



Identifying Barriers to the Safe Use of Artificial Intelligence in Emergency Clinical Decision-Making and Developing a Targeted Training Program for Emergency Physicians in Baghdad: A Mixed-Methods Study in a Private Hospital

Dr. Muntaha Kadhem Mejbil

Dijlah university

Email: muntaha.kadhem@duc.edu.iq

Abstract

Background: Artificial intelligence tools are being deployed in emergency departments quicker than systems of governance, training and validation can put in place to ensure their safe use. There is practically no evidence on the barriers experienced by emergency medicine practitioners in Iraq and there is no developed curriculum for AI safety training for those working in this specialty.

Objectives: The purpose of this study is to identify the barriers to safe application of AI in emergency clinical decision making encountered by doctors in a private clinic in Baghdad, and to develop, implement and assess training activities aimed at overcoming these barriers.

Methods: The study was institutional three-phase multi-method research conducted between January and July 2026 in the emergency department of a private tertiary care hospital in Baghdad. Phase 1 consisted of surveys using cross-sectional established questionnaire (census of all 84 eligible emergency medicine practitioners) complemented with 16 semi-structured interviews. Phase 2 was based on applying Kern's six-step strategy to create a curriculum based on the barriers identified. Phase 3 included educational program assessment based on a single-group pre-post design at 1st and 2nd Kirkpatrick stages with an -week follow-up and measurements for the effect of automation bias.

Results: Seventy-six doctors took part in the study (90.5%). Whereas 63.2% of doctors claimed having practical experience in working with AI technology, only 11.8% reported having received any formal training. The barriers were grouped into the domains rated highest: knowledge and skills gaps (4.12/5), legal and liability issues (3.98) and data constraints/validity (3.94). Prior training and digital literacy negatively correlated with the barrier score. Five qualitative themes emerged with the most important ones being trust issues. In the study with 39 participating physicians, the mean knowledge score rose from 9.8 on a scale of 20 to 15.6 ($p < 0.001$), while the percentage of errors detected was reported as increasing from 41.0% to 79.5%; on the other hand, the percentage of physicians inclined to accept incorrect AI recommendations fell from 51.3% to 17.9%. Nevertheless, results persisted after eight weeks.

Major conclusions regarding the barriers preventing safe AI use in this field are that they basically relate to educational, regulatory and contextual issues rather than technological ones. A short, barrier-specific simulation-based intervention caused considerable progress of physicians' ability to detect and override unsafe AI recommendations.

Keywords:

artificial intelligence; emergency medicine; clinical decision-making; patient safety; automation bias; medical education; Iraq.

Introduction

AI has transformed in less than a decade from a research curiosity to a means for clinicians to access applications on a smartphone. Deep learning systems are now able to match and even exceed specialists' performance on very well-defined tasks like reading ECGs or retinal pictures. While authorities^{1 2 3} argue that the merging of human and machine forms of intelligence is likely to change the way medicine is practiced, emergency medicine has been distinguished among the specializations that will be impacted most due to the necessity to deal with a heavy workload, time-sensitive decisions, diagnostic uncertainties, and to interpret clinical information quickly in emergency situations.

Currently, the scope of AI applications in emergency medicine is already wide enough. A scoping review of the use of AI in emergency departments has identified a number of studies on triage, prediction of admissions and mortality, early detection of sepsis, interpretation of imaging and ECGs, as well as patient flow prediction. However, it should be noted that most applications have been retrospectively evaluated and have not been tested in real clinical practice. The use of machine learning-based electronic triage systems showed better accuracy in predicting critical condition of patients when compared to the Emergency Severity Index (4 5). Moreover, the model based on electronic triage data was able to predict hospitalization and ICU outcomes at a higher level of accuracy in comparison with traditional models (6). Since 2023, with the advent of large language models such as GPT-which is capable of responding to any medical question asked in plain language and, thus, is capable of passing medical qualification tests successfully, the number of tools that can potentially be used in medicine has significantly increased (7 8 4 9).

The appeal of artificial intelligence in the Emergency Department is in part due to the inherent susceptibility of emergency decision-making to cognitive errors. Emergency doctors are compelled to practice often in times of fatigue, interruption, and information overload, making quick cognitive assumptions likely. Cognitive errors can include premature closure and anchoring, and are significant causes of diagnostic errors in medicine (10). In theory, a properly developed decision support tool could be a mental backstop. However, in practice, converting AI from good theoretical operation to yielding real clinical outcomes has been difficult. Some important challenges mentioned in literature include discrepancies between retrospective accuracy and prospective effectiveness, dataset shift when the model is applied to a new group of patients, difficulty of comparing different algorithms, and absence of reliable post-deployment monitoring (11). These are not merely theoretical problems. An external validation study of a very popular proprietary software utilized in hospital sepsis detection showed that that it does not perform nearly as well as claimed by its developers and misses the vast majority of people suffering from sepsis while generating numerous alerts (12). Algorithms can also encode existing biases in health care: one of the widely used algorithm used to allocate care management underestimated the needs of Black patients because it takes data on healthcare spending as a proxy of medical condition (13).

However, in addition to performance of algorithms, safety is associated with how they are used by practitioners. Challen et al. identified various AI-specific safety hazards, including

distribution shift, ignorance of consequences, black-box decision-making, and automation complacency, stating that new safety procedures are necessary instead of relying on systems that are well known in traditional software (14). One of the common phenomena affecting the medical field is automation bias, that refers to the excessive trust in automation and the belief without the necessary verification (15). The degree of complacency and automation bias is conditioned by the workflow load, the complexity of tasks, and the trust in the system used in the ED (16). Evidence suggests that the experiments showed the tendency of doctors, especially less experienced ones, to be misled by the artificial intelligence systems that provided wrong recommendations (17). What contributes to the problem is the fact that many of the recommendations derived from different models are beyond explanation, which evokes the necessity of establishing explanatory standards on legal, technical, medical, and other levels (18).

The situation with the legislation and ethics in the field is not easier. Under the current tort law, if the doctor follows the advice of artificial intelligence but this advice does not meet the medical standards, he bears the responsibility; however, the doctor who bypasses the correct recommendation provided by AI will also experience pressure as AI usage becomes standard regardless of the fact this scenario prevents the implementation of AI (19). The World Health Organization developed ethical guidelines for the application of artificial intelligence in medicine, stating that certain regulations should be respected while implementing these technologies (20). Clinical decision support systems in general possess numerous risks, in which the doctors experience alert fatigue, lose their professional skills, and find it hard to integrate into the working process (21).

To be able to understand why some medical experts use the technology safely and others do not, it is advisable to have a theoretical background. The Technology Acceptance Model suggests that the decision to use a specific technology is based on its perceived usefulness and ease of usage (22), whereas the unified theory of acceptance and use of technology extends it by introducing the concepts of social influence and surroundings in which the AI is introduced (23). These theories have been widely described in many researches in medicine, although many authors argue that they need certain correction before applying them to the medical sphere (24). The Non-adoption, Abandonment, Scale-up, Spread and Sustainability (NASSS) framework is even more complicated since it provides deeper explanation of the implementation of health technologies from the viewpoint of such aspects as condition, technology, and other components (25). Overall, upon evaluating all the theories, it becomes clear that the barriers to the safe usage of AI might be multiple and complex.

Studies of clinicians' opinions show there is a large measure of complexity, especially as regards their proposals for implementation of AI in healthcare. For instance a systematic review utilized a cross-sectional survey to explore physicians' and medical students' attitudes toward clinical AI, revealing that in general respondents believed positively toward AI but were concerned specifically about issues such as accuracy, accountability, confidentiality and the lack of financial information (26). A survey done with specialists in fields such as ophthalmology, dermatology, radiology and radiation oncology showed that while specialists were optimistic that AI will help their field in the future, most specialists did not feel prepared, and thus they proposed that there should be AI training in medical education (27). Regarding qualitative studies, research done on healthcare leaders in Sweden revealed there are concerns aside of technology; and it emphasized on the new competencies required, the changing professional roles, the uncertainty regarding the existing law governing the matter and absence

of guidelines for the implementation process (28). The studies indicate that education is the critical factor in addressing these issues. Some scholars are of the opinion that curricula should introduce the use of AI in medical practice, thereby enabling practitioners to evaluate and apply AI in healthcare(29). Other scholars point out that it is essential to restructure medical education so that practitioners learn to handle the knowledge generated through AI (30). At the same time, some authors have reviewed the importance of teaching practitioners how to address moral dilemmas arising from the introduction of AI in health setting(31). However, all the proposals in literatures mainly dwell on undergraduate education programs in first world countries without touching on the aspect of implementing structured AI safety education for qualified emergency physicians.

Iraq is a very significant and under-researched context. Due to the years of sanctions, war and dislocations, different problems emerged in the country's health care system, such as shortage of health care professionals, destruction of infrastructure and fragmentation of services (32). The long-term effects of war continue to impact the health workforce and the ability of institutions to sustain modern health systems (33). In Baghdad, private hospitals have developed rapidly and often adopt new technologies faster than public ones do, yet they are functioning with limited countrywide regulation of digital health, heterogeneous eHealth structures, and emergency departments operated by a mixture of emergency medicine physicians, board residents, and general practitioners. In such a setting, AI devices, especially free of charge LLM-based chatbots or vendors' software for interpretation, can be used ad hoc without validation by the local population, lack of institutional policy, and untrained users. Until now, we are unaware of any research on the obstacles to safe use of AI by emergency medical practitioners in Iraq and the absence of any evidence-based educational program for this group.

Solving this lack requires providing a method that connects diagnosis of the issue with a tested solution. Kern's six-step procedure of curriculum development provides that link by starting from a problem analysis and targeted needs assessment of learners to setting goals, choosing educational strategies, implementation, and evaluation (34). Kirkpatrick's model is a widespread framework for evaluating education at the stages of reaction/learning/behavior/result (35). The use of such frameworks helps convert identified obstacles into learning objectives and strategies, creating a possibility for measuring training impact instead of assuming it.

Therefore, the current research had three main objectives: first, to spot and measure barriers to the proper application of AI in making emergency clinical decisions by doctors of a private hospital in Baghdad; second, to understand the subjective experience of doctors in relation to the barriers; and third, to devise, implement, and assess a targeted training initiative aimed at overcoming the barriers amenable to educational intervention, especially the automation bias and the issue of detecting incorrect AI results.

Methodology

Study design. A three-phase sequential exploratory-explanatory mixed-methods design was deployed in the current study. Phase 1 (January–March 2026) dealt with identifying barriers through a cross-sectional survey together with semi-structured interviews, qualitative data being employed in clarifying and enhancing the quantitative results. Phase 2 (April 2026) was aimed at developing the training initiative based on the integrated findings. Phase 3 (May–July

2026) involved the provision of the training and its assessment using a single-group prepost design with a follow-up. This design was chosen since this study was aimed at discovering barriers and applying them to conduct a needs assessment for the intervention, just like it can be performed following the steps described by Kern (34).

Setting. The study was conducted in the emergency department of a private hospital in Baghdad, Iraq. The department is operational 24 hours a day and serves approximately 38,000 visits annually, including cases of trauma, cardiovascular events, respiratory problems, and surgical emergencies. The department report both emergency specialists as well as the patients in training of the Iraqi and Arab Boards of Emergency Medicine. The Ministry of Health employed general practitioners in the hospital. As of the study time, it had a partially implemented electronic patient record system, and a picture archiving and communication system equipped with an AI module for triaging images. No policies on AI application had been developed in the institution by the time of the study was conducted. The name of the hospital is confidential to maintain the anonymity of participants.

Participants and sampling. The researchers used a census approach for the survey, including physicians in the study who had direct patient care experience of months or more ($n = 84$). Physicians who were on prolonged leave and study investigators were left out of the survey study. At the same time, they used maximum-variation purposive sampling for the interviews in order to sample physicians of different ranks, experience, sex and levels of AI use (as per the survey results). The interview process continued until thematic saturation, which was achieved after conducting 16 interviews (when interviews produced no new codes). For the purpose of conducting the training evaluation, all people who filled out the survey were invited to participate in the training evaluation and the sample size was calculated to allow for a calculation of knowledge scores for paired comparisons. participants were required to achieve a medium effect (Cohens $d = 0.5$) at two-tailed $\alpha 0.05$ and power of 80% 34 ; thus, 42 participants were selected to allow for attrition up to approximately 20%.

Survey instrument. A self-administered questionnaire was developed from the literature and structured around the Technology Acceptance Model, the Unified Theory of Acceptance and Use of Technology and the NASSS framework (22–25). It had three parts: demographic and professional characteristics; current use of AI tools and prior training; and 35 barrier items grouped into seven a priori domains (knowledge and skills; trust, explainability and automation bias; data quality and local validity; infrastructure and technical factors; legal, ethical and liability issues; organisational governance; and workflow and time pressure). Barrier items were rated on a five-point Likert scale from 1 (strongly disagree) to 5 (strongly agree), with higher scores indicating a greater perceived barrier. A single item measured self-rated digital literacy on a five-point scale. The questionnaire was prepared in English, the language of medical training in Iraq, with Arabic clarification of technical terms. Content validity was assessed by a panel of six experts (three emergency physicians, one medical informatician, one medical educator and one health lawyer); items with an item-level content validity index below 0.78 were revised or removed. The instrument was piloted with 10 physicians from another Baghdad hospital, whose responses were excluded from analysis. Internal consistency was assessed with Cronbach's alpha, interpreted following established guidance (36). Questionnaires were distributed in paper and electronic form, with two reminders at two-week intervals.

Interviews. The semi-structured interviews were conducted in person in a separate facility outside the clinical duties in Arabic and/or English depending on which option was preferred

by the participant for the duration of 30–55 minutes. The guideline examined the use of AI in emergency care, when AI results have been questioned, understood obligations and perils, given some aid from the organization and request for training. Permission was given to record the interviews and phrases transcribed without making changes, with the inclusion of translations from Arabic to English that have been checked by a researcher known in both languages.

Implementation of a training program (Phase 2). Following Kern's six steps (34) for the identification of problems and understanding the needs, the literature was investigated while for the identification of particular needs, Phase 1 findings were used. For creation of the learning objectives, each of the barrier domains having the average score of 3.5 or higher and the qualitative themes related were used. The barriers that were not educationally solvable such as the problem of internet access were communicated to the administrators of the hospital. Drafting of objectives, their content and education methodologies was finalized by the expert panel and ambulatory care physicians not involved in interviews. Learning strategies corresponded to the learning goals: public lectures were planned for knowledge attainment, case studies were drafted to analyze the information critically, and employment of LLM technologies was stipulated during workshops to understand the practice of running a simulatory scenario in stressful situations.

Evaluation of the program (Phase 3). The evaluation addressed Kirkpatrick levels 1 (reaction) and 2 (learning), being replaced by an adequate substitute in the level 3 (behavior) analysis (35). Evaluation was done at each of the three stages: prior to the program (T0), upon its finishing (T1) and after eight weeks (T2). The program outcomes included knowledge tests with the number of questions equal to 20 10 0 100 1 1 covering such topics as AI principles, limitations, biases, verification, privacy and medico-legal issues. There were secondary indicators defined that reflected the level of self-perception in regard to the applicability of AI and special simulation scenarios with four cases included.

Analysis of data. SPSS version 26 was used for the analysis of quantitative variables in this research. Categorical variables were characterized by frequency and percentage values, while means and standard deviations were used to describe continuous variables. The mean score for the domain was computed based on all items of that domain. Multiple linear regression was performed on factors associated with the overall barrier score by including the variables that had a p-value of less than 0.20 0.05 37 in the bivariate analysis; model assumptions had to be verified with respect to linearity check, normality of the residuals and for checking multicollinearity. Multiple repeated-measures ANOVA, followed by Bonferroni-adjusted pairwise comparisons was applied to analyze pre-and post-tests. There are many tools for the analysis of qualitative data, one of which is thematic analysis where two independent researchers analyzed the first five transcripts and developed one codebook utilized in analyzing the remaining transcripts in NVivo 12. Themes were corrected based on the full data set. < <

Principles of ethics. The researchers obtained approval from the hospital's ethics board. Participation was voluntary and informed consent was given in writing, with participants able to withdraw at any stage without a penalty. Survey responses were anonymous and interview transcripts were de-identified using self-generated codes to link pre- and post-test data. The researchers did not gather any patient information; the simulation cases were all imaginary. Furthermore, the training was conducted outside the normal work schedule and did not interfere with job performance evaluation.

Results

In total, out of 84 doctors, 76 participated in the survey (response rate: 90.5%). The majority of participants were males (67.1%) with an average age of 34.6 ± 7.9 years. Even though 63.2% of the respondents claimed to use some AI tools, mainly chatbots, only 11.8% of the participants reported receiving any instruction in AI. (Refer to Table 1.)

Table 1. Characteristics of survey respondents (n = 76)

Characteristic	n (%) or mean $\pm$ SD
Age, years	34.6 $\pm$ 7.9
Male sex	51 (67.1)
Emergency medicine specialist	18 (23.7)
Board trainee (Iraqi/Arab Board)	31 (40.8)
General practitioner in ED	27 (35.5)
ED experience < 5 years	34 (44.7)
ED experience 5–10 years	24 (31.6)
ED experience > 10 years	18 (23.7)
Any clinical use of AI tools	48 (63.2)
General-purpose LLM chatbots	41 (53.9)
Automated ECG interpretation relied on	22 (28.9)
Imaging AI module	11 (14.5)
AI use weekly or more	29 (38.2)
Prior formal AI training	9 (11.8)

The barrier instrument demonstrated adequate internal consistency (Cronbach's α 0.74–0.88 per domain; 0.91 overall). All seven domains were scored above the scale's average of The highest scores were given to the gap in knowledge and skills domain, whereas the lowest in data quality and local validity domain (See Table 2).

Table 2. Barrier domain scores (1–5; higher = greater barrier)

Domain	Mean $\pm$ SD	Agree/strongly agree, %	Cronbach's α
Knowledge and skills gap	4.12 $\pm$ 0.61	81.6	0.86
Legal, ethical and liability	3.98 $\pm$ 0.74	77.6	0.83
Data quality and local validity	3.94 $\pm$ 0.68	76.3	0.79
Trust, explainability and automation bias	3.87 $\pm$ 0.72	72.4	0.88
Organisational governance	3.79 $\pm$ 0.77	71.1	0.81
Infrastructure and technical	3.71 $\pm$ 0.80	68.4	0.77
Workflow and time pressure	3.56 $\pm$ 0.85	61.8	0.74

At the item level, the barriers faced were, more often than not, the lack of an institutional policy regarding AI (85.5%), lack of formal training (84.2%), absence of validation in Iraqi or Arabic populations (78.9%), lack of clarity on medico-legal responsibilities (77.6%), and inability to ascertain when AI output is erroneous (73.7%).

As per the findings of multivariable analysis, it was found that prior exposure to AI training and better computer literacy corresponded with lower scores on barriers. On the other hand, having ED experience for over 10 years corresponded with higher barrier scores.

Table 3. Multiple linear regression of overall barrier score (n = 76; adjusted R² = 0.31)

Variable	β (95% CI)	p
Prior formal AI training (yes vs no)	-0.41 (-0.66 to -0.16)	0.002
Self-rated digital literacy (per point)	-0.19 (-0.31 to -0.07)	0.003
ED experience > 10 vs < 5 years	0.22 (0.01 to 0.43)	0.041
General practitioner vs specialist	0.18 (-0.04 to 0.40)	0.110
Female sex	0.06 (-0.14 to 0.26)	0.548

Thematic analysis of 16 responses resulted in five themes (Table 4). Mostly, these themes corresponded with the survey areas but also contributed to the understanding of how barriers function in transitions.

Table 4. Qualitative themes (n = 16 interviews)

Theme	Description	Linked survey domain
Uncalibrated trust	Trust swung between uncritical acceptance under pressure and blanket rejection after a single error	Trust and automation bias
Accountability vacuum	No one could say who is responsible when AI-informed decisions cause harm	Legal and liability; governance
Imported tools, local patients	Doubt that tools trained abroad fit local disease patterns, language and records	Data quality and local validity
Infrastructure fragility	Power and internet interruptions and fragmented records made tools unreliable	Infrastructure and technical
Learning by trial and error	Skills came from colleagues and social media, not structured teaching	Knowledge and skills gap

The end result of the program consisted of modules that lasted 12 hours of the mixed method of teaching them. These include AI fundamentals and limitations, critical assessment of the output of AI and automation bias, safe usage of large language models (verification, prompting, privacy), validated local emergency department tools, ethics, responsibility and documentation, and a simulation of using faulty AI advice.

Of 42 enrolled physicians, 39 (92.9%) completed all three assessments. Knowledge, self-efficacy and simulated error detection improved significantly, and overreliance on incorrect AI advice fell by about two thirds; most gains persisted at eight weeks (Table 5). Mean satisfaction was 4.5 ± 0.5 of 5.

Table 5. Training outcomes (n = 39)

Outcome	T0 (baseline)	T1 (post)	T2 (8 weeks)	p
Knowledge score (0–20), mean $\pm$ SD	9.8 $\pm$ 2.9	15.6 $\pm$ 2.4	14.7 $\pm$ 2.6	< 0.001*
Self-efficacy (0–100), mean $\pm$ SD	42.3 $\pm$ 15.1	71.8 $\pm$ 12.6	68.4 $\pm$ 13.2	< 0.001*
Overall barrier score (1–5), mean $\pm$ SD	3.91 $\pm$ 0.52	3.18 $\pm$ 0.58	3.27 $\pm$ 0.61	< 0.001*
Detected AI error in simulation, n (%)	16 (41.0)	31 (79.5)	—	< 0.001†
Acted on incorrect AI advice, n (%)	20 (51.3)	7 (17.9)	—	0.001†

*Repeated-measures ANOVA; T0 vs T1 and T0 vs T2 significant after Bonferroni correction, T1 vs T2 not significant. †McNemar's test, T0 vs T1.

Discussion

According to the research, AI has already been included in the regular practice of emergencies in a private Hospital in Baghdad but entirely out of any effective procedure. About two-thirds of doctors working in an emergency setting claim to be using AI, particularly free online chatbots. Very few of them, however, have received any formal training. The issues with using AI in that context seem to have been mainly connected with human beings as well as regulations and uncertainty: lack of knowledge, skills, uncertainty associated with the accountability and faith in AI products developed for patients in other countries. Despite that, a short training program showed a great success in increasing knowledge and improving the ability to recognize wrong AI suggestions in practice, as well as lowering the over-reliance on AI.

The gap between use and preparation is the central finding and is consistent with international evidence. Surveys of physicians and trainees in several countries have repeatedly found favourable attitudes towards AI combined with limited understanding of how it works and a strong demand for training (26,27). Studies of medical students in Germany and the United Kingdom showed that learners were largely unconcerned about being replaced by AI but felt that their education had not prepared them to work with it (38,39). What distinguishes our findings is that informal adoption has preceded both training and policy in a clinical, time-critical environment. When general-purpose LLMs are used to support emergency decisions, the risks described in the literature, including fluent but fabricated content and a lack of source transparency, are transferred directly to patient care (8,9). The high proportion of physicians who reported being unable to judge when an AI output is wrong suggests that the safety net that should sit between an unreliable tool and a patient, namely a clinician able to evaluate the output critically, was weakest exactly where it was most needed.

Trust, explanation and automation bias are important issues that need to be highlighted in both the survey and the interviews. The theme regarding uncalibrated trust indicated an occurrence in which healthcare providers would either blindly accept AI suggestions under stressful circumstances or reject them altogether after making just one error. Both possibilities are dangerous. The first relates closely to the automation bias and complacency discussed in the literature on human factors, which become exacerbated at times of stress (15,16), conditions common in emergency work. The second possibility, sometimes referred to as algorithm aversion, disregards any potential merits of a technology that has been validated. The basic simulation results corroborated the accounts given by interviewees: almost of individuals acted upon deliberately wrong advice before training took place. This finding is consistent with studies showing that even when doctors are given inferior advice, labelled as AI-provided, and evaluated highly, they adopt it (17). Explainability is often proposed as a way out (18), while in reality, participants' difficulties stem from their lack of a practical know-how to check AI results against the clinical evidence, guidelines and their own assessment. Hence, training concentrated on calibrating trust, which involved making doctors treat AI recommendations only as one of the pieces of evidence to check and recognising situations where automation bias can arise, alongside documenting the reasons for their refusal to comply with recommendations. The significant drop in overreliance on the AI suggestions after training means that such an undertaking can be carried out within a short period of time, which is good news for developing countries.

Legal and governance barriers are manifestations of broader international uncertainty, which, in turn, is exacerbated in Iraq. Participants reported an accountability gap: no one knew how to

deal with the consequences of an AI-assisted decision leading to a patient harm, so some opted for a strategy of keeping their usage to themselves. Such situation mirrors the findings of Price et al., who argue that there are no problems in case one invokes AI only to support conventional treatment, in which case there will be no gains from using it but liability will not arise if following the wrong advice (19). It also reflects a much wider debate regarding whether AI can or should be ever treated as an alternative to a doctor, which raises the question of whether responsibility can ever be transferred to the system that does not bear moral and ethical obligations to the patient (40). Lack of an organizational policy was the most often reported obstacle. However, that is the issue that cannot be handled solely by educating doctors. WHO guidelines on ethics and governance of AI in health care require clear mechanisms of accountabilities, transparency, and patient data protection (20 25,28).

Concerns regarding the quality of data and the local validity have received a strong endorsement, which is based on sound reasoning. As illustrated by the evidence from high-income settings, the performance of models during deployment is usually worse than during their development (12). Furthermore, the issue of dataset shift hampers the transfer of algorithm to other populations (11) and the presence of systematic bias against the population that is usually not representative in the training data (13) cannot be overlooked either. Patients from Iraq differ from the populations used for training the vast majority of the commercial and research models with regard to prevalence of diseases, demographic structure, the trauma burden, language, and record-keeping techniques. What is more, the long history of violence and lack of investments in this area has created gaps in the relevant IT infrastructures and record-keeping technologies (32,33) which complicates the operation of AI systems. Thus, scepticism of physicians in relation to the above-mentioned tool is justified and should not be considered as some sort of resistance to innovations. The program has offered the physicians training based on the idea that they need to learn how to verify claims concerning the validity of tools that they use. Priorities should include conducting local validity trips aimed at the modules that were already implemented in the department.

The findings of the regression analysis deepen our understanding of the phenomenon. Previous training and increased levels of digital literacy were each linked to a reduction in perceived obstacles, thus supporting the assertion that many of them can be overcome via education (29–31). It can be stated that there are two possible explanations for the link between the length of time in practice and the level of barrier ratings. More experienced physicians could be less familiar with modern technologies because of the nature and timing of their training programs, which is in line with the theory of acceptability based on the idea of perceived ease of use and facilitating conditions (22,23). In addition, those who have been in practice for a long time might have become more realistic in their assessment of risks owing to their experience and allow them to identify any inconsistencies between a procedure or treatment and the nature of the patient's condition basing on their clinical knowledge and experience. The latter assumption is supported by the notion of task expertise being responsible for the ability to withstand inaccurate advice (17), which shows that good clinicians can play an important role in developing safe practices rather than just being persuaded to accept digital tools.

The creation of the program shows how important it is to connect needs assessment with curriculum design. Kern's method provided a model that was created in response to one of the issues (34), while using simulation made it possible to measure the effects of learning not only through self-reports, but also by finding the indicator of learning effectiveness at the Kirkpatrick level 3 (35). Those who observe the issue argue that teaching AI should change

from having students know the facts about algorithms to being capable of assessing and working with machine-generated data (29,30). The results of our research confirm this assumption. The improvements in applied knowledge of errors are noticeably bigger than those in theoretical knowledge. The decline in scores from the immediate reaction to the eight-week assessment is small. Since the ED labor force in Iraq has many general practitioners and residents who often switch places, the 12-hour program is practical for scaling and can be modified for other private and public hospitals in Baghdad and other places in the country. Regular reminders, for instance through occasional reviews of AI-based decisions made in practice, can support these achievements.

The research had certain restrictions. It was carried out in only one private hospital and since private and public hospitals in Bagdad vary in many aspects, including resources, human resources and patient populations, the ability to generalize the findings is limited. The questionnaire used in the survey was based on self-assessment, which may impact the results due to social desirability. In addition, the training design included only one group without a control group and therefore there can be a chance the improvement was not caused by the training itself but rather can be regarded as the training effects, maturity and increased awareness due to participation in Phase 1; to avoid the risks, both parallel tests and blinded simulation raters were used. Simulated testing cannot give the exact picture of the real life behavior and it was not tested for real clinical outcomes. The follow-ups were conducted only weeks after the training. Lastly, the knowledge and self-efficacy test were created especially for this research, but although the content validity was checked, the need in the further psychometric testing is obvious.

Despite these limitations, the study provides the first evidence from Iraq on how emergency physicians experience AI and demonstrates a practical, low-cost response. The main message for hospital leaders is that the safe use of AI in emergency care depends less on acquiring new tools than on preparing the clinicians who use them and establishing the governance within which they work.

Conclusion

Doctors in Iraq are using AI systems to assist them with clinical decisions and although it is a private institution, many of them do not have any formal training. From a legal perspective, there are many ethical questions and through a systematic approach we will deal with all the major problems affecting the institution. The aim of this research is to help practitioners incorporate the use of AI tools in their work by focusing on six main barriers: lack of training, legal issues, local validity problems, and trust issues. The results of a quick 12-hour training program show an increase in the efficacy of the physicians and their appropriate usage of AI applications. Given these results, it is crucial to conduct multicenter, controlled studies which will help evaluate additional ethical parameters affecting the current study.

References

- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.
- Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-58.
- Stewart J, Sprivulis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine. *Emerg Med Australas*. 2018;30(6):870-4.
- Kirubarajan A, Taher A, Khan S, Masood S. Artificial intelligence in emergency medicine: a scoping review. *J Am Coll Emerg Physicians Open*. 2020;1(6):1691-702.
- Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the Emergency Severity Index. *Ann Emerg Med*. 2018;71(5):565-74.e2.
- Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA Jr, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care*. 2019;23(1):64.
- Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-80.
- Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-40.
- Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233-9.
- Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med*. 2003;78(8):775-80.
- Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195.
- Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. 2021;181(8):1065-70.
- Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-53.
- Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28(3):231-7.
- Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc*. 2012;19(1):121-7.

- Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. *Hum Factors*. 2010;52(3):381-410.
- Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lerner E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med*. 2021;4(1):31.
- Amann J, Blasimme A, Vayena E, Frey D, Madai VI; Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1):310.
- Price WN 2nd, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA*. 2019;322(18):1765-6.
- World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021.
- Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med*. 2020;3:17.
- Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q*. 1989;13(3):319-40.
- Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. *MIS Q*. 2003;27(3):425-78.
- Holden RJ, Karsh BT. The technology acceptance model: its past and its future in health care. *J Biomed Inform*. 2010;43(1):159-72.
- Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *J Med Internet Res*. 2017;19(11):e367.
- Chen M, Zhang B, Cai Z, Seery S, Gonzalez MJ, Ali NM, et al. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. *Front Med (Lausanne)*. 2022;9:990604.
- Scheetz J, Rothschild P, McGuinness M, Hadoux X, Soyer HP, Janda M, et al. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Sci Rep*. 2021;11(1):5193.
- Petersson L, Larsson I, Nygren JM, Nilsen P, Neher M, Reed JE, et al. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. *BMC Health Serv Res*. 2022;22(1):850.
- Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. *JMIR Med Educ*. 2019;5(2):e16048.

- Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. *Acad Med*. 2018;93(8):1107-9.
- Masters K. Artificial intelligence in medical education. *Med Teach*. 2019;41(9):976-80.
- Al Hilfi TK, Lafta R, Burnham G. Health services in Iraq. *Lancet*. 2013;381(9870):939-48.
- Lafta RK, Al-Nuaimi MA. War or health: a four-decade armed conflict in Iraq. *Med Confl Surviv*. 2019;35(3):209-26.
- Thomas PA, Kern DE, Hughes MT, Chen BY, editors. Curriculum development for medical education: a six-step approach. 3rd ed. Baltimore: Johns Hopkins University Press; 2016.
- Kirkpatrick DL, Kirkpatrick JD. Evaluating training programs: the four levels. 3rd ed. San Francisco: Berrett-Koehler; 2006.
- Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ*. 2011;2:53-5.
- Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3(2):77-101.
- Pinto dos Santos D, Giese D, Brodehl S, Chon SH, Staab W, Kleinert R, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. *Eur Radiol*. 2019;29(4):1640-6.
- Sit C, Srinivasan R, Amlani A, Muthuswamy K, Azam A, Monzon L, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. *Insights Imaging*. 2020;11(1):14.
- Karches KE. Against the iDoctor: why artificial intelligence should not replace physician judgment. *Theor Med Bioeth*. 2018;39(2):91-110.