Ann Acad Med Singap 2024;53:187-207
https://doi.org/10.47102/annals-acadmedsg.2023113

REVIEW ARTICLE

Clinical performance of automated machine learning: A systematic review

Arun James Thirunavukarasu[*1,2] *MB BChir*, Kabilan Elangovan[1] *BEng*, Laura Gutierrez[1] *MD*, Refaat Hassan[2] *MB BChir*, Yong Li[1,3] *MD*, Ting Fang Tan[1] *MBBS*, Haoran Cheng[1,3,4] *MPH*, Zhen Ling Teo[5] *FRCOphth*, Gilbert Lim[1] *PhD*, Daniel Shu Wei Ting[*1,3,5] *PhD*

ABSTRACT

Introduction: Automated machine learning (autoML) removes technical and technological barriers to building artificial intelligence models. We aimed to summarise the clinical applications of autoML, assess the capabilities of utilised platforms, evaluate the quality of the evidence trialling autoML, and gauge the performance of autoML platforms relative to conventionally developed models, as well as each other.

Method: This review adhered to a prospectively registered protocol (PROSPERO identifier CRD42022344427). The Cochrane Library, Embase, MEDLINE and Scopus were searched from inception to 11 July 2022. Two researchers screened abstracts and full texts, extracted data and conducted quality assessment. Disagreement was resolved through discussion and if required, arbitration by a third researcher.

Results: There were 26 distinct autoML platforms featured in 82 studies. Brain and lung disease were the most common fields of study of 22 specialties. AutoML exhibited variable performance: area under the receiver operator characteristic curve (AUCROC) 0.35–1.00, F1-score 0.16–0.99, area under the precision-recall curve (AUPRC) 0.51–1.00. AutoML exhibited the highest AUCROC in 75.6% trials; the highest F1-score in 42.3% trials; and the highest AUPRC in 83.3% trials. In autoML platform comparisons, AutoPrognosis and Amazon Rekognition performed strongest with unstructured and structured data, respectively. Quality of reporting was poor, with a median DECIDE-AI score of 14 of 27.

Conclusion: A myriad of autoML platforms have been applied in a variety of clinical contexts. The performance of autoML compares well to bespoke computational and clinical benchmarks. Further work is required to improve the quality of validation studies. AutoML may facilitate a transition to data-centric development, and integration with large language models may enable AI to build itself to fulfil user-defined goals.

Ann Acad Med Singap 2024;53:187-207

Keywords: AI, artificial intelligence, automated machine learning, autoML, machine learning, deep learning

CLINICAL IMPACT

What is New

- This systematic review identified 26 distinct autoML platforms that have been trialled and/or applied in a clinical context.
- AutoML exhibited variable performance; in head-to-head comparisons, AutoPrognosis and Amazon Rekognition performed the strongest with unstructured and structured data, respectively.

Clinical Implications

- The performance of autoML compares well to bespoke computational and clinical benchmarks across clinical tasks ranging from diagnosis to prognostication.
- Exemplar use cases include identifying pathology on common imaging modalities (e.g. chest X-ray) and predicting hospitalisation and mortality based on tabulated demographics and blood test results.
- AutoML may facilitate a transition from model-centric to data-centric development, and integration with large language models may enable automated development of AI applications to fulfil user-defined goals.

[1] Artificial Intelligence and Digital Innovation Research Group, Singapore Eye Research Institute, Singapore
[2] University of Cambridge School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom
[3] Duke-NUS Medical School, National University of Singapore, Singapore
[4] Rollins School of Public Health, Emory University, Atlanta, Georgia, USA
[5] Singapore National Eye Centre, Singapore
Correspondence: A/Prof Daniel Ting, Singapore Eye Research Institute, The Academia, 20 College Road, Level 6 Discovery Tower, Singapore 169856.
Email: daniel.ting@duke-nus.edu.sg
Dr Arun Thirunavukarasu, University of Cambridge School of Clinical Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Rd, Cambridge CB2 0SP, United Kingdom.
Email: ajt205@cantab.ac.uk

INTRODUCTION

In medicine, machine learning (ML) has been applied in a wide variety of contexts ranging from administration to clinical decision support, driven by greater availability of healthcare data and technological development.[1-5] Automated ML (autoML) enables individuals without extensive computational expertise to access and utilise powerful forms of artificial intelligence (AI) to develop their own models.[6] AutoML thereby enables developers to focus on curating high-quality data rather than optimising models manually, to facilitate a transition from model-driven to data-driven workflows.[7] AutoML has been posited as a means of improving the reproducibility of ML research, and even generating superior model performance relative to conventional ML techniques.[8]

AutoML technologies aim to automate some or all of the ML engineering process, which otherwise requires advanced data or computer science skills.[6] The first stage is data preparation, involving data integration, transformation and cleaning. Next is feature selection, where aspects of the data to be utilised in designing the ML model are decided; this may involve data imputation, categorical encoding and feature splitting.[9] Model selection, training and optimisation are then executed, with model performance evaluation being critical for identification of an optimal solution. AutoML systems use various methods and optimisation techniques to achieve state-of-the-art performance in some or all of the engineering process, such as Bayesian optimisation, random search, grid search, evolutionary based neural architecture selection, and meta-learning.[8,10] The optimised model may then be outputted for further work, such as clinical deployment, explainability analysis or external validation.

AutoML exhibits 4 major strengths, which may support its application in clinical practice and research. First, studies have reported comparable performance of autoML to conventionally developed models.[11] This raises the possibility of clinical deployment of autoML models and use in pilot studies preceding further model development. Second, autoML may improve the reproducibility of ML research by reducing the influence of human technicians who currently engage with an idiosyncratic process of tuning until a satisfactory result is achieved; this supports a transition toward more reproducible data-centric development.[7] Third, the reduction in computational experience and hardware conferred by autoML adoption should act as a major democratising force, providing a much larger number of clinicians with access to AI technology.[10] Last, the time spent on developing models is significantly reduced with autoML, as manual tuning is abolished—this improves efficiency and facilitates an acceleration of exploratory research to establish potential applications of AI.[10,12]

With the myriad of available autoML tools, democratisation of AI beyond those with clinical and computational expertise is feasible, and potential applications are diverse.[10,11] However, rigorous validation is necessary to justify deployment. Here, a systematic review was conducted to examine the performance of autoML in clinical settings. We aimed to evaluate the quality of result reporting; describe the specialties and clinical tasks in which autoML has been applied; and compare the performance of autoML platforms with conventionally developed models, as well as each other.

METHOD

The reporting of this study adheres to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidance, and the systematic review protocol was prospectively registered on international prospective register of systematic reviews (PROSPERO) (identifier CRD42022344427).[13,14] The protocol was amended to use a second quality assessment tool, the Prediction model Risk Of Bias ASsessment Tool (PROBAST) in addition to the Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence (DECIDE-AI) reporting guidelines, as described below.

Data sources and searches

The Cochrane Library, Embase (via Ovid), MEDLINE (via Ovid) and Scopus were searched from inception until 11 July 2022, with no initial restrictions on publication status or type. Our search strategy isolated autoML in clinical contexts with the use of Boolean operators, as detailed in Supplementary Appendix S1. Before screening, duplicates were removed using Zotero version 6.0.14 (Corporation for Digital Scholarship, Vienna, VA, US) and Rayyan.[13,15]

Study selection

Abstract screening was conducted in Rayyan by 2 independent researchers, with a separate third arbitrator possessing autoML expertise who resolved cases of disagreement.[15] Full-text screening was similarly conducted by 2 researchers with a separate arbitrator, in Microsoft Excel for Mac version 16.57 (Microsoft Corporation, Redmond, WA, US). The explicit, hierarchical criteria for inclusion during abstract

and full-text screening are listed below in descending order of hierarchy, with full details provided in Supplementary Appendix S2:

(1) Written in the English language
(2) Peer-reviewed primary research article
(3) Not a retracted article
(4) Utilises autoML
(5) AutoML is applied in a clinical context

Data extraction and quality assessment

For articles satisfying the inclusion criteria, data extraction was conducted by 2 researchers: a clinical researcher extracting data first, with subsequent verification by a second computational researcher. Quality assessment was conducted by a single researcher, using implicit criteria based on the DECIDE-AI framework.[16] Risk of bias and concerns regarding applicability were assessed similarly by 2 researchers using the PROBAST framework and guidance questions.[17]

Other data collected include citation details; the trialled autoML platform/s; processing location (cloud or local); code intensity of the autoML platform; technical features of the autoML platform; clinical task autoML was applied towards; medical or surgical specialty defined anatomically where possible; sources of data used to train and test models; training and validation dataset size; dataset format (i.e. structured or unstructured); evaluation metrics used to gauge performance; and benchmark figures if presented, such as with comparisons to expert clinician or conventional ML performance. Specifically, figures for area under the receiver operator characteristic curve (AUCROC), F1-score, and area under the precision-recall curve (AUPRC) were collected. If F1-score was not provided but precision (positive predictive value) and recall (sensitivity) were, F1-score was calculated as the harmonic mean of the 2 metrics. If metrics were not stated in text form but were clearly plotted in graphical form, figures were manually interpolated using WebPlotDigitizer v4.6.0 (Ankit Rohatgi, Pacifica, CA, US). Metrics were excluded if the source or modality of the tested model was unclear.[18] Where researchers disagreed, resolution was achieved through discussion or arbitration by a third researcher.

Data synthesis and analysis

A narrative synthesis was conducted because meta-analysis was precluded by heterogeneity of datasets, platforms and use cases. Quantitative comparisons of autoML models were based on performance metrics (F1-score, AUCROC and AUPRC) to judge the clinical utility of applied autoML.[19] AutoML platforms were compared using the same metrics where platforms were applied to an identical task with the same training and testing data. A statistically significant difference in metrics was defined as featuring non-overlapping 95% confidence intervals. To establish the congruence between studies' conclusions and their presented data, the discussion and conclusion sections of each study were appraised by a single researcher to identify if autoML was compared with conventional techniques, and if so whether the comparison favoured autoML, conventional techniques or neither. AutoML platforms were compared in terms of their requirements and capabilities, with researchers contacted to clarify any questions regarding code intensity, processing location or data structure. Figures were produced with R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria),[20-22] and Affinity Designer version 1.10.4 (Pantone, Carlstadt, NJ, US).

RESULTS

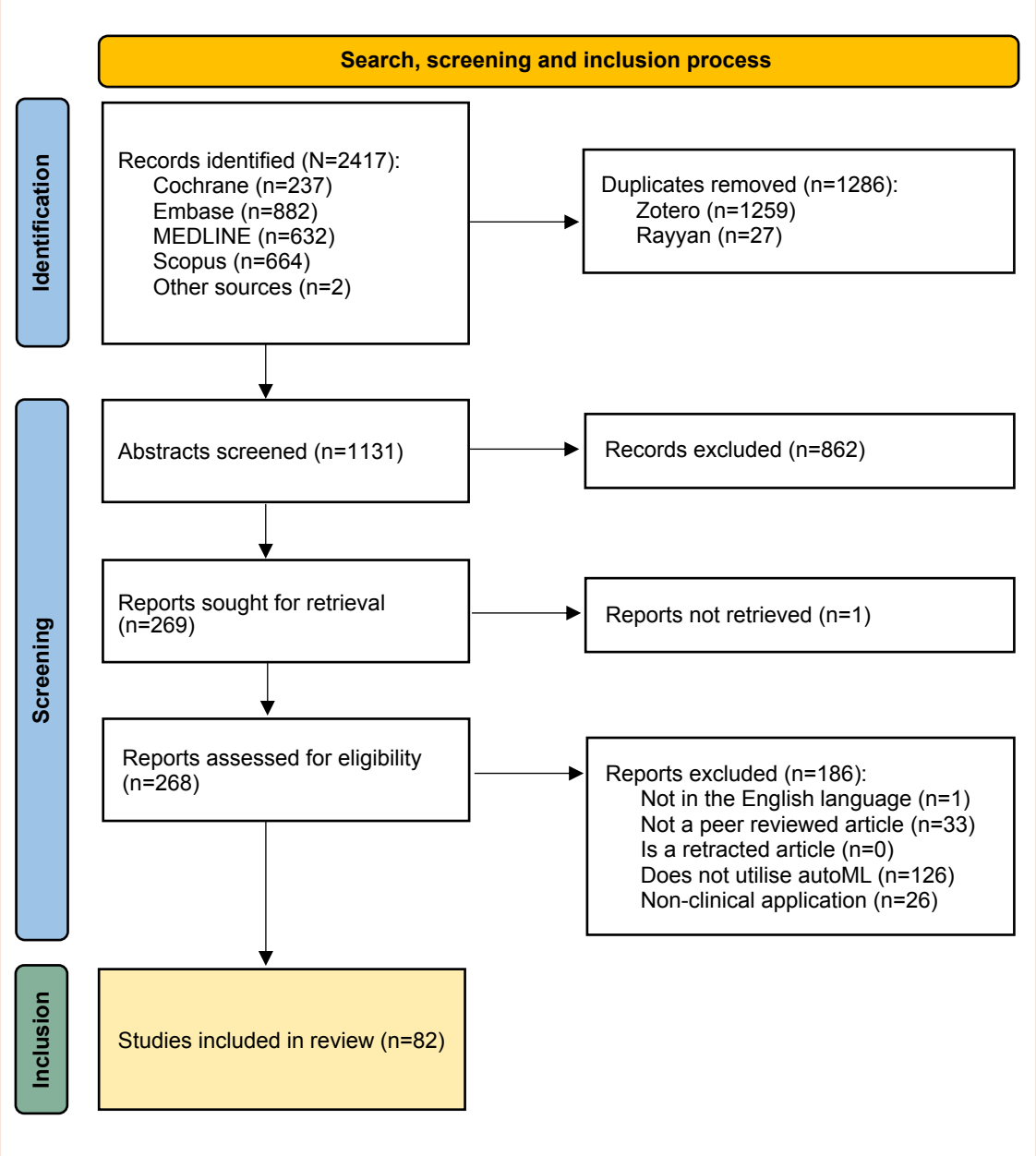
Record inclusion

Of 2417 records initially identified, 82 were included in the final analysis (Fig. 1 and Supplementary Appendix S3). In rare cases, research reports referred to autoML or similar terms in the broader context of "ML that automates", despite not utilising autoML technology: these articles were excluded for not meeting inclusion criterion 4.[23,24] Other borderline cases considered to be outside the scope of this review based on inclusion criterion 5 involved uses of autoML in clinical contexts, but without contributing to patient diagnosis, management or prognosis. These included a classifying surgical performance exhibited on video recordings and prediction of biological sex from medical images.[25,26]

Characteristics of included studies

The characteristics of the 82 included studies are summarised in Fig. 2 and Supplementary Appendix S3. To our knowledge, AutoML first entered medical literature in 2018 and has been growing in impact ever since: 1 paper in 2018, 7 in 2019, 21 in 2020, 35 in 2021, and 18 by 11 July 2022. Use cases are diverse, but diagnostic tasks (53 studies) were more common than management (4 studies) or prognostic (25 studies) tasks. The most common specialties in which autoML was used were pulmonology and neurology. Structured (e.g. tabulated) and unstructured (e.g. imaging) data were used similarly commonly. Dataset size varied widely, from 31 to 2,185,920 for training; 8 to 2,185,920 for internal validation; and 27 to 34,128 for external validation.

Fig. 1. PRISMA flow chart depicting the search, screening and inclusion process of this review.



Quality of reporting is summarised in Fig. 2F, with individual scores reported in Supplementary Appendix S4. The median number of fulfilled DECIDE-AI criteria was 14 out of 27, with the highest score being 19 out of 27. Nine criteria were fulfilled by over 90% of included studies. Thirteen criteria were not fulfilled in over half of the included studies, their DECIDE-AI criteria item numbers and items are as follows: (III) Research governance, (3) Participants, (5) Implementation, (6) Safety and errors in the results, (7) Human factors, (8) Ethics, (VI) Patient involvement, (9) Participants, (10) Implementation, (11) Modifications, (13) Safety and errors in results, (14) Human factors, and (16) Safety and errors in the discussion. Of these, 3 criteria were not fulfilled by any of the 82 included studies: (8) Ethics; (VI) Patient involvement; and (13) Safety and errors in the results.

Risk of bias and concerns regarding applicability are summarised in Fig. 2G. The most common sources of bias were retrospective study design that often used publicly available datasets, rather than testing autoML models in prospective trials to validate clinical performance and establish generalisability; and failure to provide an appropriate bespoke computational or clinical benchmark to demonstrate the performance of autoML—conferring unclear or high risk of bias in PROBAST

Fig. 2. Collectively summarised characteristics of included studies: (A) Date of publication bar chart. (B) Continent of corresponding author bar chart. (C) Country of corresponding author bar chart. (D) Funding source bar chart. (E) Clinical specialty bar chart. (F) DECIDE-AI score histogram. (G) PROBAST evaluation bar chart. (H) Dataset size histogram with logarithmic X-axis. (I) Data nature bar chart. (J) Data source bar chart.

CoA: concerns of applicability; RoB: risk of bias

appraisal (Supplementary Appendix S5). In many cases, this was because autoML was used as a tool rather than the study being a trial of autoML technology. However, a statement was made in the discussion or conclusion regarding the effectiveness of autoML in 27 of 47 studies (57%) judged to have a high or unclear risk of bias.

AutoML performance relative to other modalities

The reporting of performance metrics varied widely between papers, likely representing the inherent limitations of applied autoML platforms. There were 79 studies (96%) that provided AUCROC (Fig. 3), F1-score (Fig. 4), or AUPRC (Supplementary Appendix S6) as a measure of performance. Of these, 35 studies (44%) reported a computational or clinical benchmark to compare autoML performance against, and 21 studies (27%) provided 95% confidence intervals for estimates of performance metrics. Of 12 studies (15%) with benchmark comparisons and confidence intervals, autoML exhibited statistically significantly superior AUCROC in 6 of 17 trials (35%); significantly superior F1-score in 0 of 1 trial; and significantly superior AUPRC in 0 of 2 trials. In studies with benchmark comparisons and confidence intervals, autoML did not exhibit the lowest AUCROC, F1-score, or AUCPR in any trial. In all studies comparing modalities, autoML exhibited the highest AUCROC in 28 of 37 trials (76%), the highest F1-score in 11 of 26 trials (42%), and the highest AUPRC in 10 of 12 trials (83%). AutoML exhibited the lowest AUCROC in 5 of 37 trials (14%); the lowest F1-score in 6 of 26 trials (23%); and the lowest AUPRC in 2 of 12 trials (17%). For autoML models, AUCROC ranged from 0.346–1.000 (scores of 0.5 are equivalent to chance; maximum score=1); F1-score ranged from 0.128–0.992 (maximum score=1); and AUPRC ranged from 0.280–1.000 (maximum score=1).

There were 57 studies (70%) that compared autoML to other conventional modelling methods in the prose of their discussion or conclusion. Of these, 28 suggested that autoML was superior to conventional methods; 29 suggested that autoML was comparable to conventional methods; and none suggested that autoML was inferior to conventional methods. Only 35 studies provided a quantitative comparison in their results, as described above (Fig. 3, Fig. 4 and Supplementary Appendix S6). Conclusions of comparable effectiveness were justified by congruence with reported performance metrics in 16 of 29 studies (55%); conclusions of superior effectiveness of autoML were justified in 11 of 28 studies (39%).

Comparative performance of autoML platforms

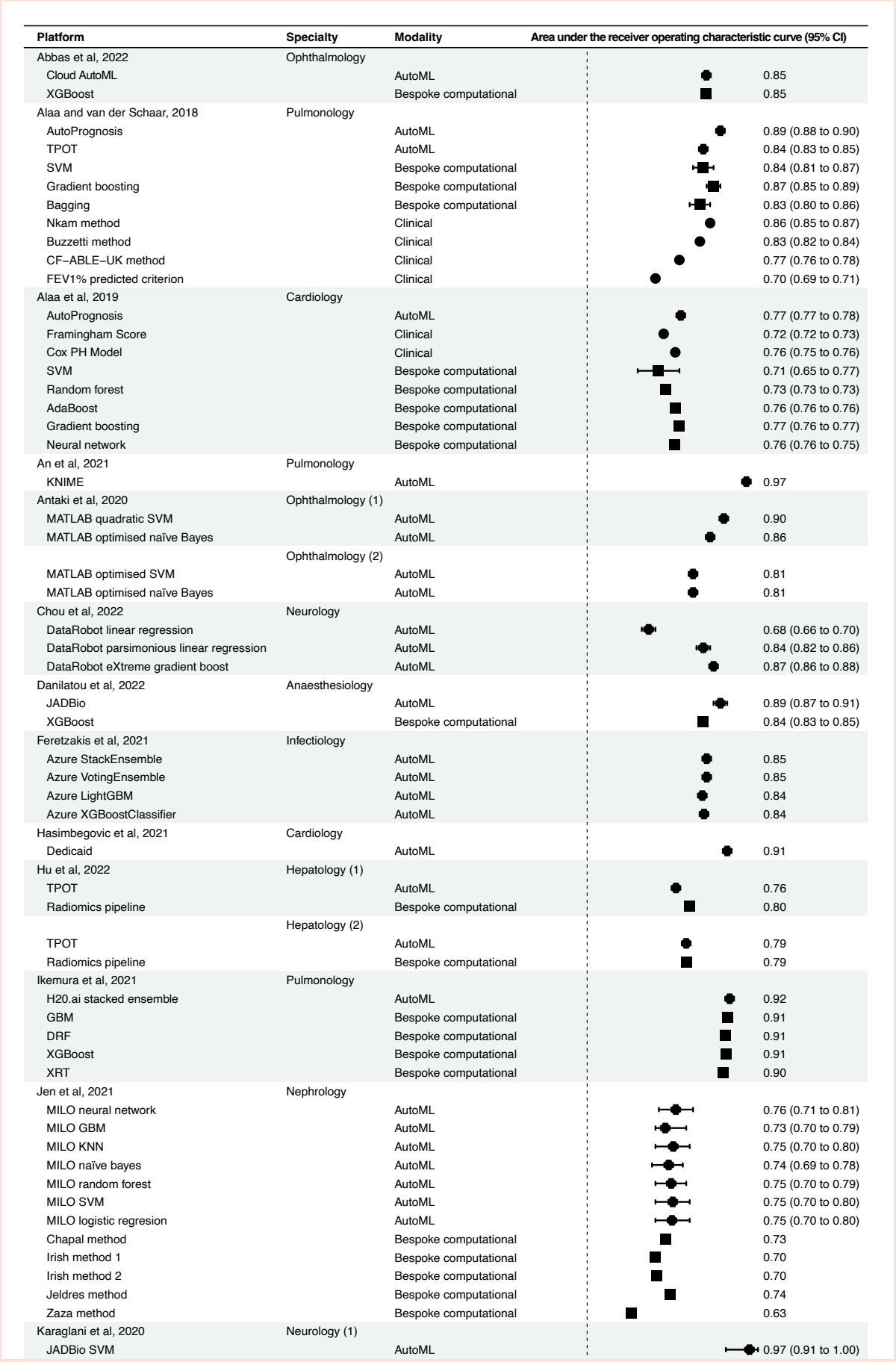
A comparative summary of the autoML platforms validated in the literature is presented in Table 1. Platforms vary greatly in their accessibility, technical features and portability. While performance in different tasks cannot be compared, 5 studies directly compared distinct autoML platforms in the same task. Of these, 1 study (20%) provided AUCROC metrics, which favoured AutoPrognosis over Tree-based Pipeline Optimization Tool (TPOT) to prognosticate mortality in cystic fibrosis.[27] Four studies (80%) provided F1-score metrics for a total of 9 trials (Fig. 5) on: prognosticating mortality in cystic fibrosis; predicting invasion depth of gastric neoplasms from endoscopic photography; diagnosing referrable diabetic retinopathy from fundus photography; diagnosing age-related macular degeneration, central serous retinopathy, macular hole and diabetic retinopathy from optical coherence tomography (OCT); diagnosing choroidal neovascularisation, diabetic macular oedema and drusen from OCT; and classifying spine implants from lumbar spine radiographs.[27-30]

AutoPrognosis (structured data) and Rekognition (unstructured data) exhibited the strongest performance as they were superior to every platform they were compared with, although this was only to TPOT for AutoPrognosis, and Rekognition was compared with fewer platforms than Cloud AutoML. Two studies (40%) reporting 5 trials provided AUPRC metrics for prognosticating mortality in cystic fibrosis and classifying electrocardiogram traces.[27,31] Here, performance favoured AutoPrognosis over TPOT; and AutoDAL-SOAR over USDM, AER, Auto-Weka, Auto-Sklearn and ASSL+US. While not all platforms can be compared against one another due to incompatibility with data structure, many possible combinations were not trialled and the number of comparative trials was small, making it difficult to establish comparative performance.

Confidence in conclusions

Confidence in conclusions is tempered by high risk of bias, particularly in retrospective study design and limited metrics facilitating statistical comparisons. However, as autoML did not exhibit statistically significantly worse performance than conventional techniques in any trial and exhibited lower performance metrics than conventional trials in a minority of studies, there is high confidence in the conclusion that autoML technology facilitates production of models with comparable performance to conventional techniques such as bespoke computational approaches. Given the low number of studies providing confidence intervals to enable statistical comparison of models' performance within trials, conclusions regarding the superiority of autoML relative to conventional techniques have low confidence. In addition, conclusions cannot be assumed to be generalisable to all use cases and datasets, given that performance is highly context-specific as demonstrated by the large variability observed in AUCROC (Fig. 3), F1-score (Fig. 4) and AUPRC (Supplementary Appendix S6). Confidence in the superior performance of AutoPrognosis with

Fig. 3. Forest plot depicting reported AUCROC metrics.



Platform	Specialty	Modality	Area under the receiver operating characteristic curve (95% CI)
Abbas et al, 2022	Ophthalmology		
Cloud AutoML		AutoML	0.85
XGBoost		Bespoke computational	0.85
Alaa and van der Schaar, 2018	Pulmonology		
AutoPrognosis		AutoML	0.89 (0.88 to 0.90)
TPOT		AutoML	0.84 (0.83 to 0.85)
SVM		Bespoke computational	0.84 (0.81 to 0.87)
Gradient boosting		Bespoke computational	0.87 (0.85 to 0.89)
Bagging		Bespoke computational	0.83 (0.80 to 0.86)
Nkam method		Clinical	0.86 (0.85 to 0.87)
Buzzetti method		Clinical	0.83 (0.82 to 0.84)
CF–ABLE–UK method		Clinical	0.77 (0.76 to 0.78)
FEV1% predicted criterion		Clinical	0.70 (0.69 to 0.71)
Alaa et al, 2019	Cardiology		
AutoPrognosis		AutoML	0.77 (0.77 to 0.78)
Framingham Score		Clinical	0.72 (0.72 to 0.73)
Cox PH Model		Clinical	0.76 (0.75 to 0.76)
SVM		Bespoke computational	0.71 (0.65 to 0.77)
Random forest		Bespoke computational	0.73 (0.73 to 0.73)
AdaBoost		Bespoke computational	0.76 (0.76 to 0.76)
Gradient boosting		Bespoke computational	0.77 (0.76 to 0.77)
Neural network		Bespoke computational	0.76 (0.76 to 0.75)
An et al, 2021	Pulmonology		
KNIME		AutoML	0.97
Antaki et al, 2020	Ophthalmology (1)		
MATLAB quadratic SVM		AutoML	0.90
MATLAB optimised naïve Bayes		AutoML	0.86
	Ophthalmology (2)		
MATLAB optimised SVM		AutoML	0.81
MATLAB optimised naïve Bayes		AutoML	0.81
Chou et al, 2022	Neurology		
DataRobot linear regression		AutoML	0.68 (0.66 to 0.70)
DataRobot parsimonious linear regression		AutoML	0.84 (0.82 to 0.86)
DataRobot eXtreme gradient boost		AutoML	0.87 (0.86 to 0.88)
Danilatou et al, 2022	Anaesthesiology		
JADBio		AutoML	0.89 (0.87 to 0.91)
XGBoost		Bespoke computational	0.84 (0.83 to 0.85)
Feretzakis et al, 2021	Infectiology		
Azure StackEnsemble		AutoML	0.85
Azure VotingEnsemble		AutoML	0.85
Azure LightGBM		AutoML	0.84
Azure XGBoostClassifier		AutoML	0.84
Hasimbegovic et al, 2021	Cardiology		
Dedicaid		AutoML	0.91
Hu et al, 2022	Hepatology (1)		
TPOT		AutoML	0.76
Radiomics pipeline		Bespoke computational	0.80
	Hepatology (2)		
TPOT		AutoML	0.79
Radiomics pipeline		Bespoke computational	0.79
Ikemura et al, 2021	Pulmonology		
H20.ai stacked ensemble		AutoML	0.92
GBM		Bespoke computational	0.91
DRF		Bespoke computational	0.91
XGBoost		Bespoke computational	0.91
XRT		Bespoke computational	0.90
Jen et al, 2021	Nephrology		
MILO neural network		AutoML	0.76 (0.71 to 0.81)
MILO GBM		AutoML	0.73 (0.70 to 0.79)
MILO KNN		AutoML	0.75 (0.70 to 0.80)
MILO naïve bayes		AutoML	0.74 (0.69 to 0.78)
MILO random forest		AutoML	0.75 (0.70 to 0.79)
MILO SVM		AutoML	0.75 (0.70 to 0.80)
MILO logistic regresion		AutoML	0.75 (0.70 to 0.80)
Chapal method		Bespoke computational	0.73
Irish method 1		Bespoke computational	0.70
Irish method 2		Bespoke computational	0.70
Jeldres method		Bespoke computational	0.74
Zaza method		Bespoke computational	0.63
Karaglani et al, 2020	Neurology (1)		
JADBio SVM		AutoML	0.97 (0.91 to 1.00)

Fig. 3. Forest plot depicting reported AUCROC metrics. Cont'd

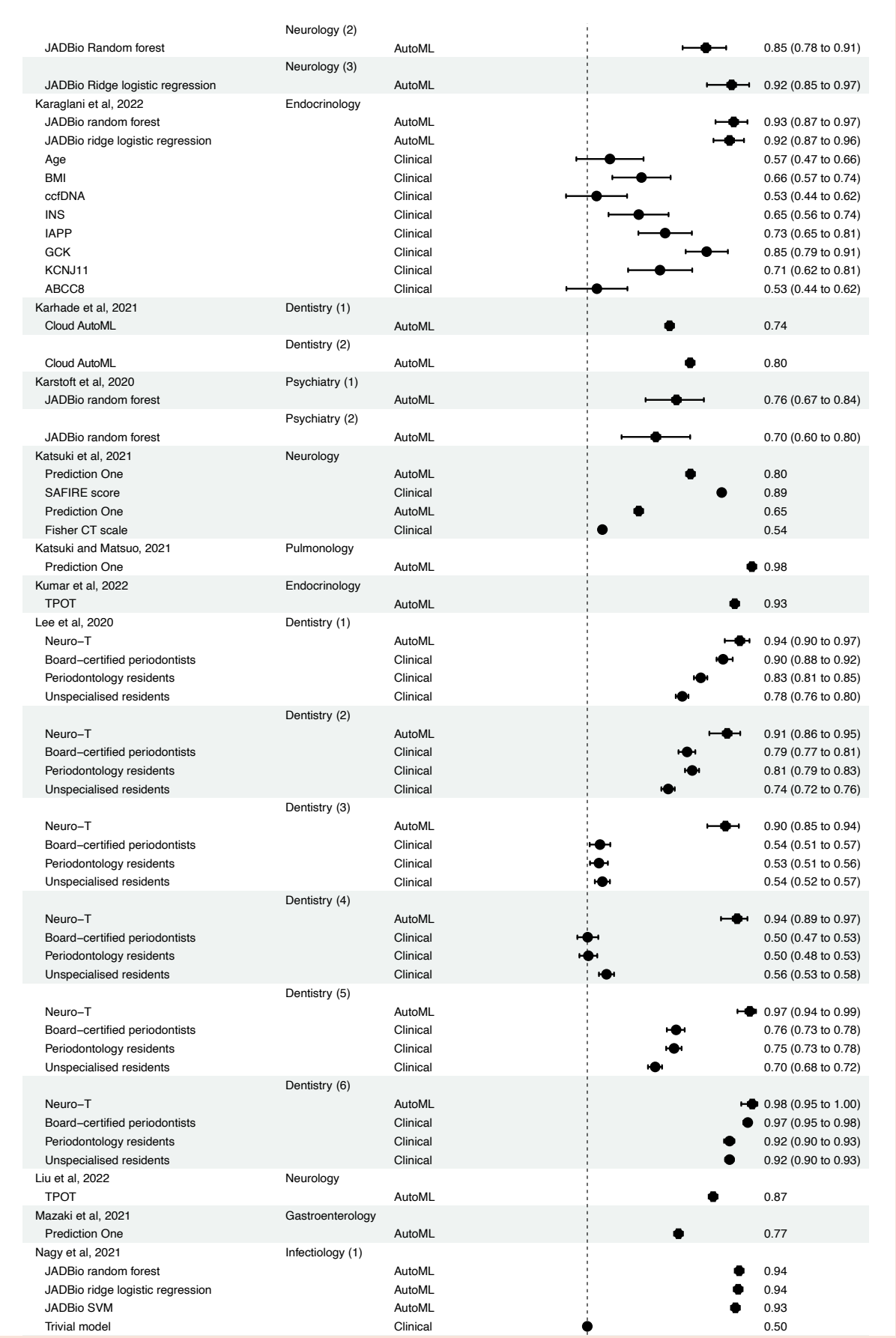


Fig. 3. Forest plot depicting reported AUCROC metrics. Cont'd

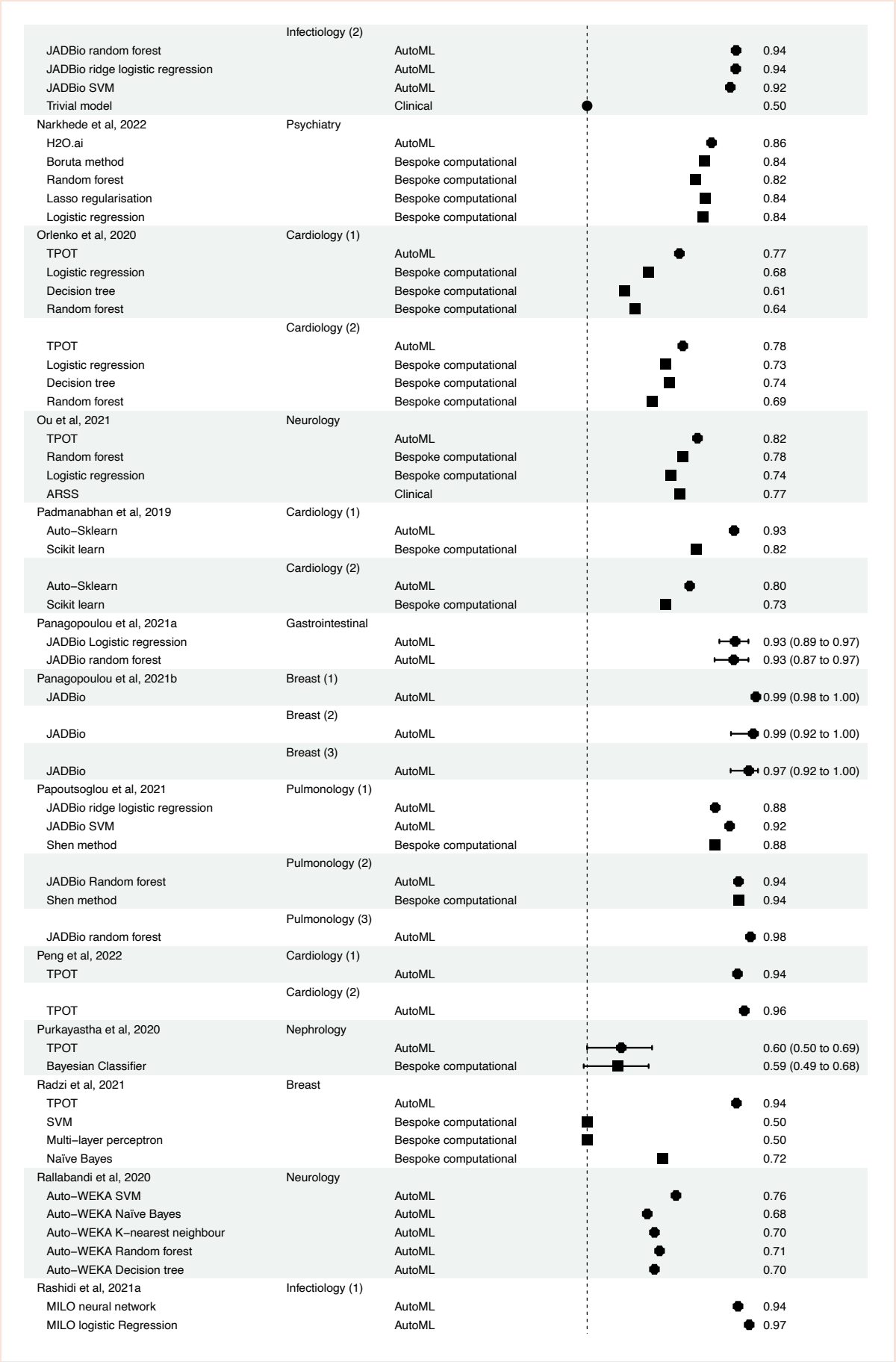


Fig. 3. Forest plot depicting reported AUCROC metrics. Cont'd

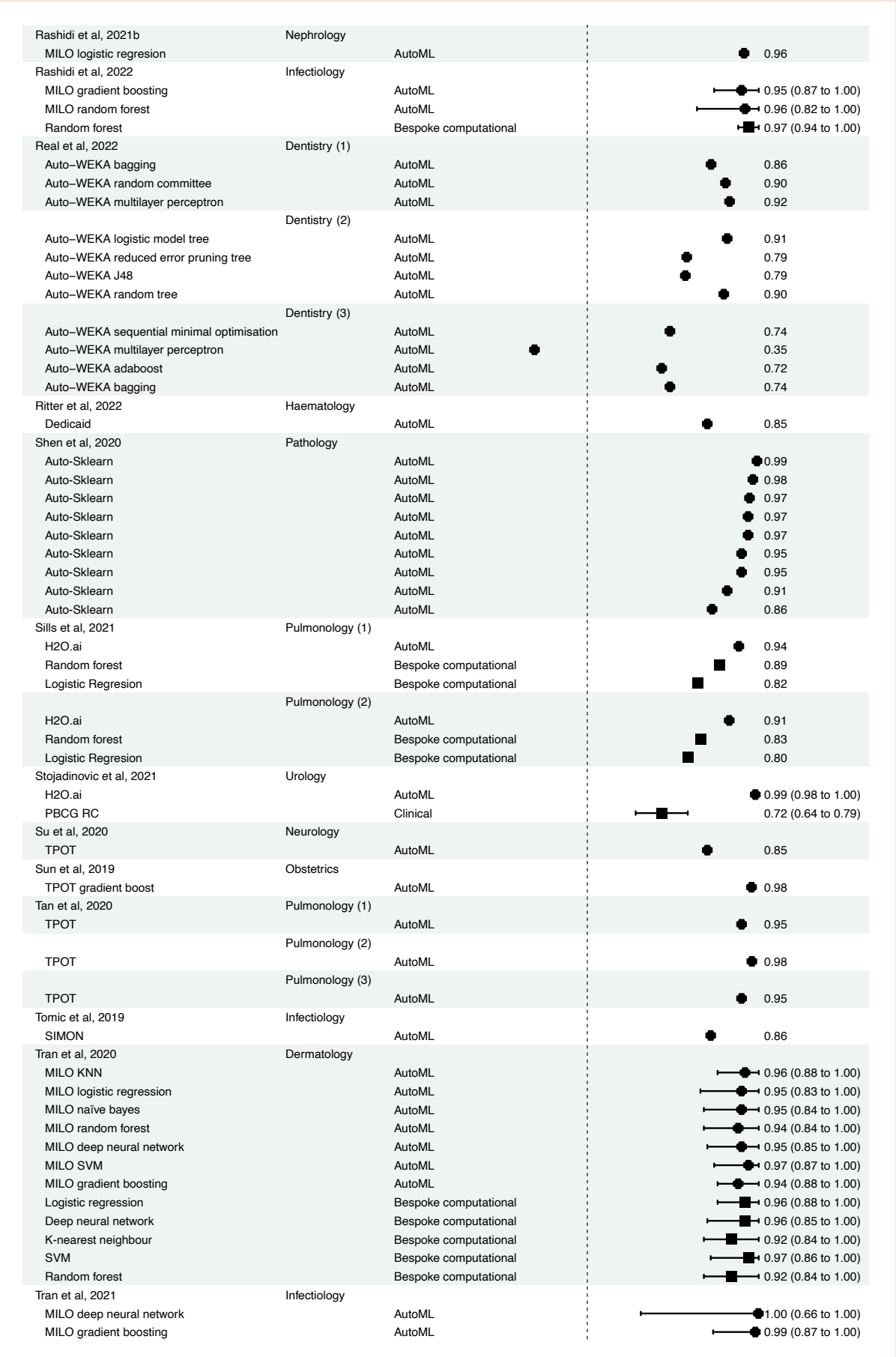
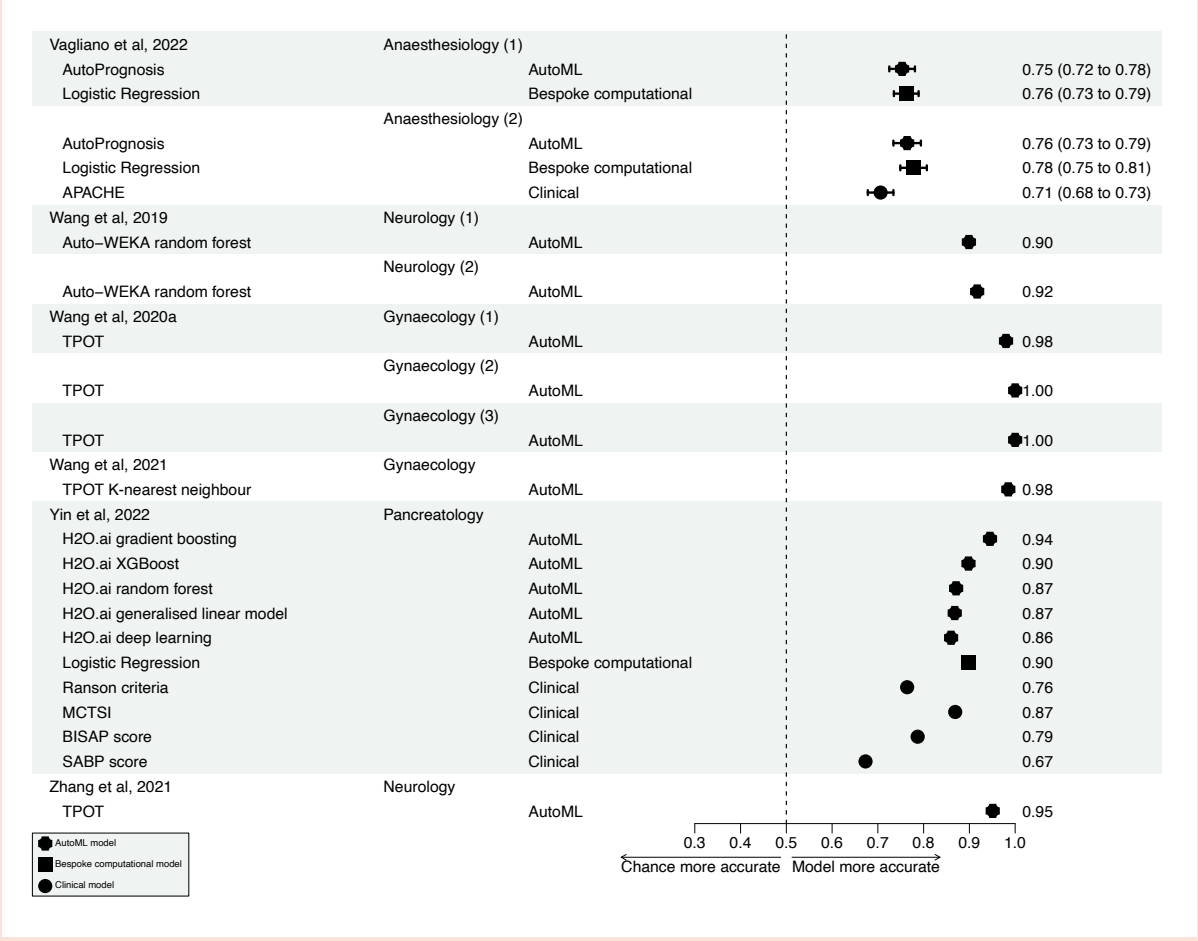


Fig. 3. Forest plot depicting reported AUCROC metrics. Cont'd



ARSS: Aneurysm Recanalization Stratification Scale; BMI: body-mass index; ccf-DNA: circulating cell-free DNA; CT: computerised tomography; DRF: distributed random forest; FEV1: forced expiratory volume in 1 second; GBM: gradient boosting machine; MCTSI: Modified Computed Tomography Severity Index; PH: proportional hazards; SVM: support vector machine; XRT: extremely randomised tree

structured data is very low, as there were very few comparative trials; and also low for the superior performance of Rekognition with unstructured data, as the number of comparative trials was low—though not as low as for structured data—and as there were no data for many possible platform comparisons.

DISCUSSION

This study shows that autoML has been trialled in a wide variety of diagnostic, patient management and prognostic tasks. AutoML has been used in many clinical specialties, most commonly in brain and lung imaging. Performance of autoML models generally compares well to bespoke computational and clinical benchmarks, often exhibiting superior performance. However, available studies and appraised risk of bias preclude conclusion that autoML provides universally superior performance to conventional modelling, as relative and absolute performance vary widely with the applied platform, use case, and data source. The strength of the evidence base supporting use of different autoML platforms is highly heterogenous, with some platforms exhibiting results more supportive of equivalence or superiority to conventional techniques than others. Few studies compared different autoML platforms to determine which provide optimal performance for a given task. Despite these knowledge gaps, a high number of non-comparative studies suggests that autoML is already being applied as a statistical tool, comparable to bespoke machine learning coding packages or statistical software.

There are 5 main deficiencies in the quality of the autoML evidence base. First, inconsistency in performance metrics may be a consequence of restrictions imposed by autoML platforms but

Fig. 4. Forest plot depicting reported F1-score metrics.

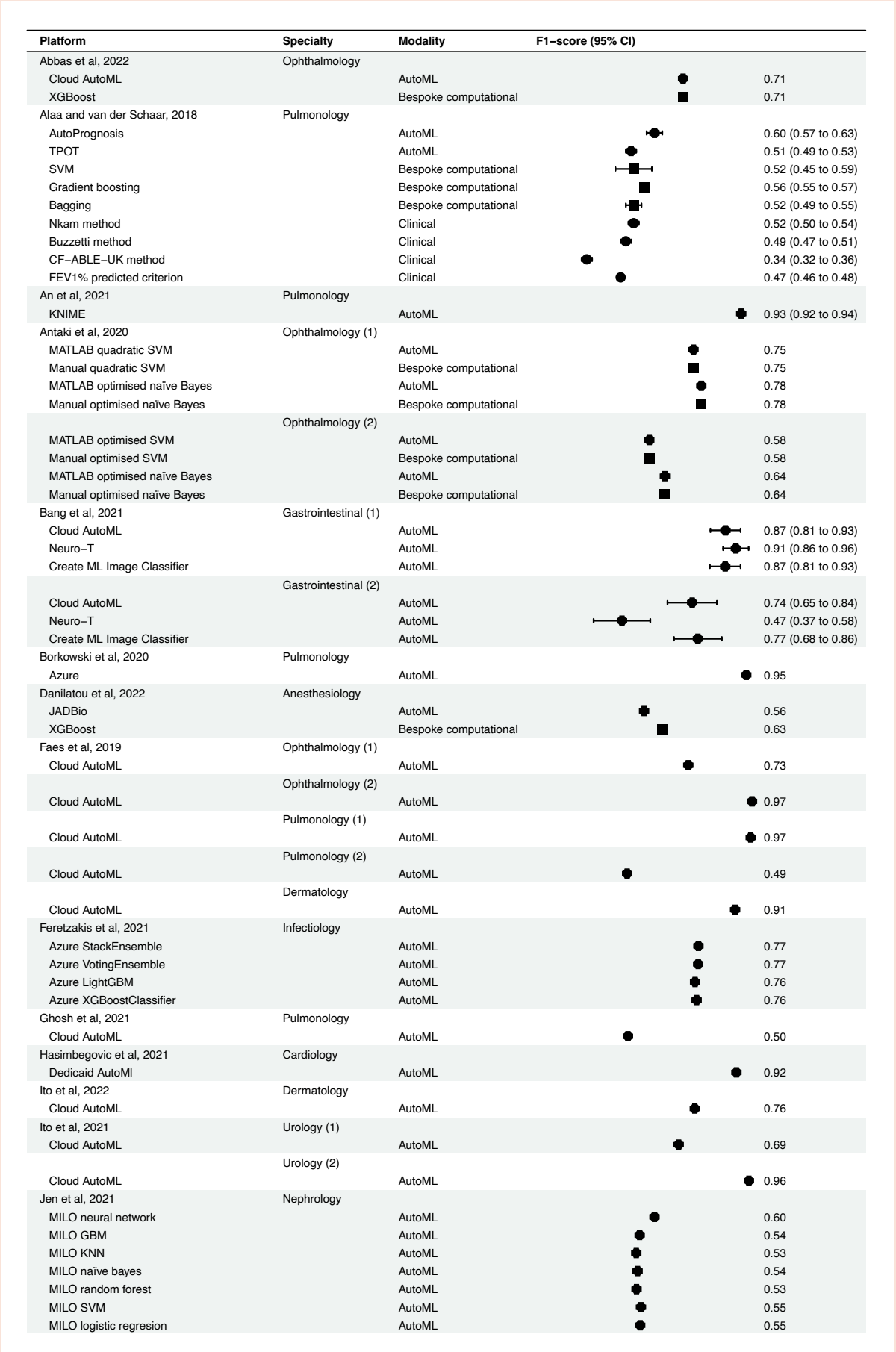


Fig. 4. Forest plot depicting reported F1-score metrics. Cont'd

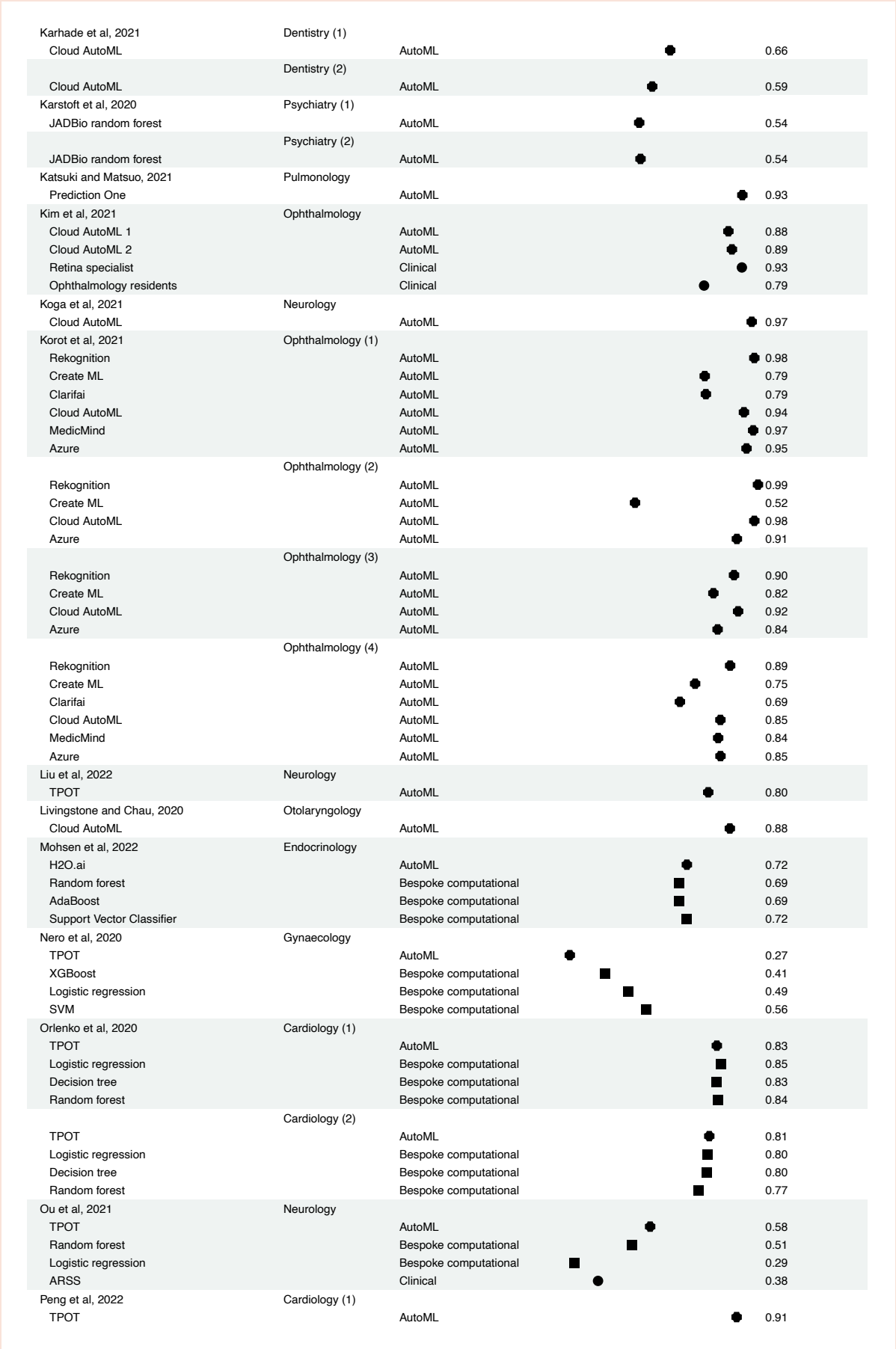


Fig. 4. Forest plot depicting reported F1-score metrics. Cont'd

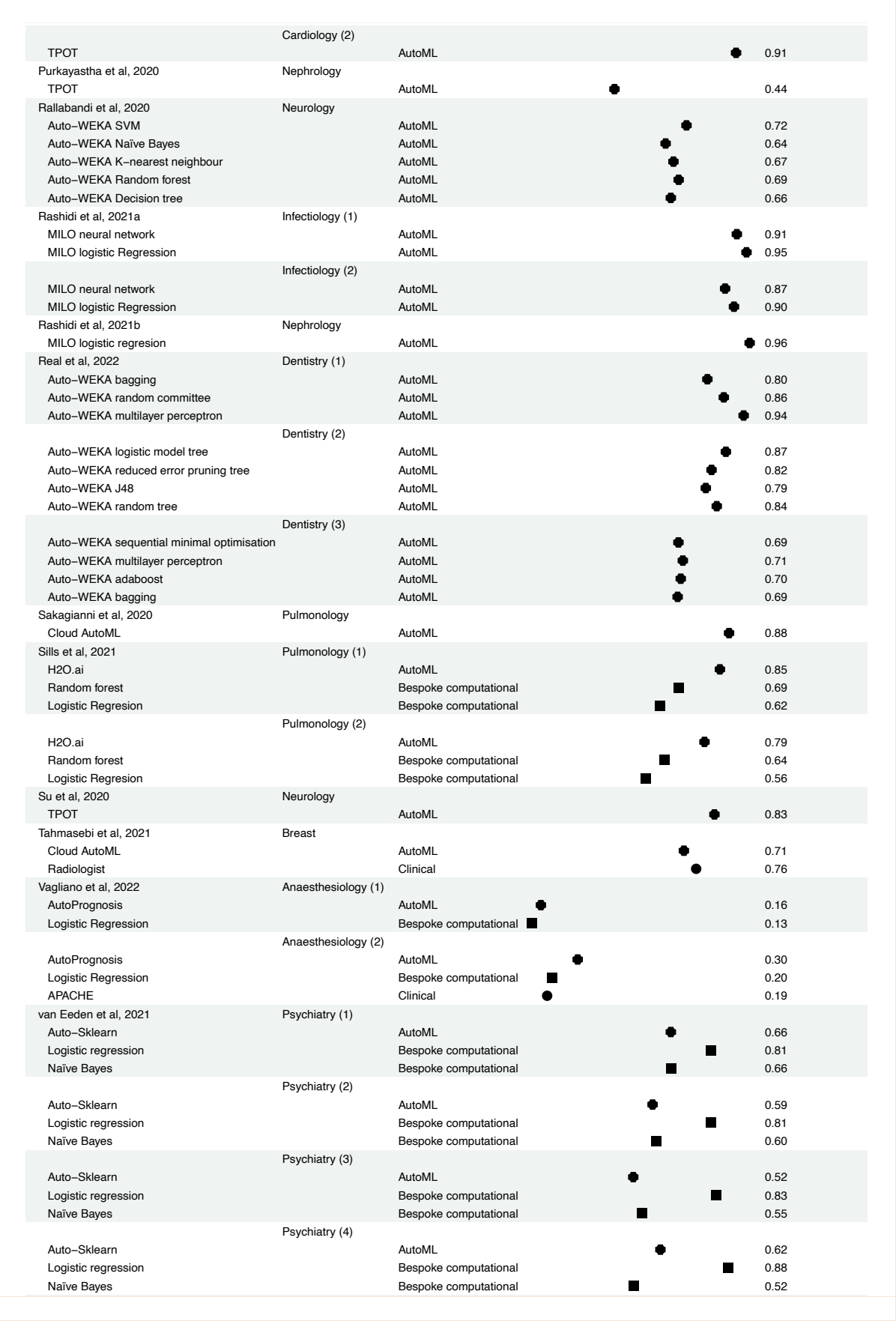
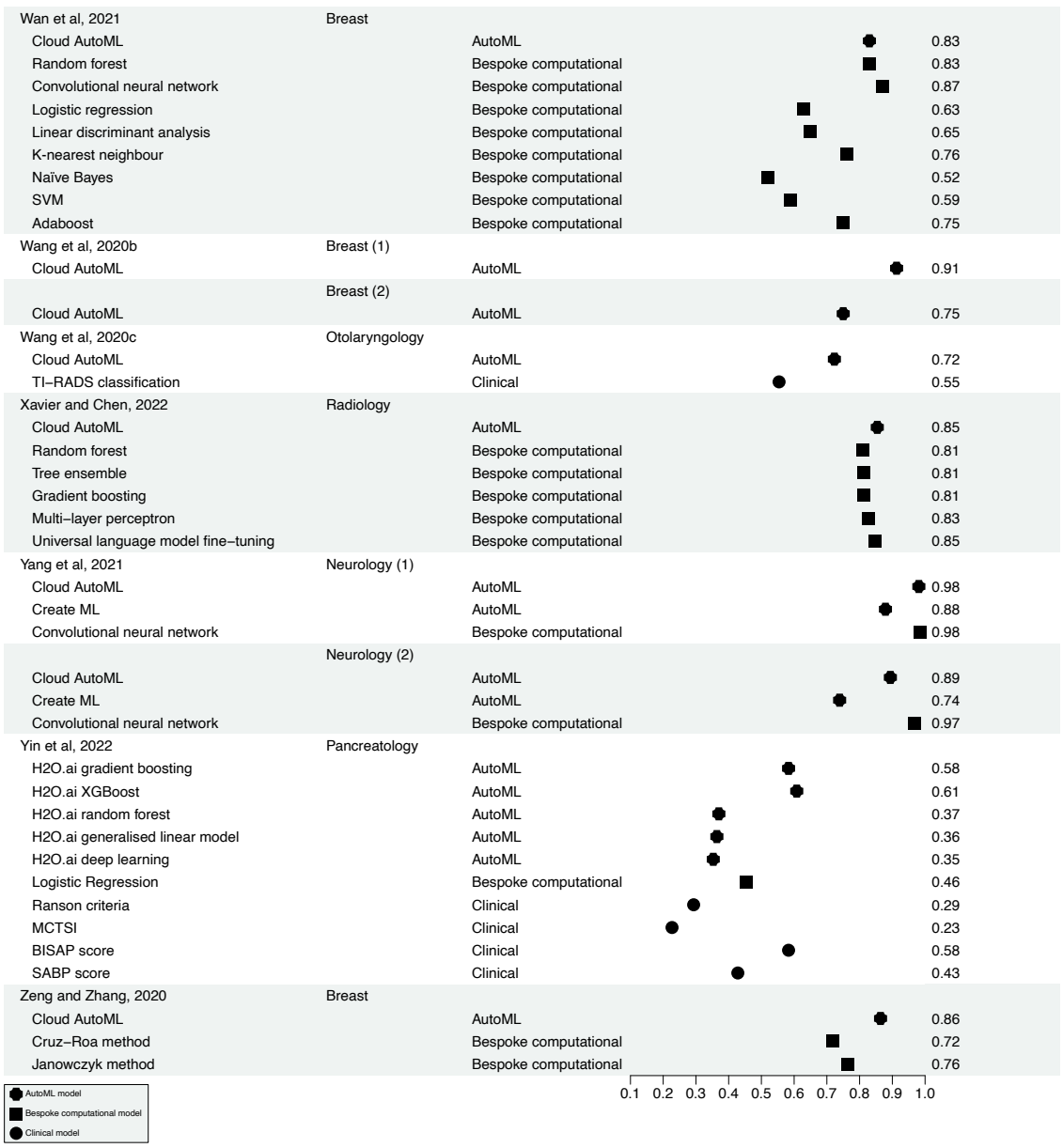


Fig. 4. Forest plot depicting reported F1-score metrics. Cont'd

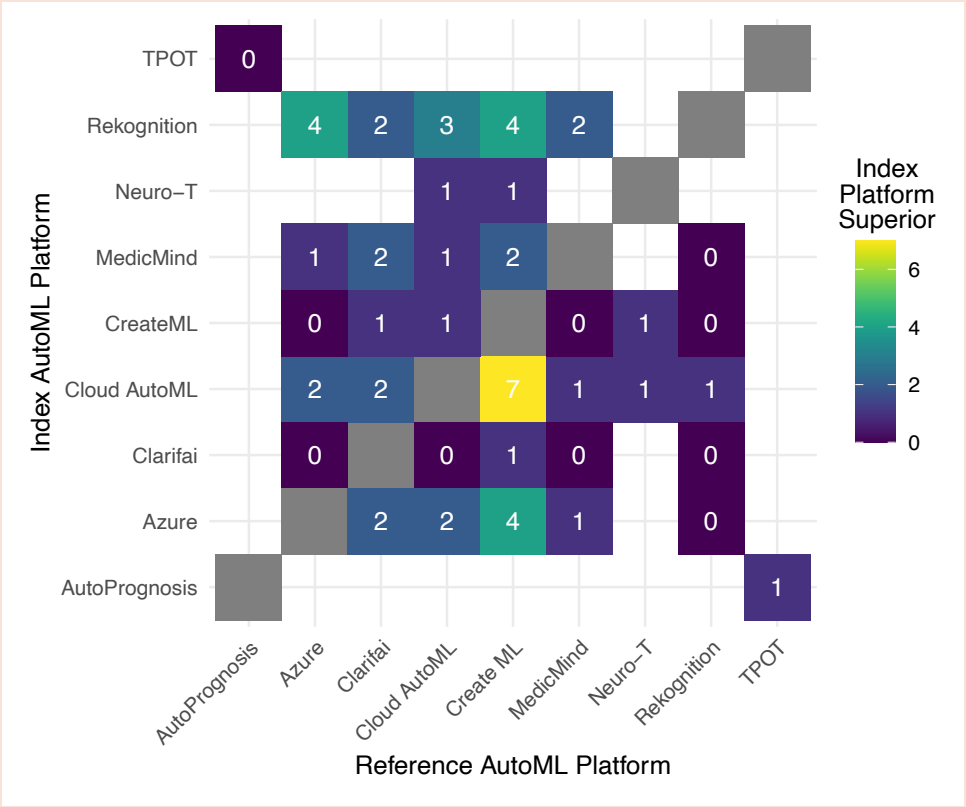


ARSS: Aneurysm Recanalization Stratification Scale; CI: confidence interval; FEV1: forced expiratory volume in 1 second; GBM: gradient boosting machine; SVM; support vector machine

observed variation between studies using similar platforms also suggests that selective reporting is common. Reporting comprehensive metrics is essential, particularly in the context of diagnostic algorithms, as some metrics are a function of prevalence or model threshold.[19] Second, explainability analysis is challenging for similar reasons to portability, but is possible with emerging technological solutions.[32] In addition, some platforms incorporate inbuilt explainability, such as by providing salience maps for deep learning models.[12] Issues regarding "black box" algorithms are accentuated in autoML research, leading to a third limitation: a lack of ethical consideration—such as regarding algorithmic fairness—by all the included studies.

Fourth, inconsistent use of benchmarking represents a form of publication bias leading to erroneous conclusions of equivalent or superior autoML performance relative to conventional bespoke computational methods or clinicians. Many studies relied on historical controls or provided no benchmark at all. To confidently conclude that autoML performance compares well to bespoke

Fig. 5. Heat map depicting the comparative performance of autoML platforms as applied to the same clinical tasks in terms of F1-score. Shading and numbers correspond to the number of superior performances exhibited by the index platform with respect to the reference platform.



models—and particularly to state-of-the-art techniques—a researcher with computational aptitude should have an opportunity to maximise performance. Finally, models should be deployed on separate datasets which were not used in testing or training, for external validation to demonstrate generalisability, which is a critical component of clinical potential.[6] Without external validation, overfitting to the datasets provided may lead to inflated estimates of performance.[6,33] External validation is limited on many autoML platforms by a lack of ability to batch test on new data, or to export models for analysis and deployment.

Limitations

This systematic review was limited by 3 issues: (1) PROBAST had to be adapted to apply it in non-diagnostic applications of autoML—we employed DECIDE-AI as a domain-specific quality indicator to mitigate this limitation, and utilised PROBAST in the context of trialling autoML technology rather than in validating models for clinical application. Development of more domain-specific tools to optimise AI-related systematic reviews is underway and will be a welcome development.[34,35] (2) Confidence in conclusions was affected by high risk of bias, a common theme in AI research more broadly.[36] We provide comprehensive indicators of quality, risk of bias and concerns regarding applicability to facilitate contextualisation of performance metrics. (3) It is difficult to draw conclusions for autoML as a modality because platforms are variable in their features, performances and requirements—future reviews may focus on individual platforms, although the number of studies featuring most platforms is very small.

Implications

Researchers applying a platform without providing benchmark comparators for the purposes of primary research or clinical work should justify their decision with validation data demonstrating that their approach is acceptable. Evidence should be contextually relevant, preferably pertaining to the same clinical task. While it is apparent that autoML has already begun to be applied in clinical research as a statistical tool, it is important that these tools are demonstrated to produce accurate, reliable and fair models. Studies purported as evidence of validation of autoML are often limited by retrospective design, high risk of bias, and unfulfill-

Table 1. Technological comparison of autoML platforms applied in the studies included in this review.

AutoML platform name	Accessibility			Technical features					Portability	
	Cost	Code requirement	Computing location	Dataset format	Feature extraction or selection	Model selection or training	Hyperparameter optimisation	Evaluation	Model exportability	Explainability
AER	Free	Coding required	Local	Structured/ Unstructured	Yes	Yes	Yes	No	Yes	No
Amazon Rekognition	Chargeable	None	Cloud	Unstructured	Yes	Yes	Yes	Yes	No	No
Apple Create ML	Free on Apple devices	None	Local	Structured/ Unstructured	Yes	Yes	Yes	Yes	Yes	No
ASSL	Free	Coding required	Local	Strucured/ Unstructured	Yes	Yes	Yes	No	Yes	No
Auto-Sklearn	Free	Coding required	Local	Structured/ Unstructured	Yes	Yes	Yes	Yes	Yes	No
Auto-WEKA	Free	Coding required	Local	Structured/ Unstructured	Yes	Yes	Yes	Yes	Yes	No
AutoDAL	Free	Coding required	Local	Structured	No	Yes	Yes	Yes	Yes	No
AutoDC	Free	Coding required	Local	Structured	Yes	Yes	Yes	No	Yes	No
AutoPrognosis	Free	Coding required	Local	Structured	Yes	Yes	Yes	Yes	Yes	Yes
Clarifai	Chargeable	None	Cloud	Unstructured	Yes	Yes	Yes	Yes	No	No
Google Cloud AutoML	Chargeable	None	Cloud	Structured/ Unstructured	Yes	Yes	Yes	Yes	No	No
DataRobot AutoML	Chargeable	None	Cloud	Structured	Yes	Yes	Yes	Yes	Yes	Yes
Dedicaid AutoML	Restricted to collaborators	None	External private server	Structured	Yes	Yes	Yes	Yes	No	No
H2O.ai R/Python Packages	Free	Coding required	Local/Cloud	Structured/ Unstructured	Yes	Yes	Yes	Yes	Yes	Yes
H2O.ai Driverless AI	Chargeable	None	Local/Cloud	Structured/ Unstructured	Yes	Yes	Yes	Yes	Yes	Yes

Table 1. Technological comparison of autoML platforms applied in the studies included in this review. (Cont'd)

AutoML Platform Name	Accessibility			Technical Features					Portability	
	Cost	Code requirement	Computing location	Dataset format	Feature extraction or selection	Model selection or training	Hyperparameter optimisation	Evaluation	Model exportability	Explainability
JADBio	Chargeable	None	Cloud	Structured	Yes	Yes	Yes	Yes	Yes	Yes
KNIME	Chargeable	None	Local	Structured	No	Yes	Yes	Yes	Yes	No
MATLAB	Chargeable	Coding required	Local	Structured/ Unstructured	Yes	Yes	Yes	Yes	Yes	No
MedicMind	Free	None	Cloud	Unstructured	Yes	Yes	Yes	Yes	Yes	Yes
Microsoft Azure AutoML	Chargeable	None	Cloud	Structured	Yes	Yes	Yes	Yes	No	No
MILO	Chargeable	None	Local	Structured	No	Yes	Yes	Yes	Yes	No
Neuro-T	Chargeable	None	Local	Unstructured	Yes	Yes	Yes	Yes	Yes	No
SIMON	Free	None	Local	Strcutured	Yes	Yes	Yes	Yes	Yes	No
Sony Prediction One	Chargeable	None	Local	Structured/ Unstructured	Yes	Yes	Yes	Yes	Yes	No
TPOT	Free	Coding required	Local	Structured	Yes	Yes	Yes	Yes	Yes	No
USDM	Free	Coding required	Local	Structured/ Unstructured	Yes	Yes	Yes	No	Yes	No

AER: approximated error reduction; ASSL: automated semi-supervised learning; AutoDAL: automated distributed active learning; AutoDC: automated data-centric processing; JADBio: Just-Add-Data Bio; KNIME: Konstanz Information Miner; MILO: Machine Intelligence Learning Optimizer; ML: machine learning; TPOT: Tree-based Pipeline Optimization Tool; USDM: uncertainty sampling with diversity maximization; WEKA: Waikato Environment for Knowledge Analysis

ment of conventional reporting standards—comparable to research regarding other AI technologies.[37] Future comparative studies should address the limitations discussed above to convince researchers, clinicians and policy makers that autoML platforms may be applied in lieu of bespoke modelling.[38]

When reporting AI algorithms tasked with a certain clinical job, it would be helpful to avoid ambiguity in terminology. We would suggest a complete restriction of the terms "automated machine learning", or "autoML" to those algorithms built with technology that automates some or all parts of the engineering process—all conventional ML models process data without human guidance, so description of these technologies as automated is redundant. Similar terms such as "automated artificial intelligence", "automated machine learning" and "automated deep learning" are also redundant in the context of bespoke computational models. A simple alternative term for conventional ML projects would be "automatic"—these systems may automate a particular task, but their development is not automated, the defining feature of autoML.

The reduced barrier to entry in terms of computational expertise and hardware requirements conferred by many autoML platforms makes them a powerful contributor to democratisation of AI technology—a far greater number of clinicians and scientists are capable of ML development through use of these platforms. AutoML could be an invaluable resource for teaching, as individuals can more rapidly develop hands-on experience, learn by trial-and-error, and thereby develop intuitive understanding of the capabilities and limitations of ML.[39] AutoML could also be applied in pilot studies, enabling clinicians with domain-specific expertise to explore possibilities for ML research—facilitating prioritisation of allocation of scarce resources such as access to graphics processing units and expert computer scientists.[6] Validated platforms may be applied more broadly, including in patient care. Moreover, autoML is well placed to respond to calls to inculcate data-centric AI as opposed to model-centric development; focusing effort on curating high-quality data, which limits development more often than code or model infrastructure.[7] Acceleration in this process may be facilitated by large language models as their emerging capability to leverage plugins will allow autoML to facilitate AI building itself to fulfil user-defined aims.[40]

Further work is indicated to improve validation of autoML platforms, either by allowing models to be exported, or by providing more comprehensive internal metrics. Other work should focus on improving the functionality of autoML, specifically on reducing the trade-offs currently implicit in selecting a platform with a given code intensity and computing locus. Using automation to reduce human error to optimise engineering and improve performance is one ideal—this has been demonstrated with structured data by AutoPrognosis.[27,41] Increased functionality of code-free platforms while retaining the customisability of code-intense solutions is another ideal—H2O.ai Driverless AI offers the same functionality as the H2O.ai R and Python packages, but with a code-free graphical user interface.[42-48] Alternatively, maximising accessibility by automating the whole engineering process may be desirable—Dedicaid is a platform requiring just data, with no customisable parameters, but has an "ethical compass" which flags inappropriate datasets.[49,50]

CONCLUSION

AutoML performance is often comparable to bespoke ML and human performance. Many autoML platforms have been developed in academia and industry, with variable strengths and limitations. AutoML may prove especially useful in pilot studies and education, but potential use cases include primary research and clinical deployment if platforms are rigorously validated.[6] Future autoML research must be more transparently reported, adhere to reporting guidelines and provide appropriate benchmarks for performance comparisons. Further autoML development should seek to minimise the trade-offs currently inherent in selecting any given platform.

Data availability statement

The raw data from this review may be provided upon request.

Disclosure

AJT is supported by The Royal College of Surgeons in Edinburgh (RCSED Bursary 2022), Royal College of Physicians (MSEB 2022), and Corpus Christi College, University of Cambridge (Gordon Award 1083874682). DSWT is supported by the National Medical Research Council, Singapore (NMCR/HSRG/0087/2018; MOH-000655-00; MOH-001014-00), Duke-NUS Medical School (Duke-NUS/RSF/2021/0018; 05/FY2020/EX/15-A58), and Agency for Science, Technology and Research (A20H4g2141; H20C6a0032). These funders were not involved in the conception, execution or reporting of this study. All authors declare no competing interests.

Supplementary materials

Appendix S1. Systematic review search strategy.
Appendix S2. Inclusion and exclusion criteria, as provided to researchers conducting abstract and full-text screening.
Appendix S3. Tabulated study characteristics including citation details and study identifiers used elsewhere in the article.
Appendix S4. Study-level data exhibiting fulfilment of DECIDE-AI reporting standards.
Appendix S5. Study-level data exhibiting appraisal of risk of bias (RoB) and concerns regarding applicability (CrA) using PROBAST.
Appendix S6. Forest plot depicting reported AUPRC metrics.

REFERENCES

1. Pianykh OS, Guitron S, Parke D, et al. Improving healthcare operations management with machine learning. Nat Mach Intell 2020;2:266-73.
2. Park JY, Hsu TC, Hu JR, et al. Predicting Sepsis Mortality in a Population-Based National Database: Machine Learning Approach. J Med Internet Res 2022;24:e29982.
3. Car J, Sheikh A, Wicks P, et al. Beyond the hype of big data and artificial intelligence: building foundations for knowledge and wisdom. BMC Med 2019;17:143.
4. Dash S, Shakyawar SK, Sharma M, et al. Big data in healthcare: management, analysis and future prospects. Journal of Big Data 2019;6:54.
5. Tan TF, Thirunavukarasu AJ, Jin L, et al. Artificial intelligence and digital health in global eye health: opportunities and challenges. Lancet Glob Health 2023;11:e1432-43.
6. Thirunavukarasu AJ, Elangovan K, Gutierrez L, et al. Democratizing Artificial Intelligence Imaging Analysis With Automated Machine Learning: Tutorial. Journal of Medical Internet Research 2023;25:e49949.
7. Khang A, Rana G, Tailor RK, et al. Data-Centric AI Solutions and Emerging Technologies in the Healthcare Ecosystem. 1st Ed. Boca Raton: CRC Press; 2023.
8. Hutter F, Kotthoff L, Vanschoren J (Eds). Automated Machine Learning: Methods, Systems, Challenges. The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing; 2018.
9. Rawat T, Khemchandani V. Feature Engineering (FE) Tools and Techniques for Better Classification Performance. IJIET 2017;8:169-79.
10. Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial Intelligence in Medicine 2020;104:101822.
11. Faes L, Wagner SK, Fu DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. Lancet Digit Health 2019;1:e232-42.
12. Thirunavukarasu A, Elangovan K, Gutierrez L, et al. Comparative analysis of diagnostic imaging models built with automated machine learning. Future Healthc J 20231;10(Suppl 3):21-3.
13. Thirunavukarasu A, Gutierrez L, Elangovan K, et al. The applications of automated machine learning in clinical contexts. PROSPERO 2022 CRD42022344427. https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42022344427. Accessed 7 January 2024.
14. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
15. Ouzzani M, Hammady H, Fedorowicz Z, et al. Rayyan—a web and mobile app for systematic reviews. Systematic Reviews 2016;5:210.
16. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022;377:e070904.
17. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019;170:51-8.
18. Shen H, Liu T, Cui J, et al. A web-based automated machine learning platform to analyze liquid biopsy data. Lab Chip 2020;20:2166-74.
19. Erickson BJ, Kitamura F. Magician's Corner: 9. Performance Metrics for Machine Learning Models. Radiol Artif Intell 2021;3:e200126.
20. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments. Research Synthesis Methods 2020;12:55-61.
21. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. Journal of Open Source Software 2019;4:1686.
22. Dayim A. forestploter. https://github.com/adayim/forestploter. Accessed 3 January 2023.
23. Cho BH, Kaji D, Cheung ZB, et al. Automated Measurement of Lumbar Lordosis on Radiographs Using Machine Learning and Computer Vision. Global Spine J 2020;10:611-8.
24. Adaszewski S, Dukart J, Kherif F, et al. How early can we predict Alzheimer's disease using computational anatomy?. Neurobiol Aging 2013;34:2815-26.
25. Smith R, Julian D, Dubin A. Deep neural networks are effective tools for assessing performance during surgical training. J Robot Surg 2022;16:559-62.
26. Korot E, Pontikos N, Liu X, et al. Predicting sex from retinal fundus photographs using automated deep learning. Sci Rep 2021;11:10286.
27. Alaa AM, van der Schaar M. Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. Sci Rep 2018;8:11242.
28. Bang CS, Lim H, Jeong HM, et al. Use of Endoscopic Images in the Prediction of Submucosal Invasion of Gastric Neoplasms: Automated Deep Learning Model Development and Usability Study. J Med Internet Res 2021;23:e25167.
29. Korot E, Guan Z, Ferraz D, et al. Code-free deep learning for multi-modality medical image classification. Nature Machine Intelligence 2021;3:288-98.
30. Yang HS, Kim KR, Kim S, et al. Deep Learning Application in Spinal Implant Identification. Spine 2021;46:E318-24.
31. Chen X, Wujek B. A Unified Framework for Automatic Distributed Active Learning. IEEE Trans Pattern Anal Mach Intell 2022;44:9974-86.
32. Abbas A, O'Byrne C, Fu DJ, et al. Evaluating an automated machine learning model that predicts visual acuity outcomes in patients with neovascular age-related macular

degeneration. Graefes Arch Clin Exp Ophthalmol 2022; 260:2461-73.

33. Ying X. An Overview of Overfitting and its Solutions. J Phys: Conf Ser 2019;1168:022022.
34. Cacciamani GE, Chu TN, Sanford DI, et al. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. Nat Med 2023;29:14-5.
35. Collins GS, Dhiman P, Navarro CLA, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 20211;11:e048008.
36. Navarro CLA, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021;375:n2281.
37. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689.
38. Thirunavukarasu AJ. How Can the Clinical Aptitude of AI Assistants Be Assayed? Journal of Medical Internet Research 2023;25:e51603.
39. Ng FYC, Thirunavukarasu AJ, Cheng H, et al. Artificial intelligence education: An evidence-based medicine approach for consumers, translators, and developers. CR Med 2023;4:101230.
40. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med 2023;29:1930-40.
41. Alaa AM, Bolton T, Di Angelantonio E, et al. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One 2019;14:e0213653.
42. Ikemura K, Bellin E, Yagi Y, et al. Using Automated Machine Learning to Predict the Mortality of Patients With COVID-19: Prediction Model Development Study. J Med Internet Res 2021;23:e23458.
43. Luna A, Bernanke J, Kim K, et al. Maturity of gray matter structures and white matter connectomes, and their relationship with psychiatric symptoms in youth. Hum Brain Mapp 2021;42:4568-79.
44. Mohsen F, Biswas MR, Ali H, et al. Customized and Automated Machine Learning-Based Models for Diabetes Type 2 Classification. Stud Health Technol Inform 2022; 295:517-20.
45. Narkhede SM, Luther L, Raugh IM, et al. Machine Learning Identifies Digital Phenotyping Measures Most Relevant to Negative Symptoms in Psychotic Disorders: Implications for Clinical Trials. Schizophr Bull 2022;48:425-36.
46. Sills MR, Ozkaynak M, Jang H. Predicting hospitalization of pediatric asthma patients in emergency departments using machine learning. Int J Med Inf 2021;151:104468.
47. Stojadinovic M, Milicevic B, Jankovic S. Improved predictive performance of prostate biopsy collaborative group risk calculator when based on automated machine learning. Comput Biol Med 2021;138:104903.
48. Yin M, Zhang R, Zhou Z, et al. Automated Machine Learning for the Early Prediction of the Severity of Acute Pancreatitis in Hospitals. Front Cell Infect Microbiol 2022;12:886935.
49. Ritter Z, Papp L, Zambo K, et al. Two-Year Event-Free Survival Prediction in DLBCL Patients Based on In Vivo Radiomics and Clinical Parameters. Front Oncol 2022;12:820136.
50. Hasimbegovic E, Papp L, Grahovac M, et al. A Sneak-Peek into the Physician's Brain: A Retrospective Machine Learning-Driven Investigation of Decision-Making in TAVR versus SAVR for Young High-Risk Patients with Severe Symptomatic Aortic Stenosis. J Pers Med 2021;11:1062.