AI Approaches for Grading and Assessing Hydronephrosis Severity in Pediatric Patients: Protocol for a Systematic Review and Meta-Analysis

Nicholas Jason Wijaya, Mukhlis Akmal Taher, Jansen Jayadi, Rawan AlSaad, Putu Angga Risky Raharja, Tariq Abbas

Submitted to: JMIR Research Protocols
on: March 08, 2025

Table of Contents

AI Approaches for Grading and Assessing Hydronephrosis Severity in Pediatric Patients: Protocol for a Systematic Review and Meta-Analysis

Nicholas Jason Wijaya[1]; Mukhlis Akmal Taher[1]; Jansen Jayadi[1]; Rawan AlSaad[2]; Putu Angga Risky Raharja[3]; Tariq Abbas[2, 4]

[1] Undergraduate Medical Program, Faculty of Medicine, University of Indonseia, Jakarta, Indonesia Jakarta ID
[2] College of medicine, Qatar University, Doha, Qatar Doha QA
[3] Department of Urology, Faculty of Medicine, University of Indonesia, Cipto Mangunkusumo Hospital, Jakarta, Indonesia Jakarta ID
[4] Urology Division, Sidra Medicine, Doha, Qatar Doha QA

Corresponding Author:
Putu Angga Risky Raharja

Department of Urology, Faculty of Medicine, University of Indonesia, Cipto Mangunkusumo Hospital, Jakarta, Indonesia
Jl. Diponegoro no. 71
Jakarta
ID

Abstract

Background: Hydronephrosis is a condition characterized by the swelling of one or both kidneys due to urine buildup, often resulting from an obstruction in the urinary tract. In pediatric patients, grading and assessing the severity of hydronephrosis are crucial for determining appropriate treatment and predicting outcomes.

Objective: The objective of this study is to systematically review and analyze the use of AI-based models for grading and assessing hydronephrosis severity in pediatric patients, evaluating their methodologies, diagnostic performance, and potential for clinical integration.

Methods: This systematic review and meta-analysis will systematically search for studies published up to 1 March 2025 in databases including MEDLINE, Cochrane, IEEE Xplore, Scopus, Google Scholar, and Taylor & Francis, as well as grey literature sources like ProQuest, OpenGrey, and conference proceedings. Eligible studies must involve AI-based models for segmentation, classification, or prediction of hydronephrosis severity in patients aged 0–18 years, utilizing imaging modalities such as ultrasound, CT, or MRI. Studies will be assessed for risk of bias using a modified version of QUADAS-2, and a narrative synthesis will be conducted. If sufficient data homogeneity exists, a meta-analysis will be performed using random-effects models.

Results: The study began in March 2025 with completion expected by July 2025.

Conclusions: This systematic review and meta-analysis will be the first review to provide insights into the potential of AI in pediatric hydronephrosis assessment, supporting its integration into clinical practice or identifying limitations in its application.

(JMIR Preprints 08/03/2025:73629)
DOI: https://doi.org/10.2196/preprints.73629

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

AI Approaches for Grading and Assessing Hydronephrosis Severity in Pediatric Patients: Protocol for a Systematic Review and Meta-Analysis

Nicholas Jason Wijaya[1], Mukhlis Akmal Taher N[1], Jansen Jayadi[1], Rawan AlSaad[2], Putu Angga Risky Raharja[3,*], Tariq Abbas[2,4,*]

[1] Undergraduate Medical Program, Faculty of Medicine, University of Indonseia, Jakarta, Indonesia

[2] College of medicine, Qatar University, Doha, Qatar

[3] Department of Urology, Faculty of Medicine, University of Indonesia, Cipto Mangunkusumo Hospital, Jakarta, Indonesia

[4] Urology Division, Sidra Medicine, Doha, Qatar

Corresponding Author:
Putu Angga Risky Raharja
Email: anggariskyraharja@gmail.com

Tariq Abbas
Email: Tariq2c@hotmail.com

Abstract

Background: Hydronephrosis is a condition characterized by the swelling of one or both kidneys due to urine buildup, often resulting from an obstruction in the urinary tract. In pediatric patients, grading and assessing the severity of hydronephrosis are crucial for determining appropriate treatment and predicting outcomes.

Objective: The objective of this study is to systematically review and analyze the use of AI-based models for grading and assessing hydronephrosis severity in pediatric patients, evaluating their methodologies, diagnostic performance, and potential for clinical integration.

Methods: This systematic review and meta-analysis will systematically search for studies published up to 1 March 2025 in databases including MEDLINE, Cochrane, IEEE Xplore, Scopus, Google Scholar, and Taylor & Francis, as well as grey literature sources like ProQuest, OpenGrey, and conference proceedings. Eligible studies must involve AI-based models for segmentation, classification, or prediction of hydronephrosis severity in patients aged 0–18 years, utilizing imaging modalities such as ultrasound, CT, or MRI. Studies will be assessed for risk of bias using a modified version of QUADAS-2, and a narrative synthesis will be conducted. If sufficient data homogeneity exists, a meta-analysis will be performed using random-effects models.

Results: The study began in March 2025 with completion expected by July 2025.

Conclusion: This systematic review and meta-analysis will be the first review to provide insights into the potential of AI in pediatric hydronephrosis assessment, supporting its integration into clinical practice or identifying limitations in its application.

Keywords: Hydronephrosis; Diagnosis; Pediatric Urology; Artificial intelligence; Machine learning

Introduction

Hydronephrosis is a condition characterized by the swelling of one or both kidneys due to urine buildup, often resulting from an obstruction in the urinary tract.[1,2] Antenatally, 1-5% patient

detected with hydronephrosis.[3,4] In pediatric patients, grading and assessing the severity of hydronephrosis are crucial for determining appropriate treatment and predicting outcomes. Hydronephrosis is commonly detected by prenatal ultrasonography (US). On the other hand, neonates or young children might present with a palpable abdominal mass or urinary tract infections. Until now, US remain the gold standard on confirming and grading the severity of hydronephrosis. However, traditional methods of assessment can be subjective and vary among clinicians.[5] Nowadays, artificial Intelligence (AI) offers a potential solution by providing objective, consistent, and potentially a more accurate grading system.[6] Therefore, this systematic review and meta-analysis aims to evaluate different AI models developed to grade and assess the severity of hydronephrosis in pediatric patients.

Methods

Study Registration

This systematic review and meta-analysis will be conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement.[7] The study is registered in PROSPERO (International Prospective Register of Systematic Reviews) under the registration number CRD42024618680.

Literature Search

A comprehensive search will be conducted using keywords and Medical Subject Headings (MeSH) related to AI, paediatrics, and hydronephrosis which will be combined using the boolean operator “AND”, as shown in table 1. We will search for studies published up to 1 January 2025 across multiple databases, including PubMed/MEDLINE, Cochrane, IEEE Xplore, Scopus, Google Scholar, and Taylor & Francis. Grey sources such as ProQuest Dissertations & Theses, OpenGrey, and conference proceedings will also be searched thoroughly. Furthermore, the reference lists of included studies will be manually reviewed to identify any further relevant articles.

Table 1. Keywords used for literature search

Terms	Keywords
Artificial Intelligence Terms	("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Neural Network*" OR "Artificial Neural Network*" OR "Predictive Analytics" OR "Algorithm*" OR "Logistic Regression" OR "Support Vector Machine*" OR "Random Forest*" OR "Decision Tree*" OR "K-Nearest Neighbor*" OR "Multilayer Perceptron" OR "Convolutional Neural Network*" OR "Recurrent Neural Network*")
Pediatrics Terms	("Pediatrics" OR "Child" OR "Adolescent" OR "Infant" OR "Neonate" OR "Newborn")
Hydronephrosis Terms	("Hydronephrosis" OR "Kidney Swelling" OR "Urinary Obstruction" OR "Pediatric Urology" OR "Renal Pelvis Dilatation" OR "Ultrasound Hydronephrosis*" OR "SFU Grading" OR "Urinary Tract Obstruction")

Inclusion and Exclusion Criteria

Studies meeting the following inclusion criteria will be included:

1. Studies that develop or validate AI models for grading or assessing hydronephrosis severity in pediatric patients (segmentation, classification, and prediction).
2. Studies involving patients aged 0-18 years.
3. All types of imaging modalities (e.g., ultrasound, CT, MRI) used in the assessment of hydronephrosis.
4. Studies published in peer-reviewed journals.
5. Studies published in English.
6. Studies from all geographical locations.

Reviews, case reports, conference abstracts, guidelines, editorials, commentaries, and editorials will be excluded. Additionally, studies will be excluded if they do not involve AI models, focus solely on adult populations, do not address the grading or assessment of hydronephrosis, or are published in languages other than English.

Studies Selection

Study selection will be conducted in two stages, an initial screening of titles and abstracts which is followed by a full-text review. Two reviewers will independently screen the titles and abstracts of all identified studies to determine their eligibility for inclusion. Studies will be classified as either “retrieve” or “do not retrieve.” Full texts of studies marked as “retrieve” will undergo an independent review by two reviewers to confirm eligibility. Discrepancies between reviewers will be resolved through consensus, and unresolved disagreements will be settled by a third reviewer. The study selection process will be presented in a flowchart according to the guidelines of the PRISMA 2020 statement (Figure 1).

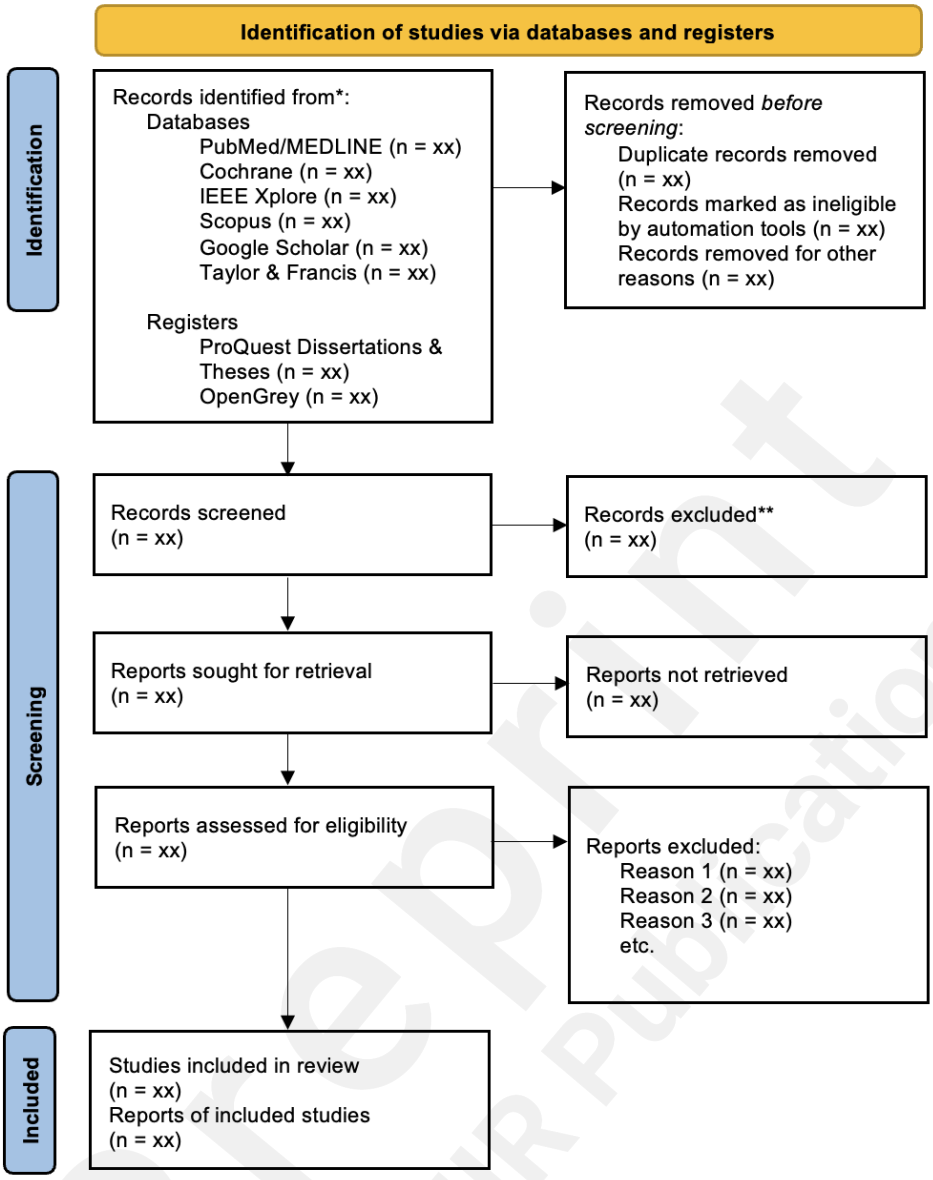


Figure 1. PRISMA Flowchart

Data Extraction

Following the full-text screening, data will be extracted from each study across several categories, including study characteristics, hydronephrosis characteristics, and AI model characteristics. A comprehensive list of data to be extracted is provided in Table 2.

Table 2. Data to be extracted

Data Category	Collected Data
Study characteristics	- Authors - Year of publication - Study period - Country of origin - Study design

	- Sample size - Average age - Gender distribution - Number of cases and controls
Hydronephrosis characteristics	- Grading system used (e.g., SFU, APD measurement) - Imaging modality employed (e.g., ultrasound, CT, MRI) - Application (e.g., segmentation, classification, prediction) - Ground truth
AI model characteristics	- Type of model (e.g., logistic regression, random forest, support vector machines, XGBoost, deep neural network, convolutional neural network, recurrent neural network) - Input features (e.g., demographics, clinical history, images, others) - Training and validation methods (e.g. internal or external validation). - Validation techniques (e.g., hold-out cross validation, K-fold cross validation, leave-one-out cross validation). - Evaluation metrics (e.g., accuracy, sensitivity, specificity, AUC)

Risk of Bias Assessment

The quality of the included studies will be assessed using an adapted version of QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2).[8] Each study will be evaluated for risk of bias and applicability based on the several domains including patient selection, index test, reference standard, flow and timing. Two independent authors will evaluate the risk of bias for each study. Discrepancies will be resolved with a third author. The risk of bias for each study will be classified into one of three categories: “low,” “high,” or “some concerns.” A graphical summary of the QUADAS-2 assessments for all included studies will be presented. Additionally, a sensitivity analysis will be performed by excluding studies identified as having a high risk of bias.

Statistical Analysis

A narrative qualitative synthesis will be performed for all included studies. If sufficient homogeneity exists, a meta-analysis will subsequently be conducted using random-effects models. The pooled estimates of model performance metrics will be calculated using the ReviewManager (RevMan) software version 5.4. For binary outcomes, odds ratios will be calculated, while weighted mean differences and relative standard deviation differences will be determined for continuous data. The 95% confidence intervals, heterogeneity, and statistical significance for each outcome will be reported, with statistical significance defined as a p-value of <0.05, Heterogeneity will be assessed using the I^2 statistic and potential sources of heterogeneity will be explored through subgroup analyses. A funnel plot will be constructed using Cochrane software to assess publication bias, given that there are more than 10 studies.

Results

Database searches will begin in March 2025. Data extraction from individual studies will be finalized by April 2025. The processes of data appraisal, preparation, summarization, and analysis be finished in June 2025, with the meta-analysis expected to be completed by July 2025.

Discussion

Hydronephrosis, which is a condition characterized by kidney swelling, is commonly diagnosed via ultrasonography. In pediatric patients, accurate grading of hydronephrosis is critical for guiding treatment, but traditional methods can be subjective and clinician-dependent.[9] AI models have the potential to provide an objective, consistent, and potentially more accurate grading of hydronephrosis severity in children.[10] This systematic review and meta-analysis will summarize various AI models for grading and assessing hydronephrosis severity in pediatric patients.

This study has several limitations. Firstly, studies rely on traditional human grading as the reference standard, which might not always hold high sensitivity and specificity, potentially influencing the evaluation of AI model accuracy. Additionally, the variability in reporting methods, definitions of hydronephrosis severity, and imaging modalities across studies may contribute to heterogeneity. Finally, due to the relatively novel application of AI in clinical situations, there will also be a limited number of high-quality randomized controlled trials. Despite these limitations, this study has several strengths. The use of a well-structured protocol, adhering to PRISMA guidelines and rigorous quality assessments such as QUADAS-2 ensure the reliability and transparency of the findings. Moreover, this study includes a wide range of study sources, including peer-reviewed publications and grey literature which will enhance the comprehensiveness of the search.

To the best of our knowledge, this will be the first systematic review and meta-analysis to provide novel insights into the diagnostic accuracy of AI for grading and assessing hydronephrosis severity in pediatric patients. The findings from this review may support the integration of AI in specific clinical applications for pediatric hydronephrosis or highlight limitations of AI in replacing human evaluators. Additionally, this protocol outlines the use of AI-based software to assist with article screening, offering a potential framework for enhancing efficiency in future systematic reviews.

Authors' Contributions

TA and PARR was responsible for developing the research question and designing the main concept, as well as drafting the manuscript. NJW, MATN, and JJ conducted the research, performed data analysis, and interpreted the findings. TA and PARR provided feedback on the manuscript drafts. All authors reviewed and approved the final version of the manuscript.

Data Availability

The datasets used and analyzed in this study is available upon reasonable request.

Conflicts of Interest

None declared.

References

1. Nguyen HT, Herndon CDA, Cooper C, Gatti J, Kirsch A, Kokorowski P, et al. The Society for Fetal Urology consensus statement on the evaluation and management of antenatal hydronephrosis. J Pediatr Urol. 2010 Jun;6(3):212–31. https://doi.org/10.1016/j.jpurol.2010.02.205
2. Kim SW. Management of hydronephrosis: a comprehensive review in pediatric urology perspective. Childhood Kidney Diseases. 2024 Jun 28;28(2):59–65. https://doi.org/10.3339/ckd.24.011
3. EK S, LIDEFELDT K, VARRICIO L. Fetal hydronephrosis; prevalence, natural history and postnatal consequences in an unselected population. Acta Obstet Gynecol Scand. 2007 Dec 31;86(12):1463–6. https://doi.org/10.1080/00016340701714802
4. Mallik M, Watson AR. Antenatally detected urinary tract abnormalities: more detection but less action. Pediatric Nephrology. 2008 Jun 1;23(6):897–904. https://doi.org/10.1007/s00467-008-0746-9
5. Onen A. Grading of Hydronephrosis: An Ongoing Challenge. Front Pediatr. 2020 Aug 27;8. https://doi.org/10.3389/fped.2020.00458
6. Krishnan G, Singh S, Pathania M, Gosavi S, Abhishek S, Parchani A, et al. Artificial intelligence in clinical medicine: catalyzing a sustainable global healthcare paradigm. Front Artif Intell. 2023 Aug 29;6. https://doi.org/10.3389/frai.2023.1227091
7. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;n71. https://doi.org/10.1136/bmj.n71
8. Huang Q-X, Huang X-W. QUADAS-2 tool for quality assessment in diagnostic meta-analysis. Ann Palliat Med. 2022 May;11(5):1844–5. https://doi.org/10.21037/apm-22-204
9. Onen A. Grading of Hydronephrosis: An Ongoing Challenge. Front Pediatr. 2020 Aug 27;8. https://doi.org/10.3389/fped.2020.00458
10. Lorenzo AJ, Rickard M, Braga LH, Guo Y, Oliveria J-P. Predictive Analytics and Modeling Employing Machine Learning Technology: The Next Step in Data Sharing, Analysis, and Individualized Counseling Explored With a Large, Prospective Prenatal Hydronephrosis Database. Urology. 2019 Jan;123:204–9. https://doi.org/10.1016/j.urology.2018.05.041

Appendix 1. The modifed version of QUADAS-2

1. Participants	**Signaling questions**	**Explanation**
	1.1 Was a consecutive or random sample of patients enrolled?	-**Yes**: if a consecutive or random sample of eligible patients was enrolled. - **No**: if patients were selected by convenience; - **Unclear**: if the study did not report the manner in which participants were enrolled.
	1.2 Did the study avoid inappropriate exclusions?	- **Yes**: If inclusion and exclusion of participants were appropriatea, so participants correspond to unselected participants of interest. - **No**: If participants are included who would already have been identified as having the outcome and so are no longer participants at suspicion of disease (diagnostic studies), or if specific subgroups are excluded that may have altered the performance of the prediction model for the intended target population. - **Unclear**: When there is no information on whether inappropriate inclusions or exclusions took place.
	1.3 Was the sample size sufficient?	- **Yes**: either having a minimum of 100 participants, or, in studies where multiple samples are recorded for each participant, having at least 100 samples. The latter condition is applicable when the input to the machine learning algorithm is derived from individual samples, rather than from entire subjects. - **No**: if there are less than 100 participants, or in studies where multiple samples are recorded for each participant, if there are fewer than 100 samples. - **Unclear**: no information on the number of features or number of participants.
	1.4 Was there a balance in the number of	- **Yes**: if the percentage of participants in any group is 66.7% or less of the total

	patients between the subgroups?	sample (≤2/3). - **No**: if the percentage of participants in any group is more than 66.7% of the total sample (>2/3). - **Unclear**: If no information was provided regarding the number of participants in the groups.
	Risk-of-bias assessment: Could the selection of participants have introduced bias?	- **Low risk of bias**: If the answer to all signaling questions is 'Yes' then the risk of bias can be considered low. If one or more of the answers is 'No', the judgment could still be low risk of bias, but specific reasons why the risk of bias can be considered low should be provided. - **High risk of bias**: If the answer to any of the signaling questions is "No" there is a potential for bias, except if defined at low risk of bias above. - **Unclear risk of bias**: If relevant information is missing for all or some of the signaling questions, and none of the answers to signaling questions is judged to put this domain at high risk of bias.
	Concerns regarding applicability: Are there concerns that the included participants and setting do not match the review question?	- **Low concern for applicability**: If the spectrum of participants (in- and exclusion criteria, setting, prior testing) matches the pre-stated requirements in the review question - **High concern for applicability**: If the spectrum of participants does not fully match the pre-stated requirements in the review question - **Unclear concern for applicability**: If there is insufficient information available to make a judgment about the applicability
2. Index test (AI algorithms)	2.1 Were the AI models described in detail?	-**Yes:** If the model details were well described. -**No:** if the model details were not well described
	2.2 Were all features (predictors) used in the	-**Yes:** If all features used in each model were clearly reported.

	model clearly identified?	-**No:** If any features used in any model were not reported. Or all features used in the paper were identified but it was not clear which features were used with each model.
	2.3 Were features assessed in the same way for all participants?	-**Yes:** If the assessment of features assessment were similar for all participants. -**No:** If different definitions were used for the same predictor or if predictors requiring subjective interpretation were assessed by differently experienced assessors. **Unclear:** If there is no information on how predictors were defined or assessed.
	2.4 Were features collected without knowledge of outcome data (hydronephrosis)?	- **Yes**: If outcome information was stated as not used during feature assessment or was clearly not (yet) available to those assessing features. - **No**: If it is clearly stated that outcome information was used when assessing predictors. - **Unclear**: No information on whether features were assessed without knowledge of outcome information.
	Risk-of-bias assessment: Could the conduct or interpretation of the index test have introduced bias?	- **Low risk of bias:** If the answer to all signaling questions is 'Yes' then the risk of bias can be considered low. If one or more of the answers is 'No', the judgment could still be low risk of bias, but specific reasons why the risk of bias can be considered low should be provided e.g., the use of objective predictors not requiring subjective interpretation. - **High risk of bias:** If the answer to any of the signaling questions is "No" there is a potential for bias, except if defined at low risk of bias above. - **Unclear risk of bias:** If relevant information is missing for all or some of the signaling questions, and none of the

		answers to signaling questions is judged to put this domain at high risk of bias.
	Concerns regarding applicability: Are there concerns that the definition, assessment, or timing of the index test in the model does not match the review question?	- **Low concern for applicability:** Definition, assessment, and timing of predictors match the review question. - **High concern for applicability:** Definition, assessment, or timing of predictors were different from the review question. - **Unclear concern for applicability:** If relevant information about the predictors is not reported.
3. Reference Standard (Ground truth)	3.1 Was the reference standard likely to correctly classify the outcome (e.g., classification or segmentation of hydronephrosis)	Is the used tool appropriate? Were the assessors/annotators qualified? **Yes**: If the study: - Followed the Society of Fetal Urology (SFU) to classify hydronephrosis using ultrasound. - Used an appropriate experimental setup to assess the classification / segmentation / prediction of hydronephrosis -**No**: If an unknown reference test, with unverified validity and reliability, was used, or if the hydronephrosis classification / segmentation / prediction was conducted by unqualified assessors. - **Unclear**: If no information was provided about the reference standard.
	3.2 Was the outcome defined and determined in a similar way for all participants?	- **Yes**: If outcomes were defined and determined in a similar way for all participants. - **No**: If outcomes were clearly defined and determined in a different way for some participants. - **Unclear**: No information on whether outcomes were defined or determined in a similar way for all participants.
	3.3 Was the outcome determined without knowledge of predictor information?	- **Yes:** If predictor information was not known when determining the outcome status, or outcome status determination is clearly reported as determined without knowledge of predictor information. - **No:** If it is clear that predictor

		information was used when determining the outcome status. - **Unclear:** No information on whether the outcome was determined without knowledge of predictor information.
	Risk-of-bias assessment: Could the reference standard, its conduct, or its interpretation have introduced bias?	- **Low risk of bias**: If the answer to all signaling questions is 'Yes' then the risk of bias can be considered low. If one or more of the answers is 'No', the judgment could still be low risk of bias, but specific reasons why the risk of bias can be considered low should be provided e.g., when the outcome was determined with knowledge of predictor information but the outcome assessment did not require much interpretation by the assessor (e.g., death regardless of cause). - **High risk of bias**: If the answer to any of the signaling questions is "No" there is a potential for bias, except if defined at low risk of bias above. - **Unclear risk of bias**: If relevant information is missing for all or some of the signaling questions, and none of the answers to signaling questions is judged to put this domain at high risk of bias.
	Concerns regarding applicability: Are there concerns that the outcome definition, timing, or determination do not match the review question?	- **Low concern for applicability**: Outcome definition, timing, and method of determination defines the outcome as intended by the review question. -**High concern for applicability**: Choice of outcome definition, timing, and method of outcome determination defines another outcome as intended by the review question. - **Unclear concern for applicability**: If relevant information about the outcome, timing, and method of determination is not reported.
4. Analysis	4.1 Were all participants included in the analysis?	- **Yes**: If all participants enrolled in the study are included in the data analysis. - **No**: If some or a subgroup of

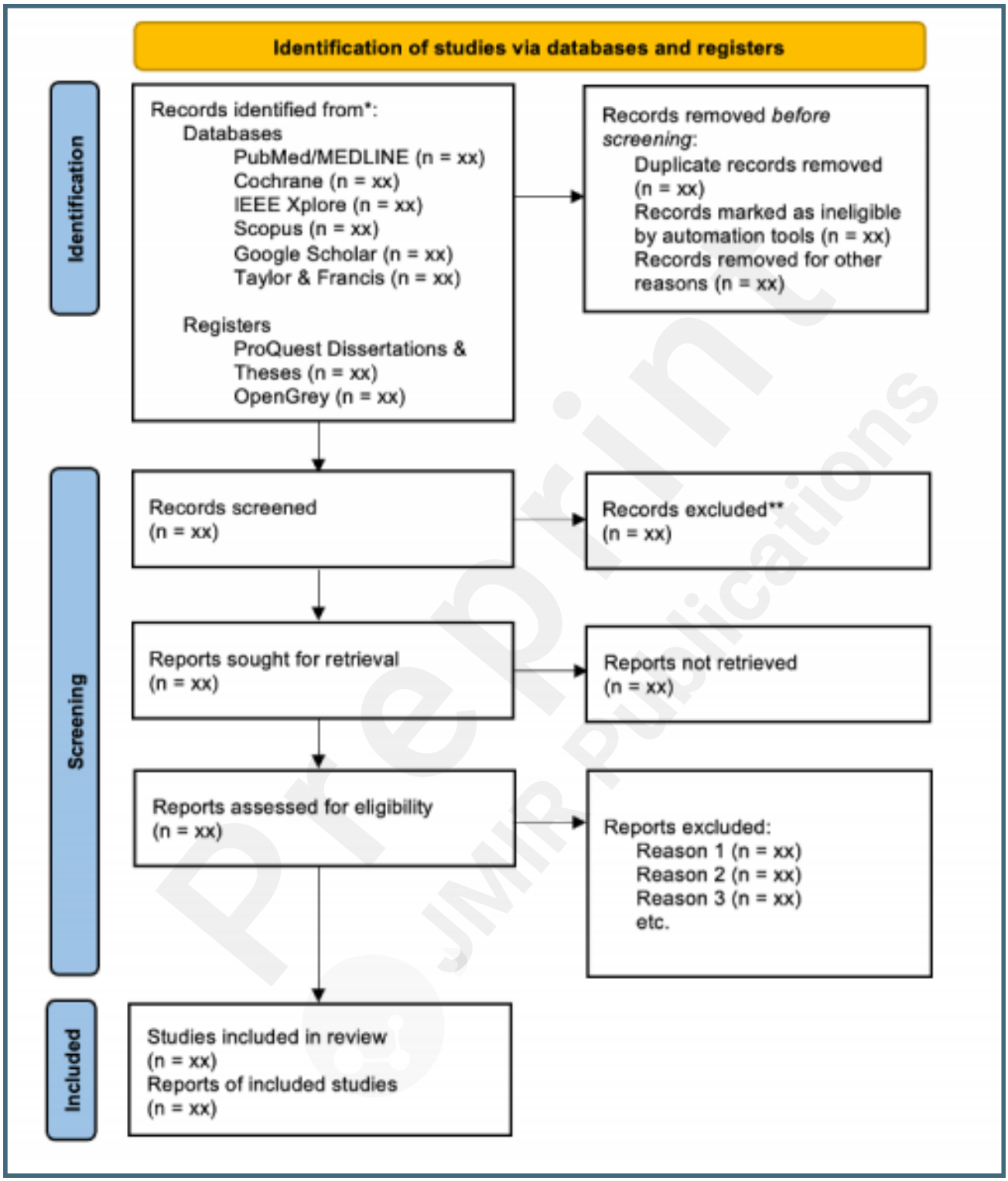
		participants are inappropriately excluded from the analysis. - **Unclear**: No information on whether all enrolled participants are included in the analysis.
	4.2 Was data preprocessing carried out appropriately?	- **Yes**: If there are no missing values of predictors or outcomes and the study explicitly reports that participants are not excluded on the basis of missing data, or if missing values are handled using multiple imputation. - **No**: If participants with missing data are omitted from the analysis, or if the method of handling missing data is clearly flawed, e.g., missing indicator method or inappropriate use of last value carried forward, or if the study had no explicit mention of methods to handle missing data. - **Unclear**: If there is insufficient information to determine if the method of handling missing data is appropriate.
	4.3 Was the breakdown of the training, validation, and test sets appropriate?	- **Yes**: • If there's a clear rationale for the chosen distribution and it aligns with best practices in the field (e.g., 70% training, 15% validation, 15% test). - **No**: • If the sizes of the validation and test sets are too small to reliably evaluate the model's performance. • If there's a lack of justification or clear rationale for how the sets were divided, or if the division contradicts standard practices without a valid reason - **Unclear**: If the study does not provide specific information on how the training, validation, and test sets were divided.
	4.4 Was the performance of the model evaluated	- **Yes**: If the evaluation metrics (Sensitivity, specificity, accuracy, AUC, AUROC, etc) was clearly presented, or

	appropriately?	more than one measure was used and the selected measures were appropriate. - **No**: If the evaluation metrics (Sensitivity, specificity, accuracy, AUC, AUROC, etc) was not presented, and only one measure was reported, or the selected measures were not appropriate. - **Unclear**: If no information was provided on the performance measures
	Risk-of-bias assessment: Could the analysis, its conduct, or its interpretation have introduced bias?	- **Low risk of bias**: If the answer to all signaling questions is 'Yes' then the risk of bias can be considered low. If one or more of the answers is 'No', the judgment could still be low risk of bias, but specific reasons why the risk of bias can be considered low should be provided. - **High risk of bias:** If the answer to any of the signaling questions is "No" there is a potential for bias, except if defined at low risk of bias above. - **Unclear risk of bias**: If relevant information is missing for all or some of the signaling questions, and none of the answers to signaling questions is judged to put the analysis at high risk of bias.

Supplementary Files

Figures

PRISMA flowchart.



Multimedia Appendixes

The modified version of QUADAS-2.
URL: http://asset.jmir.pub/assets/7cb041ea6ea73c2779834c8a0d7c3b05.docx