

Review

Quality Assessment and Model Training Impact of Synthetic Images in Medical Diagnostic Imaging: Systematic Review

Júlia Folguera-Profítos^{1,2*}, MSc; Marina Ramiro-Pareta^{2,3*}, PhD; Gerard Carot-Sans^{1,2}, PhD; Alba Jiménez-Rueda^{2,3}, MSc; Júlia Aznar-Roig⁴, MSc; Didier Domínguez-Herrera³, MSc; Maria Mercedes Nogueras^{2,3}, PhD; Esmaeil Nadimi⁵, PhD; Ferran Torres⁶, PhD; Jordi Piera-Jiménez^{2,3}, MBA, PhD

¹Information Systems Department, Catalan Health Service (CatSalut), Barcelona, Spain

²Digitalization for the Sustainability of the Healthcare System (DS3) Research Group, Barcelona, Spain

³Fundació Tic Salut i Social (FTSS), Barcelona, Spain

⁴Process Unit, Directorate of Quality, Processes and Innovation, Vall d'Hebron University Hospital, Barcelona, Spain

⁵Faculty of Engineering, Maersk Mc-Kinney Møller Institute, University of Southern Denmark, Odense, Denmark

⁶Biostatistics Unit, Medical School, Universitat Autònoma de Barcelona, Bellaterra, Spain

*these authors contributed equally

Corresponding Author:

Júlia Folguera-Profítos, MSc
Information Systems Department
Catalan Health Service (CatSalut)
Gran Via de les Corts Catalanes, 587
Barcelona 08007
Spain
Phone: +34 93 403 85 85
Email: julia.folguera@catsalut.cat

Abstract

Background: AI has increasingly been used to generate synthetic images to overcome the scarcity of large, regulatory-compliant medical image datasets. However, their quality and actual utility for algorithm development across deep learning architectures, medical settings, and image types have been poorly assessed.

Objective: This study aimed to review the use of synthetic images for algorithm development and to assess their quality and effects on downstream algorithms trained on these images, according to AI architecture, medical setting, and image type.

Methods: The authors conducted a systematic review of studies published between January 2020 and September 2025. PubMed, Scopus, IEEE Xplore, and Embase were searched for peer-reviewed studies evaluating synthetic images generated from human medical imaging data. Data were extracted for imaging modalities, generative architectures, image-quality assessment methods, downstream model performance, explainability approaches, reproducibility practices, and risk of bias. Given substantial heterogeneity and incomplete statistical reporting, findings were synthesized descriptively.

Results: A total of 160 studies were included. Generative adversarial network-based architectures dominated image generation (119 studies, 74.4%), although hybrid and diffusion-based approaches became increasingly represented in recent years. Image-quality evaluation was highly heterogeneous: 50 studies included human-reader evidence, and 101 reported an automated image-quality assessment. Synthetic-only training generally underperformed real-data training, including a median Dice similarity coefficient change of -10.2% for segmentation. Adding synthetic data to real training sets was associated with median improvements in classification precision (+7.8%), recall (+5.3%), F_1 -score (+5.5%), accuracy (+4.3%), and area under the curve (+4.0%), while the median Dice coefficient improvement was +1.2%. Explainability evidence was reported in 68 (42.5%) studies and was rarely validated against expert annotations. Reproducibility reporting was inconsistent, and only 38% (60/160) of studies were judged to have low overall risk of bias.

Conclusions: Unlike previous reviews restricted to particular diseases, imaging modalities, or generative architectures, this review provides a cross-domain quantitative synthesis of both synthetic-image quality assessment and downstream model performance. Synthetic images currently appear more valuable as a complement to real training data than as a replacement, with modest improvements in classification performance, smaller gains in segmentation, and generally poorer performance with synthetic-only training. For researchers, developers, and regulators, these findings highlight the need for standardized

multimethod image-quality evaluation, complete reporting, and rigorous external validation before synthetic-image pipelines are translated into clinical AI applications.

Trial Registration: PROSPERO CRD42023369125; <https://www.crd.york.ac.uk/PROSPERO/view/CRD42023369125>

JMIR AI 2026;5:e81472; doi: [10.2196/81472](https://doi.org/10.2196/81472)

Keywords: synthetic images; synthetic data; medical imaging; diagnostic imaging; AI; generative models; image quality assessment; data augmentation; downstream model performance; systematic review

Introduction

AI is a potential game changer in diagnostic imaging, with deep learning models already assisting radiologists in making earlier and more accurate decisions [1,2]. However, the development of robust models for image recognition requires large datasets of medical images that are often unavailable for research due to scarcity (eg, in low-prevalence conditions), strict privacy regulations, fragmented storage, and the burden of annotation.

Several strategies have been proposed to overcome data scarcity, including classical augmentation, self-supervised learning, federated learning, and other privacy-preserving techniques [3-5]. In this context, generative AI offers a powerful tool for generating synthetic images that are realistic and, at the same time, deidentified, enabling multicenter sharing without exposing patient information [6-8]. Early successes came from simple geometric or color transformations [9,10], but the field advanced rapidly with generative adversarial networks (GANs) [11,12] and is now expanding to more complex architectures, hybrid models, variational autoencoders (VAEs), and diffusion models [13].

Despite this momentum, the real-world quality of synthetic images and the downstream impact of these images on model performance have been poorly evaluated. Existing reviews primarily focus on specific diseases or imaging modalities, limiting the generalizability of their conclusions across medical imaging domains [14-22]. Others have addressed synthetic data in the broader context of medical AI, combining synthetic images with other data modalities such as electronic health records, time series, and clinical text rather than specifically evaluating synthetic imaging [23]. Some reviews have considered classical image augmentation techniques or image-to-image translation together with generative image synthesis, making it difficult to isolate the contribution of synthetic-image generation itself [24, 25]. Conversely, other reviews have restricted their scope to specific generative paradigms, particularly GAN-based methods [18,26], or to publicly available imaging datasets, overlooking the broader range of data sources and evaluation strategies currently used [27]. Consequently, there is still no comprehensive synthesis spanning imaging modalities, clinical applications, generative architectures, image-quality assessment, downstream model performance, explainability, reproducibility, and methodological quality. Moreover, no review has quantitatively synthesized the reported values of image-quality metrics or the changes in downstream performance associated with synthetic-only and synthetic-augmented training. This lack of comprehensive

benchmarking limits the ability of researchers, clinicians, and regulators to determine when synthetic images provide meaningful benefits, where their limitations lie, and which evaluation strategies are most appropriate across different clinical and technical settings.

In this systematic review, we comprehensively evaluate the use of synthetic images in medical imaging by characterizing the generative architectures, imaging modalities, clinical applications, and evaluation strategies reported in the literature. We provide a quantitative cross-study synthesis of image-quality metrics and downstream model performance, comparing synthetic-only and synthetic-augmented training across different tasks and imaging modalities. We also examine explainability, reproducibility, and methodological quality to identify current strengths, limitations, and priorities for the responsible development and evaluation of synthetic-image technologies in health care.

Methods

Information Sources, Search Strategy, and Eligibility Criteria

This review aimed to assess peer-reviewed studies evaluating de novo synthetic medical images used as training data for algorithms supporting diagnosis or clinical follow-up, irrespective of the clinical condition, imaging modality, or generative model architecture. The review addressed two primary objectives: (1) to characterize the assessment of synthetic-image quality and (2) to evaluate the impact of using synthetic images on downstream model performance. Because downstream utility was central to the review question, studies were required to report at least one quantitative measure of downstream algorithm performance to be eligible.

We searched PubMed/MEDLINE and Scopus on November 28, 2025, for peer-reviewed articles published between January 1, 2020, and September 30, 2025. IEEE Xplore was subsequently searched on December 8, 2025, and Embase was searched on July 9, 2026, using the same publication-date limits. Database-specific search strategies used title- and abstract-based free-text terms for synthetic-image generation, together with database-specific publication, language, subject-area, and content-type restrictions where applicable. The complete search strategies for all databases are provided in [Multimedia Appendix 1](#).

Studies were eligible if they met all of the following criteria: (1) evaluated the effect of synthetic images on the training of algorithms intended to support diagnostic or

follow-up decisions, (2) reported at least one quantitative measure of downstream algorithm performance, (3) used human medical imaging data, (4) generated images de novo rather than through image-to-image transformation, (5) were available as full-text articles, and (6) were published in English. The English-language restriction was applied to ensure consistent data extraction and interpretation across the review team.

We excluded the following: (1) studies using synthetic images exclusively for nondiagnostic purposes, such as education or technical simulation, because they did not report clinically relevant performance outcomes; (2) studies focused on microscopic or histopathologic imaging, because their spatial scale, image-generation pipelines, and evaluation methods differ substantially from those used in radiological imaging and would introduce considerable methodological heterogeneity; (3) studies using image-to-image transformation, defined as the generation of an output image from a corresponding source image, including cross-modality, cross-domain, or image-contrast translation, because the review focused specifically on de novo synthetic-image generation rather than the transformation of existing images; and (4) nonprimary or incompletely reported publications, including reviews, editorials, case reports, conference abstracts, preprints, letters, and guidelines, because these publication types did not provide the peer-reviewed, fully reported quantitative evidence required for data extraction and risk-of-bias assessment.

Study Selection

Two independent reviewers conducted the selection process by screening titles, abstracts, and full texts. Any discrepancies were resolved through group discussion involving the 2 reviewers and at least 2 additional authors.

Reporting Guidelines and Registration

The review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guideline [28], the search strategy was reported according to PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension) guideline [29], and the synthesis without meta-analysis was reported using the Synthesis Without Meta-Analysis (SWiM) guideline [30]; the PRISMA 2020, PRISMA-S, and SWiM checklists are provided in [Checklist 1](#). The review protocol, detailing the objectives, eligibility criteria, and planned analyses, was registered in the PROSPERO (International Prospective Register of Systematic Reviews) database (ID: CRD42023369125) [31].

Data Extraction and Outcomes

Overview

General data on the selected studies included years of publication, clinical domain, imaging modality, image dimensions and resolution, and data source (open access vs in-house). Analysis outcomes included architecture and strategy for image generation, quality of the generated images (metrics used for quality assessment and results),

and downstream impact on the performance of the model trained with synthetic images (model architecture, performance metrics used, and results), whenever data were available.

Strategy for Image Generation

The architecture for image generation was classified into GANs, variational autoencoder generative adversarial networks (VAE-GANs), diffusion models, and scientific/physical simulation models. GAN-based architectures were further categorized based on structural modifications, including progressive-growing generative adversarial network (PGGAN), style-based generative adversarial network (StyleGAN), conditional GANs, attention-based models (eg, BigGAN [large-scale generative adversarial network]-Deep and SAGAN [Self-Attention Generative Adversarial Network]), and loss-specific GANs (eg, HingeGAN and LSGAN [least squares generative adversarial network]). Hybrid approaches integrating multiple generative paradigms, including VAE-GANs, diffusion-GAN hybrids, and GANs combined with convolutional neural networks (CNNs) or domain-specific models, were documented separately. Table S1 in [Multimedia Appendix 1](#) provides a detailed taxonomy of generative models.

Image Quality

Two primary approaches to the assessment of image quality were considered: human-reader assessments and automated metrics of image quality. Human-reader assessments were extracted and classified into task categories (ie, what readers were asked to evaluate) listed in Table S2 in [Multimedia Appendix 1](#). For each category, we extracted the assessment metrics, which included Turing accuracy, synthetic authenticity rate, fooling rate, Likert realism or quality scores, clinical plausibility scores, reader diagnostic accuracy, sensitivity, specificity, area under the curve (AUC), feature-identification accuracy, text-image matching accuracy, preference or ranking metrics, and agreement statistics (Table S3 in [Multimedia Appendix 1](#)). Table S4 in [Multimedia Appendix 1](#) provides a guide for interpretation of each of the extracted human-reader metrics for assessing image quality.

Automated metrics for image-quality assessment were grouped into predefined families, including distributional similarity or fidelity, pixel/structural similarity, diversity or coverage, automated realism/no-reference quality, intensity/texture/image-statistical descriptors, reconstruction error, perceptual similarity, privacy/memorization similarity, and other automated image-quality metrics. The complete lists of approaches and metrics included in each of these groups are summarized in Table S5 and Table S6 in [Multimedia Appendix 1](#), respectively.

Quantitative values (means, SDs, CIs, and sample sizes, where available) were recorded for each metric.

Impact on Model Performance

To assess the impact of synthetic images on downstream model performance, we extracted detailed information on model architectures and performance metrics from each study. Real-to-synthetic data ratios were extracted and

classified by analytical purpose. Given their heterogeneous use across studies, these ratios were not pooled. For ratio-frequency summaries, only ratios related to downstream model development or evaluation were considered. When a study tested several ratios for the same analytical use, the ratio identified in the reported results as producing the best downstream performance was selected as the primary study-level ratio. This rule limited each study to one primary ratio per analytical use and prevented studies testing many ratios from contributing repeatedly to the compact frequency summaries; all tested ratios were retained in the study-level supplementary data.

Models were broadly categorized into deep learning and classical machine learning approaches. Among deep learning models, CNNs, with common variants including segmentation networks (eg, UNet [U-shaped convolutional neural network], UNet++, and Attention UNet), residual neural networks (eg, ResNet18 and ResNet50), inception-based architectures (eg, InceptionV3), densely connected convolutional network models (eg, DenseNet121), and standard feature extractors (eg, visual geometry group network [VGG16], MobileNet, and EfficientNet). GAN-based architectures, hybrid models, and, in rare cases, recurrent neural networks were also noted.

We documented the types of tasks addressed (eg, classification and segmentation) and extracted the reported performance metrics. For classification, these included accuracy, F_1 -score, recall (sensitivity), precision, specificity, and area under the receiver operating characteristic curve. For segmentation tasks, common metrics were the Dice similarity coefficient and Jaccard index. For each metric, we recorded available summary statistics, including mean values, SDs, CIs, and sample sizes. Explainability methods (eg, saliency-based maps) were also included in this section.

To facilitate comparative analysis, we stratified model performance by the dataset type used during training: original, traditionally augmented, synthetic-only, or synthetic-augmented. This stratification enabled the examination of performance trends across various imaging modalities and data augmentation strategies.

Explainability Methods

Because synthetic images may improve performance while also altering the visual features used by a model, we extracted explainability evidence as a separate assessment domain. Eligible methods included saliency and activation maps, gradient-weighted class activation mapping (Grad-CAM)/class activation mapping (CAM)-based approaches, attention mechanisms, feature-attribution methods, perturbation analyses, counterfactual explanations, and latent-space analyses. We recorded the purpose of each analysis (eg, inspection of clinically relevant regions, comparison of real and synthetic representations, artifact detection, class separability, or interpretation of model decisions) and noted whether outputs were clinically interpreted, expert-reviewed, or quantitatively validated against reference annotations. Conventional image-quality metrics and realism assessments were not considered evidence of explainability unless they

explicitly evaluated model reasoning. For method-purpose summaries, all reported method-purpose combinations within each study were counted.

Risk-of-Bias Assessment

Due to the lack of validated tools for assessing the risk of bias in studies on synthetic-image generation, we developed a set of criteria based on established reporting guidelines. These included the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI) extension for prediction model reporting [32,33], and selected domains from the ROBINS-I (Risk of Bias in Non-Randomized Studies of Interventions) tool [34] for evaluating bias in nonrandomized studies. Each study was assessed using the following predefined methodological domains: sampling strategy, population description, quality control, architecture description, synthetic-image generation and validation, outcome definition and reporting, and data completeness. For each domain, studies were categorized as having low concern, some concerns, or high concern based on these criteria. Table S7 in [Multimedia Appendix 1](#) provides detailed scoring guidelines that specify the conditions for classifying studies into different risk levels.

Risk of bias was assessed independently by 2 reviewers. Interrater reliability before consensus was evaluated at the domain level using exact percentage agreement and Cohen κ . Because ratings were ordinal, quadratically weighted Cohen kappa was used as the primary reliability statistic, with unweighted and linearly weighted kappa calculated as sensitivity analyses. Ratings of “not applicable,” when present, were excluded from ordinal kappa calculations and reviewed separately.

Disagreements were resolved by consensus. Overall risk-of-bias judgments were assigned after consensus, ignoring not-applicable domains. Studies were classified as high concern if 2 or more domains were rated high concern, or if 1 high-concern domain co-occurred with 3 or more domains rated as having some concerns. Studies were classified as some concerns if they had 1 high-concern domain with less than 3 some-concern domains, or 3 or more some-concern domains and no high-concern domains. All remaining studies were classified as low concern.

Synthesis Without Meta-Analysis

All data processing and visualization were performed in R (version 4.4.0; R Foundation for Statistical Computing). Results were synthesized descriptively and presented graphically, stratified by imaging modality. The synthesis was organized around the 2 review objectives: synthetic-image quality and downstream model performance. Image-quality evidence was first grouped as human-reader assessment or automated assessment and then by assessment task, metric family, specific metric, and imaging modality. Downstream evidence was grouped by training comparison (synthetic-only vs original/real or synthetic-augmented vs original/real), downstream task, harmonized performance metric, and

imaging modality. These groupings separated outcomes with different constructs and directions of interpretation. The protocol specified the 2 broad objectives but not all detailed analytic subgroups; the detailed grouping scheme was developed during extraction and analysis in response to the observed outcome heterogeneity, without changing the review eligibility criteria or objectives.

During data extraction, we noted, for each study, whether effect estimates included both the sample size and a measure of variance (SD, SE, or CI). These details were missing or incomplete in more than 80% of the eligible studies. Together with heterogeneity in populations, imaging modalities, generative models, outcome definitions, and analytic settings, this prevented the calculation of comparable effect estimates and inverse-variance weights; meta-analysis was therefore considered inappropriate. We used a structured synthesis without meta-analysis, based on study-level tabulation, frequencies, and unweighted distributions of reported values or within-study changes. We did not conduct vote counting based on statistical significance and did not calculate a pooled effect estimate or pooled CI.

Metrics were standardized only when their constructs and scales were compatible. Human-reader rates and proportions were expressed on a 0% to 100% scale; AUC values were retained on a 0 to 1 scale; and Likert, ranking, agreement, and other scores remained on their original scales. Automated image-quality metrics were retained in their native units and summarized separately by metric; Structural Similarity Index Measure (SSIM) or Multiscale SSIM (MS-SSIM) values reported as percentages were divided by 100, and metrics were separated according to whether higher or lower values indicated better quality. For downstream performance, strict within-study percentage change was calculated as $100 \times (\text{target-arm value} - \text{original/real-arm value}) / \text{absolute original/real-arm value}$. The sign was reversed for lower-is-better metrics, so that positive values consistently indicated improved performance and negative values indicated worsening.

Downstream values were paired only when the study, harmonized metric, downstream task, aggregation level, class or subgroup, operating point, comparison context, generative

model, downstream algorithm, and dataset matched. When duplicate source rows represented the same strict pairing unit and dataset arm, their reported means were averaged before pairing. Precomputed percentage-change outcomes were excluded from re-pairing to avoid calculating the same contrast twice. The paper-facing contrasts were restricted to synthetic-only versus original/real and synthetic-augmented versus original/real; synthetic-augmented versus synthetic-only was excluded because it addressed a different review question. For display, automated image-quality box plots required at least 7 studies and 7 eligible values per metric, while downstream metric plots required at least 20 strict paired-comparison data points within each contrast. These thresholds affected only the graphical display; all extracted results were retained in the supplementary tables of [Multimedia Appendix 1](#) and in the public study-level dataset.

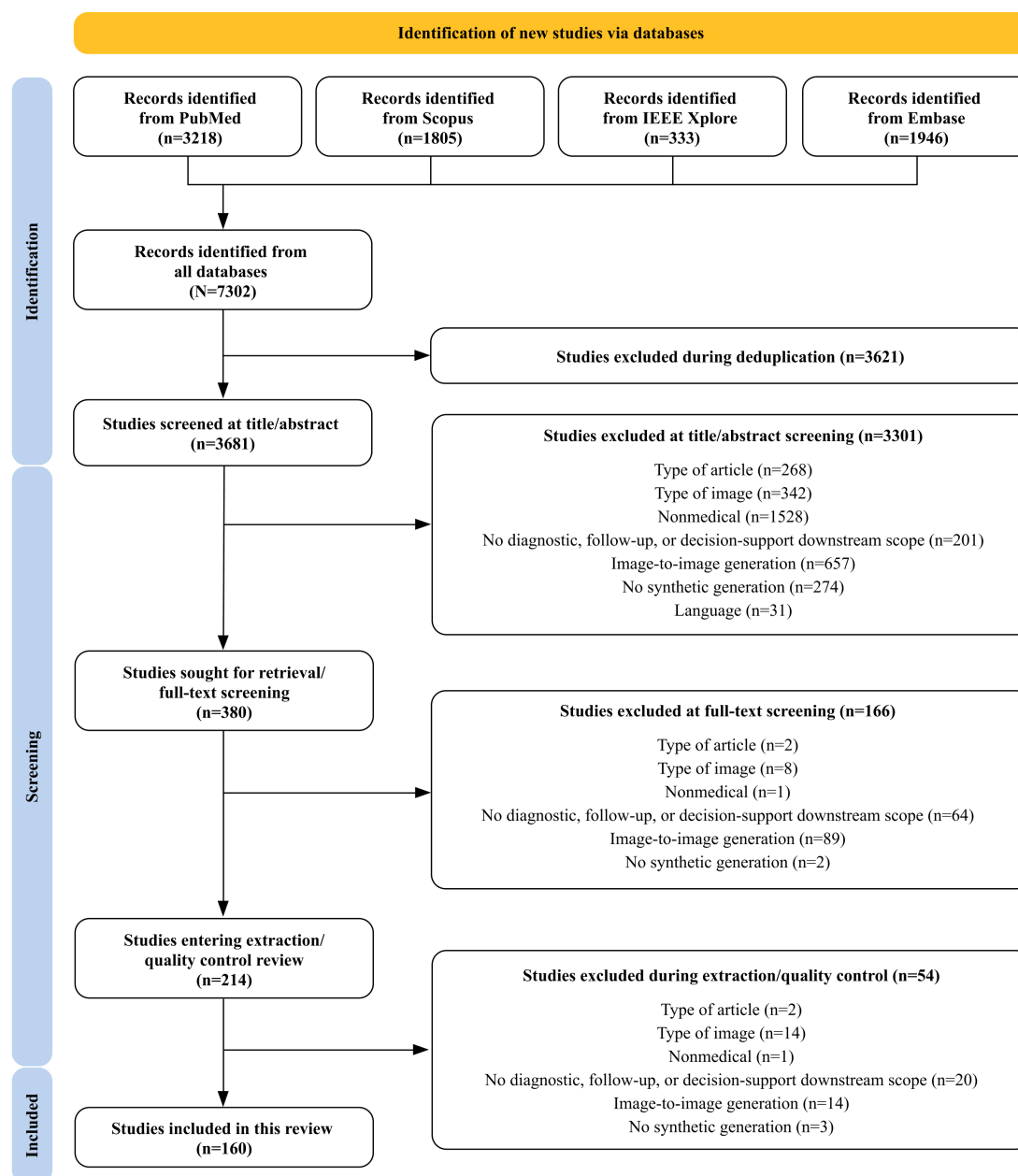
Potential heterogeneity was examined descriptively by comparing the direction, magnitude, and dispersion of findings across imaging modalities, clinical settings, generative-model families, image-quality assessment tasks, downstream tasks, dataset arms, and harmonized metrics. Risk-of-bias patterns and reporting completeness were considered when interpreting differences. Formal subgroup meta-analysis or meta-regression was not undertaken because comparable effect estimates and adequate uncertainty data were unavailable. Findings were presented in study-characteristic tables, frequency tables, bar plots, UpSet plots, heatmaps, and box plots showing medians, IQRs, and individual study-level data points.

Results

Study Selection and Characteristics

Our search identified 7302 records related to the use of synthetic images in human diagnosis, across PubMed, Scopus, IEEE Xplore, and Embase. After deduplication, screening, full-text assessment, and extraction and quality-control review, 160 studies met all eligibility criteria and were included in the systematic review ([Figure 1](#)). The main characteristics of the included studies are summarized in Table S8 in [Multimedia Appendix 1](#).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram summarizing study identification, screening, eligibility assessment, and inclusion in the systematic review.



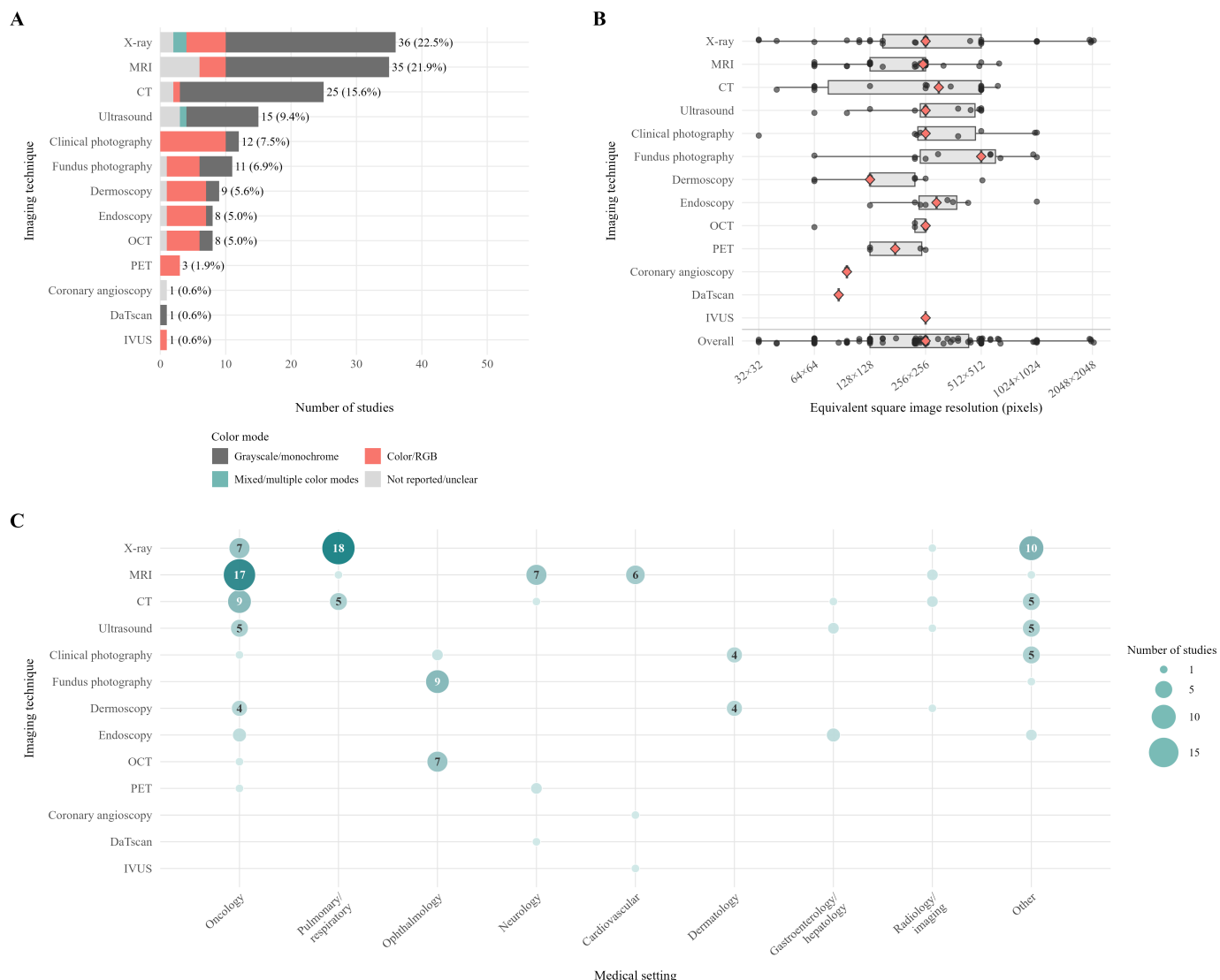
As shown in Figure 2A, x-ray, magnetic resonance imaging (MRI), and computed tomography (CT) were the leading techniques used for synthetic-image generation. Grayscale was the most used color modality in these leading image modalities, as well as in ultrasound; it was followed by RGB (red, green, and blue), particularly used in clinical and fundus photography, dermoscopy, endoscopy, and optical coherence tomography (OCT). By contrast, positron-emission tomography (PET), coronary angiography, DaTscan, and intravascular ultrasound (IVUS), together accounted for only about 4% (6/160) of studies.

Figure 2B shows that the resolutions of the generated synthetic images varied considerably, with the most common

size across studies clustering around 256×256 pixels, and a few ranging from as small as 32×32 to as large as 1024×1024.

Synthetic images were primarily applied in oncology, respiratory diseases, neurology, ophthalmology, and cardiovascular diseases. The bubble plot in Figure 2C further illustrates these associations, showing that MRI and CT were frequently used in oncology and neurology. In contrast, fundus imaging and optical coherence tomography images were mostly used in ophthalmology studies. Conversely, endocrinology, gynecology, nephrology, odontology, and otorhinolaryngology applications appeared in approximately 1% of studies, underscoring clinical areas in which the utility of synthetic images remains largely unexplored.

Figure 2. Overview of imaging techniques, color modes, image resolutions, and clinical settings in studies using synthetic images. (A) Number of studies by imaging technique and color mode; labels show n values and percentages for all 160 included studies. Studies reporting multiple imaging techniques contribute to each applicable technique. (B) Distribution of reported equivalent square image resolutions by imaging technique; points represent reported resolutions, boxes show the IQRs, and red diamonds show medians. The overall distribution is shown at the bottom. (C) Bubble plot showing imaging-technique and medical-setting co-occurrence. Bubble area represents the number of studies; medical settings with fewer than 5 studies are grouped as Other. CT: computed tomography; DaTscan: dopamine transporter scan; IVUS: intravascular ultrasound; MRI: magnetic resonance imaging; OCT: optical coherence tomography; PET: positron-emission tomography; RGB: red, green, and blue.



Architectures for Image Generation

As shown in Figure 3A, GAN-based architectures remained the dominant generative model family and were reported in 160 (n=119, 74.4%) studies. This family included heterogeneous GAN subtypes rather than a single architecture, including deep convolutional generative adversarial network (DCGAN), StyleGAN, conditional or auxiliary-classifier GAN, PGGAN, Wasserstein GAN, CycleGAN, Pix2Pix, and other task-specific variants. Diffusion models were reported in 28 (17.5%) studies and became more visible in recent publication years. Hybrid, procedural/statistical, scientific/physical simulation, registration/deformation, VAE/autoencoder, and other neural

synthesis approaches were less frequent. A detailed taxonomy and study-level architecture table are provided in Table S9 in Multimedia Appendix 1.

Across imaging modalities, GAN-based approaches were most commonly observed in MRI, CT, and x-ray studies, which were also the most frequently represented modalities overall (Figure 3C). They were also reported in less common modalities, such as PET, endoscopy, ultrasound, and photographic imaging. Less commonly used architecture families, including diffusion-based, hybrid, simulation-based, and VAE-based approaches, appeared across several imaging types, although with lower study counts.

Figure 3. Generative model architectures used for synthetic medical image generation across the 160 included studies. (A) Study-family reports by publication year; categories are not mutually exclusive because a study can report more than one family. (B) Most frequently reported architecture subtypes, with n values and percentages for all 160 included studies. (C) Bubble plot showing co-occurrence between broad generative model families and imaging techniques. Bubble area represents the number of studies, and color identifies the broad model family. CT: computed tomography; DaTscan: dopamine transporter scan; DCGAN: deep convolutional generative adversarial network; DDPM: denoising diffusion probabilistic model; IVUS: intravascular ultrasound; GAN: generative adversarial network; MRI: magnetic resonance imaging; OCT: optical coherence tomography; PET: positron-emission tomography; StyleGAN: style-based generative adversarial network; VAE: variational autoencoder; VAE-GAN: variational autoencoder generative adversarial network.

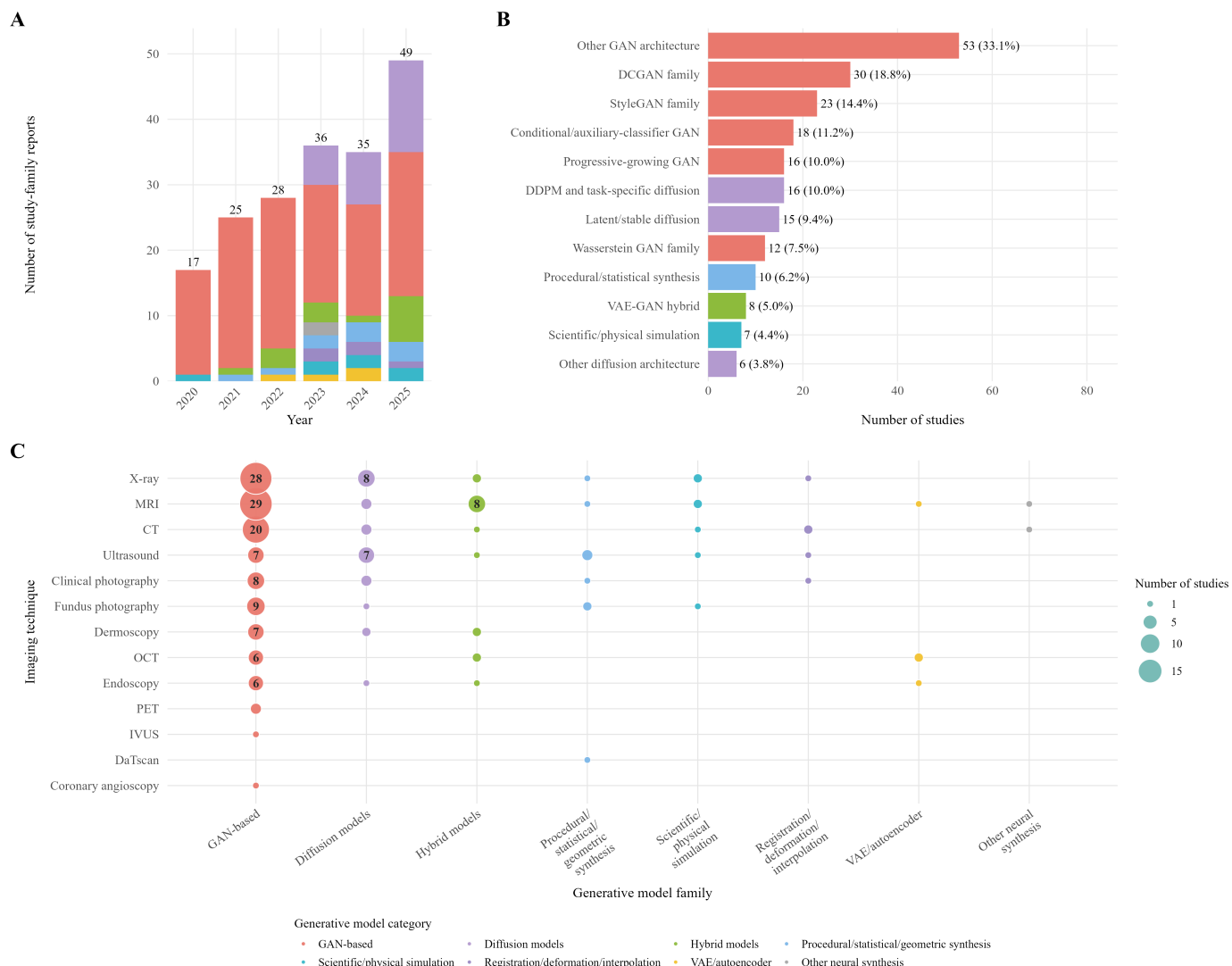


Image Quality

Overall, 160 (n=113, 70.6%) studies assessed synthetic-image quality using at least one human-reader or automated approach. Human-reader evidence was reported in 50 (31.2%) studies, automated image-quality assessment in 101 (63.1%) studies, and both approaches in 38 (23.8%) studies.

The most frequent task category among human-reader assessments was real-versus-synthetic discrimination or Turing-style assessment, followed by realism or image-quality ratings, qualitative expert review, diagnostic interpretation, feature identification or annotation, clinical plausibility or usefulness ratings, and preference or ranking tasks (Figure

4A). Most studies using human-reader approaches to assess image quality missed relevant methodological information (Figure 4B). At the metric level, the most frequently extracted end points were Turing-style discrimination metrics, followed by interrater agreement, authenticity rates, Likert or rating scores, and diagnostic performance metrics (Figure 4C). Quantitative rate/proportion end points that could be harmonized to a common 0% to 100% scale showed heterogeneous distributions across tasks and metrics (Figure 4D). A quantitative synthesis of tasks and metrics used for human-reader assessment of image quality is provided in Table S10 and Table S11 in Multimedia Appendix 1, respectively.

Figure 4. Human-reader study evidence and reporting completeness (50/160 studies). (A) Frequency of human-reader task categories, counted across unique reader studies. (B) Reporting completeness for key reader-study design items among the 50 studies with human-reader evidence. (C) Frequency of specific extracted human-reader study metrics. (D) Quantitative rate/proportion end points harmonized to a common 0% to 100% scale; studies could contribute more than one value when results were reported by reader group, image group, dataset, model, or end point. Categories are not mutually exclusive. Boxes represent the IQR (25th-75th percentile), vertical lines with orange diamonds indicate the median. Right-side labels show the median reported value and the IQR for each metric. AUC: area under the curve.

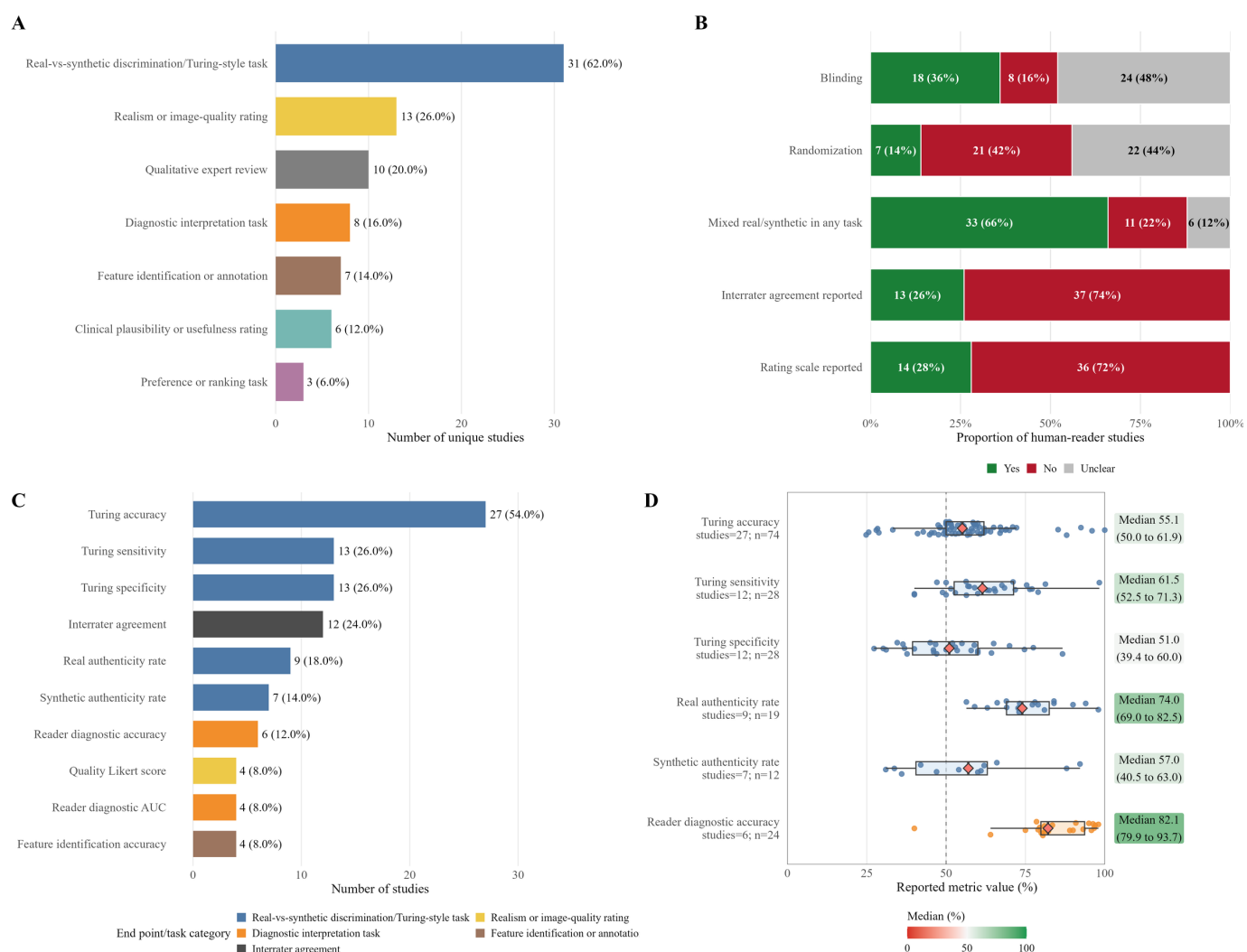


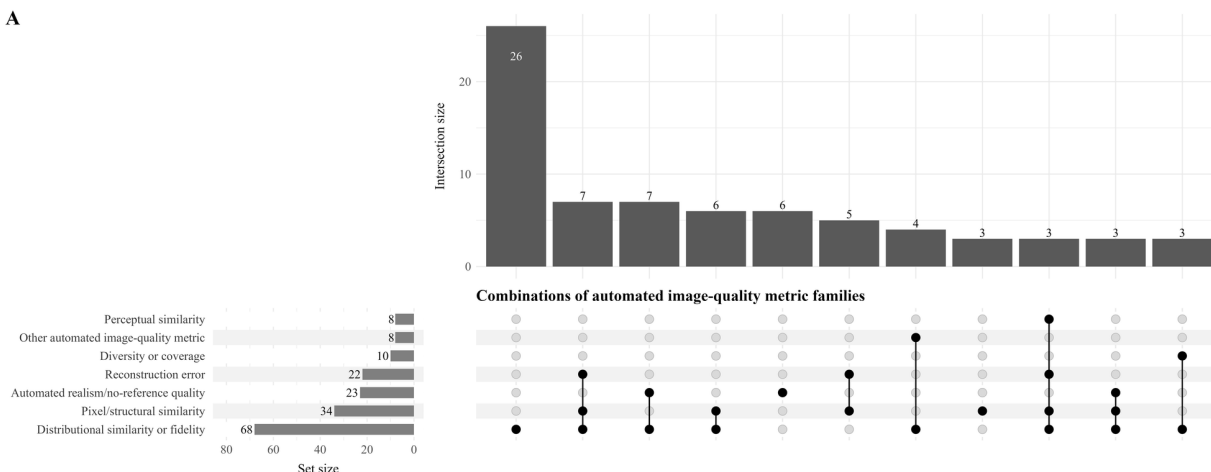
Figure 5 summarizes the use and results of automated metrics for assessing image quality. The most common trend among studies assessing image quality with automated metrics was the use of a single metric within either the distributional similarity or fidelity group (Figure 5A), primarily the Fréchet inception distance (FID) (Figure 5B). The quantitative synthesis analysis of automated metrics used for assessing image quality is summarized in Table S12 (metric groups) and Table S13 (specific metrics) in Multimedia Appendix 1.

Figure 5C-D show the distribution of reported values for the most frequently used automated image-quality metrics.

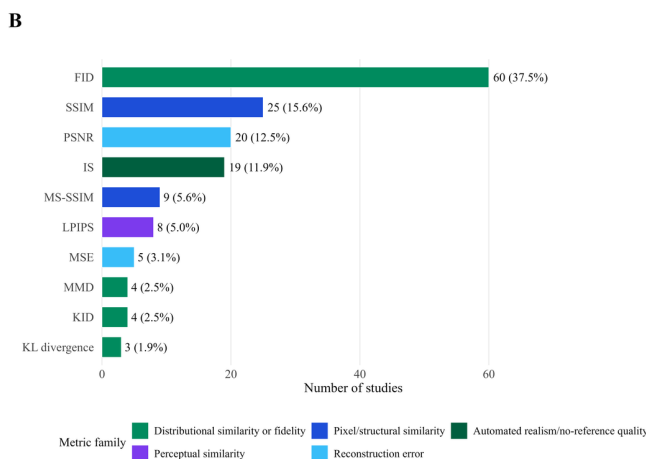
Among metrics for which lower values indicate better image quality, FID showed substantial variability across studies, a pattern also observed for learned perceptual image patch similarity (LPIPS). Among metrics where higher values indicate better image quality, SSIM was the most frequently reported, followed by peak signal-to-noise ratio (PSNR) and inception score. Overall, the wide dispersion of values across metrics highlights the substantial heterogeneity of automated image-quality assessment and limits direct comparability across studies.

Figure 5. Summary of automated image-quality metrics used to assess synthetic images. (A) The UpSet plot illustrates combinations of automated quality assessment strategies reported across studies (≥ 3 studies). Each dot-and-line intersection indicates a specific combination of strategies used within the same study. Bar heights represent the number of studies using each combination. (B) Frequency of individual automated image-quality metrics (reported in ≥ 3 studies), grouped by overarching metric family. (C-D) Box plots of numeric values for higher-count metrics (reported in ≥ 7 studies), separated according to whether lower values indicate better image quality (C) or higher values indicate better image quality (D). Boxes represent the IQR (25th-75th percentile), horizontal lines with orange diamonds indicate the median. FID: Fréchet inception distance; IS: inception score; KID: kernel inception distance; KL: Kullback-Leibler; LPIPS: learned perceptual image patch similarity; MMD: maximum mean discrepancy; MSE: mean squared error; MS-SSIM: Multiscale Structural Similarity Index Measure; PSNR: peak signal-to-noise ratio; SSIM: Structural Similarity Index Measure.

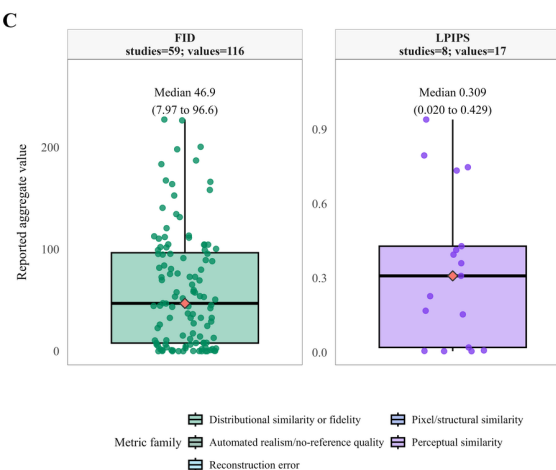
A



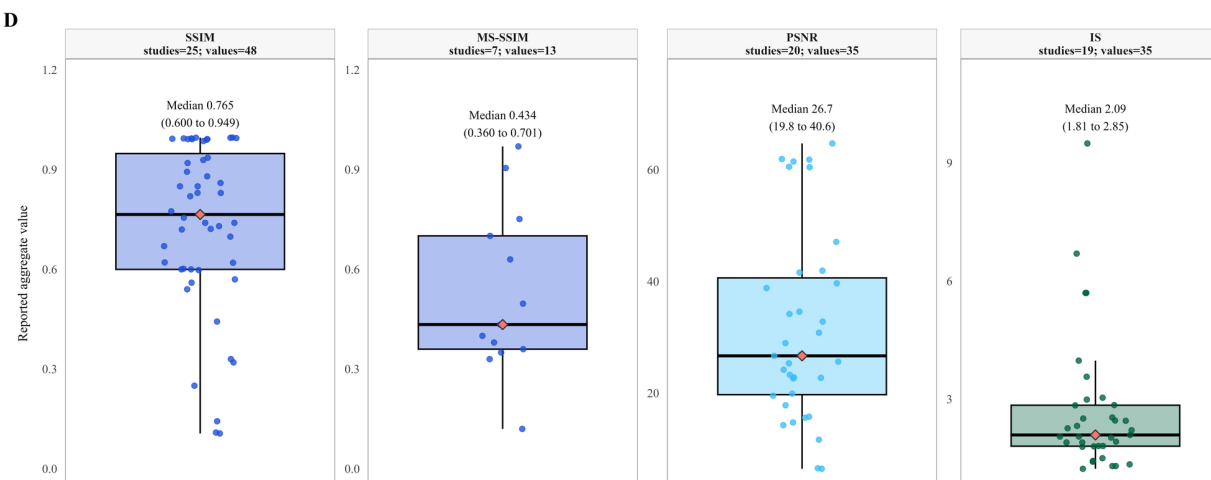
B



C



D



Impact of Synthetic Data on Downstream Model Performance

Most studies (110/160, 68.8%) reported the real-to-synthetic data ratio; among these studies, balanced designs such as 1:1

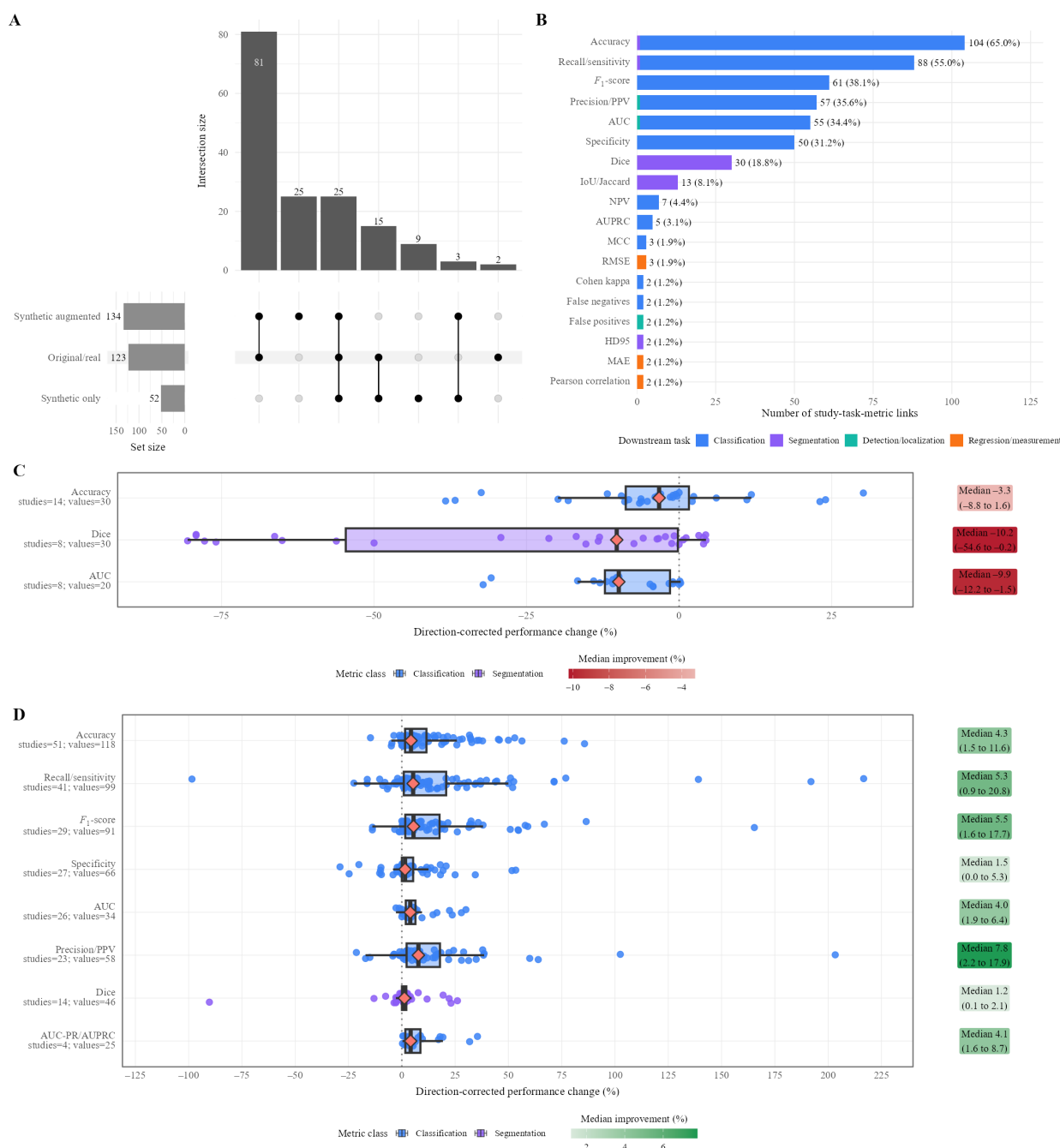
were common, while several studies used synthetic-enriched configurations such as 1:2, 1:5, or 1:10; among studies with unbalanced designs, they tended to include more synthetic than real images (Figure S1, [Multimedia Appendix 1](#)).

Detailed frequencies of ratios at the study level are provided in Figure S2 in [Multimedia Appendix 1](#).

Figure 6 summarizes the downstream evaluation strategies and performance outcomes reported across studies. The most frequent dataset-arm pattern was the joint reporting of

original/real and synthetic-augmented evidence, either alone or as part of studies that also included a synthetic-only arm. Synthetic-only evidence was less frequent and was reported either alone or in combination with original/real or with synthetic-augmented arms (Figure 6A).

Figure 6. Downstream dataset-arm evidence, evaluation metrics, and performance changes associated with synthetic images. (A) The UpSet plot showing the study-level overlap of downstream dataset arms reported across studies, including original/real, synthetic-only, and synthetic-augmented arms. (B) Frequency of reported downstream evaluation metrics, counted as study-task-metric links and stacked by downstream task; metrics reported in at least 2 study-task combinations are shown. (C) Direction-corrected percentage change in downstream performance for synthetic-only vs original/real datasets. (D) Direction-corrected percentage change in downstream performance for synthetic-augmented vs original/real datasets. Positive values indicate improved performance relative to the original/real-data baseline. Plots show harmonized metrics with at least 20 strict paired comparison data points. Boxes represent the IQR (25th-75th percentile), vertical lines with orange diamonds indicate the median, and points represent individual paired comparison data points. Right-side labels show the median (Q1-Q3) improvement for each metric. AUC: area under the curve; AUC-PR/AUPRC: area under the precision-recall curve; IoU: intersection over union; MAE: mean absolute error; MCC: Matthews correlation coefficient; NPV: negative predictive value; PPV: positive predictive value; RMSE: root mean square error.



Classification was the predominant downstream task, and accuracy was by far the most frequently reported metric (104 study-task-metric links), followed by recall/sensitivity (88), F_1 -score (61), precision/positive predictive value (PPV) (57), area under the curve (AUC) (55), and specificity (50) (Figure 6B). Segmentation studies most frequently reported the Dice coefficient (30) and IoU/Jaccard index (13), whereas regression and localization outcomes were comparatively uncommon.

Across comparisons of synthetic-only versus original/real datasets, performance tended to decline for the plotted metrics, particularly the Dice coefficient and AUC, whereas accuracy showed a smaller median decrease (Figure 6C). In contrast, studies evaluating synthetic augmentation generally reported improved downstream performance relative to original/real datasets alone (Figure 6D). Median improvements were observed for most classification metrics, including precision/PPV (+7.8%), recall/sensitivity (+5.3%), F_1 -score (+5.5%), accuracy (+4.3%), area under the precision recall curve (AUPRC or AUC-PR)(+4.1%), and AUC (+4.0%), whereas specificity showed smaller gains (+1.5%). Improvements in segmentation performance were limited, with a median improvement in the Dice coefficient of only +1.2 %.

Explainability or model-interpretability evidence was reported in 68 of the 160 (42.5%) studies. The frequencies of the different methods, the validation of explanations, and their purposes are summarized in Figure S3 in [Multimedia Appendix 1](#). Overall, latent-space or representation analyses were the most common approach, followed by Grad-CAM/CAM methods and saliency maps. Validation of explanation outputs was uncommon: only about half of the studies assessed clinical relevance, whereas quantitative validation against annotations, expert review, and explicit clinical expert involvement were rarely reported. Latent-space analyses were primarily used to examine representation structure and class separability, while Grad-CAM/CAM and saliency methods were mainly used to assess whether model attention aligned with clinically relevant regions. Overall, explainability was more often used as supportive evidence than as a formal validation of model behavior.

Risk of Bias

Figure 7 summarizes the risk-of-bias assessment across 7 methodological domains: sampling strategy, source population description, image quality and preprocessing control, generative architecture reporting, generation-validation data separation, outcome definition and reporting, and data completeness and statistical reporting.

Before consensus, exact agreement across all risk-of-bias domain ratings was 55.1%, and the pooled, quadratically

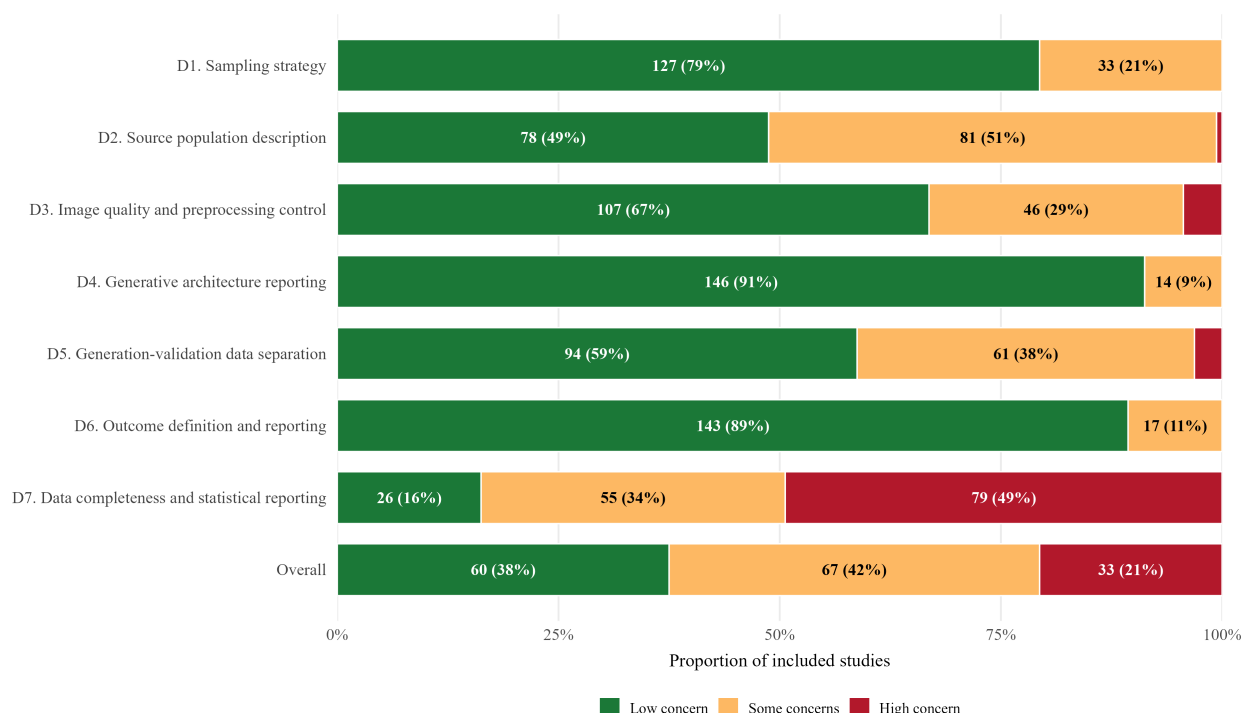
weighted Cohen κ was 0.322. Unweighted and linearly weighted κ values were 0.172 and 0.242, respectively. Domain-specific reliability estimates are provided in Table S14 in [Multimedia Appendix 1](#).

After reaching consensus, 60 of the 160 (38%) included studies were classified as having low concern, 67 (42%) as having some concerns, and 33 (21%) as having high concern overall (Figure 7). The strongest methodological domains were generative architecture reporting (n=146, 91% low concern) and outcome definition and reporting (n=143, 89% low concern). In contrast, data completeness and statistical reporting represented the main methodological weaknesses, with only 16% (n=26) of studies rated as low concern and 49% (n=79) classified as high concern. Additional concerns were observed with source population description and generation-validation data separation. Individual study ratings and the full risk-of-bias matrix are provided in Figure S4 in [Multimedia Appendix 1](#).

Reproducibility and openness reporting was heterogeneous across studies (Figure S5 in [Multimedia Appendix 1](#)). Evaluation protocols and training details were commonly reported, whereas source code, synthetic images, random seeds, model checkpoints, and pretrained models were rarely made available. Public datasets were more frequently accessible than generative-model artifacts. The most commonly used public/community datasets among included studies were International Skin Imaging Collaboration (ISIC) Archive, Lung Image Database Consortium—Image Database Resource Initiative (LIDC-IDRI), Alzheimer's Disease Neuroimaging Initiative (ADNI), Brain Tumor Segmentation Challenge (BraTS), and Human Against Machine With 10,000 Training Images (HAM10000), while ImageNet was frequently used for model development or pretraining (Figure S6 in [Multimedia Appendix 1](#)). [Multimedia Appendix 2](#) provides a complete catalog of datasets used in the reviewed studies.

We did not assign GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) or another formal certainty rating because the evidence comprised heterogeneous AI development and validation studies rather than a common intervention-effect question. We therefore calibrated our interpretation according to the number and directness of contributing studies, consistency of findings, completeness of statistical reporting, availability of strict within-study comparisons, and risk of bias. Descriptive prevalence patterns supported by many studies were considered more robust than comparative performance findings, for which certainty remained limited.

Figure 7. Risk-of-bias assessment across included studies on synthetic medical image generation and validation (N=160). Stacked horizontal bars show the proportion and number of studies classified as low concern, some concerns, or high concern for each of the 7 predefined methodological domains and for the overall judgment. The domains assessed included reporting of sampling strategy, source population characteristics, image quality and preprocessing control, generative architecture, separation between generation and validation data, outcome definition and reporting, and data completeness and statistical reporting. Final judgments were derived after consensus adjudication of reviewer disagreements.



Discussion

Principal Findings

Our systematic review of 160 studies reporting the generation of synthetic images and their adequacy or quality in the context of diagnostic imaging reveals both the promise and current limitations of this rapidly evolving technology in health care. The findings highlight significant heterogeneity in methodological approaches, evaluation strategies, and performance outcomes that collectively point to an urgent need for standardization in this field.

Grayscale-based imaging modalities, particularly MRI, x-ray, and CT, were the most frequently represented techniques for synthetic-image generation, especially in oncology, respiratory medicine, and neurology. However, our review was not designed to determine the reasons for this modality imbalance. Future studies should examine whether it reflects differences in data availability, the maturity of established AI workflows, or modality-specific challenges in image generation and validation. Photograph-based diagnostic modalities, including dermoscopy, endoscopy, fundus photography, and other clinical photography applications, were comparatively underrepresented. These modalities therefore remain important areas for future research, particularly because image color, texture, illumination, and fine surface features may require evaluation approaches that differ from those used for cross-sectional or projection imaging.

Regarding generative architecture, GAN-based methods remained dominant and were reported in 119 of the 160 included studies. However, this family comprised a heterogeneous array of architectures, including DCGANs, StyleGANs, conditional GANs, PGGANs, Wasserstein variants, CycleGANs, Pix2Pix models, and hybrid approaches. Diffusion models were less frequent overall but became increasingly represented in recent years, including applications in controllable 3D image generation, text-guided CT synthesis, pathology-conditioned generation, low-dose PET reconstruction, and fairness-aware synthesis [35-39]. These developments suggest that the field is evolving toward a more diverse ecosystem of generative approaches, although evaluation methods have not developed at the same pace.

One of the most important findings was the heterogeneity in synthetic-image quality assessment. Overall, 113 studies assessed image quality using at least one human-reader or automated approach. Human-reader evidence was reported in 50 studies, automated assessments in 101, and 38 studies combined both approaches.

Automated assessments included distributional metrics such as FID, maximum mean discrepancy, and kernel inception distance; structural metrics such as SSIM and PSNR; no-reference measures such as inception score; perceptual metrics such as LPIPS; and measures of reconstruction error, diversity, texture, and intensity. Differences in the constructs, directions, and reported values of these metrics limited direct comparisons across studies. As no single metric captures all clinically relevant dimensions of quality, evaluation should combine distributional fidelity, diversity, preservation of anatomical and pathological

features, detection of artifacts or memorization, task-based utility, and structured human-reader assessment.

Recently proposed frameworks, including the Synthetic Medical Data Scorecard and the Clinical Evaluation of Medical Image Synthesis protocol, provide useful starting points for developing more consistent and clinically meaningful evaluation practices [40,41]. Our analysis also showed that many human-reader assessments lacked essential methodological information. Future studies should report reader expertise, blinding, randomized or shuffled image presentation, rating-scale definitions, and interrater agreement, in line with broader reporting recommendations for medical-imaging AI [40,42].

The downstream-performance synthesis showed even greater heterogeneity because studies differed in clinical tasks, generative and downstream architectures, dataset composition, real-to-synthetic ratio, comparators, operating points, and reported metrics. Several studies lacked a clearly matched original-data baseline, limiting assessment of the isolated contribution of synthetic data. In strict within-study comparisons, models trained exclusively on synthetic images generally performed worse than those trained on real images, particularly for AUC and the Dice coefficient. These findings suggest that synthetic images cannot currently be assumed to be an interchangeable replacement for real training data.

By contrast, adding synthetic images to real training datasets was associated with modest median improvements in most classification metrics, including precision, recall, F_1 -score, accuracy, AUC, and AUPRC. Improvements in specificity were smaller, and the median improvement in the Dice coefficient was only 1.2%. The smaller gains in segmentation may reflect the need to preserve precise image-label correspondence, anatomical structure, and lesion boundaries, although this interpretation should be evaluated directly in future studies. Importantly, these changes are unweighted descriptive summaries rather than pooled causal effects and should not be interpreted as evidence that synthetic augmentation will improve performance in every modality, population, or clinical setting.

Despite growing interest in explainable AI, fewer than half of the studies used explainability methods. Latent-space and representation analyses were most common, followed by Grad-CAM, CAM, and other saliency-based methods. Formal validation against reference annotations, quantitative localization measures, or expert assessment was uncommon. Previous medical-imaging research has shown that commonly used saliency methods may have limited localization utility, repeatability, and reproducibility [43]. Explainability outputs should therefore not be considered reliable evidence of clinically appropriate model behavior without quantitative or expert validation.

Beyond performance and explainability, safety, privacy, and governance should also be considered. Although synthetic data are often presented as a privacy-preserving alternative to real patient data, generative models may memorize training examples or reveal information through nearest-neighbor similarity and membership-inference attacks

[44,45]. The provenance, consent basis, licensing, intended use, population representation, and limitations of the source data should therefore be clearly documented. Structured tools such as model cards, datasheets for datasets, and the STANDING Together recommendations may improve transparency regarding model performance, dataset provenance, intended use, limitations, and potential risks [46-48]. Explicit disclosure is also needed when synthetic images are used for model development, training, or evaluation.

Because no widely accepted tool is available for assessing risk of bias in synthetic-image generation studies, our criteria were pragmatically derived from relevant reporting guidelines, including STROBE and TRIPOD+AI [32,33], and should be interpreted as a structured methodological appraisal rather than a validated measurement instrument. After consensus, 60 studies were classified as having low overall concern, 67 as having some concerns, and 33 as having high concern. Generative architecture reporting and outcome definition and reporting were the strongest domains, with 91% and 89% of studies, respectively, rated as low concern. In contrast, data completeness and statistical reporting represented the main methodological weaknesses, with only 16% rated as low concern and 49% as high concern. Additional concerns were observed for source population description and generation-validation data separation. Incomplete reporting of means, variance measures, sample sizes, and pairing information prevented inverse-variance weighting and formal meta-analysis.

Consequently, the reported medians and IQRs give each eligible study-level data point equal weight, do not account for differences in precision or sample size, and should not be interpreted as pooled or causal effects. A study could also contribute more than one data point when it reported distinct eligible tasks, datasets, models, classes, or operating points. Future research should report baseline performance, uncertainty estimates, sample sizes, and sufficient pairing information to enable more robust synthesis.

The synthesis without meta-analysis has additional limitations. Detailed synthesis groupings were refined after data extraction, and alternative grouping choices might produce different patterns. Selecting the best-performing real-to-synthetic ratio for compact frequency summaries may overstate favorable performance, although all tested ratios remain available in the study-level data. Minimum-count thresholds excluded sparsely reported metrics from the main figures, formal certainty ratings were not assigned, and multiple data points from individual studies may not be statistically independent. These factors limit the strength and generalizability of the conclusions.

In summary, our findings indicate that synthetic images may help address data scarcity and imbalance, but their benefits cannot be universally assumed. Synthetic-only training generally underperformed real-data training, whereas augmentation was associated with modest classification improvements and smaller segmentation gains. Harmonized criteria and standardized methodologies are therefore needed to assess image quality, downstream utility, privacy, and

reporting completeness. A consensus reporting guideline, potentially developed in alignment with the EQUATOR Network, would help improve the validity, reproducibility, and comparability of studies in this field [49].

This review has several limitations. The rapid evolution of generative AI means that new architectures and applications may have emerged after the prespecified publication period. We included only peer-reviewed articles and excluded preprints, workshops, and conference proceedings. Although this may have omitted recent developments, it reduced publication-related heterogeneity and ensured that sufficient information was available for extraction and risk-of-bias assessment. We also excluded microscopic and histopathologic imaging and image-to-image transformation studies, limiting generalizability to these areas.

Finally, the search used a focused set of title- and abstract-based free-text terms rather than a PICO (Population-Intervention-Comparator-Outcome)-structured framework. This reflected the broad methodological nature of the review, which covered multiple specialties, modalities, generative architectures, downstream applications, and outcomes. Nevertheless, this approach may have missed studies that

did not explicitly describe synthetic-image generation in their titles or abstracts.

Conclusions

Our systematic review provides a comprehensive quantitative assessment of synthetic-image quality evaluation and downstream model impact across diagnostic medical imaging studies. Unlike previous reviews that focused on specific diseases, modalities, or generative architectures, it jointly examines image quality, downstream performance, explainability, reproducibility, and methodological quality across imaging domains.

Synthetic images show promise for addressing data scarcity and imbalance, particularly when used to complement real training datasets. However, synthetic-only training generally provides limited value, whereas synthetic augmentation is associated with modest improvements in classification performance and smaller gains in segmentation. Standardized multimethod evaluation, privacy assessment, transparent reporting, and rigorous external validation are needed to support the reliable translation of synthetic-image technologies into clinical AI applications.

Acknowledgments

Generative AI (GenAI) tools, including ChatGPT (OpenAI), were used to assist with language editing and with drafting and rephrasing selected passages of this paper. All AI-assisted text was critically reviewed and edited by the authors, who take full responsibility for the final content. GenAI was not used to perform statistical analyses or to independently interpret the findings. Grammarly was also used for language editing.

Funding

This study was conducted as part of the project AICE (AI-Supported Image Analysis in Large Bowel Camera Capsule Endoscopy), funded by the European Union under the HORIZON program (number: 101057400) and co-funded by the UK Government.

Data Availability

The study-level extraction dataset and analysis scripts are publicly available on GitHub [50].

Authors' Contributions

JF-P and MR-P share first authorship. JF-P, GC-S, JA-R, and DD-H extracted articles and selected them according to inclusion and exclusion criteria. JF-P, MR-P, AJ-R, JA-R, and DD-H participated in the investigation, data extraction, and curation of the selected articles. JF-P, MR-P, and GC-S led the conceptualization, methodology, and formal analysis. EN and FT provided methodological input and validation. The original draft was written by JF-P, MR-P, and GC-S, and subsequently revised and edited by all authors. Figures were prepared by MR-P. GC-S and JP-J provided supervision. All authors agreed to the submitted version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Methods and results (figures and tables), including search strategies, methodological definitions and taxonomies, risk-of-bias criteria, study characteristics, quantitative summaries, and supplementary figures on downstream data composition, explainability, risk of bias, reproducibility, and dataset use.

[DOCX File (Microsoft Word File), 3395 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Study-level data, including human-reader assessment evidence, a catalog of datasets used across the included studies, and detailed risk-of-bias assessment and reviewer-agreement records.

[[XLSX File \(Microsoft Excel File\), 382 KB-Multimedia Appendix 2](#)]

Checklist 1

PRISMA 2020, PRISMA-S, and SWiM checklists.

[[DOCX File \(Microsoft Word File\), 52 KB-Checklist 1](#)]

References

- Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: scoping review. *J Med Internet Res*. Oct 5, 2022;24(10):e40238. [doi: [10.2196/40238](#)] [Medline: [36197712](#)]
- Alowais SA, Alghamdi SS, Alsuehaby N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*. Sep 22, 2023;23(1):689. [doi: [10.1186/s12909-023-04698-z](#)] [Medline: [37740191](#)]
- Goceri E. Medical image data augmentation: techniques, comparisons and interpretations. *Artif Intell Rev*. Mar 20, 2023;56(11):1-45. [doi: [10.1007/s10462-023-10453-z](#)] [Medline: [37362888](#)]
- Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digit Med*. Apr 26, 2023;6(1):74. [doi: [10.1038/s41746-023-00811-0](#)] [Medline: [37100953](#)]
- Koutsoubis N, Yilmaz Y, Ramachandran RP, Schabath M, Rasool G. Privacy preserving federated learning in medical imaging with uncertainty estimation. *arXiv*. Preprint posted online on Jun 18, 2024. [doi: [10.48550/arXiv.2406.12815](#)]
- Koetzier LR, Wu J, Mastrodicasa D, et al. Generating synthetic data for medical imaging. *Radiology*. Sep 2024;312(3):e232471. [doi: [10.1148/radiol.232471](#)] [Medline: [39254456](#)]
- Mendes JM, Barbar A, Refaie M. Synthetic data generation: a privacy-preserving approach to accelerate rare disease research. *Front Digit Health*. 2025;7:1563991. [doi: [10.3389/fdgth.2025.1563991](#)] [Medline: [40171526](#)]
- McNulty JR, Kho L, Case AL, Slater D, Abzug JM, Russell SA. Synthetic medical imaging generation with generative adversarial networks for plain radiographs. *Appl Sci*. Aug 5, 2024;14(15):6831. [doi: [10.3390/app14156831](#)]
- Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. Dec 2019;6(1):60. [doi: [10.1186/s40537-019-0197-0](#)]
- Shi G, Wang J, Qiang Y, et al. Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification. *Comput Methods Programs Biomed*. Nov 2020;196:105611. [doi: [10.1016/j.cmpb.2020.105611](#)] [Medline: [32650266](#)]
- Krithika alias Anbu Devi M, Suganthi K. Review of medical image synthesis using GAN techniques. *ITM Web Conf*. 2021;37:01005. [doi: [10.1051/itmconf/20213701005](#)]
- Abdusalomov AB, Nasimov R, Nasimova N, Muminov B, Whangbo TK. Evaluating synthetic medical images using artificial intelligence with the GAN algorithm. *Sensors (Basel)*. Mar 24, 2023;23(7):3440. [doi: [10.3390/s23073440](#)] [Medline: [37050503](#)]
- Islam T, Hafiz MS, Jim JR, Kabir MM, Mridha MF. A systematic review of deep learning data augmentation in medical imaging: recent advances and future research directions. *Healthc Anal*. Jun 2024;5:100340. [doi: [10.1016/j.health.2024.100340](#)]
- Sorin V, Barash Y, Konen E, Klang E. Creating artificial images for radiology applications using generative adversarial networks (GANs)—a systematic review. *Acad Radiol*. Aug 2020;27(8):1175-1185. [doi: [10.1016/j.acra.2019.12.024](#)] [Medline: [32035758](#)]
- Jose L, Liu S, Russo C, Nadort A, Di Ieva A. Generative adversarial networks in digital pathology and histopathological image processing: a review. *J Pathol Inform*. 2021;12(1):43. [doi: [10.4103/jpi.jpi_103_20](#)] [Medline: [34881098](#)]
- Saeed AQ, Sheikh Abdullah SNH, Che-Hamzah J, Abdul Ghani AT. Accuracy of using generative adversarial networks for glaucoma detection: systematic review and bibliometric analysis. *J Med Internet Res*. Sep 21, 2021;23(9):e27414. [doi: [10.2196/27414](#)] [Medline: [34236992](#)]
- Dimitriadis A, Trivizakis E, Papanikolaou N, Tsiknakis M, Marias K. Enhancing cancer differentiation with synthetic MRI examinations via generative models: a systematic review. *Insights Imaging*. Dec 12, 2022;13(1):188. [doi: [10.1186/s13244-022-01315-3](#)] [Medline: [36503979](#)]
- Jeong JJ, Tariq A, Adejumo T, Trivedi H, Gichoya JW, Banerjee I. Systematic review of generative adversarial networks (GANs) for medical image classification and segmentation. *J Digit Imaging*. Apr 2022;35(2):137-152. [doi: [10.1007/s10278-021-00556-w](#)] [Medline: [35022924](#)]
- Liu J, Li K, Dong H, Han Y, Li R. Medical image processing based on generative adversarial networks: a systematic review. *Curr Med Imaging*. 2024;20:e15734056258198230920042358. [doi: [10.2174/0115734056258198230920042358](#)] [Medline: [37881080](#)]
- Ng CKC. Generative adversarial network (generative artificial intelligence) in pediatric radiology: a systematic review. *Children (Basel)*. Aug 10, 2023;10(8):1372. [doi: [10.3390/children10081372](#)] [Medline: [37628371](#)]

21. Sherwani MK, Gopalakrishnan S. A systematic literature review: deep learning techniques for synthetic medical image generation and their applications in radiotherapy. *Front Radiol.* 2024;4:1385742. [doi: [10.3389/fradi.2024.1385742](https://doi.org/10.3389/fradi.2024.1385742)] [Medline: [38601888](https://pubmed.ncbi.nlm.nih.gov/38601888/)]
22. de Vries JJ, Laan DM, Frey F, Koenderink GH, de Maat MPM. A systematic review and comparison of automated tools for quantification of fibrous networks. *Acta Biomater.* Feb 2023;157:263-274. [doi: [10.1016/j.actbio.2022.12.009](https://doi.org/10.1016/j.actbio.2022.12.009)] [Medline: [36509400](https://pubmed.ncbi.nlm.nih.gov/36509400/)]
23. Ibrahim M, Khalil YA, Amirrajab S, et al. Generative AI for synthetic data across multiple medical modalities: a systematic review of recent developments and challenges. *Comput Biol Med.* May 2025;189:109834. [doi: [10.1016/j.compbiomed.2025.109834](https://doi.org/10.1016/j.compbiomed.2025.109834)] [Medline: [40023073](https://pubmed.ncbi.nlm.nih.gov/40023073/)]
24. Garcea F, Serra A, Lamberti F, Morra L. Data augmentation for medical imaging: a systematic literature review. *Comput Biol Med.* Jan 2023;152:106391. [doi: [10.1016/j.compbiomed.2022.106391](https://doi.org/10.1016/j.compbiomed.2022.106391)] [Medline: [36549032](https://pubmed.ncbi.nlm.nih.gov/36549032/)]
25. McNaughton J, Fernandez J, Holdsworth S, Chong B, Shim V, Wang A. Machine learning for medical image translation: a systematic review. *Bioengineering (Basel).* Sep 12, 2023;10(9):1078. [doi: [10.3390/bioengineering10091078](https://doi.org/10.3390/bioengineering10091078)] [Medline: [37760180](https://pubmed.ncbi.nlm.nih.gov/37760180/)]
26. Chen Y, Yang XH, Wei Z, et al. Generative adversarial networks in medical image augmentation: a review. *Comput Biol Med.* May 2022;144:105382. [doi: [10.1016/j.compbiomed.2022.105382](https://doi.org/10.1016/j.compbiomed.2022.105382)]
27. Fu X, Li X, Delamare E, Dunn AG, Bi L, Kim J. A systematic review of generative artificial intelligence techniques for synthetic medical image datasets: quality, models, public availability and applications. *Comput Methods Programs Biomed.* Jun 2026;280:109331. [doi: [10.1016/j.cmpb.2026.109331](https://doi.org/10.1016/j.cmpb.2026.109331)] [Medline: [41895216](https://pubmed.ncbi.nlm.nih.gov/41895216/)]
28. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Syst Rev.* Jun 2022;18(2):e1230. [doi: [10.1002/cl2.1230](https://doi.org/10.1002/cl2.1230)] [Medline: [36911350](https://pubmed.ncbi.nlm.nih.gov/36911350/)]
29. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Syst Rev.* Jan 26, 2021;10(1):39. [doi: [10.1186/s13643-020-01542-z](https://doi.org/10.1186/s13643-020-01542-z)] [Medline: [33499930](https://pubmed.ncbi.nlm.nih.gov/33499930/)]
30. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ.* Jan 16, 2020;368:l6890. [doi: [10.1136/bmj.l6890](https://doi.org/10.1136/bmj.l6890)] [Medline: [31948937](https://pubmed.ncbi.nlm.nih.gov/31948937/)]
31. Carot-Sans G, Folguera J, Torres F, et al. Synthetic images for diagnosis and follow-up decision-support algorithms: a systematic review of image generation techniques and validation strategies. PROSPERO (International Prospective Register of Systematic Reviews). 2023. URL: https://www.crd.york.ac.uk/prospERO/display_record.php?RecordID=369125 [Accessed 2025-02-10]
32. STROBE - strengthening the reporting of observational studies in epidemiology. 2025. URL: <https://www.strobe-statement.org/> [Accessed 2026-09-16]
33. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* Apr 16, 2024;385:e078378. [doi: [10.1136/bmj-2023-078378](https://doi.org/10.1136/bmj-2023-078378)] [Medline: [38626948](https://pubmed.ncbi.nlm.nih.gov/38626948/)]
34. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ.* Oct 12, 2016;355:i4919. [doi: [10.1136/bmj.i4919](https://doi.org/10.1136/bmj.i4919)] [Medline: [27733354](https://pubmed.ncbi.nlm.nih.gov/27733354/)]
35. Wang H, Liu Z, Sun K, Wang X, Shen D, Cui Z. 3D MedDiffusion: a 3D medical latent diffusion model for controllable and high-quality medical image generation. *IEEE Trans Med Imaging.* Dec 2025;44(12):4960-4972. [doi: [10.1109/TMI.2025.3585372](https://doi.org/10.1109/TMI.2025.3585372)] [Medline: [40601457](https://pubmed.ncbi.nlm.nih.gov/40601457/)]
36. Xu Y, Sun L, Peng W, et al. MedSyn: text-guided anatomy-aware synthesis of high-fidelity 3-D CT images. *IEEE Trans Med Imaging.* Oct 2024;43(10):3648-3660. [doi: [10.1109/TMI.2024.3415032](https://doi.org/10.1109/TMI.2024.3415032)] [Medline: [38900619](https://pubmed.ncbi.nlm.nih.gov/38900619/)]
37. Krishna A, Wang G, Mueller K. Guided synthesis of annotated lung CT images with pathologies using a multi-conditioned denoising diffusion probabilistic model (mDDPM). *Phys Med Biol.* Mar 6, 2025;70(6):065007. [doi: [10.1088/1361-6560/adb9b3](https://doi.org/10.1088/1361-6560/adb9b3)] [Medline: [39993375](https://pubmed.ncbi.nlm.nih.gov/39993375/)]
38. Huang B, Liu X, Fang L, Liu Q, Li B. Diffusion transformer model with compact prior for low-dose PET reconstruction. *Phys Med Biol.* Feb 7, 2025;70(4):045015. [doi: [10.1088/1361-6560/adac25](https://doi.org/10.1088/1361-6560/adac25)] [Medline: [39832449](https://pubmed.ncbi.nlm.nih.gov/39832449/)]
39. Luo Y, Khan MO, Wen C, et al. FairDiffusion: enhancing equity in latent diffusion models via fair Bayesian perturbation. *Sci Adv.* Apr 4, 2025;11(14):eads4593. [doi: [10.1126/sciadv.ads4593](https://doi.org/10.1126/sciadv.ads4593)] [Medline: [40184460](https://pubmed.ncbi.nlm.nih.gov/40184460/)]
40. Gatoula P, Diamantis DE, Koulaouzidis A, et al. Clinical evaluation of medical image synthesis: a case study in wireless capsule endoscopy. *Sci Rep.* Oct 8, 2025;15(1):35068. [doi: [10.1038/s41598-025-14359-4](https://doi.org/10.1038/s41598-025-14359-4)] [Medline: [41062527](https://pubmed.ncbi.nlm.nih.gov/41062527/)]
41. Zamzmi G, Subbaswamy A, Sizikova E, Margerrison E, Delfino JG, Badano A. Scorecard for synthetic medical data evaluation. *Commun Eng.* Jul 21, 2025;4(1):130. [doi: [10.1038/s44172-025-00450-1](https://doi.org/10.1038/s44172-025-00450-1)] [Medline: [40691520](https://pubmed.ncbi.nlm.nih.gov/40691520/)]

42. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell*. Jul 2024;6(4):e240300. [doi: [10.1148/ryai.240300](https://doi.org/10.1148/ryai.240300)] [Medline: [38809149](https://pubmed.ncbi.nlm.nih.gov/38809149/)]
43. Budding C, Eitel F, Ritter K, Haufe S. MedNeurips_2021. GitHub. URL: https://github.com/cebudding/MedNeurips_2021 [Accessed 2025-07-14]
44. Hayes J, Melis L, Danezis G, De Cristofaro E. Logan: membership inference attacks against generative models. *Proc Priv Enhanc Technol*. Jan 1, 2019;2019(1):133-152. [doi: [10.2478/popets-2019-0008](https://doi.org/10.2478/popets-2019-0008)]
45. Carlini N, Hayes J, Nasr M, et al. Extracting training data from diffusion models. Presented at: SEC '23: Proceedings of the 32nd USENIX Conference on Security Symposium; Aug 9-11, 2023; Anaheim CA, USA. URL: <https://dl.acm.org/doi/abs/10.5555/3620237.3620531> [Accessed 2026-09-16]
46. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. Presented at: FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency; Jan 29-31, 2019; Atlanta GA, USA. [doi: [10.1145/3287560.3287596](https://doi.org/10.1145/3287560.3287596)]
47. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *Commun ACM*. Dec 2021;64(12):86-92. [doi: [10.1145/3458723](https://doi.org/10.1145/3458723)]
48. Alderman JE, Palmer J, Laws E, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. *Lancet Digit Health*. Jan 2025;7(1):e64-e88. [doi: [10.1016/S2589-7500\(24\)00224-3](https://doi.org/10.1016/S2589-7500(24)00224-3)] [Medline: [39701919](https://pubmed.ncbi.nlm.nih.gov/39701919/)]
49. EQUATOR Network. URL: <https://www.equator-network.org/> [Accessed 2025-06-17]
50. 2507-synthetic-images-sysrev. GitHub. URL: <https://github.com/DS3-SisCat/2507-synthetic-images-sysrev/> [Accessed 2026-09-16]

Abbreviations

ADNI: Alzheimer's Disease Neuroimaging Initiative
AUPRC/AUC-PR: area under the precision-recall curve
BigGAN: large-scale generative adversarial network
BraTS: Brain Tumor Segmentation Challenge
CAM: class activation mapping
CNN: convolutional neural network
CT: computed tomography
DCGAN: deep convolutional generative adversarial network
FID: Fréchet inception distance
GAN: generative adversarial network
Grad-CAM: gradient-weighted class activation mapping
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HAM10000: Human Against Machine With 10,000 Training Images
ISIC: International Skin Imaging Collaboration
LIDC-IDRI: Lung Image Database Consortium—Image Database Resource Initiative
LPIPS: learned perceptual image patch similarity
LSGAN: least squares generative adversarial network
MRI: magnetic resonance imaging
MS-SSIM: Multiscale Structural Similarity Index Measure
PET: positron-emission tomography
PGGAN: progressive-growing generative adversarial network
PICO: Population-Intervention-Comparator-Outcome
PPV: positive predictive value
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses—Search Extension
PROSPERO: International Prospective Register of Systematic Reviews
PSNR: peak signal-to-noise ratio
RGB: red, green, and blue
ROBINS-I: Risk of Bias in Non-Randomized Studies of Interventions
SAGAN: Self-Attention Generative Adversarial Network
SSIM: structural similarity index measure
SSIM: Structural Similarity Index Measure
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
StyleGAN: style-based generative adversarial network
SWiM: Synthesis Without Meta-Analysis
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI

UNet: U-shaped convolutional neural network

VAE: variational autoencoder

VAE-GAN: variational autoencoder generative adversarial network

VGG: visual geometry group network

Edited by Khaled El Emam; peer-reviewed by Gaurav Kumar Gupta, Gilbert Lim, Jiacheng Xie, Meghal Gandhi, Uday Kiran Chilakalapalli; submitted 29.Jul.2025; final revised version received 28.Aug.2026; accepted 08.Sep.2026; published 15.Oct.2026

Please cite as:

Folguera-Profítos J, Ramiro-Pareta M, Carot-Sans G, Jiménez-Rueda A, Aznar-Roig J, Domínguez-Herrera D, Nogueras MM, Nadimi E, Torres F, Piera-Jiménez J

Quality Assessment and Model Training Impact of Synthetic Images in Medical Diagnostic Imaging: Systematic Review
JMIR AI 2026;5:e81472

URL: <https://ai.jmir.org/2026/1/e81472>

doi: [10.2196/81472](https://doi.org/10.2196/81472)

© Júlia Folguera-Profítos, Marina Ramiro-Pareta, Gerard Carot-Sans, Alba Jiménez-Rueda, Júlia Aznar-Roig, Didier Domínguez-Herrera, Maria Mercedes Nogueras, Esmail Nadimi, Ferran Torres, Jordi Piera-Jiménez. Originally published in JMIR AI (<https://ai.jmir.org>), 15.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.