Artificial Intelligence in the Diagnosis and Management of Appendicitis in Pediatric Departments: A Systematic Review

Robin Rey1 Renato Gualtieri2 Giorgio La Scala3 Klara Posfay Barbe4

1 Department of Human Medicine, Faculty of Medicine, University of Geneva, Genève, Switzerland
2 Department of Pediatrics, Gynecology and Obstetrics, University of Geneva, Genève, Switzerland
3 Division of Pediatric Surgery, Hôpital des enfants, Geneva University Hospitals, Genève, Switzerland
4 Division of General Pediatrics, Hôpital des enfants, Geneva University Hospitals, Genève, Switzerland

Address for correspondence Renato Gualtieri, MD, Department of Pediatrics, Gynecology and Obstetrics, University of Geneva, Genève, Switzerland (e-mail: renato.gualtieri@unige.ch).

Introduction

Acute appendicitis (AA) is a very frequent diagnosis in any pediatric emergency department (ED), as it is diagnosed in 1 to 8% of children presenting with acute abdominal pain1 and is the most common pediatric surgical emergency worldwide.2–6 Early detection of AA is crucial since delayed diagnosis increases the risk of perforated appendicitis and its associated morbidities (e.g., peritonitis, sepsis).1,7 Of note, while a conservative management with antibiotics rather than appendectomy is increasingly reported, an early diagnosis is nevertheless required.8–10

Diagnosis of AA can be challenging, particularly in the pediatric population.2,3,6,11 A large body of research has been conducted to improve the early accurate diagnosis of AA, but no optimal strategy has been established. Clinical signs, such

Abstract

Introduction Artificial intelligence (AI) is a growing field in medical research that could potentially help in the challenging diagnosis of acute appendicitis (AA) in children. However, usefulness of AI in clinical settings remains unclear. Our aim was to assess the accuracy of AIs in the diagnosis of AA in the pediatric population through a systematic literature review.

Methods PubMed, Embase, and Web of Science were searched using the following keywords: “pediatric,” “artificial intelligence,” “standard practices,” and “appendicitis,” up to September 2023. The risk of bias was assessed using PROBAST.

Results A total of 302 articles were identified and nine articles were included in the final review. Two studies had prospective validation, seven were retrospective, and no randomized control trials were found. All studies developed their own algorithms and had an accuracy greater than 90% or area under the curve >0.9. All studies were rated as a “high risk” concerning their overall risk of bias.

Conclusion We analyzed the current status of AI in the diagnosis of appendicitis in children. The application of AI shows promising potential, but the need for more rigor in study design, reporting, and transparency is urgent to facilitate its clinical implementation.
as anorexia, vomiting, fever, and abdominal pain are non-specific, and clinical evaluation is difficult, particularly in young and preverbal children.\textsuperscript{2,6,11} No inflammatory markers or other laboratory tests have been able to identify alone AA.\textsuperscript{12} The most widely used scores are the Alvarado and Pediatric Appendicitis Score, but they have no sufficient predictive values, limiting their clinical impact.\textsuperscript{2,6,13} Ultrasound and computed tomography are part of the imaging strategy, but they also have some limitations: ultrasound is operator-dependent, it can confirm but not rule out AA, reducing its diagnostic efficiency.\textsuperscript{14} On the other hand, computed tomography has shown great accuracy,\textsuperscript{15} but requires radiation exposure that is best avoided in the pediatric population, and is more costly.

Further research and proper application of new technologies are needed to improve the diagnosis of AA.\textsuperscript{8} Recently, the increased amount of computerized data in the medical field has created a strong impetus to develop new artificial intelligence (AI) algorithms.\textsuperscript{15} AI is defined by the World Health Organization as the ability of algorithms encoded in technology to learn from data so that they can perform automated tasks without every step in the process having to be programmed explicitly by a human.\textsuperscript{16} AI has shown great promise in different fields (e.g., radiology, dermatology, pathology), and great diagnostic accuracy in other settings.\textsuperscript{16–18} AI in the diagnosis of AA uses data already available during the clinical assessment, is noninvasive, and has no direct interaction with patients, therefore being potentially a great tool for pediatric medicine. However, there are barriers to AI integration in the clinical workflow, and the lack of evidence and transparency around AI creates a Blackbox that decreases health care providers’ trust.\textsuperscript{19,20} This systematic review aims to assess the accuracy of AI in the diagnosis of AA in children, and its potential usefulness in a clinical setting.

Methods

This study was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, and a corresponding checklist is available in the Supplementary Materials.

Study Identification and Inclusion Criteria

We extensively searched the PubMed, Embase, and Web of Science databases in September 2023 without restriction of publication year, leveraging Boolean operators to link the keywords “pediatric,” “artificial intelligence,” “standard practices,” and “appendicitis.” Further details regarding our search strategy are provided in the Supplementary Materials section. Additional articles were identified by analyzing the reference lists of relevant publications.

The inclusion criteria for selecting a publication for review were as follows: a peer-reviewed scientific report of original research with the aim of using AI for predicting the absolute risk of appendicitis or classification into diagnostic groups (e.g., appendicitis or other diseases); English language; evaluation of an AI algorithm applied to the diagnosis of appendicitis in pediatric patients (<18 years old); cohorts of AA patients used to create algorithms were diagnosed based on clinical features validated by a medical expert, appendectomy, or anatomopathological analysis.

The exclusion criteria were informal publications (such as commentaries, letters to the editor, editorials, and meeting abstracts).

Study Selection and Extraction of Data

After automatic identification and removal of duplicates using Endnote, two authors (R.R. and R.G.) independently screened titles and abstracts for potentially eligible studies; therefore, each record was reviewed by at least two individuals. Full-text reports were then assessed for eligibility, and disagreements were resolved by consensus. Two authors (R.R. and R.G.) extracted data from the reports independently and in duplicate for each eligible study, and disagreements were resolved by consensus, or by a third reviewer.

The data extracted from the selected studies encompassed various parameters such as the size of the dataset, study design, country of origin, patients’ clinical data included in the study, type of AI used, proportion of the dataset used for the AI’s development and validation, and the AI’s accuracy in diagnosing or determining the severity of AA.

Risk of Bias

To evaluate the risk of bias in the selected studies, we used the Prediction model Risk Of Bias Assessment Tool (PROBAST)\textsuperscript{21}; it contains 20 signaling questions across four domains: participants, predictors, outcomes, and analysis.\textsuperscript{22}

Data Synthesis

At the time of study planning, we decided not to perform formal quantitative syntheses because of the expected heterogeneity of the algorithms and predictors used.

Patient and Public Involvement

Patients were not involved in any aspect of this study.

Results

Studies Selection

A total of 417 articles were identified from three databases (PubMed, Embase, and Web of Science). Using the previously indicated search strategy, the PRISMA flowchart 2020 shows the process from the initial search to the final included articles (\textsuperscript{Fig. 1}). One hundred fifteen duplicate records were excluded; 265 articles did not meet the inclusion criteria after title and abstract screening and 24 after full-text screening, leaving 9 articles in the final selection.\textsuperscript{23–31}

Studies Characteristics

The characteristics of the included studies are shown in Table 1. Most of the studies were recent, as eight out of nine were published in the last 4 years (2019–2023). The most frequent country of origin was Germany (n = 4), followed by Turkey (n = 3), the United States (n = 1), and Brunei Darussalam (n = 1). All studies discussed the development and application of their own and new AI algorithms using
different parameters to improve the diagnostic accuracy of AA. Each study used different combinations of demographics and clinical, laboratory, and imaging results to create their algorithm. Two studies\textsuperscript{23,29} had a prospective validation of the data, while all the others were retrospective. No randomized controlled studies were identified.

**Prospective Studies**

Akgül et al\textsuperscript{23} enrolled 320 patients suspected of having appendicitis in an ED in a prospective single-center study. A total of 190 cases of appendicitis were confirmed using histopathological analysis. The study combined physical examinations, laboratory tests (white blood cells [WBC]), absolute neutrophil count [ANC], C-reactive protein [CRP], procalcitonin, calprotectin), and ultrasonography using an artificial neural network for analysis. The authors produced a receiver operating characteristic curve with an area under the curve (AUC) of 0.91, a sensitivity of 89.8%, and a specificity of 81.2%. Shikha and Kasem\textsuperscript{29} in a study with an AI model created on retrospective data with prospective validation enrolled 305 patients (retrospectively and prospectively). The authors first developed a decision tree algorithm based only on retrospective clinical and laboratory (WBC and percentage of neutrophils) findings. They then prospectively validated this algorithm using a sample of 139 patients suspected of having appendicitis including 61 cases with the diagnosis confirmed through histopathological analysis. The authors reported an accuracy rate of 97.1%, with a sensitivity of 96.7% and a specificity of 97.4%.

**Retrospective Studies**

The remaining seven studies\textsuperscript{24–28,30,31} developed an algorithm with retrospective data, and some used k-nearest neighbor models to validate their algorithm. All studies used demographics and laboratory results (WBC, ANC, and CRP most frequently, but also urine analysis, hemoglobin, hematocrit, mean corpuscular volume, platelet, mean platelet volume, lymphocyte), and in five studies clinical and/or imaging results were used. All studies reported an accuracy or AUC >90%. A variety of algorithms have been used or tested, and most reports presented results for more than one algorithm. The most representative algorithms were random forest (5/7), support vector machine (3/7), and logistic regression (3/7). The largest dataset size was 11,384 patients (256 AA), followed by 7,244 patients (with 2,831 AA), 692 patients (with 45 AA), and 590 patients (473 AA). The remaining three datasets included 400 to 500 patients.
<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Country</th>
<th>Study design</th>
<th>Dataset size (cases of appendicitis)</th>
<th>Patient inclusion criteria</th>
<th>Predictors</th>
<th>Primary outcome</th>
<th>Artificial intelligence technique</th>
<th>Primary outcome metrics and performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Akgül</td>
<td>2021</td>
<td>Turkey</td>
<td>Prospective</td>
<td>320 (190)</td>
<td>Appendicitis suspicion</td>
<td>Clinical, laboratory, and abdominal ultrasound</td>
<td>Diagnosis accuracy</td>
<td>ANN</td>
<td>AUC: 0.91; sensitivity: 89.8%; specificity: 81.2%</td>
</tr>
<tr>
<td>2 Akmese</td>
<td>2020</td>
<td>Turkey</td>
<td>Retrospective</td>
<td>428 (214)</td>
<td>Appendicitis suspicion</td>
<td>Demographic and laboratory</td>
<td>Diagnosis accuracy</td>
<td>GB</td>
<td>Accuracy: 95.3%</td>
</tr>
<tr>
<td>3 Aydin</td>
<td>2020</td>
<td>Turkey</td>
<td>Retrospective</td>
<td>7,244 (2,831)</td>
<td>Acute abdomen (patient group) and 3 control groups</td>
<td>Demographic and laboratory</td>
<td>Diagnosis accuracy</td>
<td>RF, GB, CART, SVM, LR, K-NN, ANN</td>
<td>Accuracy: 97.45%</td>
</tr>
<tr>
<td>4 Grigull</td>
<td>2012</td>
<td>Germany</td>
<td>Retrospective</td>
<td>692 (45)</td>
<td>Children admitted in the ED</td>
<td>Demographics, clinical, and laboratory</td>
<td>Diagnosis accuracy</td>
<td>SVM, ANN, Fuzzy Logic, all combined by a voting algorithm</td>
<td>Accuracy: 97% (37/38)</td>
</tr>
<tr>
<td>5 Marcinkevics</td>
<td>2021</td>
<td>Germany</td>
<td>Retrospective</td>
<td>430 (247)</td>
<td>Appendicitis suspicion</td>
<td>Demographic, clinical, laboratory, and abdominal ultrasound</td>
<td>Diagnosis accuracy</td>
<td>RF, GB, CART, SVM, LR</td>
<td>AUROC: 0.96; AUPR: 0.94</td>
</tr>
<tr>
<td>6 Reismann</td>
<td>2019</td>
<td>Germany</td>
<td>Retrospective</td>
<td>590 (473)</td>
<td>Appendicitis suspicion</td>
<td>Demographic, laboratory, and abdominal ultrasound</td>
<td>Diagnosis accuracy</td>
<td>Supervised learning linear model</td>
<td>AUC: 0.91; accuracy: 90%</td>
</tr>
<tr>
<td>7 Shikha</td>
<td>2022</td>
<td>Brunei Darussalam</td>
<td>Retrospective</td>
<td>166 (69)</td>
<td>Appendicitis suspicion</td>
<td>Demographics, clinical, and laboratory</td>
<td>Diagnosis accuracy</td>
<td>DT</td>
<td>Accuracy: 93.3%; sensitivity: 92.8%; specificity: 93.8%</td>
</tr>
<tr>
<td>8 Stiel</td>
<td>2020</td>
<td>Germany</td>
<td>Retrospective</td>
<td>463 (336)</td>
<td>Appendicitis suspicion</td>
<td>Demographics, clinical, laboratory, and abdominal ultrasound</td>
<td>Diagnosis accuracy</td>
<td>CART (AI Score)</td>
<td>AUC: 0.92; sensitivity: 86.6%; specificity: 70.9%</td>
</tr>
<tr>
<td>9 Su</td>
<td>2022</td>
<td>United States of America</td>
<td>Retrospective</td>
<td>11,384 (2,56)</td>
<td>Appendicitis suspicion</td>
<td>Demographic, clinical, laboratory, imaging, a and unstructured data b</td>
<td>Diagnosis accuracy</td>
<td>LR</td>
<td>AUC: 0.87; accuracy: 93%</td>
</tr>
</tbody>
</table>

Abbreviations: ANN, artificial neural network; AUC, area under the curve; AUPR, area under the precision-recall curve; AUROC, area under the receiver operating characteristic; CART, classification and regression trees; DT, decision tree; GB, gradient boosting; GBM, generalized boosted regression model; GLM, generalized linear model; K-NN, K-nearest neighbors; LR, logistic regression; NB, naive Bayes; RF, random forest; SVM, support vector machine.

aAny imaging tests provided.
bData that are not readily available in predefined structured formats, such as tabular formats (e.g., free text data).
The fast-paced development and the vast potential of AI in patient care have generated a compelling need to incorporate these algorithms into clinical practice.\(^{35}\) Especially in diagnostics, AI can improve accuracy, reduce cost, time, and could be used in countries with insufficient health care workers.\(^{15,36}\) But the premature implementation of AI without a rigorous evidence-based foundation is likely to have many biases and many challenges need to be considered to enable efficient and useful implementation in clinical settings. Challenges in AI include ethical considerations, data bias and processing, security and data privacy, and personnel training, collaboration and adherence.\(^{36-38}\)

A high-quality, large dataset is required to ensure generalizability and reproducibility, accurate outcomes, and reduce overfitting and overlapping risk.\(^{36,37,39}\) The majority of the studies that we examined were retrospective, monocentric, and none underwent an external validation of the proposed scores making it difficult to compare the actual diagnostic performance between AI and clinicians. The lack of validated prospective studies could lead to an overestimation of the potential improvement in diagnostic accuracy without an appropriate assessment of potential undesired consequences, such as a high percentage of false positives, limiting the applicability of the results in a clinical setting.\(^{35,40}\)

The limited availability of codes used for creating algorithms makes it difficult to evaluate the reproducibility of AI research.\(^{40}\) The description of the hardware used was often missing or vague.\(^{37}\) An important criticism that is also applied to the studies we included is that many models have not been evaluated with the same thoroughness as we expect for other medical diagnostic tools.\(^{39}\)

Despite evident limitations in design, reporting, transparency, and risk of bias, these aspects are not adequately highlighted in the discussion of most studies, and are never mentioned in the abstracts, suggesting a general trend to underestimate the limitations of this approach.

This underscores the relevance of our findings, which focus on the need to improve the design, communication, transparency, and interpretation of studies using AI. Our analysis noted the presence of a high risk of bias in various aspects of these studies, underlining the need for further research that implements more rigorous bias control.

This will help to ensure the reliability and applicability of the results in the field of AI for the diagnosis of appendicitis.

### Study Limitations
Our study has several limitations. Although we tried to be as comprehensive as possible, our search may have missed some relevant studies. We assessed the risk of bias in the studies using guidelines (PROBAST) designed for traditional predictive modeling studies, which may not be appropriate for the evaluation of this research. Therefore, the levels of adherence we identified must be interpreted in this context. Some authors suggest that specialized guidelines for assessing the risk of bias in these types of studies are urgently required.\(^{37,41}\)

As we specifically examined AI for the diagnosis of appendicitis, our results cannot be extrapolated to other types of AI or other medical conditions.

Finally, the evaluation of bias risk involves a degree of subjective judgment, and people with different experiences of AI performance might have different perceptions.

### Conclusions
Our systematic review provides a comprehensive analysis of the current status of AI in the diagnosis of appendicitis in the pediatric population. While the application of AI shows
promises in enhancing diagnostic accuracy, we underline
the need for a more rigorous study design, reporting, and
transparency. The relatively high risk of bias observed across
studies highlights the urgency for more stringent bias con-
trol in future investigations. Given the groundbreaking and
unprecedented application of AI in human medicine, there is
a pressing need to develop methodological recommenda-
tions tailored specifically for the reporting of diagnostic
studies using AI as well as adaptive guidelines for assessing
the risk of bias.

Ethical Approval
No ethical approval was required as the review concerns
data from previously published studies.

Consent for Publication
The manuscript does not contain any individual's data in
any form.

Availability of Data and Materials
The datasets used and analyzed during the current study
are available from the corresponding author upon reason-
able request.

Competing Interests
None declared.

Author Contributions
All authors contributed to the study conception and
design. R.R. and R.G. carried out the literature search,
extracted and analyzed the data, and wrote the first draft
of the manuscript; K.P.B. and G.L.S. reviewed the manu-
script for important intellectual content; all authors ap-
proved the final version of the manuscript.

Funding
None.

Conflict of Interest
None declared.

References
1 Rothrock SG, Pagane J. Acute appendicitis in children: emergency
36(01):39–51
2 Benabbas R, Hanna M, Shah J, Sinert R. Diagnostic accuracy of
history, physical examination, laboratory tests, and point-of-care
ultrasound for pediatric acute appendicitis in the emergency
department: a systematic review and meta-analysis. Acad Emerg
Med 2017;24(05):523–551
3 Rentea RM, St Peter SD. Pediatric appendicitis. Surg Clin North Am
2017;97(01):93–112
4 Reynolds SL, Jaffe DM. Diagnosing abdominal pain in a pediatric
5 Scholer SJ, Pituch K, Orr DP, Dittus RS. Clinical outcomes of
children with acute abdominal pain. Pediatrics 1996;98(4, Pt
1):680–685
6 Bhangu A, Sørseide K, Di Saverio S, Assarsson JH, Drake FT. Acute
appendicitis: modern understanding of pathogenesis, diagnosis,
and management. Lancet 2015;386(10008):1278–1287
7 Narasule CK, Kåhle EJ, Kim DS, Anderson AC, Luks FI. Effect of delay
in presentation on rate of perforation in children with appendici-
8 Lipsett SC, Monuteaux MC, Shanahan KH, Bachur RG. Nonopera-
tive management of uncomplicated appendicitis. Pediatrics 2022;
149(05):e2021054693
choice in nonoperative vs surgical management of pediatric
uncomplicated acute appendicitis. JAMA Surg 2016;151(05):
408–415
10 Minneci PC, Hade EM, Lawrence AE, et al; Midwest Pediatric
Surgery Consortium. Association of nonoperative management
using antibiotic therapy vs laparoscopic appendectomy with
treatment success and disability days in children with uncom-
pliacted appendicitis. JAMA 2020;324(06):581–593
11 Bundy DG, Byerley JS, Liles EA, Perrin EM, Katzenelson J, Rice HE.
Does this child have appendicitis? JAMA 2007;298(04):438–451
12 Yu CW, Juan LL, Wu MH, Shen CJ, Wu JY, Lee CC. Systematic review
and meta-analysis of the diagnostic accuracy of procalcitonin,
C-reactive protein and white blood cell count for suspected acute
13 Pogorelić Z, Rak S, Mrklić I, Jurić I. Prospective validation of
Alvarado score and Pediatric Appendicitis Score for the diagnosis
of acute appendicitis in children. Pediatr Emerg Care 2015;31
(03):164–168
14 Kobelein GC, Trout AT, Rigby CK, et al; Expert Panel on Pediatric
Imaging, ACR Appropriateness Criteria® expected appendicitis-
15 Schwalbe N, Wahl B. Artificial intelligence and the future of global
16 World Health Organization. Ethics and governance of artificial
Organization; 2021
17 Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of
deep learning in medical imaging: a systematic review and meta-
analysis. NPJ Digit Med 2021;4(01):65
18 Liopyris K, Gregoriou S, Dias J, Stratigos AJ. Artificial intelligence
in dermatology: challenges and perspectives. Dermatol Ther
(Heidelb) 2022;12(12):2637–2651
19 Reddy S. Explainability and artificial intelligence in medicine.
Lancet Digit Health 2022;4(04):e214–e215
20 Quinn TP, Jacobs S, Senadeera M, Le V, Coghlan S. The three ghosts
of medical AI: can the black-box present deliver? Artif Intell Med
2022;124:102158
21 Wolff RF, Moons KGM, Riley RD, et al; PROBAST Group† PROBAST:
A tool to assess the risk of bias and applicability of prediction
22 Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess
risk of bias and applicability of prediction model studies: expla-
23 Akgül F, Er A, Ulusoy E, et al. Integration of physical examination,
old and new biomarkers, and ultrasonography by using neural
networks for pediatric appendicitis. Pediatr Emerg Care 2021;37
(12):e1075–e1081
24 Akmese OF, Dogan G, Kor H, Erbay H, Demir E. The use of machine
learning approaches for the diagnosis of acute appendicitis.
Emerg Med Int 2020;2020:7306435
learning algorithm for preoperative diagnosis of acute appendi-
26 Rey et al.
27 Marcinkevics R, Reis Wolfertetter P, Wellmann S, Knorr C, Vogt
JE. Using machine learning to predict the diagnosis, management
and severity of pediatric appendicitis. Front Pediatr 2021;
9:662183