MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230975

Using Clinical Simulation to Evaluate AI-Enabled Decision Support

David LYELL[a,1], Adriaan LUSTIG[b], Kate DENYER[b], Satya VEDANTAM[a] and Farah MAGRABI[a]
[a]*Australian Institute of Health Innovation, Macquarie University, Australia*
[b]*Faculty of Medicine, Health and Human Sciences, Macquarie University, Australia*

Abstract. Clinical simulation is a useful method for evaluating AI-enabled clinical decision support (CDS). Simulation studies permit patient- and risk-free evaluation and far greater experimental control than is possible with clinical studies. The effect of CDS assisted and unassisted patient scenarios on meaningful downstream decisions and actions within the information value chain can be evaluated as outcome measures. This paper discusses the use of clinical simulation in CDS evaluation and presents a case study to demonstrate feasibility of its application.

Keywords. Clinical decision support, evaluation, clinical decision making

1. Introduction

Little is known about the effects of AI on clinical decision-making. Machine learning (ML) enabled clinical decision support (CDS) tools have the potential to transform decision-making. Most provide information as an input to clinician decision-making or recommendations that clinicians need to confirm or approve [1]. The efficacy and safety of AI CDS is dependent upon how they contribute to the decisions made by clinicians.[2] However, few studies have examined their effects on decision-making.

One way to study use of AI-enabled CDS is via clinical simulations which provide an opportunity to examine their effects on decision-making in a variety of task environments from low fidelity laboratory tasks (e.g., use of CDS in x-ray interpretation) to high fidelity simulation environments (e.g. CDS use within a massive blood transfusion protocol). In this study we examine the feasibility of using a low fidelity clinical simulation to evaluate the effects of AI-enabled CDS on human decision-making. The development of the clinical simulation is described in Section 2. Section 3 presents a case study of using the clinical simulation to evaluate an AI that assists clinicians with interpretation of chest x-rays.

2. Methods

The contribution of CDS to healthcare can be evaluated by measuring downstream changes in the information value chain [2, 3]. Figure 1 illustrates the value chain for an

[1] Corresponding Author: David Lyell, email: david.lyell@mq.edu.au.

AI that assists clinicians in diagnosis by labelling x-ray findings: *Stage 1* Clinicians interact with CDS. *Stage 2* X-ray findings are received from CDS. Importantly, clinicians receive and consider information from other sources, including history, examination, laboratory tests and imaging. *Stage 3* Using information received, decisions are made about diagnosis and management. *Stage 4* Some decisions may result in changes to healthcare interventions. *Stage 5* Some changed interventions may improve patient outcomes.

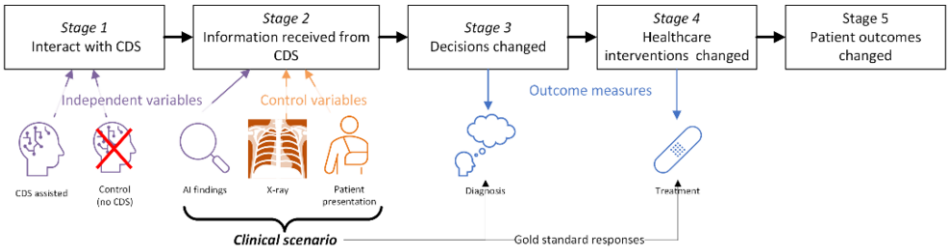

Figure 1. The information value chain. Adapted from Coiera [2].

Downstream changes in decisions (Stage 3) and actual or intended healthcare interventions (Stage 4) are outcome measures for simulation studies (Figure 1). Specific measures can include accuracy of diagnostic decisions, appropriateness of planned management and disposition. Without patients, simulation studies cannot directly assess patient outcomes. Instead, decisions and interventions can be evaluated against the current gold standard, such as clinical guidelines. An important value proposition of CDS to consider is greater efficiency despite there being no changes in downstream stages. Such measures can include resource utilization, time to decision and cognitive load [4].

The comparison is a Stage 1 manipulation, whereby CDS assisted trials are compared to current practice (as a control), where changes in decisions and healthcare interventions are the differences between intervention and control trials.

Different stages in the value chain can be manipulated depending on the study's aims. For example, how CDS is accessed can be varied between conditions in Stage 1 or the presentation of information in Stage 2, allowing different CDS designs, use cases and implementations to be compared. While varying the correctness of CDS in Stage 2 permits study of the potential for CDS to bias decisions [5].

2.1. Clinical Decision-Making Tasks

The core of any CDS study is the clinical decision-making task undertaken in providing healthcare. These are (1) key decisions that drive the provision of healthcare, and (2) made by those who have the responsibility and authority for those decisions. These key decisions are task dependent. For example, the output of examining x-rays differs by who is reading and their purpose. For radiologists, the output is their report on radiological findings, while for general physicians, it is diagnosis and patient management. Authority and responsibility for decisions are codified by regulation and sometimes by convention, which resides with human clinicians. The indications for the use of commercially available AIs provide cues by specifying how CDS fits into the task. An AI-enabled computer-assisted detection (CADe) medical device identifying

suspicious findings in screening mammograms "assist[s] interpreting physicians in identifying soft tissue densities and calcifications that may be confirmed or dismissed by the interpreting physician" [6].

2.2. Simulating Decision Making Tasks

Simulation of the decision task need only capture its essential essence, replicating as much of the task and with sufficient fidelity to be a faithful and valid representation of the task. Participants require access to the information and tools they would expect in clinical practice. For example, x-rays are read with knowledge of a patient's presenting complaint. Likewise, x-rays are read in medical image viewer software allowing manipulation of the image (e.g., zoom, pan, and adjustment of levels).

Clinical scenarios are a good method for simulating the decision-making environment and are a format that clinicians are familiar with from their clinical training. Scenarios are important Stage 2 experimental controls ensuring standardization and comparability between conditions. Likewise, scenarios can also be used to operationalize experimental variables by varying scenario characteristics between groups of equivalent scenarios, such as to test the effect of CDS on different presenting conditions or decision complexities.

The credibility of simulated decision-making studies comes from the validity of the clinical scenarios and their faithfulness to the decision task. Scenarios and the gold standard responses against which participant responses are assessed must be designed with and validated by clinical domain experts.

3. Results

We demonstrate the feasibility of clinical simulation in a student research project undertaken by medical students (AL & KD) enrolled at Macquarie University. The study examined the risk that medical students may overrely on AI CDS for identifying radiological findings in chest x-rays when making a diagnosis. Known as automation bias [7], overreliance on incorrect CDS is problematic and may lead to misdiagnosis.

Chest x-rays are one of the most ordered imaging studies, given their accessibility and usefulness in diagnostic and management decisions for a wide range of conditions.

Medical students are an interesting population as they are still gaining experience and developing expertise and are expected to receive greater benefits from CDS assistance than more experienced clinicians [8]. However, no AI model is perfectly sensitive or specific, and the expertise differential expected to benefit student diagnostic performance may also hinder their ability to recognize incorrect CDS.

3.1. Simulation Experiment

The study was within-participants, comparing three levels of CDS; correct, incorrect and control (no CDS). Automation bias is demonstrated by reduced participant diagnostic accuracy when assisted by incorrect CDS compared to a control condition.

34 students enrolled in the final two years of the Doctor of Medicine program at Macquarie University participated in the study. Their average age was 25 years, and 44% were female. Participants provided diagnoses for nine clinical scenarios with three randomly allocated to each condition. Scenarios comprised a short vignette describing

a patient presentation, a chest x-ray, and two sets of *Wizard of Oz* CDS, which is CDS generated by researchers but presented as if it comes from AI CDS. X-rays demonstrated findings consistent with pneumonia, pneumothoraces, COPD, trauma / musculoskeletal injuries, foreign body inhalation, retrosternal masses, and congestive cardiac failure. There were two normal x-rays. Students are trained to recognize these conditions in chest x-rays. Correct CDS identified findings related to the true diagnosis, while incorrect CDS did not, instead identifying plausible but false positive findings. Scenarios and the gold standard diagnoses were validated by a Fellow of the Royal Australian and New Zealand College of Radiologists.

The study was developed and deployed using Gorilla.sc [9] a platform for online behavioural science experiments. The interface was customized with the addition of two medical image viewers, one to display the chest x-ray and the second to display CDS. For control trials, the second viewer was replaced with the message "Decision support is not available for this patient." X-rays were displayed using OHIF [10], an open-source medical image viewer. Orthanc Server [11], a lightweight and open-source DICOM server, was used to serve images to OHIF using the DICOMweb protocol.

Each trial was presented in three phases; (1) a brief vignette describing a patient presentation is shown, and participants asked for their provisional diagnoses, (2) the chest x-ray and CDS are shown, and participants asked for their final diagnosis (Figure 2), (3) participants answer questions measuring their trust, reliance on CDS and cognitive load.

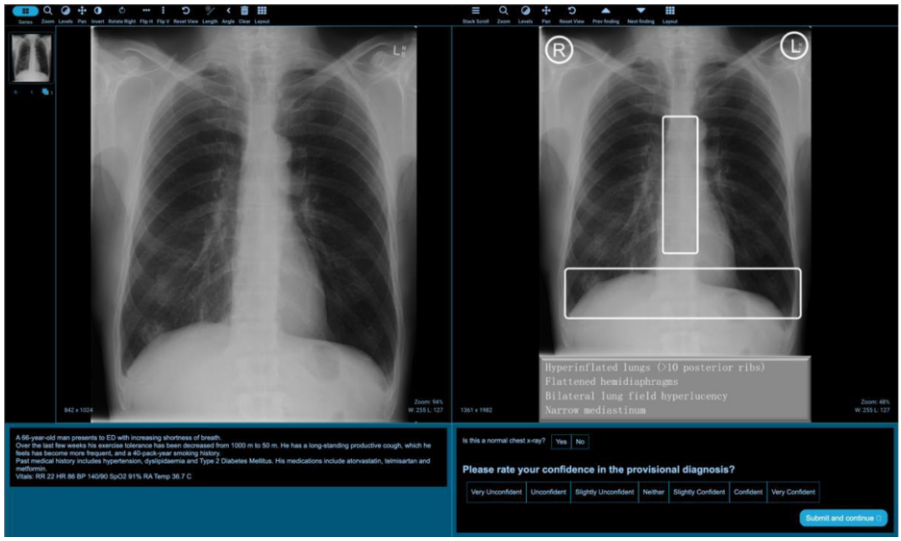

Figure 2. Phase 2 experimental task interface: Chest x-ray (top left), patient vignette (bottom left), CDS findings, (top right) and form to record participant responses (bottom right).

Participants were instructed (1) to approach each scenario as if they were treating a real patient, exercising all due care, and (2) that CDS had occasionally been incorrect and therefore CDS advice should always be double-checked. The study received ethics approval from the Macquarie University Human Research Ethics Committee, and participants were debriefed following their participation. Results (see Table 1) showed

that compared to the control condition without any CDS, correct CDS increased diagnostic accuracy by 21%, while incorrect CDS decreased accuracy by 11%.

Table 1. Experiment results (n=34).

	Control (No CDS)	Correct CDS	Incorrect CDS
Diagnostic accuracy	60%	81%	49%
Change compared to control	-	*21% increase*	*11% decrease*

4. Discussion

Simulated decision-making tasks involving clinical scenarios enable the study of CDS in a patient- and risk-free environment and provide far greater experimental control than is possible in real-world clinical studies. These studies complement rather than replace other methodologies, allowing safety to be evaluated in a simulated environment ahead of clinical deployment. Conversely, the control granted over clinician scenarios permits researchers to further investigate clinical study findings, test theory and establish causation. Especially for studying effects such as automation bias that are important for the safety and efficacy of CDS, but would not be feasible or ethical to test in real-world clinical studies.

5. Conclusions

Simulations are a valuable method to evaluate clinical decision-making tools.

References

[1] Lyell D, Coiera E, Chen J, Shah P, Magrabi F. How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices. BMJ Health Care Inform. 2021 Apr;28(1):e100301, doi: 10.1136/bmjhci-2020-100301.

[2] Coiera E. Assessing technology success and failure using information value chain theory. Stud Health Technol Inform. 2019 Jul;263:35-48, doi: 10.3233/SHTI190109.

[3] Coiera E. Guide to health informatics. Boca Raton: CRC press; 2015, doi: 10.1201/b13617.

[4] Lyell D, Magrabi F, Coiera E. The effect of cognitive load and task complexity on automation bias in electronic prescribing. Hum Factors. 2018 Nov;60(7):1008-21, doi: 10.1177/0018720818781224.

[5] Lyell D, Magrabi F, Raban MZ, Pont LG, Baysari MT, Day RO, Coiera E. Automation bias in electronic prescribing. BMC Medical Inform Decis Mak. 2017 Dec;17:1-0. doi: 10.1186/s12911-017-0425-5.

[6] U.S. Food and Drug Administration, K191994, ProFound AI Software V2.1, iCAD Inc., 2019.

[7] Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017 Mar;24(2):423-31, doi: 10.1093/jamia/ocw105.

[8] Povyakalo AA, Alberdi E, Strigini L, Ayton P. How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. Med Decis Making. 2013 Jan;33(1):98-107, doi: 10.1177/0272989X12465490.

[9] Anwyl-Irvine AL, Massonnié J, Flitton A, Kirkham N, Evershed JK. Gorilla in our midst: an online behavioral experiment builder. Behav Res Methods. 2020 Feb;52(1):388-407, doi: 10.3758/s13428-019-01237-x.

[10] Ziegler E, Urban T, Brown D, Petts J, Pieper SD, Lewis R, Hafey C, Harris GJ. Open health imaging foundation viewer: an extensible open-source framework for building web-based imaging applications to support cancer research. JCO Clin Cancer Inform. 2020 Apr;4:336-45, doi: 10.1200/CCI.19.00131.

[11] Jodogne S. The orthanc ecosystem for medical imaging. J Digit Imaging. 2018 Jun;31(3):341-52, doi: 10.1007/s10278-018-0082-y.