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

doi:10.3233/SHTI230971

How Well Do AI-Enabled Decision Support Systems Perform in Clinical Settings?

Anindya Pradipta SUSANTO[a,b,1], David LYELL[a], Bambang WIDYANTORO[b], Shlomo BERKOVSKY[a], and Farah MAGRABI[a]
[a] *Australian Institute of Health Innovation, Macquarie University, Australia*
[b] *Faculty of Medicine, Universitas Indonesia*
ORCiD ID: Anindya Pradipta Susanto https://orcid.org/0000-0001-5155-6904

Abstract. Real-world performance of machine learning (ML) models is crucial for safely and effectively embedding them into clinical decision support (CDS) systems. We examined evidence about the performance of contemporary ML-based CDS in clinical settings. A systematic search of four bibliographic databases identified 32 studies over a 5-year period. The CDS task, ML type, ML method and real-world performance was extracted and analysed. Most ML-based CDS supported image recognition and interpretation (n=12; 38%) and risk assessment (n=9; 28%). The majority used supervised learning (n=28; 88%) to train random forests (n=7; 22%) and convolutional neural networks (n=7; 22%). Only 12 studies reported real-world performance using heterogenous metrics; and performance degraded in clinical settings compared to model validation. The reporting of model performance is fundamental to ensuring safe and effective use of ML-based CDS in clinical settings. There remain opportunities to improve reporting.

Keywords. Clinical decision support, machine learning, performance

1. Introduction

Artificial Intelligence (AI) technologies, specifically machine learning (ML) models, are increasingly being embedded into clinical decision support (CDS) systems. While many ML-based CDS have been built, only a few are implemented in clinical settings and little is known about their performance in routine use [1; 2]. To address this gap, we conducted a scoping review of the use of ML-based CDS in clinical settings. The results of this scoping review have been reported in a separate publication. Here we specifically focus on examining the ML models and performance of the ML-based CDS in clinical settings.

2. Methods

We searched four bibliographic databases (PubMed, Medline, Embase, and Scopus) for original research articles describing the use of ML-based CDS in clinical settings. The search query included a combination of terms about AI/ML, CDS, clinical tasks, and clinical settings. We included studies published from January 2016 to April 2021

[1] Corresponding Author: Anindya P Susanto, email: anindya.susanto@hdr.mq.edu.au.

excluding systematic reviews, conference, and non-English papers. After removal of duplicates, titles and abstracts were screened by two independent reviewers (APS & FM).

For each included study, we extracted the CDS task, ML type, ML method, and real-world performance. CDS tasks were categorized into: (1) computerized provider order entry (CPOE) and e-prescribing, (2) diagnostic assistance, (3) therapy planning, (4) risk assessment, (5) process support systems, and (6) image recognition and interpretation including computer aided diagnosis. To obtain the ML type, method, and performance, we hand searched reference lists of retrieved articles. ML type was categorized into supervised learning, unsupervised learning, and reinforcement learning. ML performance was identified when algorithms were tested/validated on datasets and used prospectively in clinical settings to assess real-world performance. We extracted performance metrics including area under the error/loss function, receiving operator curve (AUC), precision recall curve (APR), accuracy (ACC), recall/sensitivity (SE), specificity (SP), precision/positive predictive value (PPV), and negative predictive value (NPV). The comparator and ground truth to assess real-world performance was identified.

3. Results

Of the 1,255 articles retrieved, 32 studies met the inclusion criteria (Table 1). The majority were prospective cohort studies (n=18; 56%) or randomized controlled trials (n=9; 28 %). Image recognition and interpretation (n=12; 38%) was the most common CDS task followed by risk assessment (n=9; 28%). Majority of studies reported models utilizing supervised learning (n=28; 88%) and a study used reinforcement learning (3%) [3]. Random forests (n=7; 22%) and convolutional neural networks (n=7; 22%) were the most common ML methods. Only 12 studies (37%) reported real-world performance. Of these, two compared CDS assisted decisions against a gold standard [4; 5]. The most common metrics were SE (n=10; 31%), SP (n=9; 28%), and AUC (n=5; 16%). Fifteen studies (47%) reported model performance, majority using AUC (n=9; 28%). Compared to model validation, performance of ML-based CDS degraded in the real world [6-11].

4. Discussion

Only one third of studies reported the real-world performance of ML-based CDS. Where performance was reported, data quality was poor. Heterogeneity in metrics prevented direct comparison, even for the same CDS task. While performance is assessed in development to choose the best ML method, real-world performance provides evidence about the efficacy and safety of a model for a specific CDS task. However, real-world performance is not covered by current reporting guidelines, such as DECIDE-AI [12] and CONSORT-AI [13]. As such a variety of metrics will need to be used to thoroughly examine the different clinical applications and specific tasks supported by ML-based CDS [14; 15]. For example, if the CDS task is to support chronic disease screening in healthy people, sensitivity is important. Conversely, a CDS supporting treatment planning for high-risk procedures requires high specificity. Furthermore, analysis of false positives and false negatives is necessary to support safe implementation and use. We also found problems with the reporting of ML type and method [4; 16; 17] which can help to increase transparency and enhance trustworthiness of clinical AI, and support studies to examine robustness and reproducibility.

Table 1. Studies about AI-based CDS in clinical settings by CDS task (*n*=32)

Author, Year [Reference]	**ML type; ML method**	**Model validation**	**Real-world performance**
Image recognition & interpretation (*n*=12)			
Gong, 2020 [6]	SL; Deep CNN, RF	AUC=84% to 95.24% for endoscope insertion. AUC = 90% for endoscope slipping.	Ground truth: recording video. Comparator: CDS output. Endoscope insertion; endoscope slipping. ACC=97.9%; 94.3%. SE= 95.8%; 98%. SP = 99.3%;98.8%.
Kim, 2020 [18]	SL; CNN	Not reported	Ground truth: Lab test for diagnosis. Comparator: CDS output. AUC=0.755. SE=70.2%. SP=72.7%. PPV=73.4%. NPV=61.5%.
Lin, 2019 [7]	SL; CNN	Diagnosis; treatment [19] ACC=98.87%; 97.56%	Ground truth: Expert clinician assessment. Comparator: CDS output on diagnosis; treatment. ACC=87.4%; 70.8%. SE=89.7%; 86.7%. SP=86.4%; 44.4%. PPV=74.4%. NPV=95%.
Liu, 2020 [20]	SL; CNN	Not reported	Not reported
Mori, 2020 [8]	SL; SVM	Expert; trainee. SE=93%; 95% SP=70%; 95.7% PPV=94.9%.; 94.1% NPV=63.6%; 96.4%	Ground truth: Pathology anatomy result Comparator: Clinician decision with CDS. SE, SP, PPV, NPV
Repici, 2020 [21]	SL; CNN	SE=99.7%.	Not reported
Savenije, 2020 [22]	SL; CNN	Not reported	Not reported
Tan, 2021 [23]	SL; DL, NL	Not reported	Not reported
Wang, 2019 [24] Wang, 2020 [25]	SL; CNN	AUC=0.98. SE=94.4%. SP=95.9%.	Not reported
Xiao, 2021 [26]	SL; DL	Not reported	AUC=0.74. SE=64%. SP=0.73%
Yao, 2021 [27]	SL; CNN	AUC=0.93. ACC=85.7%. SE= 86.3%. SP=85.7%. [28]	Not reported
Risk Assessment (*n*=9)			
Brennan, 2019 [5]	SL; Generalized additive model	Not reported	Ground truth: complication incidence. Comparator: Clinician assisted by CDS AUC=0.59; CDS output alone AUC=0.85.
Burdick, 2020 [29]	SL; GB	AUC=0.87 to 0.92 [30]	Not reported
Giannini, 2019 [31] Ginestra, 2019 [32]	SL; RF	AUC=0.88. PPV= 29%. SE=26%. SP=98%.	Not reported
Isma'eel, 2017 [33]	SL; ANN	Not reported	Ground truth: stress testing. Comparator: CDS output. SE=91%. SP=65%. PPV=26%. NPV=98%.
Jauk, 2020 [10] Jauk, 2021 [9]	SL; ANN	AUC=0.91	Ground truth: clinician diagnosis. Comparator: CDS output. AUC=0.86. SE=74.1%. SP=82.20%.
Sendak, 2020 [34]	SL; DL	Gain in AUC 19.4% and APR 55.5%.	Not reported
Shimabukuro [35]	SL; GB	Not reported	Not reported
Diagnostic assistance (*n*=5)			
Blomberg, 2021 [4]	Not reported	Not reported	Ground truth: diagnosis registry. Comparator: Clinician assisted by CDS. SE=85%. SP=97.4%
Grigull; 2016 [11]	SL; SVM, ANN, fuzzy, RF	AUC ranging from 0.918 to 1 for different classifiers. ACC=89.5%.	Ground truth: Test & diagnosis by specialist. Comparator: CDS output. PPV=0.83 to 1. NPV= 0.97 to 1
Marcos-P [36]	SL; RF, GB	Not reported	Not reported

Author, Year [Reference]	ML type; ML method	Model validation	Real-world performance
Rawson, 2018 [37]	SL; SVM	Not reported	Ground truth: blood culture. Comparator: CDS output. AUC=0.84. SE=89%. SP=63%.
Wintjens, 2020 [38]	SL; ANN, RF	Not reported	Ground truth: Molecular laboratory result. Comparator: CDS output. SE=86%. NPV=92%.
Therapy planning (*n*=3)			
Nicolae, 2020 [3]	RL	Not reported	Not reported
Niel, 2018 [39]	SL; Neural network	Loss function: Network error 0.00076	Not reported
Sibolt, 2021 [40]	SL; CNN	Not reported	Not reported
Diagnostic assistance (*n*=2)			
Chen, 2020 [41]	SL; GB	AUC=0.92, APR=0.56	Not reported
Romero-B [17]	Not reported	Not reported	Not reported
CPOE & e-prescribing (*n*=1)			
Segal, 2019 [16]	Not reported	Not reported	Not reported

ACC: accuracy ANN: artificial neural network, APR: area under precision-recall curve, AUC: area under receiver operating characteristic, CDS: clinical decision support, CNN: convolutional neural network, CPOE: computerized order entry, DL: deep learning, GB: gradient boosting, NL: natural language processing, NPV: negative predictive value, PPV: positive predictive value, RF: random forests, RL: reinforcement learning, SE: sensitivity, SL: supervised learning. SP: specificity, SVM: support vector machine.

5. Conclusions

This review has identified a gap in reporting about the real-world performance of ML-based CDS in clinical settings. Comprehensive performance reporting would enable clinicians to evaluate quality and safety of AI-enabled CDS for routine use.

References

[1] Coiera E, The last mile: Where artificial intelligence meets reality, J Med Internet Res 21 (2019), e16323.
[2] Hicks SA, et al., On evaluation metrics for medical applications of AI, Sci Rep 12 (2022), 5979.
[3] Nicolae A, et al., Conventional vs machine learning-based treatment planning in prostate brachytherapy: Results of a phase i randomized controlled trial, Brachytherapy 19 (2020), 470-476.
[4] Blomberg SN, et al., Effect of ML on dispatcher recognition of out-of-hospital cardiac arrest during calls to emergency medical services: A randomized clinical trial, JAMA Netw Open 4 (2021), e2032320.
[5] Brennan M, et al., Comparing clinical judgment with the mysurgeryrisk algorithm for preoperative risk assessment: A pilot usability study, Surgery (United States) 165 (2019), 1035-1045.
[6] Gong D, et al., Detection of colorectal adenomas with a real-time computer-aided system (endoangel): A randomised controlled study, The Lancet Gastroenterology and Hepatology 5 (2020), 352-361.
[7] Lin H, et al., Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: A multicentre randomized controlled trial, EClinicalMedicine 9 (2019), 52-59.
[8] Mori Y, et al., Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy a prospective study, Annals of Internal Medicine 169 (2018), 357-366.
[9] Jauk S, et al., Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: A mixed-methods study, Journal of Medical Systems 45 (2021).
[10] Jauk S, et al., Risk prediction of delirium in hospitalized patients using machine learning: An implementation and prospective evaluation study, J Am Med Inform Assoc 27 (2020), 1383-1392.
[11] Grigull L, et al., Diagnostic support for selected neuromuscular diseases using answer-pattern recognition and data mining techniques, BMC Medical Inform. Decis. Mak. 16 (2016), 31.
[12] Vasey B, et al., Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-AI, BMJ 377 (2022), e070904.

[13] Liu X, et al., Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The consort-ai extension, Nat Med 26 (2020), 1364-1374.
[14] Antoniou T, et al., Evaluation of machine learning solutions in medicine, CMAJ 193 (2021), E1425.
[15] Park SH, et al., Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction, Radiology 286 (2018), 800-809.
[16] Segal G, et al., Reducing drug prescription errors and ADEs by application of a probabilistic, ML-based CDS in an inpatient setting, J Am Med Inform Assoc 26 (2019), 1560-1565.
[17] Romero-Brufau S, et al., A lesson in implementation: A pre-post study of providers' experience with artificial intelligence-based clinical decision support, Int. J. Med. Inform. 137 (2020).
[18] Kim YJ, et al., Prospective, comparative evaluation of a deep neural network and dermoscopy in the diagnosis of onychomycosis, PLoS One 15 (2020).
[19] Long E, et al., An artificial intelligence platform for the multihospital collaborative management of congenital cataracts, Nat. Biomed. Eng 1 (2017).
[20] Liu WN, et al., Study on detection rate of polyps and adenomas in artificial-intelligence-aided colonoscopy, Saudi J Gastroenterol 26 (2020), 13-19.
[21] Repici A, et al., Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial, Gastroenterology 159 (2020), 512-520.e517.
[22] Savenije MHF, et al., Clinical implementation of mri-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy, Radiation Oncology 15 (2020).
[23] Tan JR, et al., Implementation of an artificial intelligence-based double read system in capturing pulmonary nodule discrepancy in ct studies, Curr. Probl 50 (2021), 119-122.
[24] Wang P, et al., Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study, Gut 68 (2019), 1813-1819.
[25] Wang P, et al., Lower adenoma miss rate of computer-aided detection-assisted colonoscopy vs routine white-light colonoscopy in a prospective tandem study, Gastroenterology 159 (2020), 1252-1261.e1255.
[26] Xiao W, et al., Screening and identifying hepatobiliary diseases through deep learning using ocular images: A prospective, multicentre study, The Lancet Digital Health 3 (2021), e88-e97.
[27] Yao X, et al., Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: A pragmatic, randomized clinical trial, Nat Med 27 (2021), 815-819.
[28] Attia ZI, et al., Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram, Nat Med 25 (2019), 70-74.
[29] Burdick H, et al., Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: A prospective multicentre clinical outcomes evaluation of real-world patient data from us hospitals, BMJ Health Care Inform 27 (2020).
[30] Mao Q, et al., Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu, BMJ Open 8 (2018), e017833.
[31] Giannini HM, et al., A machine learning algorithm to predict severe sepsis and septic shock: Development, implementation, and impact on clinical practice, Crit Care Med 47 (2019), 1485-1492.
[32] Ginestra JC, et al., Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock, Crit Care Med 47 (2019), 1477-1484.
[33] Isma'eel H, et al., ANN-based model enhances risk stratification and reduces non-invasive cardiac stress imaging compared to diamond–forrester and morise risk assessment models: A prospective study, J. Nucl. Cardiol 25 (2017), 1601-1609.
[34] Sendak MP, et al., Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study, JMIR Med Inform 8 (2020), e15182.
[35] Shimabukuro DW, et al., Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: A randomised clinical trial, BMJ Open Respir Res 4 (2017), e000234.
[36] Marcos-Pasero H, et al., Ranking of a wide multidomain set of predictor variables of children obesity by machine learning variable importance techniques, Sci Rep 11 (2021), 1910.
[37] Rawson TM, et al., A real-world evaluation of a case-based reasoning algorithm to support antimicrobial prescribing decisions in acute care, Clin. Infect. Dis 04 (2020).
[38] Wintjens A, et al., Applying the electronic nose for pre-operative sars-cov-2 screening, Surg Endosc 35 (2021), 6671-6678.
[39] Niel O, et al., AI outperforms experienced nephrologists to assess dry weight in pediatric patients on chronic hemodialysis, Pediatric Nephrology 33 (2018), 1799-1803.
[40] Sibolt P, et al., Clinical implementation of AI-driven cone-beam computed tomography-guided online adaptive radiotherapy in the pelvic region, Phys Imaging Radiat Oncol 17 (2021), 1-7.
[41] Chen J, et al., Development, implementation, and evaluation of a personalized ML algorithm for clinical decision support: Case study with shingles vaccination, J Med Internet Res 22 (2020), e16848.