Automation in Contemporary Clinical Information Systems: a Survey of AI in Healthcare Settings

Farah Magrabi, David Lyell, Enrico Coiera
Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia

Summary

Aim and objectives: To examine the nature and use of automation in contemporary clinical information systems by reviewing studies reporting the implementation and evaluation of artificial intelligence (AI) technologies in healthcare settings.

Method: PubMed/MEDLINE, Web of Science, EMBASE, the tables of contents of major informatics journals, and the bibliographies of articles were searched for studies reporting evaluation of AI in clinical settings from January 2021 to December 2022. We documented the clinical application areas and tasks supported, and the level of system autonomy. Reported effects on user experience, decision-making, care delivery and outcomes were summarised.

Results: AI technologies are being applied in a wide variety of clinical areas. Most contemporary systems utilise deep learning, use routinely collected data, support diagnosis and triage, are assistive (requiring users to confirm or approve AI provided information or decisions), and are used by doctors in acute care settings in high-income nations. AI systems are integrated and used within existing clinical information systems including electronic medical records. There is limited support for One Health goals. Evaluation is largely based on quantitative methods measuring effects on decision-making.

Conclusion: AI systems are being implemented and evaluated in many clinical areas. There remain many opportunities to understand patterns of routine use and evaluate effects on decision-making, care delivery and patient outcomes using mixed-methods. Support for One Health including integrating data about environmental factors and social determinants needs further exploration.

Keywords
Clinical information systems; artificial intelligence; automation; implementation; evaluation.

Yeast Med Inform 2023:115-26
http://dx.doi.org/10.1055/s-0043-1768733

1 Introduction

Artificial intelligence (AI) technologies are used for a range of tasks requiring pattern recognition, reasoning or learning [1]. While AI has been studied for more than 50 years, its current resurgence is largely driven by developments in machine learning (ML) and specifically deep learning (DL). Recently, these DL methods have achieved unprecedented levels of performance in a variety of tasks such as language and image generation, using generative AI methods, including generative pretrained transformers (GPTs). In healthcare, AI promises to transform care delivery as it has the potential to harness the vast amounts of genomic, biomarker, and phenotype data that are being generated across the health system and beyond [2, 3].

Indeed, AI technologies should play a central role in reaching the goals of One Health which seeks to balance and optimise the health of people, animals and the environment through surveillance, prevention, and mitigation at local, regional, national, and global levels [4, 5]. Adapting a One Health approach at the local level, for instance, can improve understanding of the dynamics of humans, animals, and the built-environment to inform infection control and prevention programs [6]. Another example is the integration of human as well as environmental surveillance and response systems to assist health systems in responding to and mitigating the effects of climate change [7].

Today, AI is incorporated into a variety of clinical systems for detecting findings, suggesting diagnoses and recommending treatments in data-intensive specialties like radiology, pathology and ophthalmology [3]. These AIs can aid human decision-making – from systems that acquire and analyse data, provide options for decisions, to systems with the capability of making decisions entirely on their own [1]. With time, systems are expected to become increasingly autonomous, going beyond making recommendations to autonomously performing tasks such as controlling closed loop clinical machines like ventilators or insulin pumps, triaging patients or screening referrals [8, 9]. With the public release of generative AI, their applications in assisting clinicians to create health records and generate summaries of clinical evidence are rapidly emerging [10].

To be successful in healthcare, AI must perform well in real-world clinical settings. Yet there are many complex challenges in the “last mile” of implementation that may make technically high performing algorithms perform poorly in real-world settings [11]. A fundamental challenge here is that algorithms built using ML may not necessarily generalise well beyond the data upon which they are trained, making them potentially unsafe when used on different populations. Even for restricted tasks like image interpretation, AI can make erroneous diagnoses because of differences in the training and real-world populations, as well as differences in image capture workflows [12]. Such algorithmic risks may be exacerbated with generative AI, which can produce content that is incorrect, unsafe and not grounded in scientific evidence [10]. Another last mile challenge relates to how well an AI system is embedded within
the local sociotechnical context of implementation, where an organization can be viewed as a network of people, processes, and technologies [11]. AI systems need to be seamlessly integrated into clinical workflows and existing clinical information systems (CISs) such as electronic health records (EHRs) and laboratory information systems (LISs). The performance and safety of these algorithms is highly reliant on the quality of data provided by CISs [13].

The aim of this survey is to examine automation in contemporary CISs to identify the range and impact of AI use. We review studies reporting AI implementation and evaluation in clinical settings to examine progress in digitalization and realizing the potential of AI to support clinicians in delivering patient care. Given the volume and rapidly evolving nature of the literature, we do not attempt to be comprehensive. Rather, we highlight clinical application areas and autonomy in current AI to discuss opportunities and future directions for implementing and evaluating AI in real-world settings. In keeping with the focus of this Yearbook, we sought to identify exemplars of AI systems that integrated clinical data with environmental or social factors to improve care delivery.

2 Methods

2.1 Search Strategy and Study Selection

We reviewed studies about ML systems in clinical settings published between January 2021 and December 2022. PubMed/MEDLINE, Scopus, Web of Science and EMBASE databases were searched by combining the search terms “artificial intelligence”, “machine learning”, “deep learning”, “natural language processing” with “implement*”, “evaluat*” and “clinical”. In addition, we hand-searched the table of contents of major health informatics journals including the Journal of the American Medical Informatics Association (JAMIA), JAMIA Open, Applied Clinical Informatics (ACI), Journal of Medical Internet Research (JMIR), the International Journal of Medical Informatics (IJMI), BMJ Health and Care Informatics, NPJ Digital Medicine, and BMC Medical Informatics and Decision Making. Studies about the development and validation of ML models on historical data sets were excluded. The search was limited to English language articles and grey literature was excluded.

2.2 Data Extraction, Summarising and Reporting Findings

For each included study, we extracted information about the authors, year of publication, clinical area, setting, CIS integration, study design and the effects of AI interventions. Study areas were classified by the country in which they were conducted, using the United Nations’ definition [14]. Exemplars of AI systems that integrated clinical data with social determinants or environmental factors including animal health were identified. Sociotechnical and ethical considerations for the use of AI in clinical settings were similarly identified. In addition, information about the clinical task and AI system tasks was extracted and used to examine the level of AI autonomy.

Clinical task. Clinical tasks supported by AI were categorised into [15]:

a) Diagnosis: assisting with the detection, identification or assessment of disease, or risk factors;

b) Triage: assisted with prioritising cases for clinician review, by flagging or notifying cases with suspected positive findings of time-sensitive conditions, such as stroke;

c) Procedure: assisted users performing diagnostic or interventional procedures;

d) Treatment: provided recommendations for therapy;

e) Monitoring: assisting clinicians to monitor patient trajectory over time.

Level of autonomy. The level of autonomy was examined using a previously published three-level classification based on how clinical tasks are divided between the clinician and AI [15]:

1. Autonomous information: In these systems, there is a separation between AI and clinician contributions to a task, with the AI contributing information that clinicians can then use to make a decision, e.g., an imaging system that provides a coloured imaging display to help a clinician differentiate human tissues.

2. Assistive: These AIs overlap in capability with clinicians, but clinicians provide the final decision. For example, clinicians confirm or approve AI provided information or decisions, e.g., a system assists clinicians to detect osteoarthritis from a knee X-ray image with a disclaimer that the system should not be used without a full patient evaluation.

3. Autonomous decision: Here, the AI makes the decision for a clinical task, which can then be enacted by clinicians or the AI, e.g., a system screens retinal images for diabetic retinopathy in primary care with the result used to make patient referral decisions.

To determine the level of autonomy, we examined the AI task, the stage of human information processing automated by the AI [9], and the system inputs and outputs. The clinical use case was examined to assess whether clinicians needed to verify decisions provided by the AI (assistive) or could rely on the AI information or decisions (autonomous). Classification of the stage of human information processing and level of autonomy was performed by FM and reviewed by DL.

Reported effects of AI interventions were categorised (into user experience, decision-making, care delivery and patient outcomes) based on an established framework called the information value chain [16], which separates the multiple steps from system use to impacting clinical outcomes - interacting with AI, receiving new information, decision-making, care delivery. A narrative synthesis then integrated findings into descriptive summaries.

3 Results

We identified 62 studies examining AI systems in clinical settings in 19 countries (Appendix). Most were conducted in hospitals (Figure 1; Table 1; 73%, n=45) in high- and upper middle-income nations, specifically the United States (US) (Figure
Studies largely used quantitative designs (74%, n=46) to evaluate the effects of AI in clinical settings and were predominantly focused on assessing effects on decision-making (66%, n=41); only four were randomised controlled trials (RCTs). While none of the studies explicitly addressed the goals of One Health, consideration of the United Nations Sustainable Development Goals as well as social determinants and environmental factors such as the socioeconomic status and traffic volume was evident in a few studies [17, 18]. Aside from research ethics, none of the studies reported measures to consider or address ethical issues (e.g., algorithmic fairness) around the use of AI in clinical settings.

Most studies were focused on AI for diagnostic (n=37) and triage tasks (n=10). Treatment, procedures and monitoring were less common. ML algorithms were commonly employed to assist in analysing clinical information (information analysis; n=15) and in selecting decisions (decision selection; n=39). Most systems were assistive (65%), requiring users to confirm or approve AI provided information or decisions covering 24 clinical areas. Integration with CISs including EHRs and LISs was described in some studies (n=22). In the following sections, we provide a summary of these studies by clinical area and summarise effects on decision-making, care delivery and patient outcomes.

### 3.1 Cancer

Nine studies focussed on AI for diagnosis and monitoring of different cancers. Wu et al. [19] conducted an RCT to demonstrate the effectiveness of real-time assistance for detection of early gastric cancer involving 1,050 patients at five hospitals in China. Compared with control, fewer lesions were missed in the AI group (mean 5 vs. 10). The AI correctly predicted early and advanced gastric cancers (accuracy: 85%, sensitivity: 100%, and specificity: 84%).

Another Chinese study examined effects on decision-making and care delivery. Peng et al. [20] conducted a prospective cohort study to demonstrate the efficacy of AI for detecting malignant thyroid nodules. Use of AI by 12 radiologists to interpret 366 ultrasound images and videos was shown to improve accuracy compared to unassisted interpretation (AUROC: 0.837 to 0.875), and projected to reduce the number of fine needle aspirations (62% to 35%) and decrease missed malignancy (19% to 17%).

Using mixed-methods, Calisto et al. [21] assessed the use and usability of a breast screening AI via a clinical simulation study; 45 radiologists completed three randomly selected cases from a set of 289. The study was informed by human-AI design guidelines and specifically examined the impact on clinical workflow as well as understanding, trust.

---

**Fig. 1** Geographic distribution of AI studies reviewed in this survey (n=62).
and acceptance of the AI. It found use of the system decreased false positives by 27% and false negatives by 4% while decreasing the time-to-diagnose by 3 min per patient; 91% of participants were satisfied with the AI.

The remaining six studies focussed on decision-making. Martins Jarnalo et al. [22] evaluated a commercial system for detecting pulmonary nodules in computed tomography (CT) images. When integrated into the clinical workflow of a Dutch radiology department, performance matched the vendor specification; (sensitivity: 88%, false-positive rate: 1.04 per scan and NPV: 95%). Quan et al. [23] assessed use of AI during colonoscopy. Their study, involving 300 patients found that AI assistance increased detection of adenomas and serrated polyps during colonoscopy in comparison to historical controls, but the findings were not statistically significant. Ou et al. [24] demonstrated that AI-assisted analysis of urine cytology outperformed the conventional method, with improved sensitivity (92% vs. 87%) and NPV (97% vs. 95%). Nasir-Moin et al. [25] showed that AI for interpretation of 100 colorectal polypl samples significantly improved pathologists’ classification accuracy compared with standard microscopic assessment (74% to 81%). Duan et al. [26] demonstrated improved image quality, reduced noise and processing time for CT images to assess colon cancer. Aimed at primary care physicians in Brazil, Giavina-Bianchi et al. [27] demonstrated algorithms for melanoma screening using both dermascope and smart phone images (accuracy: 89%, 85%; sensitivity: 91%, 89%; specificity: 89%, 83%).

### 3.2 Radiology

Nine studies examined systems for a variety of clinical areas in hospital and outpatient radiology departments. Taking a theory driven approach, Rabinovich et al. [28] used the Technology Acceptance Model to evaluate user satisfaction and actual use of an assistive system for chest x-ray interpretation in an Argentinian emergency department (ED) over 5-months. The system was used for 15% of studies (n=1,186), with an average of eight accesses per day. Physicians and radiology residents had similar perceptions about system usability, but differed on output quality and usefulness.

Also using mixed-methods, Chonde et al. [29] studied the use and utility of an autonomous radiology examination instruction system called RadTranslate. The AI was integrated into imaging workflows for chest radiography at a COVID-19 triage outpatient centre that served a predominantly Spanish-speaking Latino community in the US. During the 63-day test period, technologists voluntarily switched to the system to provide instructions in Spanish. The system was found to reduce strain on medical interpreters and shortened examination length.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>high-income</td>
<td>45</td>
<td>73</td>
</tr>
<tr>
<td>upper middle-income</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>lower middle-income</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>low income</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Clinical task</td>
<td></td>
<td></td>
</tr>
<tr>
<td>diagnosis</td>
<td>37</td>
<td>60</td>
</tr>
<tr>
<td>triage</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>treatment</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>procedures</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>monitoring</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Stage of human information processing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>information acquisition</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>information analysis</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>decision selection</td>
<td>39</td>
<td>63</td>
</tr>
<tr>
<td>action implementation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level of autonomy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>assistive</td>
<td>40</td>
<td>65</td>
</tr>
<tr>
<td>autonomous decision</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>autonomous information</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>Setting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hospital</td>
<td>45</td>
<td>73</td>
</tr>
<tr>
<td>primary care</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>multiple settings</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>outpatient</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>simulation centre</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Study design</td>
<td></td>
<td></td>
</tr>
<tr>
<td>quantitative</td>
<td>46</td>
<td>74</td>
</tr>
<tr>
<td>mixed-methods</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>qualitative</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Evaluation measures*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>user experience</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>actual use</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>decision-making</td>
<td>41</td>
<td>66</td>
</tr>
<tr>
<td>care delivery</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>patient outcomes</td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

*Studies used one or more evaluation measures.
Two studies examined effects on decision-making and care delivery. Duron et al. [30] reported on the utility of AI to detect missed fractures in posttraumatic radiographs. 18 radiologists and emergency physicians were asked to detect and localise 600 fractures with and without AI highlighting of fractures. AI improved the sensitivity of physicians by 9% and the specificity by 4% and reduced the average number of false-positive fractures per patient by 42% and mean reading time by 15%. Schmuelleing et al. [31] assessed the impact of a triage system that detected and alerted radiologists about ED cases with suspected pulmonary embolism on CT angiograms. While the study demonstrated good diagnostic accuracy (sensitivity 80%, specificity 95%, PPV 82%, and NPV 94%), there was no effect on report communication times and patient turnaround 9-months post-implementation.

Three studies examined assistive AI for chest x-ray and CT interpretation. Seah et al. [32] evaluated a comprehensive system that was trained to identify and highlight 127 radiological findings by asking 20 radiologists to review 2,568 cases with and without assistance in a controlled setting. The AI was shown to improve classification accuracy in 102 findings. Zhang et al. [33] investigated accuracy and efficiency in detecting rib fractures by asking radiologists to interpret CT images unassisted, assisted and with the AI as a second reader. AI as a second reader was found to improve detection accuracy (up to 6% more rib fractures) and reading efficiency (time reduced by 34-36%). Focussing on junior radiologists, Liu et al. [34] compared the diagnostic accuracy of an AI to identify rib fractures in chest CT images with and without assistance demonstrating improved sensitivity and reduced reading time by ~1 min per patient without decreasing the specificity.

Taking a qualitative approach, Lee et al. [35] describe their experiences in implementing an assistive AI for chest x-ray interpretation in a South Korean hospital. Both accuracy and immediate availability of results was reported to be critical, along with an explainable visualisation of results and the ability to configure software platforms for data presentation.

Focussing on trainees, Shiang et al. [36] surveyed 15 residents about their use of a commercial AI in a US residency curriculum. Here, residents were given access to an autonomous system that analysed CT images and notified clinicians about cases with suspected positive findings of pulmonary embolism, intracranial haemorrhage, and acute cervical spine fractures. Most residents (92%) supported incorporating AI into the curriculum and found it useful for triage (83%) and troubleshooting (67%), rather than for diagnostic purposes of speed (42%), accuracy (33%), or diagnosis determination (17%).

### 3.3 Triage

Six studies focussed on ED triage support. Hinson et al. [37] undertook a staged evaluation to assess an AI that provided a COVID-19 Clinical Deterioration Risk Level (1–10) in real-time based on EHR data. Prospective validation over 18-months at five EDs including an initial silent deployment demonstrated ML model performance for prediction of critical and inpatient care (AUC: 0.85–0.91; 0.80–0.90). Total mortality was reduced among high-risk patients.

Also focussing on COVID-19, Soltan et al. [38] evaluated a screening system in a United Kingdom ED. Automated identification using routinely collected data was reported to detect COVID-19 in 45 min, 61 min sooner than a lateral flow device, and 6 h 52 min (90%) sooner than PCR. Classification performance was high (sensitivity 87%; specificity 85%; and NPV 100%). The AI correctly excluded infection for 58% patients who were triaged by a physician to a COVID-19 suspected area but went on to test negative by PCR.

For chest pain, Wang et al. [39] demonstrated improved clinical decision-making and triage. Automated detection of ST-elevation myocardial infarction (STEMI) on electrocardiography (ECG) and clinical risk assessment (ASAP score) was reported to shorten the time to treatment (door-to-balloon time: 64 min to 53 min). Among patients with an ASAP score of 3 or higher, the median door-to-ECG time decreased (30 min to 6 min).

In a clinical simulation, Kim et al. [40] showed use of AI by ED clinicians for chest x-ray interpretation and decision-making improved their sensitivity to abnormalities regardless of experience (AUROC=0.801). Also focussed on decision-making, Ivanov et al. [41] demonstrated AI improved the accuracy of nursing triage in a US urban community hospital (paediatric: 54% to 67%; adult: 62% to 78%). Jordan et al.'s [42] qualitative examination of the impact of cultural embeddedness of this system found that although there was initial scepticism, the AI grew to be perceived as a safety net for triage decision-making among nurses.

### 3.4 Radiotherapy

Five studies examined autonomous and assistive AI for segmentation of organs at risk (OAR), treatment plans for breast cancer and risk assessment during therapy. Using mixed methods, Wong et al. [43] evaluated an AI-based auto-segmentation for central nervous system, head and neck, and prostate radiotherapy (RT) planning at two Canadian cancer centres. AI generated plans for 551 cases, required minimal edits and resulted in a positive user experience. Also using mixed-methods, Byun et al. [44] assessed an auto contouring system to delineate OAR for breast radiotherapy with an expert group. Performance of the AI was comparable with manual contouring by experts and was significantly faster (mean times for nine OAR: 37 min for manual vs. 6 min for corrected auto contours). The survey found good user satisfaction.

For prostate radiotherapy, Cha et al. [45] demonstrated clinical utility of AI for MR-based planning with 65% of cases requiring no more than minor edits, and a time saving of 12 min (30% of total contouring time) for physicians. Kneepkens et al. [46] found that although automatically generated plans resulted in slightly higher doses, they were clinically acceptable (AI: 90-95% vs. manual: 90%) and time-efficient.

Focussing on implementation, Hong et al. [47] examined the challenges with using EHR data to conduct an RCT of an AI to identify patients at high risk for ED visits and hospitalisation during cancer radiation
therapy. They found data extraction and the need for manual review required significant time for implementing RCTs. Limited data availability through the standard clinical workflow and commercial products were seen to be a barrier.

3.5 Mental Health

Four studies examined AI for diagnosis and treatment of mental health conditions. Three mixed methods studies related to an AI that operationalised Canadian guidelines for depression treatment and provided clinicians with patient-specific remission probabilities for different treatment options [48-50]. Of these, two involved a high-fidelity clinical simulation with 20 staff or residents in psychiatry or family medicine completing three 10-min clinical interactions with standardised patients portraying mild, moderate, and severe episodes of major depressive disorder. In the first, Benrimoh et al. [48] focussed on ease of use and impact on physician-patient interaction. Clinicians indicated a willingness to use the tool in real clinical practice, placed a significant degree of trust in the system’s predictions to assist with treatment selection, and its potential to increase patient understanding and trust. The second study focussed on utility; Tanguay-Sela et al. [50] reported 60% of physicians perceived the tool to be useful for treatment-selection, with family physicians perceiving the greatest utility; 50% indicated they would use the tool for all patients with depression, with an additional 35% noting that they would reserve the tool for more severe or treatment-resistant patients. The tool was also perceived to be useful in discussing treatment options with patients. Popescu et al. [49] assessed feasibility of using the tool during consultation by examining change in appointment length. Use of the tool over 11 months did not increase appointment length; most patients and physicians reported that the tool was easy to use and trustworthy but there were mixed perceptions about its impact on the patient-clinician relationship.

Focussing on suicide risk, Wilimits et al. [51] evaluated automated detection in clinical settings by combining predictions from the Columbia Suicide Severity Rating Scale with a real-time ML model. Combined models outperformed the model alone for risks of suicide attempt and suicidal ideation in a cohort study of 120,398 adult patient encounters in the US.

3.6 Cardiovascular Disease

Three quantitative studies examined assistive and autonomous AI for diagnosis and procedures. Yao et al. [52] undertook a pragmatic RCT of an AI to detect and notify clinicians about patients with suspected findings of low left ventricular ejection fraction. The study involving 22,641 patients across diverse practice settings demonstrated an increased identification of low ejection fraction within 90 days of the ECG (control: 1.6% vs. intervention: 2.1%). Edalti et al. [53] evaluated two algorithms to improve image quality and reduce noise in MRI images showing automated acquisition reduced operator dependence and was 13% faster compared to manual planning of cardiac scans. Chen et al. [54] demonstrated potential of ECG interpretation assisted by a DL algorithm to improve diagnosis of cardiovascular events in patients with heart failure.

3.7 Dermatology

Three studies examined assistive and autonomous AI for diagnosis and monitoring. Pangti et al. [55] undertook a large-scale study involving 5,014 patients across a wide variety of clinical settings in India to demonstrate the utility of a smartphone mobile app as a point-of-care tool for diagnosis of 41 skin conditions in people of colour (accuracy: overall 75%; top 3: 90%). Jain et al. [56] demonstrated another AI to help clinicians diagnose skin conditions more accurately in US primary care practices. Here, 20 physicians and 20 nurse practitioners reviewed 1,048 cases with and without assistance of an AI system that provides a differential diagnosis from images of skin conditions and medical history. AI assistance was significantly associated with higher agreement with diagnoses made by a dermatologist panel, with an increase from 48% to 58% for physicians and 46% to 58% for nurse practitioners. For monitoring fine lines, Yoelin et al. [57] examined utilisation and functionality of an AI platform to automatically measure and score fine lines by asking 71 patients to use the system over 14 days. The AI was shown to evaluate photos on a comparable level of accuracy and was more consistent than qualified raters.

3.8 Diabetic Retinopathy

Three studies examined autonomous AI for diagnosis of diabetic retinopathy in different settings. Ipp et al. [58] undertook a multicentre cross-sectional diagnostic study including 942 individuals with diabetes to demonstrate safety and accuracy. The system accurately detected both more-than-mild diabetic retinopathy (mtmDR) and vision-threatening diabetic retinopathy (vtDR) without physician oversight or need for dilation in most individuals (mtmDR sensitivity: 96%, specificity: 88% and vtDR sensitivity: 97%, specificity: 90%). Hao et al.’s [59] evaluation involved 3,933 patients in a community hospital in rural China. The AI was demonstrated to have a sensitivity of 81% and specificity of 94% and was consistent with screening by ophthalmologists. Also in China, Ming et al. [60] examined feasibility of deploying AI in primary care. The system which was capable of both detecting and grading according to the International Clinical Diabetic Retinopathy scale was demonstrated to have high specificity (98%) and acceptable sensitivity (85%).

3.9 In-hospital Deterioration

Three US studies examined assistive AI for detection of in-hospital deterioration. Martinez et al. [61] described an early warning system that combined statistical modelling with ML to identify patients at risk of deterioration. Deployment across 19 hospitals was associated with decreases in mortality (10% vs. 14%), hospital length of stay, and intensive care unit length of stay. The study estimated that more than 500 deaths could be prevented each year by the intervention. Winslow et al.’s [62] deploy-
ment in medical-surgical wards across a multicentre health system over 10 months was associated with a decrease in hospital mortality (9% vs. 14%).

Focussing on sociotechnical dimensions, Schwartz et al.'s [63] application of the human-computer trust conceptual framework to explore clinician trust is particularly noteworthy. Here, nurses and prescribers from 24 acute and intensive care units in two hospitals were interviewed about their trust in the predictive AI. The study reported that trust was influenced by clinician perceptions about being able to form a mental model and predict future system behaviour as well as the system’s technical capabilities to perform tasks accurately and correctly based on the information that is input. Trust was also influenced by actionability of system recommendations, scientific and anecdotal evidence as well as fairness in system predictions. The findings were largely similar between nurses and prescribers.

3.10 Stroke

The three studies of assistive and autonomous AI for stroke triage and diagnosis covered effects on decision-making, care delivery and patient outcomes. Gunda et al. [64] used mixed-methods to examine automated analysis of CT angiography at a primary stroke centre in Hungary. Use of AI over a 7-month period with 399 patients was reported to increase detection of thrombolysis (11%-18%) and thrombectomy (2.8%-4.8%). There was a trend towards shorter door-to-needle times (44–42 min) and CT-to-groin puncture times (174–145 min); and a non-significant trend towards improved outcomes with thrombectomy. Among physicians, the system was perceived to increase decision-making confidence and improved patient flow. Yahav-Dovrat et al. [65] evaluated the detection accuracy of an AI to detect large-vessel occlusions on CT angiograms and notify the treatment team in real-time via a dedicated mobile application at an Israeli stroke centre. The system was found to be highly accurate when used to scan all head and neck CT angiograms over a 15-month period. Hu et al.'s [82] examination of the safety and effectiveness of an AI to improve the quality of CT perfusion images reported improvements in image quality and thrombolytic therapy of acute cerebral infarct (14%).

3.11 Asthma, Sepsis, Venous Thromboembolism, Urinary Tract Infection

Four studies demonstrated effects of assistive AI on care delivery and/or patient outcomes in asthma, sepsis, and venous thromboembolism (VTE), as well as treatment of urinary tract infection. Seol et al. [18] conducted an RCT of an AI that provided a quarterly report to clinicians with relevant clinical information about asthma management along with a machine learned prediction for risk of exacerbation based on EHR data, patient-reported outcomes and non-clinical data (e.g. traffic volume and socioeconomic status). The study involving 184 patients in a US primary care paediatric practice found no difference in frequency of asthma exacerbations between the two groups (intervention: 12% vs. control: 15%), although the AI significantly reduced time for reviewing EHRs for asthma management (3 min vs. 11 min per patient). Another AI that leveraged the EHRs was evaluated by Adams et al. [66], here a real-time risk score was generated and used to alert clinicians about patients at risk of sepsis. A trial of this system at five hospitals was reported to reduce in-hospital mortality (treatment: 15% vs. comparison: 19%), organ failure and length of stay compared with patients whose alert was not confirmed by a provider within 3 hours. Taking a similar approach, Zhou et al. [67] evaluated an AI to identify and notify clinicians about patients at risk of VTE in hospital. AI-enabled automated assessment of VTE risk every 6 hours or upon EHR updates was found to reduce the rate of VTE during hospitalisation by 19% and increased anticoagulant drug use by 14%.

In primary care, Herter et al. [68] examined an AI that provided general practitioners with expected outcomes and support information about antibiotics for urinary tract infections based on the Dutch College of General Practitioners’ guidelines. A prospective observational study in 36 primary care practices over a 4-month period was associated with an increase in proportion of successful treatments from 75% to 80% in intervention practices while there was no difference in the matched controls.

3.12 Hip Repair Surgery, Dental Care

Two studies examined effects of assistive AI on decision-making and user experience in hip repair surgery and dental care. Li et al. [69] evaluated an AI to assist anaesthesiologists in assessing the risk of complications in patients after a hip surgery. The system was demonstrated to outperform the American Society of Anesthesiologist-Physical Status (ASA-PS) score, the traditional risk stratification method. The online app was user-friendly and received high satisfaction scores from anaesthesiologists. Focussing on trainees, Glick et al. [70] evaluated the performance, efficiency, and confidence level of 41 dental students on radiographic identification of furcation involvement (bone loss at branching point of the root of teeth), with and without AI assistance. While the AI did not improve decision-making speed or confidence, both groups acknowledged the role of AI in improving clinical decisions. Students also tended to over rely on AI advice.

3.13 Iron-deficient Anaemia, Chronic Inflammatory Bowel Disease, COVID-19, Wound Care, Gastrointestinal Obstruction, Chronic Kidney Disease

Six studies examined effects of AI on decision-making in iron-deficient anemia, chronic inflammatory bowel disease, COVID-19, wound care, gastrointestinal obstruction and chronic kidney disease. Kursjens et al. [71] demonstrated the utility of an algorithm that automatically assessed the risk of low body iron storage based on age, sex, a routine blood count and C-reactive protein concentration. Implementation in a hospital laboratory system
over a 1-month period resulted in one new iron deficiency diagnosis on average per day. Also using routinely collected data, Alrajhi et al. [72] demonstrated performance of a home-grown AI to predict the severity of COVID-19 infection for patients at hospital admission (recall: 78-90%; precision: 75-98%).

In imaging, Dong et al. [73] demonstrated utility of segmentation algorithm for CT images to evaluate the postoperative enteral nutrition and analyse the clinical treatment effect of high intestinal obstruction in neonates. The segmentation time of the algorithm was shorter than that of the traditional method (24 s vs. 75 s), and accuracy was higher (84% vs. 70%). Howell et al. [74] evaluated AI for wound assessment against manual assessments performed by wound care clinicians. While AI annotation algorithms performed similarly to human specialists, the degree of agreement regarding wound features among experts varied substantially, presenting challenges for defining a standard. Maeda et al. [75] undertook a prospective cohort study to show real-time use of AI during colonoscopy enabled prediction of the risk of clinical relapse in patients with ulcerative colitis in clinical remission. Chen et al. [76] demonstrated that an AI algorithm could improve image quality and reduce noise in CT images to assess nutritional management in chronic kidney disease.

4 Discussion

AI technologies are being applied in many clinical areas to improve patient care. Most contemporary systems are assistive and aimed at doctors in acute care settings in high-income nations. Studies provide evidence about AI systems being integrated and used with existing CISs including EHRs and supporting systems. Although most systems employ DL approaches, their algorithms are primarily trained on routinely collected data. Few utilised data about environmental and social factors, indicating limited support for the goals of One Health.

We found that 65% of systems were assistive, requiring users to confirm or approve AI provided information or decisions. This is consistent with the patterns observed in our analysis of ML-based medical devices approved by the US Food and Drug Administration where assistive devices made up 47% of the 49 reviewed [15]. Yet, little is known about the immediate and long-term effects of using such systems in clinical settings and is an area requiring further research [78, 79]. Immediate effects on users include the workload placed on clinicians to review and confirm AI output including the potential for errors [80, 81]. We found only one in five studies of assistive systems examined the effects on clinician time, and the two studies that assessed safety involved autonomous systems [58, 82]. The long-term effects on users include the loss of situational awareness and skill degradation which are well-documented effects of automation in other domains [9].

There is also a need to improve evaluation and reporting [83, 84]. Study designs were largely quantitative and aimed at examining effects on decision-making by comparing system performance against a gold standard (e.g., [25, 56]); or by comparing clinician performance with and without AI assistance (e.g., [20, 21, 32]). Few studies used randomised trial designs, opting instead for designs such as weaker historical case controls. Accuracy was most commonly measured, although a few studies examined safety and clinician time (e.g., [44, 46, 53]). Effects on care-delivery were assessed using a variety of measures including time to treatment (e.g., [29, 39, 67, 68]). Patient outcomes were assessed using measures like length of stay and mortality (e.g., [61, 62, 66]), but no studies examined adverse events due to AI errors. Though improvements in decision-making and care delivery are expected to improve patient outcomes, it cannot be assumed, making it essential to directly evaluate the effect of AI interventions on patient outcomes [1, 16].

While the current literature usefully provides evidence about beneficial effects on decision-making, care delivery and patient outcomes, little is known about the broader sociotechnical factors that affect the adoption and use of AI in routine settings. Few studies used mixed- or qualitative methods that can explain observed effects and examine the sociotechnical dimensions of AI [85]. There is also a need to improving report measures to identify and address ethical considerations for the use of AI in clinical settings [84].

Although a staged approach to implementation and evaluation was evident in many studies (e.g., [48, 66]), only three tracked actual use of systems by clinicians [28, 29, 75]. Evaluation of user experience was mostly confined to assessing satisfaction via surveys. A notable exception is the study by Rabinovich et al. [28] where the Technology Acceptance Model was employed to evaluate actual use and satisfaction post-implementation. Some studies used clinical simulation which permits patient- and risk-free evaluation and can inform real-world implementation [21,40,48,50]. For instance, Benrimoh et al. [48] and Tanguay-Sela et al. [50] reported high fidelity clinical simulations that informed an 11-month study to examine use of an AI for mental health in primary care [49]. Clinical simulation is particularly valuable to measure effects on decision-making including automation bias and other potentially remediable human factors issue posing safety risks [78], that are not feasible or ethical to study in clinical settings. Importantly, it enables safety and efficacy to be assessed ahead of expensive clinical deployment [86].

Other studies reported strategies to incorporate well-known considerations for the use of digital health technologies such as ensuring that AI advice was actionable (e.g., when algorithms were designed to operationalise national guidelines [48]), and integrated into clinical workflow and existing CISs including EHRs (e.g., [22, 29, 77]). In one study where the AI was not integrated into the EHRs,
general practitioners needed to enter patient characteristics into a web-based version of the AI [68]. A key challenge for system implementation is to build upon general considerations for digital health and identify specific measures required for the safe and effective use of ML algorithms. For instance, Schwartz et al. applied the human-computer trust conceptual framework to specifically examine trust around machine learned predictions of in-hospital deterioration [63]; Calisto et al. used the human-AI design guidelines to inform implementation of a breast screening AI into clinical workflow [21]. Another example is Jordan et al.’s study of the effects of cultural embeddedness on AI implementation in ED nursing triage [42].

Further work is also required to advance the goals of One Health via collaboration between the digital health and One Health communities [87]. Indeed, the One Health approach may help to move technology-driven AI beyond doctors in well-resourced acute care settings towards problem-driven systems addressing areas of specific clinical need and to improve equity in provision of health services [88]. Examples of such AI were evident in the studies we reviewed. For example, use of progressive web applications to migrate a pneumonia mortality prediction tool from a study prototype to a mobile app for a resource-constrained context was implemented in Gambia [17]. Another example is the use of non-clinical data such as traffic volume and socioeconomic status for predicting the risk of asthma exacerbations [18]. Other problem-driven examples include dermatology apps specifically developed for people of colour [27, 55], a radiology examination instruction system to support COVID-19 triage in a predominantly Spanish-speaking Latino community [29], and AI systems for trainee clinicians [36, 70].

5 Conclusions

This survey confirms that AI systems are being implemented and evaluated in many clinical areas. Most systems are assistive, requiring users to confirm or approve AI provided information or decisions. Study designs are largely quantitative measuring effects on decision-making, and there remain many opportunities to understand patterns of routine use and evaluate effects on care delivery and patient outcomes using mixed-methods. There is also a need to study the immediate and long-term effects of automation on human performance. No doubt newer generative AI that have potential to improve user interaction and reduce the burden associated with CIs need to be carefully evaluated as they come with their own risks that need to be managed. Support for One Health including better integration of data about environmental and social factors in CIs is another area for further exploration.

Funding

This research is supported by the Australian National Health and Medical Research Council Centre for Research Excellence in Digital Health (APP1134919) and Macquarie University. The funding sources did not play any role in study design, in the collection, analysis, and interpretation of data, in the writing of the report, or in the decision to submit the article for publication.

Competing Interests

The authors have no competing interests to declare.

Contributions

FM conceptualised the study, undertook the literature search, performed data analysis and drafted the article. All authors participated in writing and revising the article. All aspects of the study (including design; collection, analysis, and interpretation of data; writing of the report; and decision to publish) were led by the authors.

Acknowledgements

We thank Andre Nguyen and Ying Wang for their assistance with the searching and screening of articles.

References


Automation in Contemporary Clinical Information Systems: a Survey of AI in Healthcare Settings

JoVE (Journal of Visualized Experiments) 125

2023, Volume 2023, Issue 125, Pages 1-10


Correspondence to:
Prof. Farah Magrabi
Centre for Health Informatics
Australian Institute of Health Innovation
Macquarie University
Level 6, 75 Telopea Road, NSW 2109
Australia
Tel: +61 2 9850 2429
E-mail: farah.magrabi@mq.edu.au