of psychopathology: child and adolescent psychopathology*. Washington, DC, USA: American Psychological Association, 2018, pp. 213–230. DOI:10.1037/0000065-011
18. Essau CA, Lewinsohn PM, Lim JX, et al. Incidence, recurrence and comorbidity of anxiety disorders in four major developmental stages. *J Affect Disord* 2018; 228: 248–253.
19. Lijster JMd, Dierckx B, Utens EM, et al. The age of onset of anxiety disorders: a meta-analysis. *Can J Psychiatry* 2017; 62: 237–246.
20. Meulen WG T, Draisma S, van Hemert AM, et al. Depressive and anxiety disorders in concert–A synthesis of findings on comorbidity in the NESDA study. *J Affect Disord* 2021; 284: 85–97.
21. O'Brien D, Harvey K, Young B, et al. GPs' experiences of children with anxiety disorders in primary care: a qualitative study. *Br J Gen Pract* 2017; 67: e888–e898.
22. .Andrew A. Potential applications and implications of large language models in primary care. *Fam Med Community Health* 2024; 12: e002602.
23. Aydin S, Crone MR, Siebelink BM, et al. Recognition of anxiety disorders in children: a cross-sectional vignette-based survey among general practitioners. *BMJ Open* 2020; 10: e035799.
24. Pham KT, Nabizadeh A and Selek S. Artificial intelligence and chatbots in psychiatry. *Psychiatr Q* 2022 Mar; 93: 249–253.
25. Goodman RS, Patrinely Jr JR, Osterman T, et al. On the cusp: considering the impact of artificial intelligence language models in healthcare. *Med* 2023; 4: 139–140.
26. Biswas SS. Role of Chat GPT in public health. *Ann Biomed Eng* 2023; 51: 868–869.
27. Sng GG, Tung JY, Lim DY, et al. Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. *Diabetes Care* 2023; 46: e103–e105.
28. Elyoseph Z, Hadar Shoval D and Levkovich I. Beyond personhood: ethical paradigms in the generative artificial intelligence era. *Am J Bioeth* 2024; 24: 57–59.
29. Nowak DA, Sheikhan NY, Naidu SC, et al. Why does continuity of care with family doctors matter?: review and qualitative synthesis of patient and physician perspectives. *Can Fam Physician* 2021; 67: 679–688.
30. Dykxhoorn J, Osborn D, Walters K, et al. Temporal patterns in the recorded annual incidence of common mental disorders over two decades in the United Kingdom: a primary care cohort study. *Psychol Med* 2024; 54: 663–674.
31. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ* 2023; 23: 89.
32. Sheu YH, Magdamo C, Miller M, et al. AI-assisted prediction of differential response to antidepressant classes using electronic health records. *NPJ Digit Med* 2023; 6: 73.
33. Higgins O, Short BL, Chalup SK, et al. Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: an integrative review. *Int J Ment Health Nurs* 2023; 32: 966–978.
34. Rogan J, Bucci S and Firth J. Health care professionals' views on the use of passive sensing, AI, and machine learning in mental health care: systematic review with meta-synthesis. *JMIR Ment Health* 2024; 11: e49577.
35. Day S, Hay P, Tannous WK, et al. A systematic review of the effect of PTSD and trauma on treatment outcomes for eating disorders. *Trauma Violence Abuse* 2024; 25: 947–964.
36. Martel R, Shepherd M and Goodyear-Smith F. Implementing the routine use of electronic mental health screening for youth in primary care: systematic review. *JMIR Ment Health* 2021; 8: e30479.
37. Göndöcs D and Dörfler V. AI in medical diagnosis: AI prediction & human judgment. *Artif Intell Med* 2024; 149: 102769.