WOMEN'S HEALTH

Women's Health
Volume 20: 1–12

DOI: 10.1177/17455057241248121
journals.sagepub.com/home/whe
Sage

Assessing the Utility of artificial intelligence in endometriosis: Promises and pitfalls

Brie Dungate[1,2,3], Dwayne R Tucker[2,3,4], Emma Goodwin[2,3] and Paul J Yong[2,3,4]

Abstract

Endometriosis, a chronic condition characterized by the growth of endometrial-like tissue outside of the uterus, poses substantial challenges in terms of diagnosis and treatment. Artificial intelligence (AI) has emerged as a promising tool in the field of medicine, offering opportunities to address the complexities of endometriosis. This review explores the current landscape of endometriosis diagnosis and treatment, highlighting the potential of AI to alleviate some of the associated burdens and underscoring common pitfalls and challenges when employing AI algorithms in this context. Women's health research in endometriosis has suffered from underfunding, leading to limitations in diagnosis, classification, and treatment approaches. The heterogeneity of symptoms in patients with endometriosis has further complicated efforts to address this condition. New, powerful methods of analysis have the potential to uncover previously unidentified patterns in data relating to endometriosis. AI, a collection of algorithms replicating human decision-making in data analysis, has been increasingly adopted in medical research, including endometriosis studies. While AI offers the ability to identify novel patterns in data and analyze large datasets, its effectiveness hinges on data quality and quantity and the expertise of those implementing the algorithms. Current applications of AI in endometriosis range from diagnostic tools for ultrasound imaging to predicting treatment success. These applications show promise in reducing diagnostic delays, healthcare costs, and providing patients with more treatment options, improving their quality of life. AI holds significant potential in advancing the diagnosis and treatment of endometriosis, but it must be applied carefully and transparently to avoid pitfalls and ensure reproducibility. This review calls for increased scrutiny and accountability in AI research. Addressing these challenges can lead to more effective AI-driven solutions for endometriosis and other complex medical conditions.

Keywords

artificial intelligence, deep learning, endometriosis, machine learning, pelvic pain, research methods

Date received: 29 September 2023; revised: 29 January 2024; accepted: 29 March 2024

Introduction

Endometriosis is a chronic condition which affects up to 10% of women and an unknown proportion of gender-diverse individuals.[1] The delay in diagnosis is on average 8–12 years, and as a result, many individuals endure long spans of debilitating pain and other associated symptoms without receiving the proper medical care. Endometriosis, like other health conditions that predominantly impact individuals who are assigned female at birth, is understudied leading to knowledge gaps and limited treatment options.[2] As researchers attempt to understand this complex, high-burden disease, new methodologies and advancing technologies provide an avenue for discovery.

In the past 10 years, with advancing technological capabilities, we have seen a rise in the popularity of artificial

[1]Faculty of Medicine, The University of British Columbia, Vancouver, BC, Canada
[2]Department of Obstetrics and Gynecology, The University of British Columbia, Vancouver, BC, Canada
[3]Women's Health Research Institute, Vancouver, BC, Canada
[4]Centre for Pelvic Pain & Endometriosis, BC Women's Hospital & Health Centre, Vancouver, BC, Canada

Corresponding author:
Paul J Yong, Centre for Pelvic Pain & Endometriosis, BC Women's Hospital & Health Centre, F2-4500 Oak Street, Vancouver, BC V6H 3N1, Canada.
Email: paul.yong@vch.ca

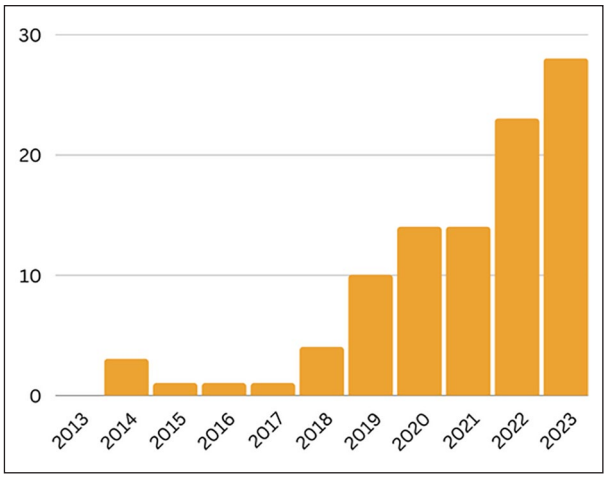

Figure 1. AI studies in endometriosis. Number of studies per year on PubMed from 2013 to 2023 using search terms ("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network") AND ("endometriosis" OR "pelvic pain" OR "endometrioma") as of 21 September 2023.

intelligence (AI), including in endometriosis research (Figure 1). AI is currently being used to aid in drug development, assist physicians in analyzing diagnostic imaging studies, and support the discovery of new genetic variants that underly diseases using precision medicine, among other applications.[3] While this method of analysis has vast capabilities in identifying previously unseen patterns in data and facilitating the analysis of large quantities of data, there are limits to the power of AI. Limitations associated with using AI predominantly revolve around the quality and quantity of data, and the expertise required to implement these algorithms.[3–5] The objective of this review is to describe the current landscape of endometriosis diagnosis and treatment and the possibilities that AI may hold to alleviate some of the burdens associated with this disease; additionally, common pitfalls when using AI algorithms will be noted. If researchers and clinicians implement AI appropriately, we hypothesize that advances in the clinical understanding and treatment of endometriosis will be immense.

Endometriosis research landscape

Research regarding conditions that affect women is disproportionately underfunded; endometriosis being no exception. Etiology, symptomology, comorbidities, and progression of this disease are poorly defined despite its effects on 1 in 10 women.[1] Delays in diagnosis increase the burden on those struggling to combat this disease and increase healthcare costs by thousands of dollars per person[6]; this is in addition to the many other tangible and intangible costs associated with endometriosis, including infertility, decreased quality of life, individual financial burden, reduced productivity, psychological trauma, and other impacts.[7]

For a common disease with a 10% prevalence, there are still significant limitations regarding diagnosis, classification, and treatment options for endometriosis.[2,7,8] Between 1973 and 2021, 22 methodologies for classification, staging, and reporting have been published exemplifying the continued need for convergence on a unifying diagnostic and treatment process; the breadth of metrics has varying utility in terms of clinical implications and patient outcomes, demonstrating the lack of a comprehensive understanding of this condition.[8]

What is AI?

The term "AI" refers to an extensive set of algorithms, encompassing machine learning (ML) and neural networks, and containing much overlap with statistical methods (Figure 2). To understand major pitfalls in the landscape of AI utilization in a clinical setting, it is important to have a common vocabulary to use and a consensus as to what is meant when terms such as "ML" are used.

AI pertains to a collection of algorithms designed to replicate human decision-making processes.[9] Within the realm of AI, there are two extreme ends; one end involves probabilistic models centered around a single variable, which arrives at decisions using predefined probabilities (e.g. decision trees) (Figure 2). The other end highlights DL algorithms, such as neural networks. Deep neural models are not only capable of replicating human-like decision-making but also possess the ability to generate novel video content—this process is referred to as "generative DL" (Figure 2).[10] ML and AI are often used interchangeably, both in research and colloquially since ML is a subset of AI. **ML** refers to a set of algorithms that can learn from data without being given explicit instructions (Figures 2 and 3).[9] ML can be further stratified into three classes of algorithms: supervised learning, unsupervised learning, and reinforcement learning (Figure 3). Supervised learning is the most widely used division: using a labeled dataset, algorithms in this class learn relationships between features in the dataset and the target class (label) (i.e. identifying factors that predict a positive diagnosis in patient data of individuals known to have an endometriosis diagnosis and control patients without an endometriosis diagnosis)[9]; supervised learning algorithms include Random Forest, Support Vector Machine (SVM), and logistic regression.[11] Unsupervised learning uses an unlabeled dataset and can identify groupings within your dataset (i.e. stratifying patient data of a group of individuals with an endometriosis diagnosis to identify similar subsets of patients and the features that distinguish these groupings)[9]; unsupervised learning algorithms include K-means, density-based spatial clustering of applications with noise (DBSCAN), and hierarchical clustering. Reinforcement learning also uses an unlabeled

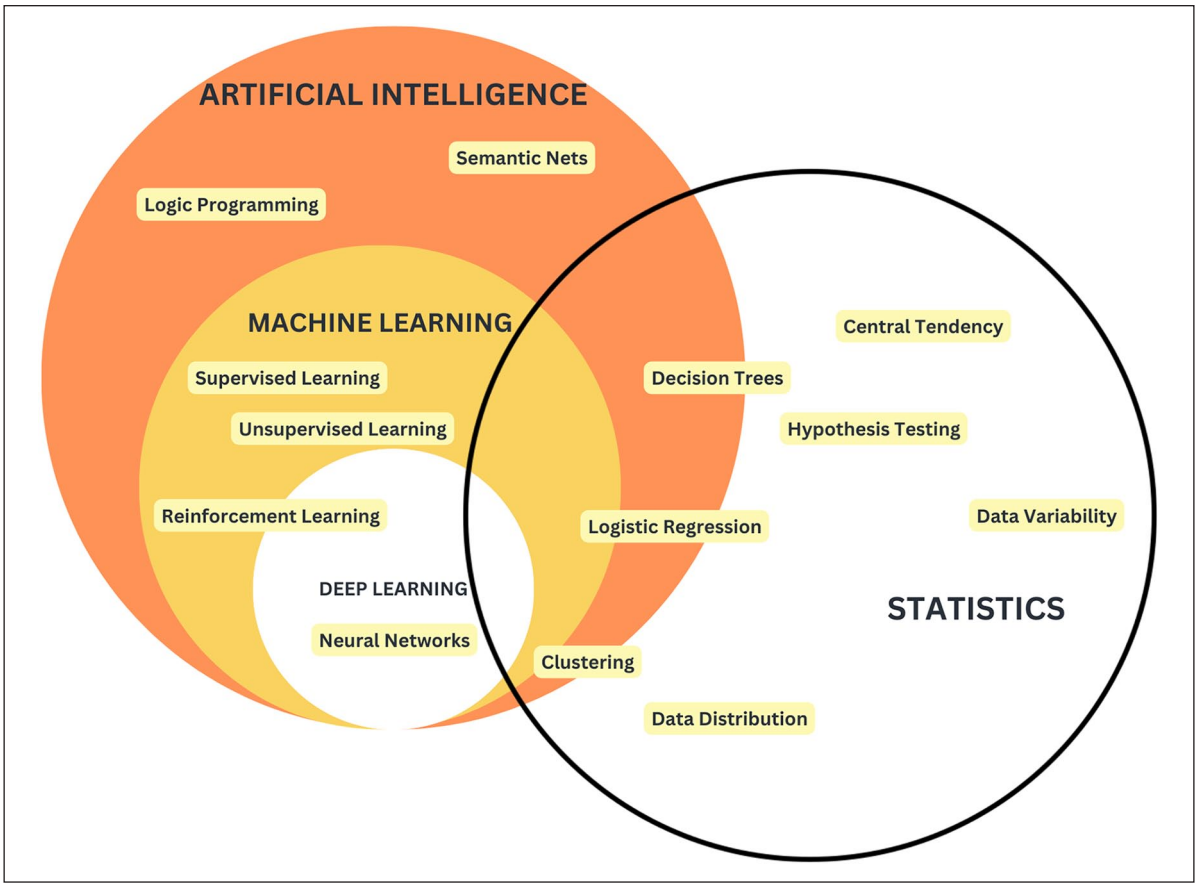

Figure 2. Venn diagram depicting the conceptual overlap between AI, ML, deep learning (DL), and statistics, with examples of algorithms within each category. Broadly, AI encompasses algorithms that aim to mimic human decision-making, with ML being a subset that learns patterns from data without explicit instructions. DL refers to neural-network-based algorithms. AI and statistics share a base in probability theory.

dataset, and this set of algorithms uses an iterative learning method to converge on an optimal behavior based on penalty and reward functions (i.e. predicting disease progression in endometriosis based on longitudinal symptom data)[9,12]; reinforcement learning algorithms include Q-learning, Hidden Markov Models, and Natural Language Processing. Unlike traditional statistical methods which require the user to give explicit instructions, this set of algorithms can learn and adapt once implemented, without continued user input.[9] Since ML algorithms learn directly from data, more data and better data translate to a higher-quality model. DL is a subset of ML which refers to algorithms that are based on a neural network architecture; neural networks are meant to replicate neurons in the human brain and consist of a series of hidden layers that transform input data into output data.[9,10]

The principle of **statistics** refers to the analysis of data to describe and make inferences about an underlying population. AI algorithms share commonalities with traditional statistical methods regarding data analysis, hypothesis testing, and particularly, concerning utilization of probability theory.[13,14] However, fundamental differences exist in the objectives, approaches, and the problems that these methods are designed to address. Notably, as AI algorithms grow in complexity, they combine probability with an exploration of the hypothesis space, a term that describes the set of all hypotheses which map an input to the desired output. In this context, the AI algorithm's objective shifts to identifying the optimal hypotheses that facilitate this mapping, serving as a predictive outcome rather than an inferential one. This leads to a key distinction between AI and statistics: statistics is model-driven, relying on the a priori assumptions, while AI is largely data-driven.[14] That is, AI algorithms rely on the scaffolding of a model, which learns patterns from data and can apply these patterns to make predictions or decisions, but its implementation does not rely on an a priori hypothesis of the relationship between features and the target variable. This flexibility also allows AI algorithms to generalize from data and perform tasks without strict a priori assumptions about data distribution or hypothesis testing, as commonly seen in statistical analysis.

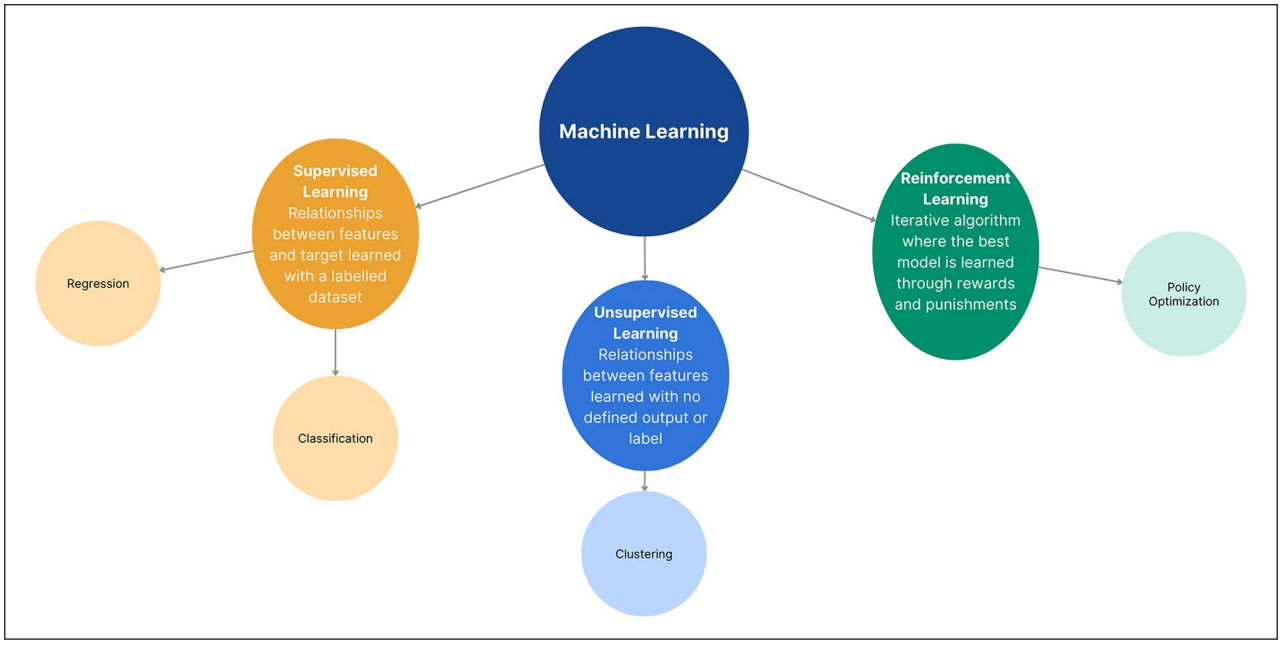

Figure 3. Subclassification schematic of ML. This schematic contains three subclasses of ML: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning refers to algorithms that detect relationships between features and a target within a labeled dataset, consisting of regression and classification algorithms. Unsupervised learning refers to algorithms that detect relationships that exist within a dataset with no defined label or pre-existing grouping, consisting of clustering algorithms. Reinforcement learning refers to algorithms that learn optimal models through an iterative process of learning through rewards and punishments, consisting of policy optimization algorithms.

Current guidelines and standards for AI usage in clinical research

The surge in popularity of AI in the medical field has necessitated the development of guidelines for use by governing bodies. Organizations that have guidelines on AI usage in healthcare include the World Health Organization (WHO) and the EQUATOR Network (Enhancing the QUAlity and Transparency Of health Research).[15,16] The WHO recognizes the immense capabilities of AI in addition to its potential for misuse and harm.[15] The WHO developed six guiding principles for their proposal: protecting human autonomy; promoting human well-being and safety and the public interest; ensuring transparency, explainability, and intelligibility; fostering responsibility and accountability; ensuring inclusiveness and equity; and promoting AI that is responsive and sustainable.[15]

The EQUATOR Network is taking the next steps and creating two tools, TRIPOD-AI (Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis) and PROBAST-AI (Prediction model Risk Of Bias ASsessment Tool), for individuals who are building prediction models using AI[16]; these tools are currently progressing through a five-stage development process which began in 2021, the goal being to produce a reporting guideline for key information regarding the application of AI in prediction models. This marks a key milestone in moving toward a research landscape where processes that contribute to research output are widely understood and research output is high quality and replicable. Having a set of clear and adherable guidelines allows both researchers and readers to have a unified understanding of the expectations regarding new research using AI.

Current applications of AI in endometriosis research (Figure 4)

In the clinical care of patients with endometriosis, diagnostic delays come at a great cost to patients and the healthcare system.[1,6] Diagnostic delays and misdiagnosis are caused by compounding issues of varied symptom presentation, limited awareness of endometriosis, and a culture of dismissing/diminishing women's pain.[17] Diagnostic imaging backlog contributes to delays in accessing care and misdiagnosis, as endometriosis patients' symptoms are then attributed to a variety of other physical and mental health conditions.[6,18] The Canadian Association of Radiologists' 2023 report for funding and expanded access to medical imaging includes the recommendation that AI be integrated into medical imaging infrastructure to address this diagnostic backlog.[18] The use of transfer learning and deep neural networks has allowed researchers to create models for diagnostic imaging.[19] Transfer learning uses models that are trained on large datasets of images that may or may not be related to the topic of interest; in

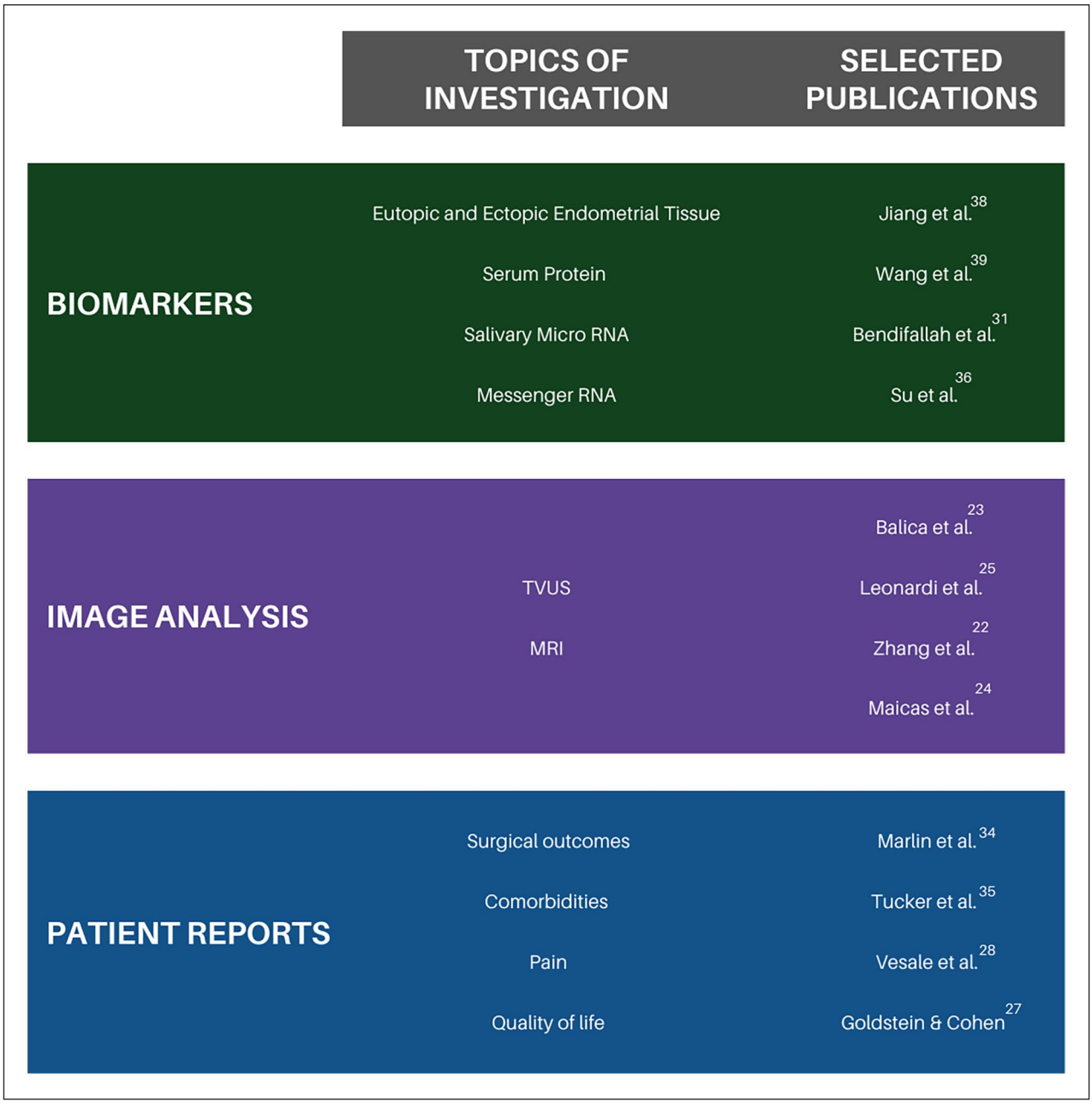

Figure 4. Overview of current research in endometriosis using AI. List of selected topics in the field of endometriosis and AI research, and selected publications to explore these topics further. For an exhaustive list of references, please refer to the "References" section.

the case of Maicas et al., a model was trained on Kinetics-400, a dataset of 306,245 videos of a wide range of human bodily actions. In transfer learning, it is best to use models that are pre-trained on a dataset that is as close as possible to the dataset of interest, in this case, transvaginal ultrasound (TVUS) videos.[19] The process of pre-training allows an algorithm to learn features from a large, labeled dataset (i.e. textures, movement, depth) and apply these general patterns to new datasets.[20] It is difficult to quantify the relationality between the Kinetics-400 dataset and the TVUS dataset since the model is a neural network and we cannot extract features used in the analysis. However, the performance reported by the researchers indicates that the model was able to provide high diagnostic performance (AUC = 0.965 and accuracy = 88.7%) for classifying Pouch of Douglas (POD) obliteration verified by assessment by two sonologists.[19]

Another benefit of transfer learning is the ability to use knowledge translation across imaging modalities. Both TVUS and magnetic resonance imaging (MRI) can be

used in the non-invasive diagnosis of endometriosis through the identification of POD obliteration, with TVUS typically providing an easier diagnosis.[21,22] However, access to clinicians with expertise concerning image-based endometriosis diagnosis is limited, and a single patient will not typically have both TVUS and MRI imaging studies done. Zhang et al. were able to leverage high-performance TVUS image analysis models (AUC = 96.9%), as a teacher to pre-train a similar model for use on MRI images, thereby increasing the model performance for identifying the same anomaly on MRI images from an AUC of 65.0% to an AUC of 90.6%.[22] Comparable studies looked at non-invasive diagnosis via TVUS or MRI, but required radiologists with > 10 years of experience reading gynecological imaging studies achieving an accuracy of 88% and 55% on TVUS and MRI, respectively.[21] A pilot study by Balica et al. presented DL models that reliably classify endometriosis patients retrospectively using ultrasound. The DenseNet Convolutional Neural Network (CNN) algorithm was the best-performing algorithm diagnosing endometriosis with an AUC of 90% and an accuracy of 80%.[23] In a study to predict rectosigmoid endometriosis using a training cohort of 222 and a testing cohort of 110, a neural network algorithm exhibited a 73% accuracy and an AUC of 0.82. This study used "soft" ultrasound markers, including age and non-bowel-related ultrasound variables as predictors, wherein the parameters and subsequent best models were primarily evaluated by accuracy.[24] Collectively, these studies validate the reliability of AI in detecting endometriosis using imaging. While the field may not be ready to fully transition to automated image analysis at this point, these results provide hope for the ability of AI models to approximate human performance. These studies, alongside others examining patient-reported measures, can act to guide screening of patients with high likelihood of endometriosis diagnosis and recommendations for further care or referral to specialists.[25,26,27]

The use of AI has extended to investigating other non-invasive correlates of laparoscopic diagnosis in endometriosis and indicators of surgical success in endometriosis lesion removal.[19,26–32] This research serves as a foundation for improving informed clinical decision-making between patients and clinicians. In the development of a screening tool based on self-reported symptoms to reduce time-to-diagnosis in endometriosis, Goldstein and Cohen found that extreme menstrual bleeding, irregular periods, and dysmenorrhea were the most important diagnostic predictors.[27] Using decision trees, Random Forest, Gradient Boosting Classifier (GBC), and Adaptive Boosting (AdaBoost), the study achieved excellent discrimination with an AUC ranging from 0.88 to 0.94.[27] However, these models lacked validation on data external to the model training process. In a study using similar predictors, model performance was evaluated with validation data using a collection of six ML algorithms; the models demonstrated excellent performances in distinguishing between endometriosis diagnosis and non-diagnosis; the best-performing algorithms for this study, Random Forest and a GBC achieved an AUC of 0.889 on the validation cohort.[33]

The prediction of endometriosis outcomes and surgical success faces challenges due to factors, such as the multifactorial nature of the disease, which impedes the ability to make generalizable predictions. Though this area of research is still evolving, including studies working on clinical prediction models for surgical success based on pain reduction and overall quality of life or health status (e.g. Creating a Clinical Prediction Model to predict Surgical Success in Endometriosis (CRESCENDO) study), AI has been a hopeful avenue for identifying reliable predictors.[34] ML, specifically least absolute shrinkage and selection operator (LASSO) regression, was employed to identify pelvic pain comorbidities associated with underlying central sensitization (PHQ-9 depression scores, abdominal wall pain, and pelvic floor myalgia) as important predictors of pain-related quality of life after endometriosis surgery. The analysis considered a variety of endometriosis-related factors, including revised American Society for Reproductive Medicine (rASRM) stage and residual endometriosis after surgery.[35] Multivariate logistic regression was used in the work by Vesale et al. to predict the occurrence of voiding dysfunction after surgical removal of deep endometriosis lesions. Both clinical characteristics and imaging were used to generate risk predictions in the model.[28] Studies using AI have provided a richer understanding of predictors of successful treatments and are part of a continual effort to change that standard and give patients autonomy in choosing their treatment path.[6,26–28,31]

As the power of AI is being harnessed for analysis of large datasets, genetic analyses, such as salivary microRNA (miRNA), are emerging as possible avenues of augmenting or bettering diagnostic processes.[30–32,36] While there is currently no blood test to diagnose endometriosis, biomarker discovery is one of the research priorities in endometriosis that is progressing with the help of AI.[17,37] Biomarkers specific to endometriosis may provide diagnostic and mechanistic information, and generate potential treatment targets; these biomarkers have been found in endometrial tissue, serum, and saliva.[31,36,38,39] Five genes (CXCL12, PDGFRL, AGTR1, PTGER3, and S1PR1) were identified in the work by Jiang et al. through feature importance in an SVM model and used to differentiate ectopic and eutopic endometrium samples in endometriosis patients. These genes were then analyzed as potential immune-mediated drug targets.[38] Using 10 binary classification algorithms, Su et al. found an mRNA signature containing nine genes to predict endometriosis diagnosis, including AGTR1 as found in the work by Jiang et al.; the best-performing model used LASSO regression (AUC = 0.791).[36] On a batch of 22 blood samples, serum

protein fingerprints were used in the work by Wang et al.[39] to assess the peripheral markers of endometriosis and found to have a sensitivity and specificity of 91.7% and 90.0%, respectively. In a study involving 200 patients, Bendifallah et al.[31] reported a salivary miRNA diagnostic signature that effectively predicts endometriosis (AUC = 0.98) using a Random Forest algorithm. These methods, pending further validation, could be a more accessible and time-efficient method to screen for and diagnose endometriosis, and provide insight into treatment targets and disease etiology.

Diagnosis is particularly important in the context of accessing care; the availability of healthcare providers experienced in recognizing and treating patients with endometriosis is highly variable across regions.[17] This problem of equitable care was further exacerbated during the COVID-19 pandemic when access to in-person care from medical specialists was restricted, especially when patients were required to travel for care.[40] Access to the proper specialists, pain medication, and support can depend on a confirmed diagnosis of endometriosis.[6] Thus, utilization of algorithms that can accurately detect endometriosis from non-invasive patient data/imaging, independent of healthcare provider training specific to endometriosis, has the potential to expedite diagnoses, avoid surgery, reduce healthcare costs, and give answers to women about the source of their pain.[6]

Common missteps using AI in medicine (Figure 5)

The main limitations to using AI in healthcare research are related to sample size, data quality, and non-representative samples.[5] With ML algorithms being data-driven, the quality of a model is dependent on the quantity and quality of data that the model is built on. Acquiring health data for conditions with low incidence or studies with low funding poses difficulties.[5,26] This can result in limited sample sizes, or samples that are not representative of the disease population, placing limitations on the types of algorithms that are applicable.[41] In supervised learning, it is standard practice to divide the dataset into a training and test set; the test set allows analysis of the performance of the model on "unseen" data, scoring metrics such as accuracy, precision, and recall describe the ability of the model to correctly identify data labels.[42] Battling with issues of data availability and algorithm complexity contributes to the problem of overfitting, particularly in the context of medical research.[42,43] Overfitting is when a model is fitted to a training dataset and rigidly adheres to nuances in the data.[43,44] Indications of overfitting include wide gaps in training and testing scores and high variance in scoring metrics. Gaps between training and testing scores indicate that the model is not generalizing well from the training set to the test set; similar discrepancies can occur if the dataset is too small to accurately characterize variance in the general population.[42–44]

Building models with non-representative samples or class imbalance, it is important to consider which scoring metric is reported; reported scoring metrics may not communicate the whole picture of the analysis.[45] Having an a priori understanding of which outcome is more "valuable" to predict (i.e. in a screening test, you want to identify as many potential cases as possible so you want high sensitivity) can help guide the choice of an appropriate metric to report, but it is important to contextualize these results in the scope of the problem you are investigating and ensure all necessary results are presented. Examples of potential metrics include precision, recall, area under the curve (AUC), and positive predictive value.[46] Hicks et al.[45] have proposed a set of five pitfalls found when analyzing sets of reported and omitted performance metrics for five studies implementing ML models; they also discuss reporting metrics to support clinician understanding of the contextual use of findings in these studies. While there is no hard and fast rule about the size of a dataset, generally the larger the dataset the better where a "rule of thumb" for a sample size of 10-fold the number of features is recommended, such that a model with 10 features requires a sample size of at least 100 samples.[47,48] In a scoping review of AI use in endometriosis studies, there are examples of ML being applied with 100 features and a sample size of < 100.[26] Other studies found in this review revealed analyses that had half of the sample being endometriosis patients, when the population prevalence of endometriosis is estimated at 5%–10%; such bias in samples introduces further issues in generalization to true population parameters. These models can still provide valuable clinical insight, but when looking at the deployment of such models in a "real world" setting, it is expected that the performance would decrease.

Strategies to mitigate overfitting when there is low data availability include cross-validation, penalizing model complexity, performing pre-analysis feature selection, and using ensemble or stacking methods.[43,49] Cross-validation follows the same premise as a train/test splitting of the data and allows reduction of the impact of noise in the training set; cross-validation further divides the training set into, for example, five segments within the training process and averages the score achieved on each of these segments to elicit a better picture of the impact of sampling within the dataset.[50] Penalizing model complexity and pre-analysis feature selection are both methods to reduce the complexity of the model by removing features from the dataset that have low predictive value for the model, such as features that have lots of missing data or features that have a low correlation with the measured outcome.[49,51] Ensemble and stacking methods combine or average multiple models to reduce variance, improve accuracy, or limit bias. While these methods are accessible and easy to implement, the lack of training regarding these methods is a roadblock to proper ML algorithm application.[4]

Figure 5. Summary table containing five examples of common missteps in AI utilization. Divided into categories of sample size, non-representative samples, model validation, data quality, and data contamination, this chart describes the issue addressed along with its impact on the performance of an AI model.

Another limitation is the insufficient training in data science, whereby ML algorithms may be applied without awareness of best practices for this type of data analysis.[52] This lack of training can lead to data contamination and errors in model validation or reporting.[46,51] The caveat of using powerful, black-box algorithms in research, such as AI, is that errors or missteps can occur without the individual implementing the algorithm understanding where these are happening. There are dangers associated with overanalyzing and claiming significance in results that do not exist due to the nature of these algorithms.[4] The higher complexity compared to traditional statistical methods can result in the application of AI algorithms without full awareness of the precise steps and contextual factors, which may affect reproducibility in research. Errors in ML can also be perpetuated without transparent reporting, despite published and in-progress guidelines.[53] Reproducibility becomes a central concern when the trial and error in the development of a model is not reported, leading to potential misrepresentations of the dependencies and other factors

affecting model performance.[54] Incremental tweaking of models throughout the building process after results have been assessed causes data contamination, violating a standard of ML, that data in the test set should not affect the training of the model.[51] It is important to report validation, test, or deployment metrics; this allows a researcher to comment on whether a model could be generalizable and limitations of the applicability of the model in practice.[42,46] Bias, similar to p-hacking associated with traditional statistical methods, can be introduced into the research process when using AI.[55] P-hacking refers to the process by which data analyses are skewed or selectively reported to favor statistically significant ($p < 0.05$) results.[55] In AI model development, researchers may try many different algorithms and evaluate model performance with multiple metrics, but report only results which support their claims. To mitigate these concerns, recommendations for reporting results for projects using AI include making both your code and your data available publicly.[56]

AI has quickly become a vital component of medical research, but the requirements to effectively implement this collection of methods and tools have lagged their implementation and publication.[3,52] Conclusions drawn from models built on small datasets and non-representative samples need further scrutiny.[57]

Applicable translation between research and clinical practice

Historically, we have seen a lag in the adoption of research into clinical care, including best treatment practices, medication choices, and risk factor identification. This lag is exacerbated by the separation between research and clinical practice and inefficiency in research coupled with increasing costs.[58,59] Utilization of available innovative technologies is crucial for the advancement of endometriosis diagnosis, treatment, and etiological understanding. Canada-wide and globally, there exists gaps in healthcare services, and with AI, the potential exists for augmentation and supplementation of clinical care services.[60,61] Organizations such as the Canadian Association for Drugs and Technologies in Healthcare (CADTH) and the National Institutes of Health (NIH) recognize the importance of adopting practices that integrate these new technological advances into clinical care.[60,61] The best practices of implementing AI into clinical care center on the idea of augmenting physician knowledge, practice, and expertise. In areas such as medical imaging, pathology, and genomics, AI has been able to expedite the analysis of scans, samples, and genomes. All of these can be useful in the diagnosis and treatment of endometriosis, in addition to help elucidate pathogenesis. Supporting individuals in training to adopt the use of AI, funding applicable necessary technology, and creating guidelines for implementing AI in a clinical setting are all crucial stepping stones to practically applying the available technology.

Limitations

This review approaches the topic of AI and endometriosis assessing general themes of potential future pathways and current missteps in application of this emerging technology. As this is a narrative review, a complete analysis of the field was not conducted. While growing, the scope of research specific to AI and endometriosis is narrow and the authors have attempted to summarize key findings. Continuing analysis of research coming out in the field of endometriosis research is pertinent to ensure the patency of changes to clinical care guidelines. It is yet unknown how the implementation of AI within endometriosis research will change the landscape of clinical care. However, following best practices, some of which are laid out in this review, will ensure the most beneficial outcome for patients and researchers alike.

Conclusion

In assessing the use of AI in endometriosis, the positive potential of adopting these algorithms into research and clinical practice can be amplified by understanding common pitfalls. AI algorithms provide a method of enhancing understanding of previously unidentified patterns in data. With increased access to analysis software, implementation of these algorithms can be adopted widely; this accessibility creates opportunities for investigating the many unknowns of endometriosis. While these tools have great power, proper implementation of AI algorithms must guide research endeavors.

Approaching the implementation of AI in endometriosis with caution and awareness of common pitfalls will improve the validity of research and clinical decision-making that relies on these models. Common pitfalls are associated with limited sample size, non-representative samples, model validation, data contamination, and data quality.[42,43,49,57] These challenges are driven by the cost and complexity of data acquisition and the need for more AI training among researchers. To harness the full potential of AI for application in endometriosis, researchers can defer to guidelines put forward by the EQUATOR network, the NIH, the WHO, and local governing bodies.[15,46,60,61]

AI algorithms have shown promise in supporting diagnostic imaging, etiological understanding, and surgical outcome prediction in endometriosis.[19,26,27,30] Adopting the use of technology with the potential to alleviate healthcare costs and improve patient outcomes is crucial for progress. A focus on data quality, transparency in reporting, and awareness of up-to-date guidelines will allow for AI to have a maximal positive impact on the care of individuals with endometriosis.

Declarations

Ethics approval and consent to participate

An ethical statement is not applicable because this study is a review and is based exclusively on published literature.

Consent for publication

The consent for publication is not applicable because this study is a review and is based exclusively on published literature.

Author contribution(s)

Brie Dungate: Conceptualization; Investigation; Visualization; Writing—original draft; Writing—review & editing.
Dwayne R. Tucker: Conceptualization; Investigation; Writing—review & editing.
Emma Goodwin: Conceptualization; Writing—review & editing.
Paul J. Yong: Conceptualization; Supervision; Writing—review & editing.

Acknowledgements

None

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Competing interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Availability of data and materials

Not applicable

ORCID iD

Brie Dungate https://orcid.org/0009-0000-1272-2340

References

1. Maddern J, Grundy L, Castro J, et al. Pain in endometriosis. *Front Cell Neurosci* 2020; 14: 590823.
2. Ellis K, Munro D and Clarke J. Endometriosis is undervalued: a call to action. *Front Glob Womens Health* 2022; 3: 902371.
3. Meskó B and Görög M. A short guide for medical professionals in the era of artificial intelligence. *NPJ Digit Med* 2020; 3: 126.
4. Cabitza F, Rasoini R and Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* 2017; 318: 517.
5. Rajput D, Wang W-J and Chen C-C. Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics* 2023; 24: 48.
6. Surrey E, Soliman AM, Trenz H, et al. Impact of endometriosis diagnostic delays on healthcare resource utilization and costs. *Adv Ther* 2020; 37(3): 1087–1099.
7. Frankel LR. A 10-year journey to diagnosis with endometriosis: an autobiographical case report. *Cureus* 2022; 14(1): e21329.
8. Vermeulen N, Abrao MS, Einarsson JI, et al. Endometriosis classification, staging and reporting systems: a review on the road to a universally accepted endometriosis classification. *Facts Views Vis Obgyn* 2021; 13: 305–330.
9. Ongsulee P. Artificial intelligence, machine learning and deep learning. In: *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*. Bangkok, Thailand, 22–24 November 2017, pp. 1–6. New York: IEEE.
10. Aldausari N, Sowmya A, Marcus N, et al. Video generative adversarial networks: a review. *ACM Comput Surv* 2023; 55: 1–25.
11. Caruana R and Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: *Proceedings of the 23rd international conference on Machine learning–ICML '06*. Pittsburgh, PA, 25–29 June 2006, pp. 161–168. New York: ACM Press.
12. Hu M, Zhang J, Matkovic L, et al. Reinforcement learning in medical image analysis: concepts, applications, challenges, and future directions. *J Appl Clin Med Phys* 2023; 24(2): e13898.
13. Faes L, Sim DA, Smeden M, et al. Artificial intelligence and statistics: just the old wine in new wineskins? *Front Digit Health* 2022; 4: 833912.
14. Ley C, Martin RK, Pareek A, et al. Machine learning and conventional statistics: making sense of the differences. *Knee Surg Sports Traumatol Arthrosc* 2022; 30(3): 753–757.
15. Ethics and governance of artificial intelligence for health, 2021, http://apps.who.int/bookorders.
16. 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* 2021; 11: e048008.
17. Allaire C, Bedaiwy MA and Yong PJ. Diagnosis and management of endometriosis. *Can Med Assoc J* 2023; 195: E363–E371.
18. Canadians need better access to medical imaging: addressing the diagnostic backlog, https://car.ca/wp-content/uploads/2022/09/CAR-PreBudgetSubmission-2023-FINAL.pdf
19. Maicas G, Leonardi M, Avery J, et al. Deep learning to diagnose pouch of Douglas obliteration with ultrasound sliding sign. *Reprod Fertil* 2021; 2(4): 236–243.
20. Hosna A, Merry E, Gyalmo J, et al. Transfer learning: a friendly introduction. *J Big Data* 2022; 9(1): 102.
21. Indrielle-Kelly T, Frühauf F, Fanta M, et al. Diagnostic accuracy of ultrasound and MRI in the mapping of deep pelvic endometriosis using the International Deep Endometriosis Analysis (IDEA) consensus. *Biomed Res Int* 2020; 2020: 3583989.
22. Zhang Y, Wang H, Butler D, et al. Distilling missing modality knowledge from ultrasound for endometriosis diagnosis with magnetic resonance images, http://arxiv.org/abs/2307.02000 (2023, accessed 27 January 2024).

23. Balica A, Dai J, Piiwaa K, et al. Augmenting endometriosis analysis from ultrasound data using deep learning. In: Bottenus N and Boehm C (eds) *Medical Imaging 2023: Ultrasonic Imaging and Tomography.* San Diego, United States: SPIE, p. 25
24. Guerriero S, Pascual M, Ajossa S, et al. Artificial intelligence (AI) in the detection of rectosigmoid deep endometriosis. *Eur J Obstet Gynecol Reprod Biol* 2021; 261: 29–33.
25. Leonardi M, Uzuner C, Mestdagh W, et al. Diagnostic accuracy of transvaginal ultrasound for detection of endometriosis using International Deep Endometriosis Analysis (IDEA) approach: prospective international pilot study. *Ultrasound Obstet Gynecol* 2022; 60: 404–413.
26. Sivajohan B, Elgendi M, Menon C, et al. Clinical use of artificial intelligence in endometriosis: a scoping review. *NPJ Digit Med* 2022; 5: 109.
27. Goldstein A and Cohen S. Self-report symptom-based endometriosis prediction using machine learning. *Sci Rep* 2023; 13: 5499.
28. Vesale E, Roman H, Abo C, et al. Predictive approach in managing voiding dysfunction after surgery for deep endometriosis: a personalized nomogram. *Int Urogynecol J* 2021; 32(5): 1205–1212.
29. Dutta M, Joshi M, Srivastava S, et al. A metabonomics approach as a means for identification of potential biomarkers for early diagnosis of endometriosis. *Mol Biosyst* 2012; 8: 3281.
30. Ghazi N, Arjmand M, Akbari Z, et al. 1H NMR-based metabolomics approaches as non-invasive tools for diagnosis of endometriosis. *Int J Reprod Biomed* 2016; 14(1): 1–8.
31. Bendifallah S, Suisse S, Puchar A, et al. Salivary microRNA signature for diagnosis of endometriosis. *J Clin Med* 2022; 11: 612.
32. Avery JC, Deslandes A, Freger SM, et al. Non-invasive diagnostic imaging for endometriosis Part 1: a systematic review of recent developments in ultrasound, combination imaging and artificial intelligence. *Fertil Steril* 2023; 121: 164–188.
33. Bendifallah S, Puchar A, Suisse S, et al. Machine learning algorithms as new screening approach for patients with endometriosis. *Sci Rep* 2022; 12: 639.
34. Marlin N, Rivas C, Allotey J, et al. Development and validation of clinical prediction models for surgical success in patients with endometriosis: protocol for a mixed methods study. *JMIR Res Protoc* 2021; 10: e20986.
35. Tucker DR, Noga HL, Lee C, et al. Pelvic pain comorbidities associated with quality of life after endometriosis surgery. *Am J Obstet Gynecol* 2023; 229(2): 147.e1–147.e20.
36. Su D, Guo Y, Yang R, et al. Identifying a panel of nine genes as novel specific model in endometriosis noninvasive diagnosis. *Fertil Steril* 2023; 121: 323–333.
37. Peter AW, Adamson GD, Al-Jefout M, et al. Research priorities for endometriosis: recommendations from a global consortium of investigators in endometriosis. *Reprod Sci* 2017; 24: 202–226.
38. Jiang H, Zhang X, Wu Y, et al. Bioinformatics identification and validation of biomarkers and infiltrating immune cells in endometriosis. *Front Immunol* 2022; 13: 944683.
39. Wang L, Zheng W, Mu L, et al. Identifying biomarkers of endometriosis using serum protein fingerprinting and artificial neural networks. *Int J Gynaecol Obstet* 2008; 101(3): 253–258.
40. Kaya C, Usta T and Oral E. Telemedicine and artificial intelligence in the management of endometriosis: future forecast considering current progress. *Geburtshilfe Frauenheilkd* 2023; 83(1): 116–117.
41. Foster KR, Koprowski R and Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research—commentary. *Biomed Eng Online* 2014; 13: 94.
42. Vabalas A, Gowen E, Poliakoff E, et al. Machine learning algorithm validation with a limited sample size. *PLoS ONE* 2019; 14(11): e0224365.
43. Kernbach JM and Staartjes VE. Foundations of machine learning-based clinical prediction modeling: part II—generalization and overfitting. *Acta Neurochir Suppl* 2022; 134: 15–21.
44. Slutsky D. Statistical errors in clinical studies. *J Wrist Surg* 2013; 02: 285–287.
45. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep* 2022; 12: 5979.
46. Yacouby R and Axman D. Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. In: *Proceedings of the first workshop on evaluation and comparison of NLP system*s, Online, November, pp. 79–91. Kerrville, TX: Association for Computational Linguistics.
47. Alwosheel A, Van Cranenburgh S and Chorus CG. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *J Choice Model* 2018; 28: 167–182.
48. Raudys SJ and Jain AK. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans Pattern Anal Mach Intell* 1991; 13: 252–264.
49. Ying X. An Overview of Overfitting and its Solutions. *J Phys Conf Ser* 2019; 1168: 022022.
50. Berrar D. Cross-Validation. In: Ranganathan S, Gribskov M and Nakai K, Schönbach C (eds) *Encyclopedia of Bioinformatics and Computational Biology*. Amsterdam: Elsevier, pp. 542–545.
51. Azuaje F and Pearson RK. Mining imperfect data: dealing with contamination and incomplete records. *BioMed Eng OnLine*. Epub ahead of print 18 July 2005. DOI: 10.1186/1475-925x-4-43.
52. Kolachalama VB and Garg PS. Machine learning and medical education. *NPJ Digit Med* 2018; 1: 54.
53. Garbin C and Marques O. Assessing methods and tools to improve reporting, increase transparency, and reduce failures in machine learning applications in health care. *Radiol Artif Intell* 2022; 4(2): 1–9.
54. Hullman J, Kapoor S, Nanayakkara P, et al. The worst of both worlds: a comparative analysis of errors in learning from data in psychology and machine learning. In: *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society*, Oxford, 19–21 May 2021, pp. 335–348. New York: ACM.
55. Adda J, Decker C and Ottaviani M. P-hacking in clinical trials and how incentives shape the distribution of results across phases. *Proc Natl Acad Sci* 2020; 117: 13386–13392.

56. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019; 25: 1337–1340.
57. Kokol P, Kokol M and Zagoranski S. Machine learning on small size samples: a synthetic knowledge synthesis. *Sci Prog* 2022; 105(1): 368504211029777.
58. Morris ZS, Wooding S and Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. *J R Soc Med* 2011; 104(12): 510–520.
59. Lamontagne F, Rowan KM and Guyatt G. Integrating research into clinical practice: challenges and solutions for Canada. *Can Med Assoc J* 2021; 193: E127–E131.
60. Mason J, Morrison A, Visintini S, et al. An overview of clinical applications of artificial intelligence CADTH issues in emerging health technologies 2, https://www.cadth.ca/grey-matters
61. Artificial intelligence initiatives. National Institutes of Health, https://datascience.nih.gov/artificial-intelligence/initiatives#