


Brief Communications

Evaluating automated machine learning platforms for use in healthcare

Ian A. Scott , MHA^{*,1,2}, Keshia R. De Guzman, PhD^{3,4}, Nazanin Falconer, PhD^{3,4}, Stephen Canaris, BSc⁵, Oscar Bonilla, BSc⁵, Steven M. McPhail, PhD^{5,6}, Sven Marxen, BPharm (Hons)⁷, Aaron Van Garderen, BPharm (Hons)^{5,7}, Ahmad Abdel-Hafez, PhD^{5,6}, Michael Barras, PhD^{3,4}

¹Centre for Health Services Research, University of Queensland, Brisbane, 4102, Australia, ²Department of Internal Medicine and Clinical Epidemiology, Princess Alexandra Hospital, Brisbane, 4102, Australia, ³Department of Pharmacy, Princess Alexandra Hospital, Brisbane, 4102, Australia, ⁴School of Pharmacy, The University of Queensland, Brisbane, 4102, Australia, ⁵Digital Health and Informatics, Metro South Health, Brisbane, 4102, Australia, ⁶Australian Centre for Health Services Innovation and Centre for Healthcare Transformation, School of Public Health and Social Work, Queensland University of Technology, Brisbane, 4059, Australia, ⁷Pharmacy Service, Logan and Beaudesert Hospitals, Logan, 4131, Australia

*Corresponding author: Ian A. Scott, MHA, Centre for Health Services Research, University of Queensland, 20 Cornwall Street, Woolloongabba, QLD 4102, Australia (i.scott@uq.edu.au)

Abstract

Objective: To describe development and application of a checklist of criteria for selecting an automated machine learning (Auto ML) platform for use in creating clinical ML models.

Materials and Methods: Evaluation criteria for selecting an Auto ML platform suited to ML needs of a local health district were developed in 3 steps: (1) identification of key requirements, (2) a market scan, and (3) an assessment process with desired outcomes.

Results: The final checklist comprising 21 functional and 6 non-functional criteria was applied to vendor submissions in selecting a platform for creating a ML heparin dosing model as a use case.

Discussion: A team of clinicians, data scientists, and key stakeholders developed a checklist which can be adapted to ML needs of healthcare organizations, the use case providing a relevant example.

Conclusion: An evaluative checklist was developed for selecting Auto ML platforms which requires validation in larger multi-site studies.

Lay Summary

Machine learning (ML) is a form of artificial intelligence whereby computers learn associations within large complex datasets and encode these into a statistical model that can then be applied to new datasets in generating predictions or classifications. In healthcare, such models can assist clinicians in making diagnostic, therapeutic, and prognostic decisions. However, developing and testing such models for different use cases take time and effort on the part of data scientists, collaborating clinicians, and informatics teams who may not have extensive data and model processing capacity. Auto ML platforms are designed to rapidly build and validate ML models by automating complex, time-consuming tasks involved in data processing and model training. Numerous Auto ML platforms now available from both open source and commercial vendors necessitate guidance in how to choose the one most appropriate to organizational needs. Using systematic methods, a multidisciplinary team formulated an evaluation checklist for objectively appraising different Auto ML platforms in making a final selection. The checklist was assessed for its utility by its application to a practical use case of a dosing model for an intravenous antithrombotic with unpredictable therapeutic effects. The checklist may prove useful to other users and can accommodate organizational procurement requirements.

Key words: machine learning; automated; artificial intelligence.

Introduction

Machine learning (ML) is a branch of artificial intelligence whereby computers “learn” patterns or associations, often within large complex data sets, and apply this knowledge to new datasets in generating predictions or classifications.^{1,2} The increasing availability of digital health data presents an opportunity to use ML platforms capable of training and testing predictive models that can optimize healthcare.^{3,4} In

hospital settings, the increasing adoption of electronic health records (EHRs), also referred to as electronic medical records, has contributed to the availability of large datasets to which ML can be applied.⁵ ML models can potentially assist with diagnosis or prognosis of clinical conditions,^{6,7} determine optimal treatment pathways and medication dosing,⁸ improve healthcare efficiency,^{9,10} and identify patients at risk of adverse outcomes.^{11–13} As the science of ML further

Received: December 21, 2023; Revised: March 6, 2024; Editorial Decision: March 29, 2024; Accepted: April 22, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the American Medical Informatics Association.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

evolves, and more EHR data becomes available, health organizations need user friendly ML platforms that can efficiently develop and validate ML models.¹⁴ This need has instigated the development of automated ML or “Auto ML” technologies.

Auto ML is a sub-field of ML which has been defined as “the intersection of machine learning and automation.”¹⁴ Auto ML platforms have the potential to rapidly build and validate ML models and reduce demand on data scientists who currently spend up to 70% of their time on the model development process.^{14,15} By using Auto ML, a wide range of ML ensemble models can be simultaneously developed and tested for a specific task. Auto ML has emerged as a growing field to automatically select, compose, and parametrize ML models to achieve optimal performance for a given task or dataset.^{14,15} Auto ML platforms simplify complex and time-consuming modelling tasks including data preparation, feature engineering, model optimization, model training, and internal validation.^{15,16} Another driving force for Auto ML has been the rise of data democratization, the ongoing process of enabling all individuals, irrespective of technical expertise, to work confidently and comfortably with data and to use it more efficiently and productively.^{17,18}

Data democratization and Auto ML also shift the development of ML models from coding and scripting languages to easy-to-use graphical interfaces or visual and interactive environments. This significantly reduces the required expertise and effort to build ML models.^{17,18} Automation also has the potential to reduce human error and bias, reinforce the replicability of the analyses, and promote collaboration between clinicians and data scientists. Auto ML has been previously used to develop ML models in medical imaging, disease diagnosis, and EHR data analysis.^{19–21} In a recent study, the authors demonstrated how Auto ML could be successfully applied to the development and validation of a model for dosing unfractionated heparin.²²

Auto ML platforms comprise commercially available programs such as DataRobot, H2O Driverless AI, Vertex AI, Azure AutoML, and Google AutoML, and open-source programs such as H2O AutoML,²³ Tree-based Pipeline Optimization Tool (TPOT),²⁴ AutoKeras,²⁵ general automated machine learning assistant (GAMA),²⁶ and Auto-Sklearn.^{27,28} As more platforms become available, determining which is best suited to the digital infrastructure and governance of a particular healthcare organization requires an objective, structured method of assessment of different platforms. In this article, we describe the development of a checklist that was applied to a set of written vendor submissions, and subsequent live demonstrations from short-listed vendors, in choosing an Auto ML platform for building and validating ML models within a large multi-hospital service. The target use case to which the chosen platform would be first applied was developing and validating a model for predicting the bolus and maintenance dosing of intravenous heparin in hospitalized adult patients presenting with acute thrombotic disorders, such as acute coronary syndrome or venous thromboembolism, which together account for around 5% of all hospital admissions.²² This anticoagulant drug has unpredictable pharmacokinetics and the current use by prescribing clinicians of weight-based dosing nomograms achieves therapeutic range in less than 50% of patients at 48 hours,²⁹ with many either under-dosed, and at risk of further thrombosis, or over-dosed, and at risk of bleeding. A more accurate ML model could allow clinicians to achieve therapeutic dosing more quickly in more patients, with less risk of complications.

Method

Using a 3-step methodology, a provisional list of selection criteria was developed (Figure 1) that reflected the needs and constraints of the digital environment of the Metro South Hospital and Health Services district in south-east Queensland, Australia. This district is comprised of 5 hospitals, including a quaternary public hospital, comprising a total of 1500 beds, managing over 300 000 acute presentations annually, employing 11 000 fulltime equivalent clinical staff and serving a population of 1.2 million.

Step 1. Identification of key requirements

In identifying key requirements of a preferred AutoML platform, a group of key stakeholders were selected comprising clinical informatics experts and data scientists from the district’s Digital Health and Informatics Directorate, clinicians from the disciplines of Medicine and Pharmacy, and all members of the multidisciplinary Metro South Clinical Artificial Intelligence Working Group. A literature scan was conducted in providing a synthesis of available evidence for automated platforms and which included recent review publications.^{14,16} This literature review, combined with stakeholder expertise, were used by the panel over a series of meetings to formulate a provisional list of key criteria for AutoML platform selection which were then formalized to align with the district’s digital technology procurement pathway. The criteria were divided into 21 functional and 6 non-functional criteria, where functional criteria were considered as universal processes that the platform should be able to perform (eg, record, calculate, display, publish), and non-functional criteria were considered as platform attributes perceived as pertinent to organizational and technical needs (eg, cyber security, easy to use). All criteria were rated as either mandatory, highly desirable, or desirable based on panel consensus. As the checklist was designed to assess utility of the Auto ML platforms, and as legal and confidentiality restrictions applicable to commercial vendors were outside the remit of the evaluators, license fees or procurement costs relevant to such submissions were not included as criteria.

Step 2. Market scan

A market scan was conducted by the clinical informatics experts using the Gartner Magic Quadrant for Data Science and Machine Learning Platform March 2021 (available at: <https://www.gartner.com/en/documents/3998753>) to identify potential vendors, yielding 20 contenders. We then developed a shortlist of vendors whose platforms met the following mandatory requirements set by local hospital regulations: (1) confirmed availability on a cloud-based server accessible within the Australian setting; (2) complete fulfilment of local health cybersecurity requirements; and (3) evidence of application to actual use cases, as verified by documentation on vendor websites and through face to face demonstration and question answering sessions with vendors at a data and analytics summit held in Sydney, Australia in August 2021 and attended by 1 of the authors (A.A.-H.). This resulted in a shortlist of 2 vendors, #9 and #12, (Table 1) who were then invited to submit a full application, using a structured template listing all the provisional criteria and requesting a detailed response, on how their platform satisfied each criterion. This was accompanied by a live, hands-on demonstration of the platform to the local assessment panel (see below).

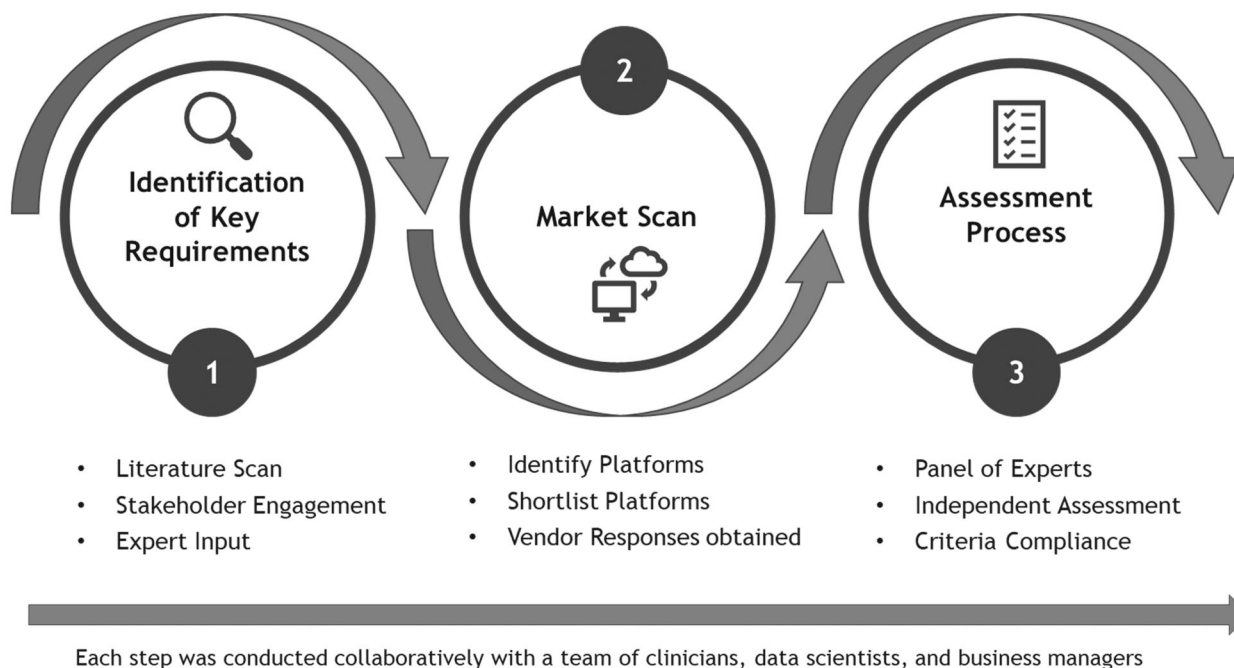


Figure 1. Key steps in developing criteria for an Auto ML checklist.

Table 1. Mandatory requirements used for short-listing Auto ML contenders.

Vendors	Confirmed availability on a cloud-based server accessible within the Australian setting	Complete fulfilment of local health cybersecurity requirements	Evidence of application to actual use cases	AutoML capabilities
1	Yes	Unknown	Unknown	No
2	No	Unknown	Unknown	No
3	Yes	Yes	Unknown	No
4	Yes	Yes	Unknown	No
5	No	Unknown	Yes	Yes
6	Yes	Yes	No	No
7	Yes	Yes	Yes	No
8	Yes	Yes	Yes	No
9	Yes	Yes	Yes	Yes
10	No	Unknown	No	No
11	Yes	Yes	Yes	No
12	Yes	Yes	Yes	Yes
13	Yes	Yes	No	No
14	No	Unknown	Yes	Yes
15	No	Unknown	Yes	Yes
16	Yes	Yes	Yes	No
17	No	Unknown	Yes	Yes
18	No	Unknown	Unknown	No
19	Yes	Yes	Unknown	No
20	Yes	Unknown	Unknown	No

Abbreviation: Unknown = sufficient information was unavailable to make a decision.

In cases where the stakeholder panel judged by consensus that responses were insufficient, additional information was requested.

Step 3. Assessment process

A panel of 4 experts, comprising the Director of Clinical Informatics, Director of Pharmacy, Director of Internal Medicine, and a senior data scientist experienced with ML as applied to healthcare was assembled. Three panelists independently assessed the responses received in writing from each of the shortlisted vendors against the listed criteria, and categorized each criterion response as fully compliant (C),

partially compliant (P), or non-compliant (N), based on a qualitative band and associated score, from 0 (unsatisfactory) to 10 (exceptional), reflecting the degree to which each response was concordant with stakeholder-defined descriptions for each criterion (Table 2). For each criterion, a final score was calculated by averaging the aggregated scores from the individual panel members. The maximum total score for the 21 functional and 6 non-functional criteria were 210 and 60, respectively, yielding a total maximum score for all criteria of 270. In response to comments and suggestions made by vendors in their responses, some modification to the wording of the criteria were made in producing the finalized list of

Table 2. Assessment process for determining vendor compliance with checklist criteria.

Compliance:		
■ Fully complies (C) ■ Partially complies (P) ■ Does not comply (N)		
Qualitative band	Score	Interpretation
Exceptional	10	Meets and exceeded requirements in <i>all respects</i> . Completely convincing and credible. Response demonstrates superior capability, capacity and experience relevant to, or understanding of, the requirements of the criterion. The Evaluation Panel is confident requirements are met to a very high standard. Low risk .
Excellent	8-9	Requirements are exceeded in <i>most key respects</i> and addressed to a very high standard in all others. Response demonstrates outstanding capability, capacity and experience relevant to, or understanding of, the requirements of the criterion. The Evaluation Panel is confident requirements are met to a high standard. Low risk .
Good	6-7	Requirements met to a <i>high standard in all respects</i> . Response demonstrates good capability, capacity and experience relevant to, or understanding of, the requirements of the criterion. The Evaluation Panel is confident requirements are met to a good standard. Low risk .
Adequate	5	Requirements addressed to a <i>consistent acceptable</i> standard. Response demonstrates acceptable capability, capacity and experience, relevant to, or understanding of, the requirements of the criterion. Some minor gaps or errors that can be easily corrected/overcome. The Evaluation Panel is reasonably confident requirements are met to a reasonable standard. Low to medium risk .
Limited	3-4	Requirements <i>poorly addressed or not fully met</i> . The platform demonstrates marginal capability, capacity and experience, relevant to, or understanding of, the requirements of the criterion. Some gaps, errors or weaknesses identified which are difficult to correct/overcome and make acceptable. The Evaluation Panel has some reservations whether vendor will be able to satisfactorily complete the requirements. Minor weaknesses/medium to high risk .
Inadequate	1-2	Requirements <i>not met or inadequately</i> dealt with in most or all respects. The platform does not demonstrate capability, capacity and experience, relevant to, or understanding of, the requirements of the criterion. Weaknesses or omissions identified which cannot be corrected/overcome to make them acceptable. The Evaluation Panel is not confident vendor will be able to satisfactorily complete the requirements. Major weaknesses/extreme risk .
Unsatisfactory	0	There is <i>insufficient</i> information to assess offer response. Offeror was not evaluated as it did not provide minimum level of requested information.

evaluative criteria. All 4 panelists subsequently attended the live demonstration and then participated in a group meeting to discuss vendor submissions, at which time the 3 scoring panelists could, if they felt necessary, adjust their scores.

Results

The finalized list of 21 functional and 6 non-functional criteria, with explanations, are shown in Table 3, including a description of how each criterion applied to the heparin dosing model use case.²² Of the total 27 criteria, 14 were considered “mandatory” to meet the core health service needs, 11 as “highly desirable” and 2 as “desirable.”

The scores assigned to each criterion by the 3 scoring panelists for each short-listed vendor submission are listed in Table 4. These scores showed little inter-rater variation for both vendors. All items attracted a score of 5 or above (ie, fully or partially compliant), although those with lower scores of 5 or 6 related to data cleansing and transformation capabilities, ability to perform multiple ML functions or natural language processing, and the ability to export developed models outside of the Auto ML platform. Among the 2 vendor submissions, the 1 with the highest score—average of the aggregated panel score of 226.34 out of total possible score of 270—was used to select the AutoML platform subsequently used for the use case.

Discussion

To our knowledge, this is the first attempt to formulate a list of criteria for selecting an Auto ML platform best suited to meet local organizational requirements on the basis of input from a diverse multidisciplinary team of clinicians, data scientists, and key stakeholders. We consider our criteria to be comprehensive, transparent and objective, and the checklist provides a structured and fair approach to selecting the most suitable platform. Importantly, while both short-listed platforms were from commercial vendors, license fees and other potential costs were not listed as criteria in our checklist as the focus was on assessing utility.

Several limitations of Auto ML methods have been recognized in past reports, which may partly explain why Auto ML has had limited application to healthcare to date.³⁰ These include lack of high quality data from EHRs, different EHR systems that are not interoperable, lack of transparency in black-box AutoML systems, complexity in their establishment and maintenance, inability to handle very large datasets, and limited customization and domain expertise for specialized use cases.^{20,31} Users of Auto ML platforms, just as much as ML model developers, need a clear understanding of the data and potential for errors in modelling, and be aware that AutoML tools usually have tailored or prespecified settings which complicates any attempt to standardize them with other modelling approaches or platforms. Auto ML websites may not provide in-depth explanations of how to apply their platforms or to interpret outputs, requiring some level of data science knowledge or oversight to ensure appropriate use.

However, Auto ML platforms compared to conventional ML techniques, do offer potential for time efficiency, greater accessibility and scalability, and reduced costs. These platforms continue to evolve with the aim of overcoming these limitations,^{32,33} with recent studies showing Auto ML tools capable of producing models superior in their performance to those developed using more traditional methods.^{34–36} Future research will likely investigate other applications of Auto ML in routine care in terms of feasibility and impact. Given these challenges, we believe a checklist of clearly defined criteria,

Table 3. Description of criteria, rating categories, and application to the heparin dosing model.

Functional criteria	Category	Application to heparin dosing model
1. Support the handling of multiple data sources (ie, structured, unstructured)	Mandatory	Comma-separated values (CSV) files or a database table of pre-processed data was extracted from EHR system.
2. Perform feature engineering on imported data sets	Mandatory	Feature engineering was performed on the original dataset; however, extra features discovery is desirable for modelling purposes.
3. Data cleansing and data transformation capabilities	Highly desirable	Data transformation was necessary for the regression model with mixture of Yeo-Johnson and min-max transformers needing to be applied.
4. Conduct supervised learning	Mandatory	Supervised learning models were required to predict therapeutic response to heparin therapy measured by activated partial thromboplastin time (aPTT).
5. Apply ensemble models or blend deep learning models with rule-based models	Mandatory	Ensemble models were desired to combine different models to improve prediction accuracy given the large dataset from 4 hospitals over 2 years. 1126 alternative models were trained, including LightGBM, XG Boost model, and ensemble models; the chosen model was an ensemble of 4x LightGMB models that were linearly blended.
6. Apply multiple models on given data sets to determine best fit-for-purpose	Mandatory	This was required as the best performing ML model was not knowable at the outset.
7. Update libraries used in developing data models	Mandatory	Continuous improvement and updating of ML models important for retraining model into the future.
8. Customize for additional tuning/optimization of developed models	Mandatory	Enabling ML model customization to local needs was a highly desirable feature in conferring transparency in the model-building process and confidence in the final model. Customization involved feature selection with input from clinicians with domain knowledge.
9. Conduct unsupervised, reinforcement and deep learning	Highly desirable	Deep learning models were also tested to predict the outcome.
10. Perform natural language processing on given data sets	Desirable	Considered desirable for further model refinement using unstructured, free-text clinical notes from EMR, although not required for heparin dosing model which relied on structured and tabulated data.
11. Apply different metrics when developing a model	Mandatory	Need to measure MAE, RMSE, and R ² for regression models, with minimization of RMSE chosen for model optimization in predicting aPTT.
12. Generate documentation detailing model findings with exportability in MS Word or PDF format	Mandatory	Automated documentation facilitates understanding of what AutoML tool has done to build the model.
13. Have a graphing function and plotting performance of a developed model	Highly desirable	Different graphs required to provide initial data analysis which influenced format of model outputs. In heparin use case, data visualization included clear heatmaps, plots and residual graphs relevant to both initial feature selection and subsequent model aPTT predictions.
14. Allow for multiple models to be deployed into production	Mandatory	Regression model for predicting aPTT as a continuous variable and classification model for categorizing aPTT as therapeutic, supra- or sub-therapeutic range required for the heparin model.
15. Scoring pipelines to make predictions on newly acquired data	Mandatory	Scoring pipelines required to test the model on an external dataset (from a different hospital) for further model validation. For use case, regression and classification models were built and deployed, with scoring pipelines used to test models on new data.
16. Export developed models for use outside the Auto ML platform	Highly desirable	Future prospective validation study using live EHR data required model to be extracted from the AutoML tool, located on local server, as separate files or have access to the Python code generated by the AutoML platform to build the model as a stand-alone tool able to be connected to the EHR reporting platform.
17. Encompass model monitoring capability to identify model drift	Highly desirable	Auto ML platform needed to accommodate monitoring and recalibration of the heparin model following deployment.
18. Have action logs and the ability to audit historic user modelling/system interactions	Mandatory	This attribute considered necessary for future implementation studies of the EHR-integrated model although not directly applicable to initial development and validation of the heparin dosing model.
19. Provide explanation for predictions made through scoring pipelines	Mandatory	Ability of the platform to indicate to clinicians most important features contributing to predictions.
	Highly desirable	In heparin use case, model predicted aPTT based on the heparin dosing inputs, with an optimization method used to enable

(continued)

Table 3. (continued)

Functional criteria	Category	Application to heparin dosing model
20. Conduct optimization process to the input variables to achieve specific outcome		the model to predict most appropriate heparin dose for individual patients in achieving a target aPTT value.
21. Customizable benefits calculation for a developed model	Desirable	This attribute considered necessary for future implementation studies of the EHR-integrated model although not applicable to initial development and validation of the heparin dosing model.
Non-functional criteria	Category	
22. The vendor provides business hours support 08.00–16.30 AEST	Highly desirable	
23. Consultation based days for teaching and improving in-house capability an option in vendor support contract	Highly desirable	
24. On call support capability an option in vendor support contract	Highly desirable	
25. User-friendly user interface	Mandatory	
26. Easy to navigate	Highly desirable	
27. Can be hosted on a cloud-based platform	Highly desirable	

Abbreviations: aPTT = activated partial thromboplastin time; EHR = electronic health record; MAE = mean absolute error; RMSE = root mean square error; AEST = Australian Eastern Standard Time.

against which any existing or new Auto ML platform can be compared, is likely to be of benefit to both potential users and vendors.

Our methodology in formulating the checklist has some limitations, principally driven by time pressures, limited administrative resources, and small sample size of vendor contenders. Our multidisciplinary group of individuals were known to each other, had existing working relationships, and reached consensus by discussion rather than using formal Delphi or other anonymized consensus methods. However, this approach allowed robust conversations about particular issues with no one individual feeling their views were unable to be expressed or considered. The group comprised experienced clinicians, informaticians, and data scientists ensuring a range of clinical and organizational perspectives were represented, although there may be additional perspectives specific to other healthcare organizations. However, our checklist is designed to be adaptable to other healthcare contexts. The number of short-listed vendors was small, and our checklist requires validation in being applied to a larger sample of vendor submissions. Finally, cost estimates and organizational willingness to pay for commercial Auto ML platforms may pragmatically influence choice of platform, irrespective of checklist scores relating to utility.

There are strengths to our methodology. The criteria development process was informed by contemporary feedback in various formats from vendor representatives in refining the list of provisional criteria. We also employed the criteria using an actual and practical use case to demonstrate its relevance and applicability, centered on a drug with direct patient safety implications. Detailed descriptions of each criterion and quantitative scoring methods used to categorize compliance reflects an emphasis on rendering the methodology explicit and transparent. However, we were unable to test the reliability or validity of the scoring method more broadly across several different use cases, and further studies are required. Nonetheless, we consider the checklist to be a starting point for further refinement and additions as the field of Auto ML evolves.

AutoML platforms may not replace most data scientists but should assist such experts in completing their assignments more quickly and support the broadening range of potential applications of ML in healthcare. In addition, AutoML will help clinical professions with less experience with ML models to participate in model development and evaluation, in our case physicians and pharmacists. Our multidisciplinary process encouraged collaboration and expression of a diversity of views involving clinicians and data scientists, with shared decision-making around the choice of Auto ML platform. We contend this checklist will be useful for those needing to select an Auto ML platform for use in their own organization by virtue of its consideration of the processes that the platform must perform to meet specific digital technology needs of the organization and potential use cases.

Author contributions

Ian A. Scott conceptualized the study, undertook data analysis, assisted in writing the first draft and wrote the final draft. Keshia R. De Guzman and Nazanin Falconer undertook literature search and assisted in writing the first draft. Stephen Canaris, Oscar Bonilla, and Ahmad Abdel-Hafez undertook the market scan and collated the vendor submissions. Stephen Canaris, Oscar Bonilla, Ahmad Abdel-Hafez, Ian A. Scott, and Michael Barras developed the draft evaluation checklist and reviewed and rated vendor written submissions and in person demonstrations. Sven Marxen, Aaron Van Garderen, and Steven M. McPhail reviewed and commented on the checklist and assisted in providing further references. All authors critically reviewed the final manuscript before submission and assisted in revising the manuscript in response to reviewers' comments. All authors read and approved the final manuscript.

Funding

None declared.

Table 4. Scoring system for short-listed vendor submissions.

No.	Description	Weight	Vendor #12			Vendor #9				
			R1	R2	R3	Average	R1	R2	R3	Average
Functional requirements (FR)										
FR01	Will support handling of multiple data sources	Mandatory	9	8	9	8.67	9	8	9	8.67
FR02	Will perform feature engineering on imported data	Mandatory	9	9	9	9.00	8	8	8	8.00
FR03	Will have data cleansing and transformation capabilities in data preparation	Highly Desirable	5	5	5	5.00	7	7	7	7.00
FR04	Will have ability to conduct supervised learning	Mandatory	8	8	9	8.33	9	9	10	9.33
FR05	Will have ability to apply ensemble models or blend deep learning models with rule-based models	Mandatory	8	8	8	8.00	9	9	10	9.33
FR06	Will have ability to apply multiple algorithms on a given data sets to determine best fit-for-purpose	Mandatory	8	8	8	8.00	9	9	10	9.33
FR07	Will update libraries used in the development of data models	Mandatory	9	9	9	9.00	9	9	9	9.00
FR08	Where a model has been developed, will be able to further customize for additional tuning/optimization	Mandatory	9	8	8	8.33	9	9	10	9.33
FR09	Will have ability to conduct unsupervised, deep and reinforcement learning	Highly Desirable	5	6	6	5.67	7	8	6	7.00
FR10	Can perform natural language processing on given data sets	Desirable	5	6	6	5.67	5	6	6	5.67
FR11	Will have ability to apply different metrics when developing a model	Mandatory	9	8	10	9.00	9	8	10	9.00
FR12	Will generate documentation detailing model findings exportable in MS Word or PDF format	Mandatory	9	10	10	9.67	8	8	9	8.33
FR13	Has a graphing function, plotting performance of a developed model	Highly Desirable	8	8	8	8.00	9	9	9	9.00
FR14	Will allow for multiple models to be deployed into production	Mandatory	8	9	9	8.67	8	8	8	8.00
FR15	Has scoring pipelines to make predictions on newly acquired data	Mandatory	9	8	9	8.67	9	8	9	8.67
FR16	Can export developed models for use outside the system	Highly Desirable	6	6	6	6.00	9	8	8	8.33
FR17	Encompasses model monitoring capability to identify model drift	Highly Desirable	8	8	9	8.33	8	8	9	8.33
FR18	Has action logs and can audit historic user modelling/system interactions	Mandatory	9	8	9	8.67	9	8	9	8.67
FR19	Shall provide explanation for predictions through scoring pipelines	Mandatory	9	8	9	8.67	9	8	9	8.67
FR20	Can conduct model optimization within the tool	Highly Desirable	8	8	9	8.33	8	8	9	8.33
FR21	Has customizable benefits calculation for a developed model	Desirable	9	8	9	8.67	9	8	9	8.67
Total Score—FR						168.33				176.67
Non-functional requirements (NFR)										
NFR01	Vendor shall provide business hours support covering 0.800–16.30 AEST	Highly Desirable	10	8	8	8.67	7	7	7	7.00
NFR02	Consultation days for improving in-house capability is available within vendor support agreement	Highly Desirable	9	8	9	8.67	9	8	9	8.67
NFR03	On call support capability is available within vendor support agreement	Highly Desirable	9	8	9	8.67	9	8	9	8.67
NFR04	Has a user-friendly user interface	Mandatory	9	9	9	9.00	8	8	8	8.00
NFR05	Easy to navigate	Highly Desirable	9	8	9	8.67	9	8	9	8.67
NFR06	Can be hosted on a cloud-based platform e.g. Azure, AWS, GCP	Highly Desirable	9	8	9	8.67	9	8	9	8.67
Total score—NFR						52.33				49.67
Combined requirements score						220.66				226.34

Abbreviation: R = reviewer.

Conflicts of interest

None declared.

Data availability

The data underlying this article are available in the article.

References

- Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216-1219. <https://doi.org/10.1056/nejmp1606181>
- Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*. 2007;2:59-77.
- Ghassemi M, Naumann T, Schulam P, et al. A review of challenges and opportunities in machine learning for health. *AMIA Jt Summits Transl Sci Proc* 2020. 2020:191-200. 20200530.
- Mitchell TM. *Machine Learning*. McGraw-Hill; 1997.
- Baumann LA, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: a systematic review. *Health Policy*. 2018;122(8):827-836. <https://doi.org/10.1016/j.healthpol.2018.05.014>
- Wong J, Murray Horwitz M, Zhou L, et al. Using machine learning to identify health outcomes from electronic health record data. *Curr Epidemiol Rep*. 2018;5(4):331-342. <https://doi.org/10.1007/s40471-018-0165-9>
- Golas SB, Shibahara T, Agboola S, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC Med Inform Decis Mak*. 2018;18(1):44. <https://doi.org/10.1186/s12911-018-0620-z>
- Falconer N, Abdel-Hafez A, Scott IA, et al. Systematic review of machine learning models for personalised dosing of heparin. *Br J Clin Pharmacol*. 2021;87(11):4124-4139. <https://doi.org/10.1111/bcp.14852>
- White NM, Carter HE, Kularatna S, et al. Evaluating the costs and consequences of computerized clinical decision support systems in hospitals: a scoping review and recommendations for future practice. *J Am Med Inform Assoc*. 2023;30(6):1205-1218. <https://doi.org/10.1093/jamia/ocad040>
- Mohammed Selim S, Kularatna S, Carter HE, et al. Digital health solutions for reducing the impact of non-attendance: a scoping review. *Health Policy Technol*. 2023;12(2):100759. <https://doi.org/10.1016/j.hlpt.2023.100759>
- Desautels T, Calvert J, Hoffman J, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform*. 2016;4(3):e28. <https://doi.org/10.2196/medinform.5909>
- Morgan DJ, Bame B, Zimand P, et al. Assessment of machine learning vs standard prediction rules for predicting hospital readmissions. *JAMA Netw Open*. 2019;2(3):e190348. <https://doi.org/10.1001/jamanetworkopen.2019.0348>
- Parsons R, Blythe RD, Cramb SM, et al. Inpatient fall prediction models: a scoping review. *Gerontology*. 2023;69(1):14-29. <https://doi.org/10.1159/000525727>
- Yao Q, Wang M, Escalante HJ, et al. 2018. Taking human out of learning applications: a survey on automated machine learning. arXiv, arXiv:1810.13306v1 [cs.AI].
- He X, Zhao K, Chu X. AutoML: a survey of the state-of-the-art. *Knowl-Based Syst*. 2021;212:106622. <https://doi.org/10.1016/j.knosys.2020.106622>
- Waring J, Lindvall C, Umeton R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *Artif Intell Med*. 2020;104:101822. <https://doi.org/10.1016/j.artmed.2020.101822>
- Shang Z, Zraggen E, Buratti B, et al. Democratizing data science through interactive curation of ML pipelines. In: *SIGMOD '19 Proceedings of the 2019 International Conference on Management of Data*. Amsterdam, Netherlands; June 30-July 5, 2019. <https://doi.org/10.1145/3299869.3319863>
- Zogaj F, Cambronero JP, Rinard MC, et al. Doing more with less. *Proc VLDB Endow*. 2021;14(11):2059-2072. <https://doi.org/10.14778/3476249.3476262>
- Frondelius T, Atkova I, Miettunen J, et al. Diagnostic and prognostic prediction models in ventilator-associated pneumonia: systematic review and meta-analysis of prediction modelling studies. *J Crit Care*. 2022;67:44-56. <https://doi.org/10.1016/j.jcrc.2021.10.001>
- Wang D, Li J, Sun Y, et al. A machine learning model for accurate prediction of sepsis in ICU patients. *Front Public Health*. 2021;9:754348-20211015. <https://doi.org/10.3389/fpubh.2021.754348>
- Li WT, Ma J, Shende N, et al. Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. *BMC Med Inform Decis Mak*. 2020;20(1):247. <https://doi.org/10.1186/s12911-020-01266-z>
- Abdel-Hafez A, Scott IA, Falconer N, et al. Predicting therapeutic response to unfractionated heparin therapy: machine learning approach. *Interact J Med Res*. 2022;11(2):e34533. <https://doi.org/10.2196/34533>
- LeDell E. H2O AutoML: scalable automatic machine learning. In: *Proceedings of the 7th ICML Workshop on Automated Machine Learning*. 2020. Accessed October 16, 2023. https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf
- Olson RS, Moore JH. TPOT: a tree-based pipeline optimization tool for automating machine learning. In: Hutter F, Kotthoff L, Vanschoren J, eds. *Automated Machine Learning: Methods, Systems, Challenges*. Springer International Publishing; 2019:151-160.
- Jin H, Song Q, Hu X. Auto-Keras: an efficient neural architecture search system. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery; 2019:1946-1956.
- Gijsbers P, Vanschoren J. GAMA: genetic automated machine learning assistant. *J Open Source Softw*. 2019;4(33):1132. <https://doi.org/10.21105/joss.01132>
- Mustafa A, Rahimi Azghadi M. Automated machine learning for healthcare and clinical notes analysis. *Computers*. 2021;10(2):24. <https://doi.org/10.3390/computers10020024>
- Feurer M, Klein A, Eggenberger K. Auto-sklearn: efficient and robust automated machine learning. In: Hutter F, Kotthoff L, Vanschoren J, eds. *Automated Machine Learning: Methods, Systems, Challenges*. Springer International Publishing; 2019:113-134.
- Cuker A. Unfractionated heparin for the treatment of venous thromboembolism: best practices and areas of uncertainty. *Semin Thromb Hemost*. 2012;38(6):593-599.
- Luo G. A review of automatic selection methods for machine learning models and hyper-parameter values. *Netw Model Anal Health Inform Bioinform*. 2016;5(1):18. <https://doi.org/10.1007/s13721-016-0125-6>
- Karmaker SB, Hassan M, Smith MJ, et al. AutoML to date and beyond: challenges and opportunities. *ACM Comput Surv*. 2021;54(8):175.
- Dinu V, Nadkarni P. Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. *Int J Med Inform*. 2007;76(11-12):769-779. <https://doi.org/10.1016/j.ijmedinf.2006.09.023>
- Luo G. Predicit-ML: a tool for automating machine learning model building with big clinical data. *Health Inf Sci Syst*. 2016;4:5. <https://doi.org/10.1186/s13755-016-0018-1>
- Paladino LM, Hughes A, Perera A, et al. Evaluating the performance of automated machine learning (AutoML) tools for heart disease diagnosis and prediction. *AI*. 2023;4(4):1036-1058.
- Imrie F, Cebere B, McKinney EF, van der Schaar M. 2022. AutoPrognosis 2.0: democratizing diagnostic and prognostic modelling in healthcare with automated machine learning. arXiv, arXiv:2210.12090 v1 [cs.LG].
- Liu G, Lu D, Lu J. Pharm-AutoML; an open-source, end-to-end automated machine learning package for clinical outcome prediction. *CPT Pharmacomet Syst Pharmacol*. 2021;10(5):478-488.

© The Author(s) 2024. Published by Oxford University Press on behalf of the American Medical Informatics Association.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

JAMIA Open, 2024, 7, 1–8

<https://doi.org/10.1093/jamiaopen/ooae031>

Brief Communications