diagnostics

MDPI

Systematic Review

Diagnostic Accuracy of AI for Opportunistic Screening of Abdominal Aortic Aneurysm in CT: A Systematic Review and Narrative Synthesis

Maria R. Kodenko [1,2,*], Yuriy A. Vasilev [1], Anton V. Vladzymyrskyy [1,3], Olga V. Omelyanskaya [1], Denis V. Leonov [1,4], Ivan A. Blokhin [1], Vladimir P. Novik [1], Nicholas S. Kulberg [5], Andrey V. Samorodov [2], Olesya A. Mokienko [1] and Roman V. Reshetnikov [1]

1 Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, Petrovka Street, 24, Building 1, 127051 Moscow, Russia
2 Department of Biomedical Technologies, Bauman Moscow State Technical University, 2nd Baumanskaya Street, 5, Building 1, 105005 Moscow, Russia
3 Department of Information and Internet Technologies, I.M. Sechenov First Moscow State Medical University (Sechenov University), Trubetskaya Street, 8, Building 2, 119991 Moscow, Russia
4 Department of Fundamentals of Radio Engineering, Moscow Power Engineering Institute, Krasnokazarmennaya Street, 14, Building 1, 111250 Moscow, Russia
5 Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Vavilova Street, 44, Building 2, 119333 Moscow, Russia
* Correspondence: m.r.kodenko@yandex.ru

check for updates

Citation: Kodenko, M.R.; Vasiliev, Y.A.; Vladzymyrskyy, A.V.; Omelyanskaya, O.V.; Leonov, D.V.; Blokhin, I.A.; Novik, V.P.; Kulberg, N.S.; Samorodov, A.V.; Mokienko, O.A.; et al. Diagnostic Accuracy of AI for Opportunistic Screening of Abdominal Aortic Aneurysm in CT: A Systematic Review and Narrative Synthesis. *Diagnostics* **2022**, *12*, 3197. https://doi.org/10.3390/diagnostics12123197

Academic Editor: Md Mohaimenul Islam

Received: 21 October 2022
Accepted: 14 December 2022
Published: 16 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract: In this review, we focused on the applicability of artificial intelligence (AI) for opportunistic abdominal aortic aneurysm (AAA) detection in computed tomography (CT). We used the academic search system PubMed as the primary source for the literature search and Google Scholar as a supplementary source of evidence. We searched through 2 February 2022. All studies on automated AAA detection or segmentation in noncontrast abdominal CT were included. For bias assessment, we developed and used an adapted version of the QUADAS-2 checklist. We included eight studies with 355 cases, of which 273 (77%) contained AAA. The highest risk of bias and level of applicability concerns were observed for the "patient selection" domain, due to the 100% pathology rate in the majority (75%) of the studies. The mean sensitivity value was 95% (95% CI 100–87%), the mean specificity value was 96.6% (95% CI 100–75.7%), and the mean accuracy value was 95.2% (95% CI 100–54.5%). Half of the included studies performed diagnostic accuracy estimation, with only one study having data on all diagnostic accuracy metrics. Therefore, we conducted a narrative synthesis. Our findings indicate high study heterogeneity, requiring further research with balanced noncontrast CT datasets and adherence to reporting standards in order to validate the high sensitivity value obtained.

Keywords: abdominal aortic aneurysm; opportunistic screening; computed tomography; artificial intelligence; QUADAS

1. Introduction

Abdominal aortic aneurysm (AAA) has no specific symptoms and can be asymptomatic at the early stages [1]. When untreated, AAA can lead to an aortic rupture, a life-threatening condition with an overall mortality of 80% [2–4]. Presently, the accepted diagnostic modality for AAA screening is ultrasonic imaging, while computed tomography angiography (CTA) remains the "gold standard" for treatment planning [2]. Compared to ultrasonography, the advantages of CT include superior image quality, a lower operator dependency, three-dimensional reconstruction, and the possibility of a retrospective data audit [2]. A CT is also more sensitive to aortic dilation than ultrasonography [5]. The radiation exposure associated with CT restricts its application as a screening

method, but CT data can be used for opportunistic AAA detection either while reporting the study or via retrospective analysis of scans with the abdominal aorta in the field of view. According to the results of such audits, non-reported AAAs ranged from 0.4% (one of 261 patients) [6] to 5.8% (187 of 3246 patients) [7]. Taking into account the high volume of accumulated CT data (for example, in the USA the number of CT examinations was 278.5 per 1000 inhabitants in 2019 [8]), opportunistic screening could yield an increase in early diagnosed aneurysms in the population. Despite its potential, opportunistic screening for AAA at the CT exam remains a challenging task. The reported radiologist diagnostic accuracy for this task depends on the aneurysm's size, with the lowest sensitivity of 0.52 for the small ones (30–39 mm) [7]. The radiologists' errors consist of false-negatives and incorrect classification due to human-based or technical reasons [9]. AAA detection is also complicated by the measurement ambiguity of the key diagnostic parameter, the aneurysmal sac maximum transverse diameter [2].

Artificial intelligence (AI) has already shown its high potential for CT image-processing automatization [5,10,11] and promises to be a powerful assistant for radiologists' practice. Automatization of diagnostic information processing has several advantages. First, it provides a tool for a retrospective audit of big data. Second, AI yields reproducible and precise measurements, addressing the ambiguity issue of human experts.

The aim of this review was to quantify the diagnostic accuracy of AI algorithms for AAA detection by noncontrast CT, regardless of the aneurysm's size.

2. Materials and Methods

This systematic review was planned, conducted, and reported in accordance with the PRISMA statement [12], and the full protocol was registered on PROSPERO on 25 July 2021, before the literature search (PROSPERO ID CRD42021264021). The target condition evaluated was abdominal aortic aneurysm, defined as a permanent localized pathological dilatation of the abdominal aorta with a diameter greater than 3 cm or more than 50% larger than the nondilated part [13]. We also defined a negative diagnosis for AAA as the absence of abdominal aortic dilatation corresponding to the criteria above. In this review, we focused on the opportunistic screening model—the interpretation of noncontrast CT studies. For that reason, we considered CT studies without intravenous contrast and containing the aorta abdominal region in the field of view.

The index test was AAA detection by an AI algorithm. AI should have provided enough information to conclude whether an AAA was present or absent from the noncontrast CT images of the abdominal area. The image processing could be of any AI type, but the segmentation should have been fully automatic.

Manual expert segmentation was considered to be the reference standard, or "ground truth" (GT). The quality of the reference standard was estimated either by the level of expertise for a single observer or by any of the agreement metrics [14].

2.1. Search Methods for the Identification of Studies

The PubMed database was used as the main data source. Additional data (including gray literature) were searched using the Google Scholar search engine. The last search date was 2 February 2022. As sources of grey literature, we explored commercial websites with AI solutions for automated aortic segmentation in noncontrast CTs, because they rarely report results in articles [15]. The key concepts were the following: artificial intelligence, CT, abdominal aortic aneurysm, and opportunistic screening. We defined the most important and specific components of the query following the method proposed by Bramer et al. [16]. The terms included medical object (AAA), technical subject (AI), and type of intervention (detection or segmentation). Despite the research question being focused on the processing of noncontrast CT images, we did not exclude the MeSH term "angiography" from the query to avoid omissions of comparative analysis or studies with mixed target datasets. Suitable MeSH terms and keywords were identified using PubMed tools [17] and the Yale MeSH Analyzer [18]. Additionally, the most repeated words (except function words) were

identified for a subset of five studies [19–23] by full-text automatic semantic analysis using an in-house developed Python script. To avoid extra bias, we did not include the MeSH term "sensitivity and specificity" in the query. The search strategies and queries are shown in Appendix A.

2.2. Data Collection and Analysis

We exported all articles identified in the database searches into the Mendeley Reference Manager [24], where duplicates were removed. Narrative review papers, commentaries, and letters to the editor were excluded. Two reviewers independently screened the titles and abstracts of all articles for eligibility. We emailed the authors if we were unable to retrieve the full paper, requesting a copy of the full publication. Authors were re-emailed after two weeks in the case of nonresponse, and if no contact had been made after three weeks, the study was excluded. The same pathway was used if the relevant data were not available in the published report. All full-text articles were independently and in duplicate screened for suitability, and reasons for exclusion were recorded. Any discrepancies in opinion between reviewers were discussed, the third reviewer was consulted in the case of disagreement.

Two reviewers independently identified and extracted the following data from each publication: study authors, country of origin, study design, sample size (including training and validation sets), dataset structure, test details and technical parameters (both for index test and reference test), and outcome measures. When possible, we extracted 2 × 2 contingency tables or summary statistics, from which they could be computed. If a study stratified the results by the aneurysm size, we divided the data into subgroups. If the number of included studies was small or of high heterogeneity, we summarized the key study-level information and synthesized the findings narratively, focusing on AI sensitivity and specificity. We also extracted Dice similarity coefficient (DSC) values, because this metric allows estimation of segmentation quality, essential for single-case studies. If there was no information about this metric, we calculated it from the presented images of the AI-segmented mask and GT mask according to the formula below [25]:

$$DSC = \frac{2|X \cap Y|}{|X| \cup |Y|}, \quad (1)$$

where X represented the coordinates of the AI-segmented mask pixels, and Y represented the coordinates of the GT mask pixels. For this task, we exported the presented images of segmentation and GT (or original image) in JPEG format. If there were no expert markups, our medical expert (a certified radiologist with experience of 3 years) segmented it manually with a stylus using the Procreate 5.2.6 application [26] on iPad Pro 11. Then both pictures were binarized: the area was white inside the mask (values equal to 1) and black outside (values equal to 0), and they were aligned and analyzed automatically with an in-house developed script (Figure 1), prepared with R 4.1.2 [27].

We assessed the risk of bias and applicability concerns independently and in duplicate. Any disagreements were resolved through discussion. We did not use the QUADAS-2 domain list, as it was shown neither to accommodate the niche terminology encountered nor to signal the sources of bias found within AI studies [28]. Instead, we developed and used AI-specialized domain questions based on the traditional QUADAS-2 [29] (detailed information is presented in Appendix B).

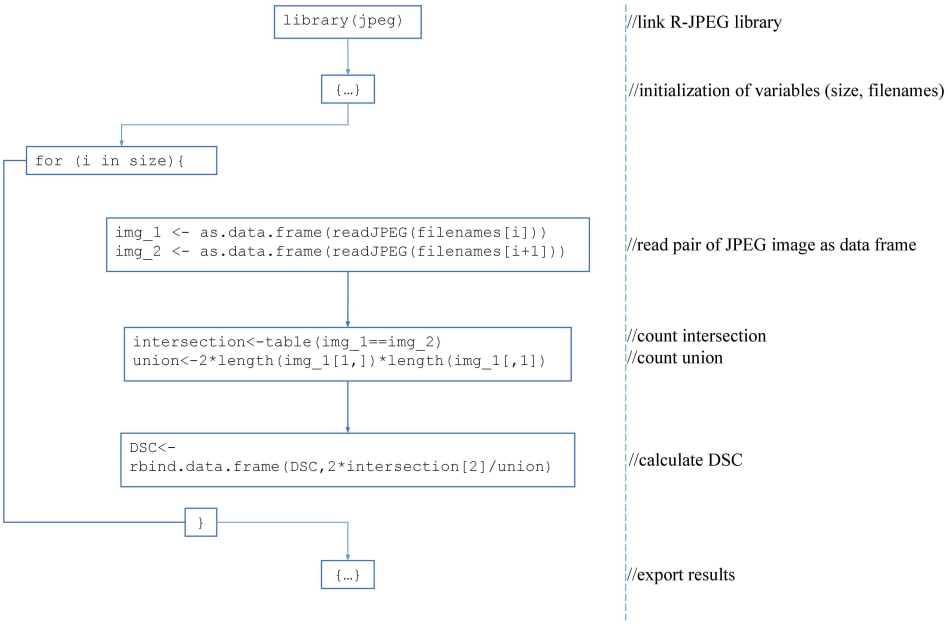

Figure 1. Scheme of R script for DSC calculation with comments.

3. Results

In total, we identified and imported 730 search results from PubMed into a Mendeley library. No additional relevant information was found in the grey literature sources. After title and abstract screening, 695 records were removed, including duplicates. Of the 35 studies selected for full-text assessment, we included eight studies in this review. Refer to Figure 2 for the PRISMA flow diagram of the search and inclusion results [12]. Exclusions were mainly due to ineligible study design (23 studies), ineligible study outcomes (three studies), or the absence of results (one study).

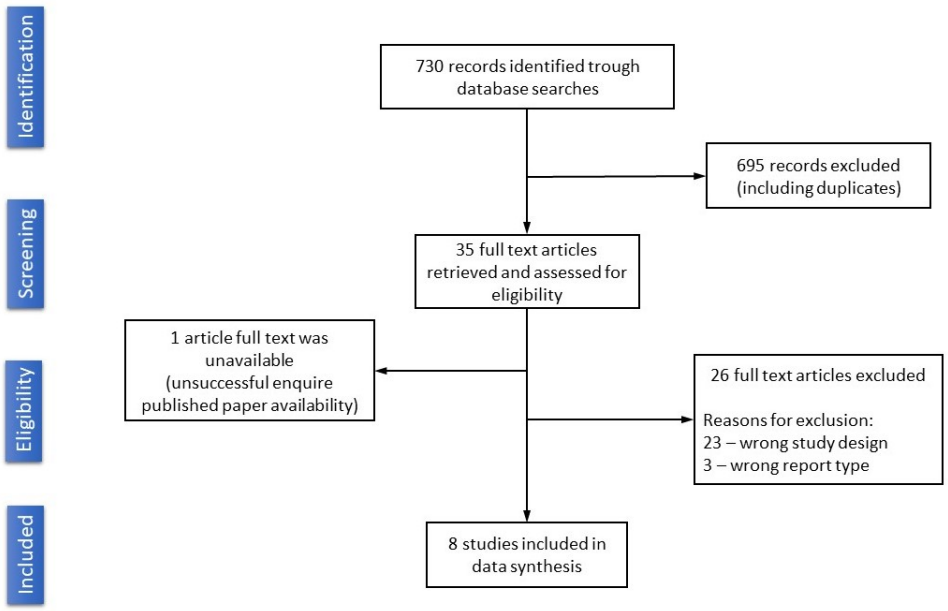

Figure 2. Flow diagram.

3.1. Description of Included Studies

We included eight studies (three journal articles and five conference papers) with a total of 355 cases, of which 273 (77%) had the diagnosis of AAA (Table 1). The studies were widely geographically distributed: three studies from the USA and one study each from

Croatia, Greece, Japan, Iran, and Malaysia. Only three studies (37.5%) reported the data origin sources. These three studies presented algorithms based on neural network (NN) approach [23,30,31], and data augmentation was used in two of these works [23,31]. Other studies proposed different non-NN models and did not report any information regarding the data source or the expertise of the data tagging specialists. The examined outcomes were variable. Four articles did not present any quantitative metrics of AI accuracy [32–35]. Two articles did not present suitable images for DSC calculation. An example of the processing for the cases with the highest and the lowest DSC is presented in Figure 3.

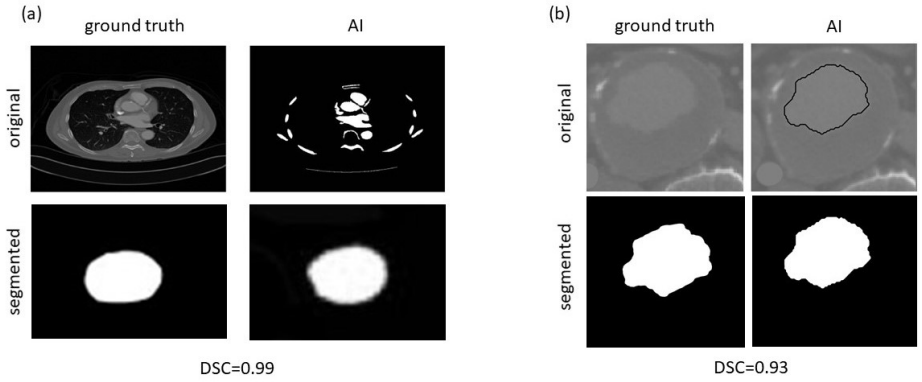

Figure 3. Image extraction for cases with the highest (**a**) and the lowest (**b**) DSC.

3.2. Dataset Characteristics

The datasets of the included studies can be grouped in several ways. More than half of the studies (62.5%) were a "single-study" (or contained a single noncontrast series). Only one study used a full noncontrast CT: a single AAA-positive case consisting of 145 slices. There were four studies that used mixed datasets (noncontrast and contrast-enhanced CT images) [23,30,32,33]. Two studies [23,30] used representative sets, consisting of 321 (with 232 slices per study on average) and 10 (each case consisted of 160 slices) studies with pathology rates of 77% and 20%, respectively. Two other studies [32,33] used single cases consisting of 170 and 40 slices, respectively. Three studies [31,34,35] did not report the contrast usage in the CT examination. The overall data contained 21 AAA-positive cases with variable slice numbers: the mean number of slices in each case was 186 [31] and was of a single-case for two others.

3.3. Findings

The mean values for the relevant outcomes were as follows: 95% (95% CI 100–87%; three studies) for the sensitivity, 96.6% (95% CI 100–75.7%; two studies) for the specificity, 95.2% (95% CI 100–54.5%; two studies) for the accuracy, and 0.91 (95% CI 0.97–0.84; two studies) for the DSC. Only two studies simultaneously reported the accuracy, DSC, and sensitivity. Four studies did not report any quantitative diagnostic metrics. It was possible to calculate the DSC for six (75%) studies, and for two of them reporting the DSC, our calculation corresponded to the author-provided values. The data on the number of TP, FP, TN, and FN cases were provided only in one study [23] (Table 2). We tried to contact other authors to clarify the missing values, but unfortunately, the necessary data were not provided (either there was no answer or the authors had no information). There were also several design drawbacks, e.g., only two publications included nonpathological cases in the testing dataset. This made the quantitative estimation of the sensitivity, specificity, and accuracy impossible. One study [36] computed the sensitivity and positive predictive value by dividing a single case into two parts. The whole set consisted of 145 noncontrast CT scans of which 111 had AAA. For training and improving accuracy, the authors used 30 and 9 manually segmented noncontrast CT scans, respectively. No other dataset was used for algorithm validation. We suppose that this approach cannot be completely satisfactory, as the near-slice connection of the ROI (aortic lumen) introduced bias to the estimates.

Table 1. Key characteristics of the studies [1].

№	1st Author (Year)	Study/Data Origin (Country)	Objectives	Type of Data Processing	Key Characteristics of Datasets	Relevant Outcomes	Calculated DSC
1	Almuntashri A. (2012) [32]	USA/-	AAA segmentation	Digital image processing algorithms	Two studies (one noncontrast case), 100% pathology rate, mixed	-	0.94
2	Fujiwara J. F. (2021) [36]	Japan/-	AAA detection and measurement	NN (not specified)	A single study, 100% pathology rate, noncontrast CT	Se 94.6%	-
3	Habijan M. (2020) [31]	Croatia/Belgium	AAA segmentation	NN (fourfold cross validation)	19 studies, 100 % pathology rate, CT type n/s	DSC 0.91 ± 0.16	0.96
4	Hosseini B. (2010) [33]	Malaysia/-	AAA detection	Non-NN (logical algorithm)	Two studies (one noncontrast case), 100% pathology rate, mixed	-	0.99
5	Kossioris G. T. (2008) [34]	Greece/-	AAA segmentation	Non-NN (level set method)	A single study, 100% pathology rate, CT type n/s	-	0.93
6	Lu J.-T. (2019) [30]	USA/USA	AAA detection	NN (fivefold cross validation)	321 studies, 77% pathology rate, mixed	Ac 92.0 %; Se 92.0%; Sp 95.0%; DSC 0.90 ± 0.05	0.99
7	Mohhamadi S. (2019) [23]	Iran/Iran	AAA segmentation and classification	Hough's algorithm and NN (fivefold cross validation)	10 studies, 20% pathology rate, mixed	Ac 98.4%; Se 98.4%; Sp 98.3%	-
8	Schei T. R. (2003) [35]	USA/-	AAA detection	Non-NN (computer algorithm)	A single study, 100% pathology rate, CT type n/s	-	0.97

[1] Note: used abbreviations: Se—sensitivity; Sp—specificity; Ac—accuracy.

Table 2. Data presence for confusion matrix arrangement [1].

№	Study First Author (Year)	Test Set Size (Images)	TP	FP	TN	FN
1	Almuntashri A. (2012) [32]	40				
2	Fujiwara J. F. (2021) [36]	9				
3	Habijan M. (2020) [31]	not stated	no information			
4	Hosseini B. (2010) [33]	170				
5	Kossioris G. T. (2008) [34]	1				
6	Lu J.-T. (2019) [30]	57				
7	Mohhamadi S. (2019) [23]	1448	357	11	1080	5
8	Schei T. R. (2003) [35]	1	no information			

[1] Note: TP—true positive, FP—false positive, TN—true negative, FN—false negative (responses).

3.4. Methodological Quality of Included Studies

The risk of bias due to the imbalanced dataset usage was high in six (75%) and low in two (25%) studies (Figure 4); the main concern was associated with the dataset imbalance in terms of the pathology and demographic ratios. The risk of bias due to concerns regarding the AI algorithm implementation was unclear in one (12.5%) and low in seven (87.5%) studies. These results were due to the fact that non-NN algorithms were used in half of the studies (thus, some QUADAS questions were irrelevant). The risk of bias due to the ground truth labeling concerns was unclear in four (50%) and low in four (50%) studies. The main concern was associated with the low clarity of the human readers' expertise. Finally, the risk of bias due to the use of heterogeneous data was unclear in two (25%) and low in six (75%) studies. The reason for the ambiguity was connected to the low detail of the data processing pathway. The weights for each study were assigned proportionally to the number of processed cases. Additional details about the risk of bias assessment are provided in Appendix B. Concerns about the applicability for all domains were low for all studies, because the authors clearly postulated the research question, and their data, index, and reference tests were prepared and performed according to the claimed task.

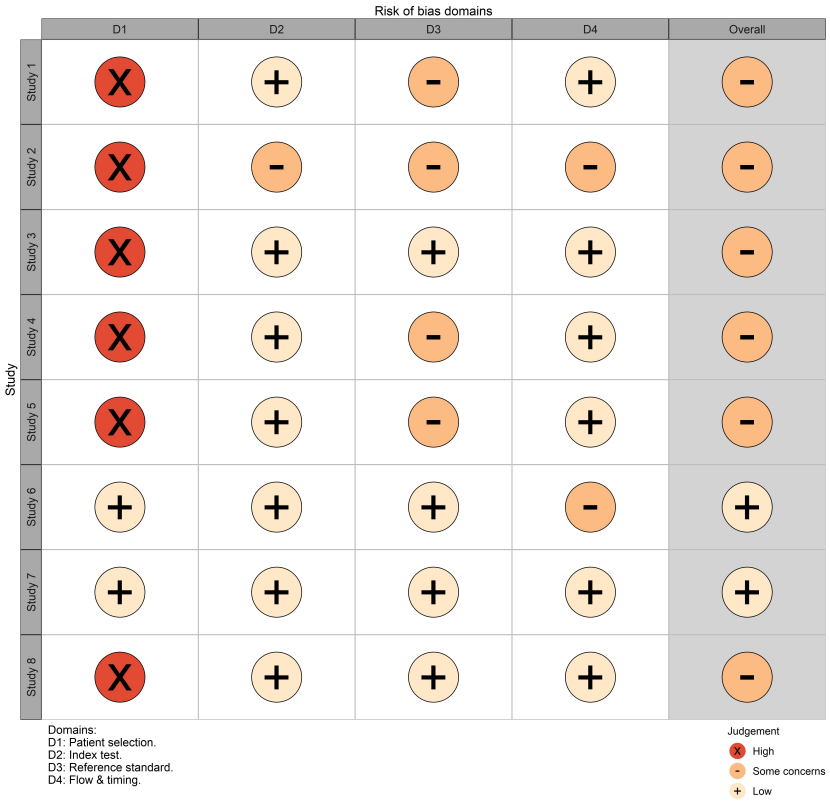

Figure 4. Risk of bias domains.

4. Discussion

This systematic review summarized the published data on the application of AI for the automatic detection of AAA on noncontrast CT images and included eight unique studies. The major findings from our review include the following:

1. The AI sensitivity for AAA detection varied from 92 to 98.4% with a mean value of 95% (95% CI 100–87%; three studies);
2. The AI specificity for AAA detection varied from 95 to 98.3% with a mean value of 96.6% (95% CI 100–75.7%; two studies);
3. The AI accuracy for AAA detection varied from 92 to 98.4% with a mean value of 95.2% (95% CI 100–54.5%; two studies);
4. The DSC for AAA segmentation varied from 0.93 to 0.99 with a mean value of 0.96 (95% CI 0.99–0.94; two studies).

Since it was possible to perform only one measurement for DSC calculation, we considered the obtained values as estimates of the mean for segmentation quality. However, we observed a discrepancy between our measurements and two provided DSC values [30,31]. For the first algorithm, the reported mean DSC value was 0.91 versus our calculated 0.96. For the second one, the reported value was 0.9 versus our 0.99. We assume that the authors may have presented the best-case scenario for their algorithms, which could differ significantly from their real-life performance. These reported estimates of segmentation accuracy could be inflated. Therefore, we encourage authors to include examples of failed or suboptimal segmentation in order to access real-world applicability of the algorithms.

The success of the application of AI for the automatic detection of AAA on CTA has been previously reported by many researchers and has been already systematically reviewed [37]. At the same time, less attention has been paid to the AI-based screening capabilities. Screening tasks are usually performed with restricted timing, without contrast enhancement, and involve big data analysis. Our purpose was to investigate whether AI was applicable for tasks of AAA detection on CT without contrast enhancement. The reported AI sensitivity (95%) for AAA detection in noncontrast CT was higher than the AAA incidental detection sensitivity by radiologists (65%) [7]. Thus, AI may have the potential for AAA opportunistic screening automatization to increase the early detection of this pathology. However, due to objective reasons, this paper was unable to conduct a complete meta-analysis of the AI diagnostic accuracy. Moreover, the methodological quality analysis revealed several significant shortcomings of the included studies, causing serious doubts about the plausibility and reproducibility of the obtained metrics. In our opinion, the lack of regulations and reporting standards may be the reasons for the AI metrics' overestimation in the original studies. To this end, STARD-AI recommendations for diagnostic accuracy studies are currently being developed [38].

4.1. Limitations of the Review

Our study had several limitations. Despite our results demonstrating the high diagnostic accuracy of AI for the automatic detection of AAA detection on a noncontrast CT, there were some concerns on the applicability and safety of the reviewed models in a clinical setting. The main reasons for the concerns were the sampling bias and the hidden stratification. Only two studies (25%) included nonpathological cases in the testing datasets. Moreover, 62.5% of included studies used a single AAA-positive CT scan to validate their algorithm, which does not allow estimation of the accuracy, specificity, and sensitivity. Because of this, we believe that the reported values of the sensitivity and specificity may be artificially high and need to be reassessed using standardized protocol and a high-quality independent testing dataset. Only a few papers met the inclusion criteria. However, the number of studies is not as important as their methodological quality: even if there were more studies, the methodological flaws and inflated diagnostic accuracy values cause doubts of the feasibility of meta-analysis. This is a well-known problem of reviews of AI studies [39] that requires regulatory attention. Perhaps consideration should be given not

only to the reporting standardization of papers on diagnostic accuracy (STARD-AI) but also to AI-specific analyses in systematic reviews of such papers.

4.2. Implications of the Results for Practice, Policy, and Future Research

Our study revealed a significant difference in the number of studies on the detection of AAA from CT images with (over 500 studies) and without (eight studies) contrast enhancement. Nevertheless, despite its objective technical complexity, we consider the task of AAA detection from noncontrast CT scans just as clinically important, and we are looking forward to obtaining the results of the pilot project on AAA opportunistic screening [40] in the Moscow Experiment on Computer Vision in Radiology [41].

5. Conclusions

The uncertainty resulting from the high or unclear risk of bias associated with the heterogeneous parameters of the datasets (pathology ratio, studies per dataset, and slices per CT scan) limit our ability to confidently draw conclusions based on our results. Moreover, all eight studies included in the analysis evaluated automated AAA detection and segmentation on noncontrast CT using different accuracy metrics. To pool the accuracy values, we developed an original approach to approximate the DSCs from the imaging data included in the studies. According to our estimates, the algorithms in the included studies demonstrated high segmentation quality (DSC 0.96 ± 0.02). However, our results overestimated the DSC values provided by the authors (0.99 versus 0.9, and 0.96 versus 0.91), indicating a trend towards showcasing only the best examples of the algorithm's performance, and the limited applicability of this approach. During the literature search, we observed an evident tendency in the published studies towards the use of contrast-enhanced scans for analysis (over 500 studies with CTA versus eight with noncontrast CT). Despite the higher task complexity of AAA detection and segmentation on noncontrast scans, it remains a promising meeting point for opportunistic screening prerequisites and practical computer vision implementation. Further studies are required, focused on balanced datasets with noncontrast CT scans and the utilization of reporting standards for satisfactory results' reproducibility.

Author Contributions: Conceptualization, M.R.K., R.V.R. and O.A.M.; methodology, M.R.K. and R.V.R.; formal analysis, M.R.K. and D.V.L.; data curation, M.R.K. and V.P.N.; software, M.R.K. and V.P.N.; writing—original draft preparation, M.R.K.; writing—review and editing, M.R.K., R.V.R., I.A.B. and D.V.L.; visualization, M.R.K.; supervision, A.V.S. and N.S.K.; project administration, Y.A.V., A.V.V. and O.V.O. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was prepared by a group of authors as a part of the medical research project (No. USIS (in the Unified State Information System for Accounting of Research, Development, and Technological Works): 122112400040-1) "Reference datasets for sustainable development of artificial intelligence-based diagnostics to minimize the long-term impact of the COVID-19 pandemic on the health of the Moscow population".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article, additional details are available upon request.

Acknowledgments: The authors would like to thank the Moscow Center for Diagnostics and Telemedicine for organizing the course on writing systematic reviews with meta-analysis.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AAA	abdominal aortic aneurysm
AI	artificial intelligence
CT	computed tomography
CTA	computed tomography angiography
DSC	Dice similarity coefficient
GT	ground truth
NN	neural network
ROI	region of interest

Appendix A. Search Queries

Appendix A.1. PubMed

"Tomography, X-ray Computed"[mh] OR ((X-Ray[tiab]) AND (Comput*[tiab]) AND (Tomograph*[tiab])) OR CT X-ray*[tiab] OR tomodensitometry[tiab] OR X-ray CT Scan*[tiab] OR Electron Beam Tomograph*[tiab] OR "Radiography, Abdominal"[mh] OR "Aortic Aneurysm, Abdominal/diagnostic imaging"[mh] OR "Dilatation, Pathologic/diagnostic imaging"[mh] OR CT[tiab]) AND ("Neural Networks, Computer"[mh] OR "Artificial Intelligence"[mh] OR "Diagnosis, Computer-Assisted"[mh] OR "Radiographic Image Interpretation, Computer Assisted"[mh] OR AI[tiab]) AND ("Aortic Aneurysm, Abdominal"[mh] OR Aortic Aneurysm, Abdominal/classification[mh] OR abdominal aortic aneurysm[tiab].

Appendix A.2. Google Scholar

allintitle:AI|deep|neural|"artificial intelligence"|algorithm segmentation|detection AAA|"abdominal aortic aneurysm".

Appendix B. QADAS-CAD

The patient selection domain questions were reoriented to data specification. The first question ("Was a consecutive or random sample of patients enrolled?") was loosely related to index test performance. Instead of this, we proposed to check whether the datasets used had any disbalance. The proposed questions were: "Were the data (training and testing sets) balanced by the severity (including the absence) of the target pathology?" and "Were the data (training and testing sets) balanced in terms of demographic factors?". The second original question ("Was a case-control design avoided?") was omitted, since the composition of a dataset is always known in retrospective AI studies. The third question ("Did the study avoid inappropriate exclusions?") remained the same, with the only appropriate exclusion being due to low image quality.

The index test domain was fully reorganized. The first question "Were the index test results interpreted without knowledge of the results of the reference standard?" was replaced by a more appropriate: "If a neural network was used, did the training and testing datasets have no intersections or resembles?". This question reflects the idea of interpretation independence, while adding AI-specific issues that can affect the outcomes of a study. We added a question on the data preparation: "If a neural network was used, was the size of each set rationalized?". The dataset size is a parameter that is significant both for the segmentation algorithm's training efficiency and for the diagnostic accuracy measurement [42]. Finally, we split the question "If a threshold was used, was it prespecified?" into two subquestions. The first subquestion was condition-specific. For example, AAA can be defined as an aortic diameter greater than 3 cm or a diameter more than 50% larger than the normal width of a healthy aorta [43]. The second subquestion was algorithm-specific. The algorithm can provide continuous data as outputs (for example, target condition probability), and in order to classify a case as positive or negative the data should be dichotomized using some threshold of test positivity. We supposed that predefinition of both mentioned thresholds was necessary to exclude bias from the study.

The reference standard domain had only one replacement. The question "Were the reference standard results interpreted without knowledge of the results of the index test?" was inconclusive, as the ground truth for AI studies is conventionally obtained before the index test performance measurements, at least in retrospective designs. Therefore, we instead proposed to check the quality of data preparation to penalize cases of inappropriate methodology: "Were the reference standard results prepared or verified with the required level of expertise?".

For the flow and timing domain, we proposed to address the properties of the data used to assess the diagnostic accuracy of the index test and reference standard. The time interval between the reference standard and index test had no significance. Both interventions require analysis of the same medical image, which is a piece of static data. However, the nontransparency of how the outcome results were obtained may raise serious doubts about both their reproducibility and their validity. That is why we proposed to include the following question, "Was there transparency in how the outcomes were generated?".

Each signal question of the four domains was expected to be answered as "yes", "no", or "unclear". The overall risk of bias for each domain was assigned as "low", "high", or "unclear". Generally, if the answer was 'yes' to all signaling questions for a domain, the risk of bias could be judged as 'low'. There were three signaling questions with the critical potential of bias introduction (Table A1 italicized). The answer "unclear" or "no" to any of these questions could render a corresponding domain as having an "unclear" or "high" risk of bias, respectively. The applicability concerns questions were left unchanged.

Table A1. QUADAS-CAD [1].

Domain	Patient Selection (D1)	Index Test (D2)	Reference Standard (D3)	Flow and Timing (D4)
Description	A description of included patient data (previous intervention, pathology rate, and severity)	A description of the index test and how it was conducted and interpreted within the context of the study	A description of the reference standard and how it was conducted and interpreted within the context of the study	A description of any difference between the index test(s) and the reference standard performance conditions
Signaling questions (yes/no/unclear)	*Were the data (training and testing sets) balanced by the severity (including the absence) of the target pathology?*	*If a neural network was used, did the training and testing datasets have no intersections or resembles?*	Is the reference standard likely to correctly classify the target condition?	Was there transparency in how the outcomes were generated?
	Were the data (training and testing sets) balanced in terms of demographic factors?	If a neural network was used, was the size of each set rationalized?	Were the reference standard results prepared or verified with the required level of expertise?	Did all patient data have the same reference standard?
	Did the study avoid inappropriate exclusions?	If a pathology threshold was used, was it prespecified? If a decision threshold (for AI) was used, was it prespecified?		Were all patient data included in the analysis?
Risk of bias (high/low/unclear)	Could the selection of patient data has introduced bias?	Could the conduct or interpretation of the index test have introduced bias?	Could the reference standard, its conduct, or its interpretation have introduced bias?	Could the patient flow have introduced bias?
Concerns regarding applicability: high/low/unclear	Are there concerns that the included patient data do not match the review question?	Are there concerns that the index test, its conduct, or interpretation differ from the review question?	Are there concerns that the target condition as defined by the reference standard does not match the review question?	

[1] Note: gray color denotes sections of proposed changes; the *italic* font denotes key questions.

Table A2. QUADAS domain questions [1].

Domain	Question	Almuntashri A. (2012) [32]	Fujiwara J. F. (2021) [36]	Habijan M. (2020) [31]	Hosseini B. (2010) [33]	Kossioris G. T. (2008) [34]	Lu J.-T. (2019) [30]	Mohhamadi S. (2019) [23]	Schei T. R. (2003) [35]
D1	*Were the data (training and testing sets) balanced by the severity (including the absence) of the target pathology?*	no	no	no	no	no	yes	yes	no
	Were the data (training and testing sets) balanced in terms of demographic factors?	no	no	unclear	no	no	unclear	yes	no
D2	Did the study avoid inappropriate exclusions?	yes	yes	yes	yes	yes	yes	yes	yes
	If a neural network was used, did the training and testing datasets have no intersections or resembles?	x	unclear	yes	x	x	yes	yes	x
	If a neural network was used, was the size of each set rationalized?	x	unclear	yes	x	x	yes	yes	x
	If a pathology threshold was used, was it prespecified?	yes	yes	yes	yes	yes	yes	yes	yes
	If a decision threshold (for AI) was used, was it prespecified?	x	unclear	unclear	x	x	unclear	unclear	x
D3	Is the reference standard likely to correctly classify the target condition?	unclear	unclear	yes	unclear	unclear	yes	yes	yes
	Were the reference standard results prepared or verified with the required level of expertise?	unclear	unclear	yes	unclear	unclear	yes	yes	yes
D4	Was there transparency in how the outcomes were generated?	yes	no	yes	yes	yes	yes	yes	yes
	Did all patient data have the same reference standard?	yes	unclear	yes	yes	yes	unclear	unclear	yes

[1] Note: for the names of the domains, see Table A1; the *italic* font denotes key questions.

Table A3. Risk of bias.

Study	1st Author (Year)	D1	D2	D3	D4	Overall	Weight (%)
Study 1	Almuntashri A. (2012) [32]	high	low	some concerns	low	high	3
Study 2	Fujiwara J. F. (2021) [36]	high	some concerns	some concerns	some concerns	high	11
Study 3	Habijan M. (2020) [31]	high	low	low	low	low	31.6
Study 4	Hosseini B. (2010) [33]	high	low	some concerns	low	high	12.9
Study 5	Kossioris G. T. (2008) [34]	high	low	some concerns	low	high	0.1
Study 6	Lu J.-T. (2019) [30]	low	low	low	some concerns	low	29.2
Study 7	Mohhamadi S. (2019) [23]	low	low	low	low	low	12.1
Study 8	Schei T. R. (2003) [35]	high	low	low	low	low	0.1

References

1. Gawenda, M.; Brunkwall, J. Ruptured abdominal aortic aneurysm: The state of play. *Dtsch. Arztebl. Int.* **2012**, *109*, 727. [CrossRef] [PubMed]
2. Erbel, R.; Aboyans, V.; Boileau, C.; Bossone, E.; Bartolomeo, R.D.; Eggebrecht, H.; Evangelista, A.; Falk, V.; Frank, H.; Gaemperli, O.; et al. 2014 ESC Guidelines on the diagnosis and treatment of aortic diseases: Document covering acute and chronic aortic diseases of the thoracic and abdominal aorta of the adult The Task Force for the Diagnosis and Treatment of Aortic Diseases of the European Society of Cardiology (ESC). *Eur. Heart J.* **2014**, *35*, 2873–2926. [CrossRef] [PubMed]
3. Mussa, F.F. Screening for abdominal aortic aneurysm. *J. Vasc. Surg.* **2015**, *62*, 774–778. [CrossRef] [PubMed]
4. Ferket, B.S.; Grootenboer, N.; Colkesen, E.B.; Visser, J.J.; van Sambeek, M.R.; Spronk, S.; Steyerberg, E.W.; Hunink, M.M. Systematic review of guidelines on abdominal aortic aneurysm screening. *J. Vasc. Surg.* **2012**, *55*, 1296–1304. [CrossRef]
5. Manning, B.J.; Kristmundsson, T.; Sonesson, B.; Resch, T. Abdominal aortic aneurysm diameter: A comparison of ultrasound measurements with those from standard and three-dimensional computed tomography reconstruction. *J. Vasc. Surg.* **2009**, *50*, 263–268. [CrossRef]
6. Tisi, P.; McLain, A.; Jeddy, T.; Ashton, H.; Scott, R. Screening for abdominal aortic aneurysm: Is opportunistic detection a realistic alternative? *Eur. J. Vasc. Endovasc. Surg.* **1998**, *15*, 532–534. [CrossRef]
7. Claridge, R.; Arnold, S.; Morrison, N.; van Rij, A.M. Measuring abdominal aortic diameters in routine abdominal computed tomography scans and implications for abdominal aortic aneurysm screening. *J. Vasc. Surg.* **2017**, *65*, 1637–1642. [CrossRef]
8. Number of Examinations with Computer Tomography (CT) in Selected Countries as of 2019. Available online: https://www.statista.com/statistics/283085/computer-tomography-examinations-in-selected-countries/ (accessed on 3 September 2021).
9. Busby, L.P.; Courtier, J.L.; Glastonbury, C.M. Bias in radiology: The how and why of misses and misinterpretations. *Radiographics* **2018**, *38*, 236. [CrossRef]
10. Cai, L.; Gao, J.; Zhao, D. A review of the application of deep learning in medical image classification and segmentation. *Ann. Transl. Med.* **2020**, *8*, 713. [CrossRef]
11. Singh, S.P.; Wang, L.; Gupta, S.; Goli, H.; Padmanabhan, P.; Gulyas, B. 3D deep learning on medical images: A review. *Sensors* **2020**, *20*, 5097. [CrossRef]
12. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gotzsche, P.C.; Ioannidis, J.P.; Clarke, M.; Devereaux, P.J.; Kleijnen, J.; Moher, D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *J. Clin. Epidemiol.* **2009**, *62*, e1–e34. [CrossRef] [PubMed]
13. Abdominal Aortic Aneurysm. Available online: https://radiopaedia.org/cases/85063 (accessed on 3 September 2021).
14. Yu, Y.; Gao, Y.; Wei, J.; Liao, F.; Xiao, Q.; Zhang, J.; Yin, W.; Lu, B. A three-dimensional deep convolutional neural network for automatic segmentation and diameter measurement of type B aortic dissection. *Korean J. Radiol.* **2021**, *22*, 168. [CrossRef]
15. ACR Data Science Institute AI Central. Available online: https://aicentral.acrdsi.org/ (accessed on 3 January 2021).
16. Bramer, W.M.; De Jonge, G.B.; Rethlefsen, M.L.; Mast, F.; Kleijnen, J. A systematic approach to searching: An efficient and complete method to develop literature searches. *J. Med Libr. Assoc. JMLA* **2018**, *106*, 531. [CrossRef] [PubMed]
17. MeSH (Medical Subject Headings). Available online: https://www.ncbi.nlm.nih.gov/mesh/ (accessed on 3 September 2021).
18. Yale MeSH Analyzer. Available online: https://mesh.med.yale.edu/ (accessed on 3 September 2021).
19. Lareyre, F.; Adam, C.; Carrier, M.; Dommerc, C.; Mialhe, C.; Raffort, J. A fully automated pipeline for mining abdominal aortic aneurysm using image segmentation. *Sci. Rep.* **2019**, *9*, 13750. [CrossRef] [PubMed]
20. Sedghi Gamechi, Z.; Bons, L.R.; Giordano, M.; Bos, D.; Budde, R.P.; Kofoed, K.F.; Pedersen, J.H.; Roos-Hesselink, J.W.; de Bruijne, M. Automated 3D segmentation and diameter measurement of the thoracic aorta on non-contrast enhanced CT. *Eur. Radiol.* **2019**, *29*, 4613–4623. [CrossRef] [PubMed]
21. Caradu, C.; Spampinato, B.; Vrancianu, A.M.; Berard, X.; Ducasse, E. Fully automatic volume segmentation of infrarenal abdominal aortic aneurysm computed tomography images with deep learning approaches versus physician controlled manual segmentation. *J. Vasc. Surg.* **2021**, *74*, 246–256. [CrossRef] [PubMed]
22. Kauffmann, C.; Tang, A.; Therasse, E.; Giroux, M.F.; Elkouri, S.; Melanson, P.; Melanson, B.; Oliva, V.L.; Soulez, G. Measurements and detection of abdominal aortic aneurysm growth: Accuracy and reproducibility of a segmentation software. *Eur. J. Radiol.* **2012**, *81*, 1688–1694. [CrossRef] [PubMed]
23. Mohammadi, S.; Mohammadi, M.; Dehlaghi, V.; Ahmadi, A. Automatic segmentation, detection, and diagnosis of abdominal aortic aneurysm (AAA) using convolutional neural networks and hough circles algorithm. *Cardiovasc. Eng. Technol.* **2019**, *10*, 490–499. [CrossRef]
24. Mendeley Reference Manager. Available online: https://www.mendeley.com/reference-management/reference-manager (accessed on 3 November 2021).
25. Dice, L.R. Measures of the Amount of Ecologic Association between Species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
26. Procreate. Available online: https://procreate.art/ (accessed on 3 November 2021).
27. RStudio: Open Source and Professional Software for Data Science Teams—RStudio. Available online: https://www.rstudio.com/ (accessed on 3 September 2021).
28. Sounderajah, V.; Ashrafian, H.; Rose, S.; Shah, N.H.; Ghassemi, M.; Golub, R.; Kahn, C.E.; Esteva, A.; Karthikesalingam, A.; Mateen, B.; et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat. Med.* **2021**, *27*, 1663–1665. [CrossRef]

29. QUADAS-2: University of Bristol. Available online: https://www.bristol.ac.uk/population-health-sciences/projects/quadas/quadas-2/ (accessed on 3 October 2021).
30. Lu, J.T.; Brooks, R.; Hahn, S.; Chen, J.; Buch, V.; Kotecha, G.; Andriole, K.P.; Ghoshhajra, B.; Pinto, J.; Vozila, P.; et al. DeepAAA: Clinically Applicable and Generalizable Detection of Abdominal Aortic Aneurysm Using Deep Learning. In *Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2019; pp. 723–731. [CrossRef]
31. Habijan, M.I.; Galić, H.L.K.R.; Babin, D. Abdominal aortic aneurysm segmentation from ct images using modified 3d u-net with deep supervision. In Proceedings of the 2020 International Symposium ELMAR, Zadar, Croatia, 14–15 September 2020; pp. 123–128. [CrossRef]
32. Almuntashri, A.; Finol, E.; Agaian, S. Automatic lumen segmentation in CT and PC-MR images of abdominal aortic aneurysm. In Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Seoul, Republic of Korea, 14–17 October 2012; pp. 2891–2896. [CrossRef]
33. Hosseini, B.; Mashak, S.V.; Majd, E.M.; Sheikh, U.U.; Abu-Bakar, S. Automatic segmentation of abdominal aortic aneurysm using logical algorithm. In Proceedings of the 2010 Fourth UKSim European Symposium on Computer Modeling and Simulation, Pisa, Italy, 17–19 November 2010; pp. 147–151. [CrossRef]
34. Kossioris, G.; Papaharilaou, Y.; Zohios, C. Detection of lumen, thrombus and outer wall boundaries of an abdominal aortic aneurysm From 2D medical images using level set methods. In Proceedings of the ASME Summer Bioengineering Conference, Marco Island, FL, USA, 25–29 June 2008; pp. 25–29. [CrossRef]
35. Schei, T.R.; Barrett, S.; Jones, D.; Krupski, W. Automated Abdominal Aortic Aneurysm segmentation using MATLAB. *Biomed. Sci. Instrum.* **2003**, *39*, 53–58. [PubMed]
36. Fujiwara, J.; Orii, O.; Araki, K.; Ogura, M.; Ito, T.; Oyamada, K.; Morino, Y.; Yoshioka, K. Fully automatic detection and measurement of abdominal aortic aneurysm using artificial intelligence. *Eur. Heart J.* **2021**, *42*, ehab724.3070. [CrossRef]
37. Lareyre, F.; Adam, C.; Carrier, M.; Raffort, J. Artificial intelligence and automatic segmentation of abdominal aortic aneurysm: Past, present, and future. *J. Vasc. Surg.* **2021**, *74*, 347–348. [CrossRef] [PubMed]
38. Sounderajah, V.; Ashrafian, H.; Golub, R.M.; Shetty, S.; De Fauw, J.; Hooft, L.; Moons, K.; Collins, G.; Moher, D.; Bossuyt, P.M.; et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: The STARD-AI protocol. *BMJ Open* **2021**, *11*, e047709. [CrossRef] [PubMed]
39. Oakden-Rayner, L.; Dunnmon, J.; Carneiro, G.; Re, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In Proceedings of the ACM Conference on Health, Inference, and Learning, Toronto, ON, Canada, 2–4 April 2020; pp. 151–159. [CrossRef]
40. The Moscow Experiment on CV in Radiology. Available online: https://mosmed.ai/ (accessed on 3 January 2021).
41. Morozov, S.; Vladzymyrskyy, A.; Ledikhova, N. Moscow experiment on computer vision in radiology: Involvement and participation of radiologists. *Vrach I Inf. Tehnol.* **2020**, *20*, 14–23. [CrossRef]
42. Orlando, N.; Gyacskov, I.; Gillies, D.J.; Guo, F.; Romagnoli, C.; D'Souza, D.; Cool, D.W.; Hoover, D.A.; Fenster, A. Effect of dataset size, image quality, and image type on deep learning-based automatic prostate segmentation in 3D ultrasound. *Phys. Med. Biol.* **2022**, *67*, 074002. [CrossRef]
43. Thresholds for Abdominal Aortic Aneurysm Repair: Abdominal Aortic Aneurysm: Diagnosis and Management: Evidence Review. Available online: https://www.ncbi.nlm.nih.gov/books/NBK556917/ (accessed on 3 January 2021).