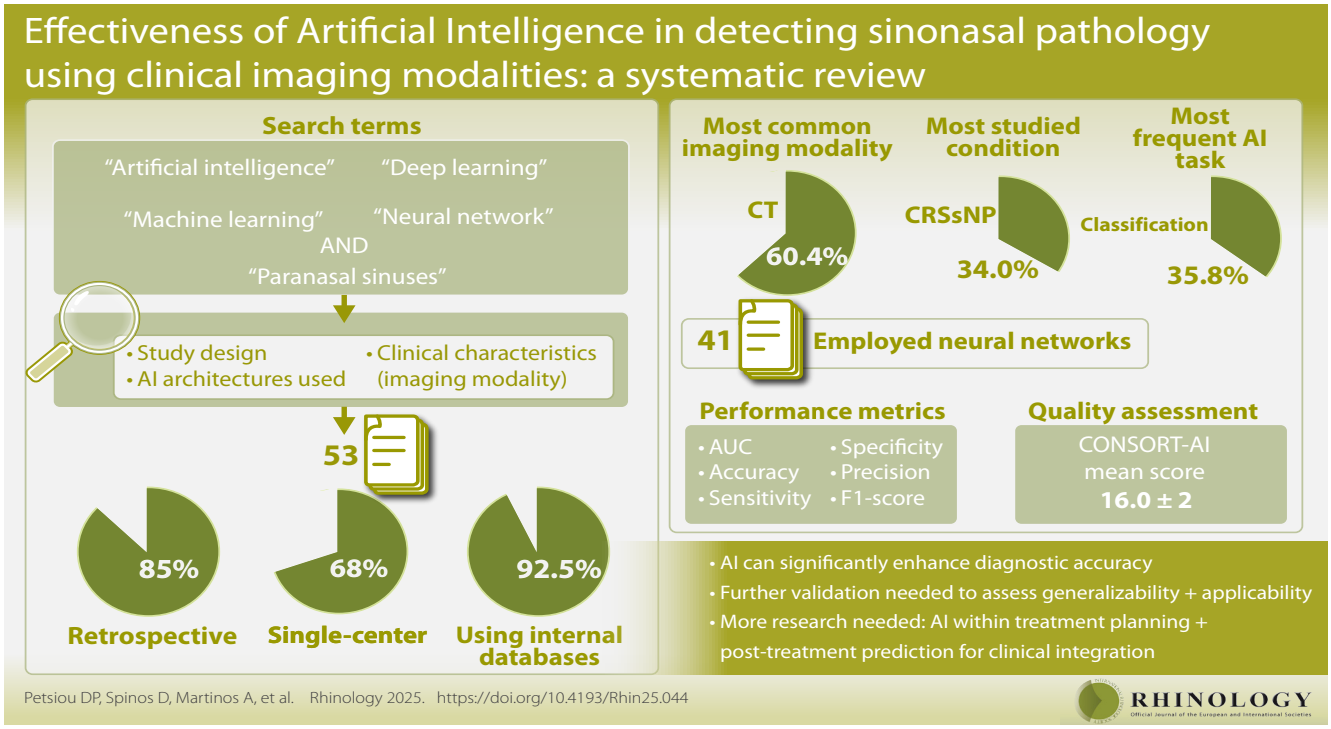
Effectiveness of Artificial Intelligence in detecting sinonasal pathology using clinical imaging modalities: a systematic review

Dioni-Pinelopi Petsiou[1], Dimitrios Spinos[2], Anastasios Martinos[1], Jameel Muzaffar[3], Georgios Garas[4,5], Christos Georgalas[6,7]

Rhinology 63: 4, 0 - 0, 2025

https://doi.org/10.4193/Rhin25.044



Abstract

Background: Sinonasal pathology can be complex and requires a systematic and meticulous approach. Artificial Intelligence (AI) has the potential to improve diagnostic accuracy and efficiency in sinonasal imaging, but its clinical applicability remains an area of ongoing research. This systematic review evaluates the methodologies and clinical relevance of AI in detecting sinonasal pathology through radiological imaging. **Methodology**: Key search terms included "artificial intelligence," "deep learning," "machine learning," "neural network," and "paranasal sinuses,". Abstract and full-text screening was conducted using predefined inclusion and exclusion criteria. Data were extracted on study design, AI architectures used (e.g., Convolutional Neural Networks (CNN), Machine Learning classifiers), and clinical characteristics, such as imaging modality (e.g., Computed Tomography (CT), Magnetic Resonance Imaging (MRI)). **Results**: A total of 53 studies were analyzed, with 85% retrospective, 68% single-center, and 92.5% using internal databases. CT was the most common imaging modality (60.4%), and chronic rhinosinusitis without nasal polyposis (CRSsNP) was the most studied condition (34.0%). Forty-one studies employed neural networks, with classification as the most frequent AI task (35.8%). Key performance metrics included Area Under the Curve (AUC), accuracy, sensitivity, specificity, precision, and F1-score. Quality assessment based on CONSORT-AI yielded a mean score of 16.0 ± 2. **Conclusions**: AI shows promise in improving sinonasal imaging interpretation. However, as existing research is predominantly retrospective and single-center, further studies are needed to evaluate AI's generalizability and applicability. More research is also required to explore AI's role in treatment planning and post-treatment prediction for clinical integration.

Key words: Sinonasal pathology, artificial intelligence, medical imaging, diagnosis, systematic review

Corrected Proof

Introduction

The sinonasal complex includes the nasal cavity and paranasal sinuses, which are further subdivided into 4 anatomic components: maxillary, ethmoid, frontal, and sphenoid sinuses [1]. The proximity of the paranasal sinuses to critical structures such as the orbit, anterior and middle cranial fossae, cranial nerves, and the internal carotid artery emphasizes the importance of accurate imaging interpretation, particularly in cases involving invasive disease. Sinonasal pathology encompasses a wide range of conditions, varying from benign inflammatory disorders, such as chronic rhinosinusitis, to malignant tumors [2]. While imaging techniques, such as CT and MRI are widely used for diagnosing such conditions, their interpretation is often prone to inter-observer variability and can lead to diagnostic delays [3-5].

Given these challenges, Artificial Intelligence (AI), particularly machine learning (ML) and deep learning (DL) models, has emerged as a potential solution to improve diagnostic accuracy and efficiency, reduce physicians' workload, and minimize diagnostic errors [5]. ML involves algorithms that learn from data to generate predictions, while the more advanced form of ML, DL, utilizes neural networks with multiple layers to automatically extract features from complex data, especially images. Among the most widely used DL models are convolutional neural networks (CNNs), which are particularly effective in image-related tasks like segmentation and classification [6].

Despite advancements in AI applications, several knowledge gaps persist in this domain, including a lack of studies focused on developing AI systems that can integrate information from multiple imaging techniques (e.g., CT, MRI, endoscopy), the absence of established regulatory frameworks for evaluating and deploying AI algorithms in clinical practice, limited multi-institutional collaborations, insufficient AI research in surgical planning and assessment for rhinology, the limited application of AI in predicting long-term patient outcomes, disease recurrence, and treatment responses, as well as the often unaddressed bias in the original data sets used for AI training [7, 8]. In addition, AI in sinonasal imaging has more room for development and remains underexplored compared to other medical fields, such as oncology, due to the region's complex anatomy and the relatively limited application of AI in sinonasal conditions compared to oncological care [7].

This systematic review aims to provide a descriptive summary and evaluation of AI methodologies designed to detect sinonasal pathology using radiological imaging. Specifically, it comprehensively analyses the clinical design of original studies in this domain and assesses their quality. The reviewed literature successfully demonstrates the potential and limitations of AI applications in supporting clinicians' work, which will be further explored throughout the manuscript.

Materials and methods

Study protocol

For this systematic review, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [9] were adhered to. The study protocol was prospectively registered in the PROSPERO database (CRD42024596954). Since no original human or animal research was conducted and thus the study relied solely on published anonymized data, ethical committee approval and patient consent were not required. The literature search was limited to human studies.

Inclusion and exclusion criteria

All original studies involving human subjects that employed AI techniques in tasks such as segmentation, interpretation, classification, or other relevant applications on sinonasal radiological images were included. For clarity, the AI tasks identified in the included studies were categorized as follows: classification, which involves assigning labels to imaging data (e.g., distinguishing between benign and malignant lesions); segmentation, which delineates anatomical structures or pathological regions within the sinonasal complex; diagnosis, which aids in detecting and characterizing sinonasal pathology; identification, which recognizes specific features or anomalies within the imaging data; treatment planning, which assists clinicians in selecting optimal therapeutic strategies based on imaging analysis; and post-treatment course prediction, which forecasts disease progression, recurrence, or treatment outcomes based on AI-driven analysis [10].

Studies were excluded if they met any of the following criteria: (a) studies not describing AI applications in human sinonasal radiological imaging; (b) abstract or full text unavailable; (c) articles not published in English; (d) narrative or systematic reviews, case reports, book chapters, preprint articles, commentaries, or conference papers.

The selection criteria for this study were structured using the PICOTS model, which considers the following elements:

- Population (P): Human studies applying AI algorithms for the interpretation and analysis of paranasal sinus radiological images.
- Intervention (I): Any AI-based methodology employed for radiological imaging interpretation and analysis.
- Comparison (C): Imaging analysis and diagnostic accuracy as performed by clinicians.
- Outcomes (O): Primary outcomes included classification and validation. Secondary outcomes encompassed identification and diagnostic accuracy, treatment planning, and predictions of the post-treatment course.
- Timing (T): No restrictions on the study period, including both retrospective and prospective studies.

Corrected Proof

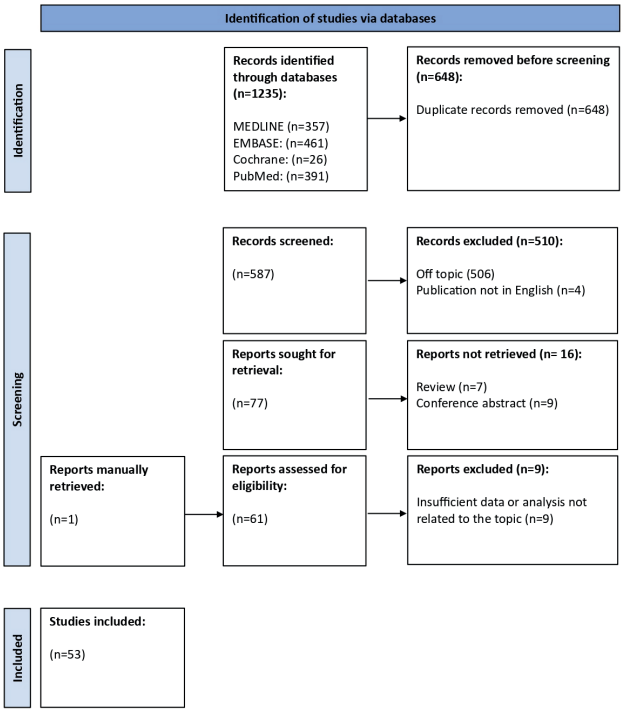


Figure 1. Flowchart of the search methodology and study selection process in adherence to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines

- Setting (S): Clinical and research settings involving radiological imaging of paranasal sinuses

Information sources and database search

A comprehensive literature search was conducted using PubMed, MEDLINE, EMBASE, and the Cochrane Library databases. The final search was performed on December 10, 2024. All retrieved articles were reviewed based on the predefined eligibility criteria. The search included both Medical Subject Headings (MeSH) and free-text terms related to AI applications in radiological imaging of sinonasal pathology, specifically: "artificial intelligence," "deep learning," "machine learning," "neural network," "paranasal sinuses," "maxillary sinus," "frontal sinus," "ethmoid sinus," "sphenoid sinus," "maxilla," "imaging," "scan," "computed tomography," "magnetic resonance imaging," and "X-ray."

Study selection and data retrieval

The full-text screening was carried out independently by two authors. Any conflicts were addressed and resolved by the senior author. A detailed description of the search methodology and the study inclusion procedure is provided in the PRISMA flowchart (Figure 1). The following data were extracted: (a) study characteristics (name of first author and year of publication, dataset, type of AI used, evaluation metrics, and main outcomes); (b) study design (data collection methods, study centres involved, database sources, and performance comparisons); (c) clinical characteristics (imaging modality, conditions or specific pathologies of interest); and (d) AI approach (AI architecture and task, learning approach, data augmentation, and validation process).

Quality assessment

To assess the quality of the studies, the CONSORT-AI extension was applied, which provides guidelines for clinical trials utilizing AI architectures [11]. Building on the approach adopted by Gray et al. in 2012 [12], who used the 2001 version of the CONSORT statement [13] to develop a 30-point checklist that assigned one point to each item to evaluate trial reporting, we developed a 20-point scoring system. Each evaluation parameter was assigned equal weight, resulting in a total score of 20 per study. All criteria were based on the CONSORT-AI extension, with one additional item included in the methods section. This extra item addressed the validation methodology of AI applications, aiming to clarify a key aspect of model performance and generalization. CONSORT-AI elements that were not relevant to the dataset were excluded. The highest possible score for a study was 20 (indicating the highest quality), and the lowest possible was 0 (indicating the lowest quality). The variability in study quality will be demonstrated by comparing the highest and lowest scores numerically, and by calculating the mean and standard deviation of all quality scores. The 20-point scoring system is outlined in Table 1. While the original CONSORT statement, first published in 1996 and periodically updated, does not recommend using the checklist to assign numerical scores [14, 15], this approach has proven valuable in providing a comprehensive review of specific reporting items and offering an overall assessment of reporting standards.

Results

Included articles

The literature search identified 1,235 records in total. After removing duplicates, two reviewers independently conducted the screening process. Title and abstract screening of the remaining 587 records led to the exclusion of 510 articles. Subsequently, 77 articles were sought for retrieval, of which 61 met the criteria for full-text evaluation. Additionally, 1 more publication was manually retrieved. Following the full-text screening, 53 studies met the predefined inclusion criteria and were included in the analysis [16 – 68]. Further details on the article selection methodology can be found in Figure 1, with a summary of included studies provided in Table 2. In Table 2, key performance differences are highlighted in bold text.

Study design

Information about the study design is summarized in Table 3. Most studies (45, 85.0%) adopted a retrospective design, while 8 (15.1%) were prospective. Regarding study center involvement,

Corrected Proof

Table 1. A 20-point scoring system derived from the CONSORT-AI extension.

Article section	Assessment item	Description	N of studies reporting item (%)
Title and abstract	Summary structure	Declare the use of the AI application within the study context in the title and/or abstract	53 (100.0)
Introduction	Background and objectives	Describe the intended use of the AI application within the clinical pathway (purpose, intended users etc.)	52 (98.1)
Methods	Eligibility criteria	Specify the inclusion and exclusion criteria for participants	46 (86.8)
		Specify the inclusion and exclusion criteria for input data	45 (84.9)
	Interventions	Specify the version of the applied AI architecture	42 (79.2)
		Explain how the input data were obtained and selected for the AI application	50 (94.3)
		Explain how low-quality or missing data were evaluated and handled	23 (43.4)
		State whether there was a human-AI collaboration in the management of the input data	43 (81.1)
		Define the output of the AI application	53 (100.0)
	Validation	Explain the validation methodology used to assess the AI architecture	51 (96.2)
	Sample size	Explain how the sample size was established and justified	18 (34.0)
	Analysis	Declare the statistical methods used for data analysis	51 (96.2)
Results	Participants	Specify the number of participants who were randomly selected and assessed for the primary outcome	33 (62.3)
		Mention the baseline demographic and clinical features of the participants	38 (71.7)
	Outcomes	Report the results for each outcome (and its precision, e.g. 95% confidence interval)	51 (96.2)
		Report any performance errors and how they were retrieved and analyzed. If not planned or performed, explain the reasons	30 (56.6)
Discussion	Limitations	Explain the limitations of the study (causes of potential bias, imprecisions etc.)	48 (90.6)
	Generalizability	Explain the generalizability (external validity, applicability) of the study outcomes	27 (50.9)
	Interpretation	Present an interpretation in accordance with the results, discuss the advantages and disadvantages and examine other relevant data	52 (98.1)
Other	Funding	Declare any sources of financial and other support	40 (75.5)

36 studies (68.0%) were single-center, 13 (24.5%) were multicenter, and 4 (7.5%) did not specify the center type. In terms of data sources, the majority (49, 92.5%) used internal databases. Only 2 studies (3.8%) relied on public databases, while 1 study (1.9%) combined internal and public databases, and another (1.9%) did not specify the data source. Performance comparisons varied across studies: 15 (28.3%) compared AI systems to other AI models, 30 (56.6%) to human performance, and 1 (1.9%) to both AI and human performance. Additionally, 4 studies (7.5%) compared AI to traditional algorithmic methods, and 3 (5.7%) did not specify the comparison type.

Clinical approach

Among the included studies (n=53), the most used imaging modality was CT, utilized in 32 studies (60.4%). MRI was employed in 15 studies (28.3%), while nasopharynx X-ray, alone or in combination with CT, was used in 6 studies (11.4%). Regarding the main conditions or pathologies of interest, 9 studies (17.0%) focused on chronic rhinosinusitis without nasal polyposis (CRSsNP), and 4 studies (7.5%) addressed chronic rhinosinusitis with nasal polyposis (CRSwNP). Eosinophilic CRS (eCRS) was investigated in 2 studies (3.8%), sinus fungal ball (mycetoma) rhinosinusitis in 3 studies (5.7%), and sphenoid sinus pneumatization in 1 study (1.9%). Inverted Schneiderian papilloma and/or malignant sinonasal neoplasms were examined in 13 studies (24.5%), maxillary sinusitis in 5 (9.4%), paranasal sinus mucosal abnormalities in 7 (13.2), and other sinonasal pathologies in 9 studies (17.0%). Table 4 provides a summary of the above information.

AI methodology

Key points of the AI methodology are presented in Table 5. Among the included studies (n=53), CNNs were the most utilized AI architecture, featured in 39 studies (73.6%). Artificial Neural Networks (ANNs) and other neural networks were each used in 1 study (1.9%), while various other deep learning algorithms were applied in 2 studies (3.8%). Machine Learning (ML) classifiers were employed in 10 studies (18.9%). Regarding AI tasks, classification was the most frequently reported, appearing in 19 studies (35.8%), followed by diagnosis in 11 studies (20.8%)

Corrected Proof

Table 2. Summary of all included studies.

Author, Year	Dataset	Type of AI	Evaluation metrics	Main outcomes
Peng et al., 2024 [16]	Axial CT imaging data from 107 patients	DL; two-stage model (VGG19 and Faster-RCNN with FPN)	Accuracy, mAP, AP50	Classification accuracy: 92.7%; Detection mAP: 0.7; AP50: 0.8;
Zou et al., 2024 [17]	CT scans and biopsy results of 192 patients	Modified ResNet	AUC, Accuracy, Recall, Precision, F1-Score, Confusion Matrix	AUC: 99.1%, Accuracy: 96.5%, F1-Score: 97.0%
Lai S et al., 2024 [18]	22,265 CT images from 192 patients	ResNet-18	Accuracy, Precision, Recall, Specificity, AUC	Accuracy: 98.4%, Precision: 98.1%, Recall: 98.1%, Specificity: 98.7%, AUC: 98.4%
Kwon et al, 2024 [19]	1080 CT images (2158 maxillary sinuses)	YOLOv8	Precision, F1-Score, mAP, IoU	Overall precision: 97.1%; F1-score: 95.4%; Precision on difficult dataset: 92.4%
Bhattacharya et al., 2024 [20]	2619 participants with cranial MRI scans	3D CNN (DenseNet)	AUC, Sensitivity, Specificity	AUC: 95.0%; Sensitivity: 85.0%; Specificity: 90.0%
Du et al., 2024 [21]	29,993 CT images from patients with CRSwNP	ResNet-18	Accuracy, Recall, Precision, Confusion Matrix, ROC Curve, AUC, Kappa Score	AUC: 99.3% (training), 96.6% (validation), 96.3% (testing); Kappa: 98.5% (training), 92.8% (validation), 92.2% (testing); Overall AUC: 96.2%
Whangbo et al., 2024 [22]	39,605 paranasal CT scans from 201 patients	3D U-Net variations	F1 score, True positive rate	F1 score: 84.3% (normal test set), 79.3% (abnormal test set)
Wang et al., 2024 [23]	MRI scans from 1711 adult patients	ML classifiers	AUC, Accuracy, NPV, PPV, Sensitivity, Specificity, Kappa value	AUC: 94.7% (training), 84.9% (validation), 87.1% (test1), 88.7% (test2) for the fusion model (T1WI + T2WI + CE-T1WI); highest AUC achieved in fusion model;
Gudapati et al., 2024 [24]	548 axial CT images from three FESS candidates	ML classifiers	IoU, DSC	Soft tissue DSC: 94.0%-98.0%; Bone DSC: 30.0-66.0%; IoU for soft tissue: 89.0-97.0%; IoU for bone: 44.0-49.0%
Cheong et al., 2024 [25]	OASIS-3 MRI dataset	AutoML	Precision, Sensitivity, Accuracy	Sensitivity: 91.3%; Precision: 92.8%; Accuracy: 92.0%
Maria Jesi et al., 2023 [26]	N/S	CNN	Accuracy, F-measure, Specificity, Sensitivity	CNN classifier: 99.0% accuracy; 98.7% Sensitivity; 98.9%
Celebi et al., 2023 [27]	298 CBCT images	Res-Swin-UNet	F1-score, Accuracy, IoU	F1-score: 91.7%, Accuracy: 99.0%, IoU: 84.7%; The model outperforms state-of-the-art models
Park et al., 2024 [28]	MRI from 68 patients	ML radiomics and classifiers	AUC	AUC for radiomics model: 83.8%, AUC for combined model (clinical and radiomics): 85.0%
Massey et al., 2024 [29]	CT scans from 84 CRS patients	CNN	AUC	AUC for OMC obstruction classification: 79.0% (left) and 77.0% (right)
Lin et al., 2024 [30]	MRI scans from 231 SNSCC patients	ML and DL models	AUC	AUC for RS-DLR in the test set: 81.7%. RS-DLR outperformed RS-DTL and RS-HC in the training cohort ($p < 0.050$)
Bhattacharya et al., 2023 [31]	Maxillary Sinuses MRI scans of 299 patients	CNN	AUPRC	Sampling and MIE improved performance by 21.9 ± 11.9% and 4.3 ± 5.0%, respectively. Sampling and MIE combined increased performance by 28.9 ± 12.8% and 9.9 ± 4.0%, respectively
Xiong et al., 2024 [32]	CT scans from 437 patients	ML classifiers	AUC, Sensitivity, Specificity, Negative Likelihood Ratio, Positive Likelihood Ratio, Calibration Curve, Brier Score	AUC: 89.0%, Sensitivity: 81.0% Specificity: 75.0%, Positive Likelihood Ratio: 3.2, Negative Likelihood Ratio: 0.3, Brier Score: 0.1
Bhattacharya et al., 2024 [33]	MRI scans from 1067 labelled and 1559 unlabeled patients	3D CAE; U-Net-inspired architecture with ResNet18 backbone	AUPRC	AUPRC: 79.0% on 10.0% of the annotated dataset; outperforming Self-Supervised Learning Methods

Corrected Proof

Table 2 *continued*. Summary of all included studies.

Author, Year	Dataset	Type of AI	Evaluation metrics	Main outcomes
Sukswai et al., 2024 [34]	CT from 1539 adult CRS patients	DL models	Accuracy, Sensitivity (Recall), Specificity, Precision, F1-score, ROC AUC, Kappa agreement	MobileNetV3: 81.0% accuracy, 47.4% sensitivity, 87.9% specificity, 66.8% precision, 67.2% F1-score. YOLOv8X-SEG: 94.1% accuracy, 85.9% sensitivity, 95.8% specificity, 88.9% precision, 89.8% F1-score. Rhinologist: 93.5% accuracy, 84.6% sensitivity, 95.3% specificity, 78.6% precision, 81.5% F1-score.
Zeng et al., 2023 [35]	CBCT images (200 test set, total dataset split 7:2:1)	DL models	AUC, AUPRC, Accuracy, Precision, Recall, Specificity, F1-score, McNemar Test, Kappa Agreement	AUC: 95.3%, AUPRC: 88.7%. Accuracy at optimal cut-off: >90%. Dentist-model comparison: Model outperforms dental students.
Taylor et al., 2023 [36]	Total 898 CT slices from 462 patients	CNN	Accuracy, CI	Overall weighted accuracy: 85.9%. Confidence intervals for each category: 87.1–97.0, 79.9–92.7, and 78.0–91.4, respectively
Lin et al., 2023 [37]	MRI scans from 265 SNSCC patients	DL segmentation model	DC, AUC, Accuracy	RS-Combined: AUC 85.4% and accuracy 74.3%; DC: 72.0% (T2WI), 72.7% (T1c), 75.6% (ADC)
Ha et al., 2023 [38]	426 maxillary sinuses (213 patients)	CNN	Accuracy, Sensitivity, Specificity	Accuracy (healthy sinuses classification): 98.0%; Accuracy (cysts or tumors classification): 90.0%
Yoo et al., 2023 [39]	CBCT images of maxillary sinuses from 67 patients	U-Net	Jaccard coefficient, DSC, Precision, Recall	Best performance for maxillary sinus lesions segmentation: Jaccard = 78.7%, Dice = 87.5%, Precision = 89.7%, Recall = 85.8%
He et al., 2023 [40]	CT images from 265 CRS patients (2 centres)	Multi-task DL	AUC, DSC	Segmentation (DSC 83.3%); Recurrence Prediction (AUC 74.2%)
Zhou et al., 2022 [41]	CT images from 109 CRSwNP patients	ANN	AUC, Sensitivity, Specificity	AUC: 97.6% (ANN 4 features), 97.0% (ANN 15 features), Better than LR models (AUC: 90.2% and 84.5%)
Hua et al., 2022 [42]	CT images from 878 CRS patients	CNN	Dice coefficient, AUC, Accuracy, Confusion matrix, Grad-CAM interpretability	AUC for classification models: 84.8% (single image), 85.3% (per patient); Mean accuracy: 76.2% (single image), 89.3% (per patient)
Zhang et al., 2023 [43]	133 MRI scans	ML models	Sensitivity, specificity, accuracy, precision, F1 score, AUC-ROC, AUC-PRC	SVM model with 7 features: specificity 90.3%, accuracy 90.0%, precision 72.7%, F1 score 80.0%, AUC-PRC 91.9%, sensitivity 88.9%. Outperformed radiology residents ($P < 0.05$), but not experienced radiologists ($P > 0.05$)
Kim et al., 2023 [44]	Pseudo-CBCT data; Internal dataset (n = 512)	CNN	Micro-average AUC, macro-average AUC, accuracy, sensitivity, precision, F1 score	Proposed method improved micro-average AUC by 7.4%, macro-average AUC by 5.6%, accuracy by 9.6%, and human diagnosis accuracy by 11%
Sun et al., 2022 [45]	MRI scans from 1048 patients	ResNet50	AUC, Accuracy, Delong test	Model combining tumor and peritumor ROIs using multimodal images: AUC: 88.4%, ACC: 78.1%
Hung et al., 2022 [46]	445 CBCT scans (890 maxillary sinuses)	CNN	AUC, DSC	Mucosal thickening: AUC 91.0% (low-dose), 89.0%(full-dose) and DSC 72.9% (low-dose), 66.3% (full-dose); Mucosal retention cysts: AUC 84.0% (low-dose), 93.0% (full-dose) and DSC 67.8% (low-dose), 78.7%(full-dose)
Kong et al., 2022 [47]	Paranasal sinus X-ray images (890 maxillary sinuses)	CNN	Accuracy, Sensitivity, Specificity, F1-score, PPV, NPV, AUC	Best performance with GAN-based data: AUC 92.4%, accuracy 83.3%, sensitivity 87.9%, specificity 78.8%, F1-score 84.1%, PPV 81.0%, NPV 86.7%
Lim et al., 2022 [48]	587 PNS series from 279 males and 308 females	CNN	AUC	AUC: 72.2% (sinusitis classification), AUC: 75.0% (left maxillary sinusitis), AUC: 70.0% (right maxillary sinusitis)

Corrected Proof

Table 2 *continued*. Summary of all included studies.

Author, Year	Dataset	Type of AI	Evaluation metrics	Main outcomes
Morgan et al., 2022 [49]	264 sinuses from CBCT images	3D U-Net	DSC, Inter-observer reliability	Automated segmentation of maxillary sinus: Time: 0.4 min (automatic) vs. 60.8 min (semi-automatic), DSC: 98.4%, Inter-observer DSC: 99.6%
Liu et al., 2022 [50]	90 patients with 446 MRI images	CNN	Sensitivity, Specificity, Accuracy, AUC	Sensitivity: 66.7%, Specificity: 81.5%, Accuracy: 77.9%, AUC: 80.0%
Serindere et al., 2022 [51]	148 healthy and 148 inflamed sinus images (PRs and CBCTs)	CNN	Accuracy, Sensitivity, Specificity, AUC, PPV, NPV	Accuracy (PRs): 75.7%, Sensitivity (PRs): 75.7%, Specificity (PRs): 75.7%; Accuracy (CBCT): 99.7%, Sensitivity (CBCT): 100%, Specificity (CBCT): 99.3%
Beswick et al., 2022 [52]	30 PwCF with CRS (25 completed study)	DL models	Change in sinus CT opacification (%SO), SNOT-22, Health utility, Productivity loss	% SO improvement: 22.9%, SNOT-22 improvement: 15.3, Health utility improvement: 0.068, Productivity loss improvement (all $p < 0.049$)
Li et al., 2022 [53]	3382 CT slices from 136 patients	CNN	Accuracy, AUC	Accuracy: 88.4%, AUC: 87.0%
Gu et al., 2022 [54]	MRI data from 247 patients	ML classifiers	AUC, Calibration Curves	AUC for T2WI-SVM model: 87.8% and 91.4% for test sets. Combined model AUC: 91.2% and 92.7%. Combined model outperformed clinical model ($P = 0.011, 0.005$)
Corino et al., 2022 [55]	T1 and T2 Weighted MRIs from 50 sinonasal cancer patients	ML classifiers	AUC, Accuracy	AUC for T1: 79.0%, T2: 76.0%, ADC: 93.0%
Kuo et al., 2022 [56]	175 CT sets, 50 participants	CNN	DC, MioU, Pixel Accuracy, ROC-AUC	Dice coefficient: 91.57%, MioU: 89.43%, Pixel accuracy: 99.75%
Nakagawa et al., 2022 [57]	168 lesions with malignant nasal or sinonasal tumors	CNN	Accuracy	CNN Accuracy: 92.0%. Radiologists' accuracy with model assistance: 94.0% and 100.0% in second reading
Qi et al., 2021 [58]	660 CT images (training and validation), 200 images (testing)	CNN	DSC	CNN: Dice improvement of 25.0% over FLS and 12.0% over CRF-FCN
Jung et al., 2021 [59]	83 CBCT dental volumes	3D nnU-Net	DSC	DL improved segmentation accuracy for air (DSCs: 92.0%, 92.5%, 93.0%) and lesion (DSCs: 77.0%, 75.0%, 76.0%)
Jeon et al., 2021 [60]	CT images from 1535 patients	CNN	AUC	AUC for maxillary sinusitis: 88.0%, for ethmoid: 78.0%, and for frontal: 71.0%. Multi-view model outperformed single view for maxillary sinusitis ($p = 0.038$)
Ramakrishnan et al., 2021 [61]	Ct images from 611 participants	ML classifiers	AUC, Sensitivity, Specificity	SVM: AUC 75.4%, highest sensitivity
Chen et al, 2021 [62]	164 patients, 271 MRI features	ML classifiers	AUC, Accuracy, Sensitivity, Specificity	AUC: 100.0% (train), 96.5% (validation), 97.9% (test). Accuracy: 89.0%, Sensitivity: 88.0%, Specificity: 92.0%
Kuo et al., 2019 [63]	79 CT scan images	BPNN	Accuracy, Sensitivity	Accuracy: 96.3%, Sensitivity: 95.1%
Humphries et al., 2020 [64]	690 CT scans	CNN	Spearman correlation with LM scores	CNN scores and LM scores showed strong positive correlation ($rho=0.82$, $p<0.001$)
Kim et al, 2019 [65]	9000 Waters' view radiographs from 60,389 patients	CNN	AUC, Sensitivity, Specificity, Cohen kappa coefficient	AUC: 93.0% for temporal test set, 88.0% for geographic test set; algorithm: superior AUC compared to radiologists
Murata et al., 2018 [66]	Imaging of 400 healthy and 400 inflamed maxillary sinuses	CNN	ROC, Accuracy, Sensitivity, Specificity, AUC	Accuracy: 87.5%, Sensitivity: 86.7%, Specificity: 88.3%, AUC: 87.5%. Performance comparable to radiologists, better than dental residents
Ramkumar et al., 2017 [67]	MRI of 46 sinonasal tumor patients	ML classifiers	Accuracy	Accuracy (90.9% for training, 84.6% for validation), Neuroradiologists' review accuracy (56.5% for ROI, 73.9% for tumors, 87.0% for entire images)

Corrected Proof

Table 2 *continued*. Summary of all included studies.

Author, Year	Dataset	Type of AI	Evaluation metrics	Main outcomes
Altun et al., 2024 [68]	CBCT from 307 patients	YOLOv5x	Recall, Precision, F1 scores	Maxillary sinus segmentation: Recall: 100.0%, Precision: 98.5%, F1: 99.2%; Sinusitis: Recall: 100.0%, Precision: 94.2%, F1: 97.0%

Abbreviations: AI = Artificial Intelligence; CT = Computed Tomography; DL = Deep Learning; mAP = mean Average Precision; AP50 = Average Precision at 50% IoU; IoU = Intersection over Union; VGG19 = Visual Geometry Group 19-layer model; Faster-RCNN = Faster Region-Based Convolutional Neural Network; FPN = Feature Pyramid Network; AUC = Area Under the Curve; ResNet = Residual Network; YOLO = You Only Look Once; CNN = Convolutional Neural Network; DenseNet = Densely Connected Networks; CRSwNP = Chronic Rhinosinusitis with Nasal Polyps; ROC = Receiver-Operating Characteristic curve; MRI = Magnetic Resonance Imaging; ML = Machine Learning; NPV = Negative Predictive Value; PPV = Positive Predictive Value; FESS = Functional Endoscopic Sinus Surgery; DSC = Dice Similarity Coefficient; N/S = Not Specified; CBCT = Cone Beam Computed Tomography; CRS = Chronic Rhinosinusitis; OMC = Ostiomeatal Complex; SNSCC = Sinonasal Squamous Cell Carcinoma; RS-DLR = Radiomics Signature-Deep Learning Radiomics; RS-DTL = Radiomics Signature-Deep Transfer Learning; RS-HC = Radiomics Signature-Hand-crafted features; AUPRC = Area Under Precision-Recall Curve; MIE = Multiple Instance Ensembling; CAE = Convolutional Autoencoder; CI = Confidence Interval; DC = Dice Coefficient; RFS = Recurrence-Free Survival; RS = Radiomics Score; T2WI = T2-Weighted Imaging; T1c = Contrast-enhanced T1-Weighted Imaging; ADC = Apparent Diffusion Coefficient; LR = Logistic Regression; Grad-CAM = Gradient-weighted Class Activation Mapping; SVM = Support Vector Machine; ROI = Region of Interest; GAN = Generative Adversarial Network; PNS = Paranasal Sinuses; PR = Panoramic Radiograph; PwCF = People with Cystic Fibrosis; SNOT-22 = SinoNasal Outcome Test; MioU = Mean Intersection over Union; FLS = Fast Level Set; CRF-FCN = Conditional Random Field - Fully Convolutional Network; BPNN = Back Propagation Neural Network; LM scores = Lund-MacKay scores.

and segmentation in 10 studies (18.9%). Identification tasks were conducted in 6 studies (11.3%), treatment planning in 4 studies (7.5%), and post-treatment course prediction in 3 studies (5.7%). Most studies (51, 96.2%) employed a supervised learning approach, while 1 study (1.9%) used a semi-supervised approach, and 1 study (1.9%) did not specify the learning approach. Data augmentation was applied in 22 studies (41.5%), whereas 30 studies (56.6%) did not use augmentation, and 1 study (1.9%) did not specify its use. For validation, holdout validation was the most frequently used method, applied in 31 studies (58.5%), while 20 studies (37.7%) utilized cross-validation. Two studies (3.8%) did not specify the validation method. Among the most used metrics in AI performance evaluation were Area Under the Curve (AUC), accuracy, sensitivity, specificity, precision, and F1-score.

Quality assessment

The studies included in the analysis had a mean quality assessment score of 16.0 ± 2.6. The highest score achieved was 19, and the lowest was 8, showing a range of 11 points (8–19). Based on their scores, they were categorized into 5 groups: Group 1 (scores 1–4), Group 2 (scores 5–8), Group 3 (scores 9–12), Group 4 (scores 13–16), and Group 5 (scores 17–20). Group 1 included no studies, while Group 2 comprised one study with a score of 8. Group 3 included 6 studies (mean score: 11.3 ± 0.8), Group 4 consisted of 16 studies (mean score: 14.7 ± 0.9), and Group 5 contained 30 studies (mean score: 17.8 ± 0.8). The results per group are shown in Figure 2, and the number of studies reporting each assessment item is provided in Table 1. Table 6 displays the numerical evaluation score of each included study, along with its assigned group based on that score.

Discussion

Overview

In the era of automation and digitalization, this systematic review aims to evaluate advancements in the integration of AI tasks within the field of sinonasal radiological imaging. These advancements have the potential to enhance the quality and efficiency of patient care, leading to faster diagnoses, more accurate treatment plans, and improved clinical outcomes. AI has already been successfully applied in various fields of otolaryngology, including head and neck radiology. Although still in its early stages, AI has been trained on diverse imaging modalities, such as CT, MRI, positron emission tomography CT (PET/CT), ultrasound, and X-ray. Studies on head and neck imaging have demonstrated that AI can achieve high accuracy in lesion detection and classification, often outperforming traditional statistical methods and human experts [69]. Emerging evidence also suggests that AI has successfully performed similar tasks in paranasal sinus radiology. Notably, the majority of studies included in our review (e.g. Sukswai et al. [34], Zeng et al. [35], Kim et al. [65], Murata et al. [66] and others) showed that, according to different evaluation metrics, the AI architecture performed strongly to exceptionally well, often yielding results comparable to, or even surpassing, those of clinicians. To assess AI's ability to enhance patient care, the models in the included studies were evaluated using metrics such as AUC and accuracy, with a focus on clinical outcomes, including diagnostic accuracy, treatment

Corrected Proof

Table 3. Study design.

		N of studies (n=53) (%)
Data collection	Retrospective	45 (85.0)
	Prospective	8 (15.1)
Study centre involvement	Single-centre	36 (68.0)
	Multi-centre	13 (24.5)
	N/S	4 (7.5)
Data source	Internal database	49 (92.5)
	Public database	2 (3.8)
	Both internal and public databases	1 (1.9)
	N/S	1 (1.9)
Performance Comparison	AI-to-AI	15 (28.3)
	AI-to-human	30 (56.6)
	Both AI-to-AI and AI-to-human	1 (1.9)
	AI to traditional algorithmic methods	4 (7.5)
	N/S	3 (5.7)

Abbreviations: N/S = Not Specified; AI = Artificial Intelligence.

Table 4. Clinical approach.

		N of studies (n=53) (%)
Imaging modality	CT	32 (60.4)
	MRI	15 (28.3)
	PNS X-ray	3 (5.7)
	Both CT and PNS X-ray	3 (5.7)
Main condition or pathology of interest	CRSsNP	9 (17.0)
	CRSwNP	4 (7.5)
	eCRS	2 (3.8)
	Sinus fungal ball rhinosinusitis	3 (5.7)
	Maxillary sinusitis	5 (9.4)
	Paranasal sinus mucosal abnormalities	7 (13.2)
	Sphenoid sinus pneumatisation	1 (1.9)
	Inverted papilloma & malignant lesions	13 (24.5)
	Other abnormalities	9 (17.0)

Abbreviations: CT = Computed Tomography; MRI = Magnetic Resonance Imaging; PNS = Paranasal Sinuses; CRSsNP = Chronic Rhinosinusitis without Nasal Polyps; CRSwNP = Chronic Rhinosinusitis with Nasal Polyps; eCRS = eosinophilic Chronic Rhinosinusitis.

optimization, and patient prognosis prediction. This underscores AI's potential as a valuable tool to support clinicians in patient management and in addressing complex clinical challenges. To the best of our knowledge, as of the date of our most recent literature search, no other systematic review has been published that specifically addresses paranasal sinus pathological entities while incorporating the most up-to-date literature.

Within the analysis of the 53 studies included in this review, the primary focus was on the scientific methodology of the articles. This was done to assess the robustness and reliability of their findings, ensuring that the conclusions drawn were based on valid data and appropriate research techniques. A strong emphasis was placed on evaluating the quality of the included studies, as highlighting their strengths and weaknesses can guide future research directions. This process helps identify areas for improvement, standardization, or the adoption of novel methods to enhance research quality. Ultimately, high-quality studies with well-defined methodologies are crucial for making evidence-based recommendations, fostering progress in the field, and ensuring that future advancements are built on solid foundations.

Table 5. AI methodology.

		N of studies (n=53) (%)
AI architecture	CNN	39 (73.6)
	ANN	1 (1.9)
	Other NN	1 (1.9)
	Various DL algorithms	2 (3.8)
	ML classifiers	10 (18.9)
AI task	Classification	19 (35.8)
	Segmentation	10 (18.9)
	Diagnosis	11 (20.8)
	Identification	6 (11.3)
	Treatment planning	4 (7.5)
	Post-treatment course prediction	3 (5.7)
Learning approach	Supervised	51 (96.2)
	Semi-supervised	1 (1.9)
	N/S	1 (1.9)
Data augmentation	Yes	22 (41.5)
	No	30 (56.6)
	N/S	1 (1.9)
Validation	Cross-validation	20 (37.7)
	Holdout validation	31 (58.5)
	N/S	2 (3.8)

Abbreviations: CNN = Convolutional Neural Network; ANN = Artificial Neural Network; NN = Neural Network; DL = Deep Learning; ML = Machine Learning; N/S = Not Specified

Study design and potential sources of bias

Most of the included studies (85.0%) utilized a retrospective design, consistent with the trend in AI research, whereby large historical datasets are often leveraged for model development. Prospective studies (15.1%) represent a smaller but important subset of research, as they offer the potential for validation in real-world, forward-looking settings. The lack of prospective studies in the field may be due to a combination of factors, including their cost, time commitment, logistical complexity, and ethical considerations, all of which make their implementation more challenging and difficult to set up. Bhattacharya et al. [20] published a large, prospective, long-term study that included 2,619 participants. The authors used cross-validation to train their AI model and then applied it to a newly collected, unlabeled dataset (N=1,550) to test its generalizability. The prospective design, with data captured over an extended period, along with

Corrected Proof

Table 6. Numerical evaluation score of each study and score group.

Author, Year	QA score	Score Group
Peng et al., 2024 [16]	19	Group 5
Zou et al., 2024 [17]	16	Group 4
Lai et al., 2024 [18]	14	Group 4
Kwon et al., 2024 [19]	14	Group 4
Bhattacharya et al., 2024 [20]	17	Group 5
Du et al., 2024 [21]	17	Group 5
Whangbo et al., 2024 [22]	18	Group 5
Wang et al., 2024 [23]	18	Group 5
Gudapati et al., 2024 [24]	13	Group 4
Cheong et al., 2024 [25]	18	Group 5
Maria Jesi et al., 2023 [26]	12	Group 3
Celebi et al., 2023 [27]	11	Group 3
Park et al., 2024 [28]	17	Group 5
Massey et al., 2024 [29]	14	Group 4
Lin et al., 2024 [30]	18	Group 5
Bhattacharya et al., 2023 [31]	18	Group 5
Xiong et al., 2024 [32]	14	Group 4
Bhattacharya et al., 2024 [33]	19	Group 5
Sukswai et al., 2024 [34]	18	Group 5
Zeng et al., 2023 [35]	19	Group 5
Taylor et al., 2023 [36]	17	Group 5
Lin et al., 2023 [37]	17	Group 5
Ha et al., 2023 [38]	12	Group 3
Yoo et al., 2023 [39]	19	Group 5
He et al., 2023 [40]	14	Group 4
Zhou et al., 2022 [41]	17	Group 5
Hua et al., 2022 [42]	17	Group 5
Zhang et al., 2023 [43]	18	Group 5
Kim et al., 2023 [44]	14	Group 4
Sun et al., 2022 [45]	17	Group 5
Hung et al., 2022 [46]	17	Group 5
Kong et al., 2022 [47]	12	Group 3
Lim et al., 2022 [48]	15	Group 4
Morgan et al., 2022 [49]	19	Group 5
Liu et al., 2022 [50]	15	Group 4
Serindere et al., 2022 [51]	14	Group 4
Beswick et al., 2022 [52]	8	Group 2
Li et al., 2022 [53]	16	Group 4
Gu et al., 2022 [54]	17	Group 5
Corino et al., 2022 [55]	18	Group 5
Kuo et al., 2022 [56]	15	Group 4
Nakagawa et al., 2022 [57]	19	Group 5
Qi et al., 2021 [58]	10	Group 3
Jung et al., 2021 [59]	19	Group 5
Jeon et al., 2021 [60]	18	Group 5
Ramakrishnan et al., 2021 [61]	16	Group 4
Chen et al., 2021 [62]	16	Group 4
Kuo et al., 2019 [63]	11	Group 3
Humphries et al., 2020 [64]	17	Group 5
Kim et al., 2019 [65]	17	Group 5
Murata et al., 2018 [66]	15	Group 4
Ramkumar et al., 2017 [67]	18	Group 5
Altun et al., 2024 [68]	18	Group 5

Abbreviations: QA = Quality Assessment

the use of a large, diverse sample, further ensured the model's robustness by evaluating its performance on data that reflects real-world conditions and evolving trends [20]. Following a similar pattern, the 611-patient prospective multicenter study of Ramakrishnan et al. [61], who studied Chronic Rhinosinusitis (CRS) patients, offered a holistic approach by continuously monitoring variables such as nasal polyps, prior surgery, smoking history, and others. This approach allowed a clearer understanding of how these factors interact over time and assisted the ML model in identifying key predictors of olfactory loss [61]. Increasing the number of prospective studies would therefore help ensure that AI models can perform consistently across a variety of clinical settings [70].

Notably, most of the studies (68.0%) were single-center. Multi-center studies, which are crucial for validating AI models across diverse populations and clinical settings, accounted for only 24.5% of the studies [71]. The conduct of more large multi-center studies in the field of AI implementation in sinonasal pathology will significantly enhance the generalizability of findings, minimize bias, and provide more robust data, making the results more applicable to real-world scenarios. Such multi-center trials also ensure consistency in procedures, data collection methods, and evaluations across different clinical centers [71]. Regarding data sources, most studies (92.5%) relied on internal databases, which are often more readily available and tailored to specific clinical settings. The reliance on public databases was minimal (3.8%), indicating a potential area for future research to standardize and share datasets for broader AI development [72]. Published in 2024, the study by Cheong et al. [25] used the publicly available OASIS-3 MRI radiological dataset. The dataset is standardized and de-identified, making it machine-readable, and it includes expert consensus labels, ensuring high data quality. The broader use of such open databases enables researchers and developers to build upon them, improving AI tools and algorithms in medical imaging and enhancing the reproducibi-

Corrected Proof

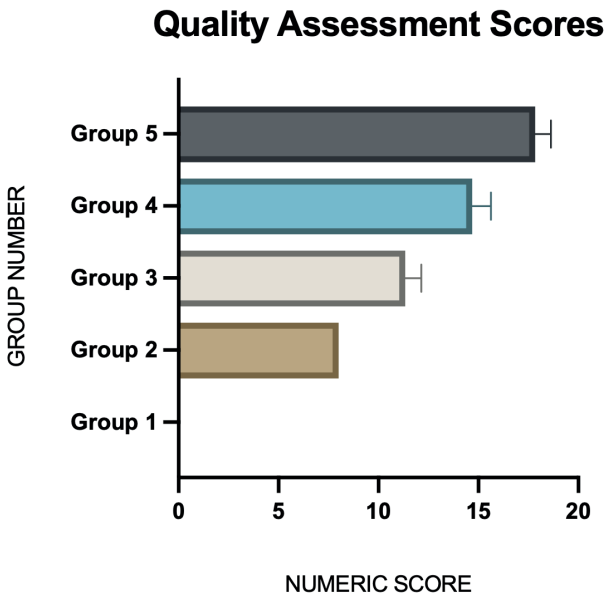


Figure 2. Quality assessment scores on a 20-point-scale, based on the CONSORT-AI extension. Categorization of studies into five groups based on their scores: Group 1 (scores 1–4, no studies), Group 2 (score 8, 1 study), Group 3 (scores 9–12, 6 studies; mean score: 11.3 ± 0.8), Group 4 (scores 13–16, 16 studies; mean score: 14.7 ± 0.9), and Group 5 (scores 17–20, 30 studies; mean score: 17.8 ± 0.8). The interval bars represent the standard deviation (SD) of numeric scores for each group.
Based on their scores, they were categorized into 5 groups: Group 1 (scores 1–4), Group 2 (scores 5–8), Group 3 (scores 9–12), Group 4 (scores 13–16), and Group 5 (scores 17–20). Group 1 included no studies, while Group 2 comprised one study with a score of 8. Group 3 included 6 studies (mean score: 11.3 ± 0.8), Group 4 consisted of 16 studies (mean score: 14.7 ± 0.9), and Group 5 contained 30 studies (mean score: 17.8 ± 0.8). The results per group are shown in Figure 2.

lity of AI research [25]. Only a small proportion of studies (5.7%) compared AI performance with traditional algorithmic methods, suggesting that most studies focused on AI-to-AI or AI-to-human comparisons, which may limit the understanding of AI's effectiveness and progress compared to traditional methods that are already widely implemented in clinical settings.

AI approaches and areas for development

Regarding the AI methodology, Convolutional Neural Networks (CNNs) were the dominant AI architecture (73.6%), reflecting their success in image-related tasks, particularly in medical imaging. AI-driven diagnosis has been the area of interest for many of the included studies, notably among those that were recently published (e.g. Peng et al. [16], Bhattacharya et al. [20], Sukswai et al. [34], Zeng et al. [35], Ha et al. [38], Kim et al. [44], Nakagawa et al. [57] and others). In most studies using CNNs, these models have outperformed traditional diagnostic methods, demonstrating their superiority in image recognition tasks [57]. Moreover, as highlighted by Kim et al. [65], CNNs maintain a strong diagnostic performance across both geographic and temporal external test sets, further confirming their robustness and generalizability (65). Machine learning classifiers were used in 18.9% of studies, but deep learning methods appear to be the preferred approach due to their ability to automatically learn hierarchical features from complex image data [73]. Classification was the most common AI task (35.8%), which is unsurprising given its importance in diagnosing various sinonasal conditions. Segmentation and diagnosis tasks were also prevalent (18.9% and 20.8%, respectively), further illustrating the varied applications of AI in this field. However, tasks related to treatment planning and post-treatment course prediction (5.7%) were less frequently explored, suggesting that AI's role in treatment decisions and monitoring remains underdeveloped in sinonasal pathology. This may be attributed to the complexity of treatment planning, which often requires a combination of clinical judgment and multidisciplinary input, making it harder to model with AI compared to diagnostic tasks. Additionally, post-treatment monitoring involves long-term data collection and subtle changes that are challenging for AI to assess accurately, especially with inconsistent follow-up.

Most studies (96.2%) employed supervised learning, which is typical in medical imaging where labeled datasets are necessary for training AI models. The use of semi-supervised learning or unspecified learning approaches was minimal, indicating that while supervised learning dominates, there is room for exploration of alternative approaches. In 2022, Kuo et al. [56] contributed to this direction by demonstrating that their semi-supervised approach outperformed existing state-of-the-art models. To address inaccuracies in predictions on unlabeled data, they applied a confidence threshold as a filter. Additionally, their model was designed to handle noisy data more effectively than traditional supervised learning [56]. Data augmentation was applied in 41.5% of studies, which is a standard technique to improve the robustness and generalizability of AI models, especially in medical imaging where dataset sizes are often limited [74]. Validation methods were also diverse, with holdout validation being the most common (58.5%), followed by cross-validation (37.7%). The prevalence of holdout validation, which involves splitting the dataset into training and testing sets, highlights its simplicity and practical application in real-world clinical settings. It is inferior however to the cross-validation, which "folds" the dataset multiple times and utilizes each subset both as a training and testing set during different iterations. As a result, cross-validation provides a more reliable estimate of the model's performance by averaging evaluation results across several folds [75]. Notably, the more validation folds there are, the clearer the picture of how well the model will perform on unseen data in real-world scenarios. 10-fold-cross validation, specifically, as used in the studies of Cheong et al. [25], Xiong et al. [32] and Gu et

Corrected Proof

al. [54] is considered optimal for minimizing errors and improving performance, leading to a more generalized AI model [25]. Moreover, while further research is needed to standardize data usage before integrating AI into real-time clinical workflows, training AI models with electronic health records (EHRs) holds great potential for real-time decision-making. This can be achieved by extracting, interpreting, and organizing large-scale patient data to enable more personalized treatment and diagnosis [76].

Research quality: current strengths and future challenges

The studies included in the analysis had a mean quality assessment score of 16.0 ± 2.6. The results were grouped into 5 categories based on their quality scores, with most studies falling into the highest quality groups (Groups 4 and 5). Notably, Group 5 contained a significantly larger number of studies compared to other groups, highlighting a strong trend towards higher-quality research. This suggests that a substantial proportion of the studies have undergone rigorous methodological review, with an emphasis on transparency and robustness. The absence of studies in the lowest quality group (Group 1) is also a positive indicator, showing that most studies adhered to appropriate research standards. Furthermore, the fact that most of the studies were published within the past 3 years (2022-2024) underscores a growing commitment to fundamental research standards. This suggests that a substantial proportion of the studies underwent rigorous methodological review, emphasizing transparency and robustness.

While the included studies generally demonstrated strong adherence to reporting standards, several areas present opportunities for improvement. A notable gap is the handling of low-quality or missing data, with only 43.4% of studies explaining how such data were evaluated and managed. Addressing this is crucial to ensure the reliability of results, and future studies should provide more detailed descriptions of their approach to data quality. Additionally, only 34.0% of studies justified their sample size, which is important for assessing the strength of their conclusions [77]. Furthermore, only 56.6% of studies reported performance errors and how these were handled, suggesting that the limitations of the automated algorithms used may not have been fully investigated. Among the various limitations reported in the studies, the most common were the inability to confirm the generalizability of the AI architecture [18, 33, 34, 42, 46, 54, 56, 64], limited sample sizes [32, 35, 37, 50, 51, 60], and the heterogeneity of the data, which made categorization and handling more difficult [29, 57, 58].The generalizability of results was discussed in only 50.9% of studies, highlighting the need for future research to focus more on how findings can be applied to broader clinical and real-world settings [78]. It is important to note that the variability in study design has made quality assessment of the research articles a challenging task. Given the need for high-quality and valid data, it is essential that study design guidelines such as CONSORT-AI be adopted and implemented in all future AI-related studies [11].

Limitations and future directions

Although this systematic review offers valuable insights, it is important to address its limitations and highlight areas that require future research. First, the substantial variation in clinical focus, the diverse nature of pathologies examined, differences in study populations and inclusion/exclusion criteria, variability in imaging modalities, AI architectures, validation methods, AI tasks, and outcome measures all precluded the feasibility of conducting a meta-analysis. The heterogeneity of the data also prevented a thorough examination of interobserver variability, differences in the performance of various AI models (e.g., CNNs and other ML architectures), and performance variations across distinct imaging modalities (e.g., CT, MRI). Although several of the included studies included more than one human annotator for the AI training dataset, the specific impact of human annotation variability on methodological transparency in AI training datasets is not directly addressed. Second, most studies were retrospective and single-center, which limits the generalizability and external validity of the findings. Third, there was no real-world validation of AI models in clinical practice, which hampers their translation to everyday healthcare settings. Fourth, the absence of standardized datasets further limits the reproducibility and comparability of AI models. Fifth, there is a need for standardized, publicly accessible sinonasal imaging datasets for training AI models to ensure consistency across studies. Sixth, multi-center and prospective studies are necessary to establish the population-wide and real-world clinical utility of AI. Seventh, AI strategies for long-term monitoring and treatment outcome prediction remain underexplored and should be investigated in future research. Additionally, while AI has shown great potential in classification and segmentation tasks, its role in treatment planning and post-treatment course prediction remains underexplored and requires further investigation. Furthermore, some relevant literature may not have been captured in this review due to the exclusion of publications in languages other than English and conference papers. Lastly, studies with unclear or controversial data reporting may have been unintentionally misclassified.

Conclusion

This systematic review highlights the innovations and overall progress of AI applications in the imaging of sinonasal pathology. Currently, AI is being used primarily as a decision-support tool, assisting clinicians in identifying pathologies more accurately and efficiently. For AI to be widely adopted in mainstream clinical practice, several steps must be taken, including rigorous validation of AI models to ensure their reliability and generali-

Corrected Proof

zability across different clinical settings. Additionally, it is crucial that AI integration remains seamless and supportive, rather than replacing clinicians' expertise. While the potential for AI to improve diagnostic accuracy is significant, the current state of AI requires rigorous testing to validate its performance across various clinical environments. This, combined with universal adherence to study design guidelines, would improve the generalizability of AI models, and ultimately pave the way for higher-quality healthcare services.

Acknowledgements

We would like to thank the University Hospitals Birmingham (UHB) Library and Knowledge Service, Sutton Coldfield, UK, for their assistance with the data collection for this study.

Authorship contribution

DPP: study design, data collection and analysis, manuscript writing; DS: study design, data collection, manuscript writing; AM: data collection and analysis, manuscript writing; JM, GG: manuscript revision; CG: manuscript revision, approving the final draft

Conflict of interest

The authors declare no conflict of interest.

Funding

No funding was utilized for this work.

References

1. Qureshi MF, Usmani A. A CT-Scan review of anatomical variants of sinonasal region and its correlation with symptoms of sinusitis (nasal obstruction, facial pain and rhinorrhea). Pak J Med Sci. 2021;37(1):195-200.
2. Whyte A, Boeddinghaus R. The maxillary sinus: physiology, development and imaging anatomy. Dentomaxillofac Radiol. 2019;48(8):20190205.
3. Gomaa MA, Hammad MS, Abdelmoghny A, Elsherif AM, Tawfik HM. Magnetic resonance imaging versus computed tomography and different imaging modalities in evaluation of sinonasal neoplasms diagnosed by histopathology. Clin Med Insights Ear Nose Throat. 2013;6:9-15.
4. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. Bioengineering (Basel). 2023;10(12):1435.
5. Paudyal R, Shah AD, Akin O, et al. Artificial intelligence in CT and MR imaging for oncological applications. Cancers (Basel). 2023;15(9):2573.
6. Petsiou DP, Martinos A, Spinos D. Applications of artificial intelligence in temporal bone imaging: advances and future challenges. Cureus. 2023;15(9):e44591.
7. Amanian A, Heffernan A, Ishii M, Creighton FX, Thamboo A. The evolution and application of artificial intelligence in rhinology: a state-of-the-art review. Otolaryngol Head Neck Surg. 2023;169(1):21-30.
8. Osie G, Darbari Kaul R, Alvarado R, et al. A scoping review of artificial intelligence research in rhinology. Am J Rhinol Allergy. 2023;37(4):438-48.
9. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
10. Bulfamante AM, Ferella F, Miller AM, et al. Artificial intelligence, machine learning, and deep learning in rhinology: a systematic review. Eur Arch Otorhinolaryngol. 2023;280(2):529-42.
11. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-74.
12. Gray R, Sullivan M, Altman DG, Gordon-Weeks AN. Adherence of trials of operative intervention to the CONSORT statement extension for non-pharmacological treatments: a comparative before and after study. Ann R Coll Surg Engl. 2012;94(6):388-94.
13. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet. 2001;357(9263):1191-4.
14. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA. 1996;276(8):637-9.
15. Cuschieri S. The CONSORT statement. Saudi J Anaesth. 2019;13(Suppl 1):S27-S30.
16. Peng L, Wu Q, Shi R, et al. A computerized tomography-based deep learning diagnostic method of maxillary sinus fungal balls. Int J Radiat Res. 2024;22(1):9-15.
17. Zou J, Lyu Y, Lin Y, et al. A multi-view fusion lightweight network for CRSwNPs prediction on CT images. BMC Med Imaging. 2024;24(1):112.
18. Lai S, Kang W, Chen Y, et al. An end-to-end CRSwNP prediction with multichannel ResNet on computed tomography. Int J Biomed Imaging. 2024;2024:4960630.
19. Kwon KW, Kim J, Kang D. Automated detection of maxillary sinus opacifications compatible with sinusitis from CT images. Dentomaxillofac Radiol. 2024;53(8):549-57.
20. Bhattacharya D, Becker BT, Behrendt F, et al. Computer-aided diagnosis of maxillary sinus anomalies: validation and clinical correlation. Laryngoscope. 2024;134(9):3927-34.
21. Du W, Kang W, Lai S, et al. Deep learning in computed tomography to predict endotype in chronic rhinosinusitis with nasal polyps. BMC Med Imaging. 2024;24(1):25.
22. Whangbo J, Lee J, Kim YJ, Kim ST, Kim KG. Deep learning-based multi-class segmentation of the paranasal sinuses of sinusitis patients based on computed tomographic images. Sensors (Basel). 2024;24(6):1933.
23. Wang Y, Han Q, Wen B, et al. Development and validation of a prediction model for malignant sinonasal tumors based on MR radiomics and machine learning. Eur Radiol. 2024.
24. Gudapati V, Chen A, Meyer S, et al. Development of a machine learning-enabled virtual reality tool for preoperative planning of functional endoscopic sinus surgery. J Neurol Surg Rep. 2024;85(3):e118-e123.
25. Cheong RCT, Jawad S, Adams A, et al. Enhancing paranasal sinus disease detection with AutoML: efficient AI development and evaluation via magnetic resonance imaging. Eur Arch Otorhinolaryngol. 2024;281(4):2153-2158.
26. Jesi P, Maria S, Kamatchi D, Madhivadhani, Lily Saron Grace. HRSHO: a hybrid rain-optimized spotted hyena optimizer for efficient feature selection in CNN-based sinusitis classification. Biomed Signal Process Control. 2024;87:105441.
27. Çelebi A, Imak A, Üzen H, et al. Maxillary sinus detection on cone beam computed tomography images using ResNet and Swin Transformer-based UNet. Oral Surg Oral Med Oral Pathol Oral Radiol. 2024;138(1):149-161.
28. Park CJ, Choi SH, Kim D, et al. MRI radiomics may predict early tumor recurrence in patients with sinonasal squamous cell carcinoma. Eur Radiol. 2024;34(5):3151-3159.
29. Massey CJ, Humphries SM, Mace JC, Smith TL, Soler ZM, Ramakrishnan VR. Multi-institutional validation of an AI-based sinus CT analytic platform with olfactory assessments. Int Forum Allergy Rhinol. 2024;14(11):1806-1809.
30. Lin N, Shi Y, Ye M, Wang L, Sha Y.

Corrected Proof

Multiparametric MRI-based radiomics approach with deep transfer learning for preoperative prediction of Ki-67 status in sinonasal squamous cell carcinoma. Front Oncol. 2024;14:1305836.
31. Bhattacharya D, Behrendt F, Becker BT, et al. Multiple instance ensembling for paranasal anomaly classification in the maxillary sinus. Int J Comput Assist Radiol Surg. 2024;19(2):223-231.
32. Xiong P, Chen J, Zhang Y, et al. Predictive modeling for eosinophilic chronic rhinosinusitis: nomogram and four machine learning approaches. iScience. 2024;27(2):108928.
33. Bhattacharya D, Behrendt F, Becker BT, et al. Self-supervised learning for classifying paranasal anomalies in the maxillary sinus. Int J Comput Assist Radiol Surg. 2024;19(9):1713-1721.
34. Sukswai P, Hnoohom N, Hoang MP, et al. The accuracy of deep learning models for diagnosing maxillary fungal ball rhinosinusitis. Eur Arch Otorhinolaryngol. 2024;281(12):6485-6492.
35. Zeng P, Song R, Lin Y, et al. Abnormal maxillary sinus diagnosing on CBCT images via object detection and 'straight-forward' classification deep learning strategy. J Oral Rehabil. 2023;50(12):1465-1480.
36. Taylor A, Habib AR, Kumar A, Wong E, Hasan Z, Singh N. An artificial intelligence algorithm for the classification of sphenoid sinus pneumatisation on sinus computed tomography scans. Clin Otolaryngol. 2023;48(6):888-894.
37. Lin M, Lin N, Yu S, et al. Automated prediction of early recurrence in advanced sinonasal squamous cell carcinoma with deep learning and multi-parametric MRI-based radiomics nomogram. Acad Radiol. 2023;30(10):2201-2211.
38. Ha EG, Jeon KJ, Choi H, Lee C, Choi YJ, Han S. Automatic diagnosis of retention pseudocyst in the maxillary sinus on panoramic radiographs using a convolutional neural network algorithm. Sci Rep. 2023;13:2734.
39. Yoo YS, Kim D, Yang S, et al. Comparison of 2D, 2.5D, and 3D segmentation networks for maxillary sinuses and lesions in CBCT images. BMC Oral Health. 2023;23:866.
40. He S, Chen W, Wang X, et al. Deep learning radiomics-based preoperative prediction of recurrence in chronic rhinosinusitis. iScience. 2023;26(4):106527.
41. Zhou H, Fan W, Qin D, et al. Development, validation, and comparison of artificial neural network and logistic regression models predicting eosinophilic chronic rhinosinusitis with nasal polyps. Allergy Asthma Immunol Res. 2023;15(1):67-82.
42. Hua HL, Li S, Xu Y, et al. Differentiation of eosinophilic and non-eosinophilic chronic rhinosinusitis on preoperative computed tomography using deep learning. Clin Otolaryngol. 2023;48(2):330-338.
43. Zhang Y, Lin N, Xiao H, Xin E, Sha Y. Differentiation of sinonasal NKT from diffuse large B-cell lymphoma using machine learning and MRI-based radiomics. J Comput Assist Tomogr. 2023;47(6):973-981.
44. Kim K, Lim CY, Shin J, Chung MJ, Jung YG. Enhanced artificial intelligence-based diagnosis using CBCT with internal denoising: clinical validation for discrimination of fungal ball, sinusitis, and normal cases in the maxillary sinus. Comput Methods Programs Biomed. 2023;240:107708.
45. Sun K, Zhang J, Liu Z, et al. A deep learning radiomics analysis for identifying sinus invasion in patients with meningioma before operation using tumor and peritumoral regions. Eur J Radiol. 2022;149:110187.
46. Hung KF, Ai QYH, King AD, Bornstein MM, Wong LM, Leung YY. Automatic detection and segmentation of morphological changes of the maxillary sinus mucosa on cone-beam computed tomography images using a three-dimensional convolutional neural network. Clin Oral Investig. 2022;26(5):3987-3998.
47. Kong HJ, Kim JY, Moon HM, et al. Automation of generative adversarial network-based synthetic data augmentation for maximizing the diagnostic performance with paranasal imaging. Sci Rep. 2022;12:18118.
48. Lim SH, Kim JH, Kim YJ, et al. Aux-MVNet: auxiliary classifier-based multi-view convolutional neural network for maxillary sinusitis diagnosis on paranasal sinuses view. Diagnostics (Basel). 2022;12(3):736.
49. Morgan N, Van Gerven A, Smolders A, Vasconcelos KdF, Willems H, Jacobs R, et al. Convolutional neural network for automatic maxillary sinus segmentation on cone-beam computed tomographic images. Sci Rep. 2022;12:7523.
50. Liu GS, Yang A, Kim D, Hojel A, et al. Deep learning classification of inverted papilloma malignant transformation using 3D convolutional neural networks and magnetic resonance imaging. Int Forum Allergy Rhinol. 2022;12(8):1025-1033.
51. Serindere G, Bilgili E, Yesil C, Ozveren N. Evaluation of maxillary sinusitis from panoramic radiographs and cone-beam computed tomographic images using a convolutional neural network. Imaging Sci Dent. 2022;52(2):187-195.
52. Beswick DM, Humphries SM, Balkissoon CD, et al. Impact of cystic fibrosis transmembrane conductance regulator therapy on chronic rhinosinusitis and health status: deep learning CT analysis and patient-reported outcomes. Ann Am Thorac Soc. 2022;19(1):12-19.
53. Li X, Zhao H, Ren T, Tian Y, Yan A, Li W. Inverted papilloma and nasal polyp classification using a deep convolutional network integrated with an attention mechanism. Comput Biol Med. 2022;149:105976.
54. Gu J, Yu Q, Li Q, et al. MRI radiomics-based machine learning model integrated with clinic-radiological features for preoperative differentiation of sinonasal inverted papilloma and malignant sinonasal tumors. Front Oncol. 2022;12:1003639.
55. Corino VDA, Bologna M, Calareso G, et al. Refining tumor treatment in sinonasal cancer using delta radiomics of multi-parametric MRI after the first cycle of induction chemotherapy. J Imaging. 2022;8(2):46.
56. Kuo CJ, Liao YS, Barman J, Liu SC. Semi-supervised deep learning semantic segmentation for 3D volumetric computed tomographic scoring of chronic rhinosinusitis: clinical correlations and comparison with Lund-Mackay scoring. Tomography. 2022;8(2):718-729.
57. Nakagawa J, Fujima N, Hirata K, et al. Utility of the deep learning technique for the diagnosis of orbital invasion on CT in patients with a nasal or sinonasal tumor. Cancer Imaging. 2022;22:52.
58. Qi X, Zhong J, Cui S. Adaptive localizing region-based level set for segmentation of maxillary sinus based on convolutional neural networks. Comput Intell Neurosci. 2021;2021:4824613.
59. Jung SK, Lim HK, Lee S, Cho Y, Song IS. Deep active learning for automatic segmentation of maxillary sinus lesions using a convolutional neural network. Diagnostics (Basel). 2021;11(4):688.
60. Jeon Y, Lee K, Sunwoo L, et al. Deep learning for diagnosis of paranasal sinusitis using multi-view radiographs. Diagnostics (Basel). 2021;11(2):250.
61. Ramakrishnan VR, Arbet J, Mace JC, et al. Predicting olfactory loss in chronic rhinosinusitis using machine learning. Chem Senses. 2021 Jan 1;46:bjab042.
62. Chen C, Qin Y, Cheng J, Gao F, Zhou X. Texture analysis of fat-suppressed T2-weighted magnetic resonance imaging and use of machine learning to discriminate nasal and paranasal sinus small round malignant cell tumors. Front Oncol. 2021;11:701289.
63. Kuo CJ, Leu Y, Hu D, Huang C, Siao J, Leon K. Application of intelligent automatic segmentation and 3D reconstruction of inferior turbinate and maxillary sinus from computed tomography and analysis of the relationship between volume and nasal lesion. Biomed Signal Process Control. 2020;57:101291.
64. Humphries SM, Centeno JP, Notary AM, et al. Volumetric assessment of paranasal sinus opacification on computed tomography can be automated using a convolutional neural network. Int Forum Allergy Rhinol. 2020;10(11):1218-1225.
65. Kim Y, Lee KJ, Sunwoo L, et al. Deep learning in diagnosis of maxillary sinusitis using conventional radiography. Invest Radiol. 2019;54(1):7-15.
66. Murata M, Ariji Y, Ohashi Y, et al. Deep-learning classification using convolutional neural network for evaluation of maxillary sinusitis on panoramic radiography. Oral Radiol. 2019;35(3):301-307.
67. Ramkumar S, Ranjbar S, Ning S, et al. MRI-

Corrected Proof

based texture analysis to differentiate sinonasal squamous cell carcinoma from inverted papilloma. AJNR Am J Neuroradiol. 2017;38(5):1019-1025.
68. Altun O, Özen DÇ, Duman ŞB, et al. Automatic maxillary sinus segmentation and pathology classification on cone-beam computed tomographic images using deep learning. BMC Oral Health. 2024;24(1):1208.
69. Mahmood H, Shaban M, Rajpoot N, Khurram SA. Artificial intelligence-based methods in head and neck cancer diagnosis: an overview. Br J Cancer. 2021;124(12):1934-1940.
70. Ramirez-Santana M. Limitations and biases in cohort studies. Cohort Studies in Health Sciences. InTech; 2018.
71. Das MK. Multicenter studies: relevance, design and implementation. Indian Pediatr. 2022;59(7):571-579.
72. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J. 2021;8(2):e188-e194.
73. Kufel J, Bargieł-Łączek K, Kocot S, et al. What is machine learning, artificial neural networks and deep learning? Examples of practical applications in medicine. Diagnostics (Basel). 2023;13(15):2582.
74. Sandfort V, Yan K, Pickhardt PJ, et al. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. Sci Rep. 2019;9:16884.
75. Spinos D, Martinos A, Petsiou DP, Mistry N, Garas G. Artificial intelligence in temporal bone imaging: a systematic review. Laryngoscope. 2024;134(10):2381-2389.
76. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5:194.
77. Eldredge JD, Weagel EF, Kroth PJ. Defining and identifying members of a research study population: CTSA-affiliated faculty members. Hypothesis (Macon). 2014;26(1):5-11.
78. Yang J, Soltan AAS, Clifton DA. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. NPJ Digit Med. 2022;5(1):69.

Dimitrios Spinos
Department of Cancer and Genomics
School of Medicine
University of Birmingham
Birmingham
United Kingdom

E-mail: Dimitrios.spinos@nhs.net

Dioni-Pinelopi Petsiou[1], Dimitrios Spinos[2], Anastasios Martinos[1], Jameel Muzaffar[3], Georgios Garas[4,5], Christos Georgalas[6,7]

[1]Department of Otolaryngology - Head and Neck Surgery, National and Kapodistrian University of Athens, School of Medicine, Athens, Greece

[2] Department of Cancer and Genomics, School of Medicine, University of Birmingham, Birmingham, United Kingdom

[3] Department of Otolaryngology, Head and Neck Surgery, University Hospitals of Birmingham, Birmingham, United Kingdom

[4] Surgical Innovation Centre, Department of Surgery and Cancer, Imperial College London, St. Mary's Hospital, London, United Kingdom

[5] Athens Medical Centre, Marousi and Psychiko Clinic, Athens, Greece

[6] Department of Head, School of Medicine, University of Nicosia, Nicosia, Cyprus

[7] Department of Head, Neck and Skull Base surgery, Hygeia Hospital, Athens, Greece

Rhinology 63: 4, 0 - 0, 2025
https://doi.org/10.4193/Rhin25.044

Received for publication:
January 19, 2025
Accepted: April 4, 2025

Associate Editor:
Ahmad Sedaghat