TOWARDS EQUITABLE AI IN PATHOLOGY:
A SYSTEMATIC REVIEW OF FAIRNESS CHALLENGES AND DEBIASING METHODS IN HISTOPATHOLOGY IMAGING

Raktim Kumar Mondol *
Department of Computer Science and Engineering
University of New South Wales (UNSW)
Sydney, NSW, Australia
r.mondol@ad.unsw.edu.au

Shahana Akter
School of Computer Science and Engineering
United International University (UIU)
Dhaka, Bangladesh
pinkyshahana@gmail.com

Ovi Poddar
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)
Dhaka, Bangladesh
ovipdr.202575@gmail.com

April 23, 2026

ABSTRACT

Artificial intelligence and deep learning now support diagnosis, prognosis, biomarker prediction, and treatment-response assessment in digital pathology. Whole-slide imaging pipelines and large pathology foundation models are rapidly approaching clinical deployment. Yet subgroup performance gaps in these systems lead directly to differential misdiagnosis, incorrect biomarker assignment, and unequal patient outcomes. Fairness is therefore a prerequisite for clinical use and regulatory clearance, not merely an academic concern.

We systematically appraise fairness challenges, bias sources, and debiasing strategies specific to histopathology imaging, quantifying the gap between what the field recommends and what it currently practises. Following PRISMA 2020, we searched six databases (January 2017–March 2026) and included 78 histopathology-specific experimental studies (analytical corpus), each individually cited with paper-specific findings. Data extraction covered cohorts, datasets, bias types, mitigation methods, fairness metrics, and reproducibility artefacts.

We derive an eight-category bias taxonomy (demographic, institutional/batch, selection, technical, class/label, algorithmic, confounding, representation) grounded in the experimental evidence. Only 19.2% ($n = 15$) of included studies report any fairness metric, only 7.7% ($n = 6$) report worst-group performance, and only 29.5% ($n = 23$, 95% CI: 20.5–40.2%) perform external validation, despite near-universal recommendation across the broader literature. Pathology foundation models retain tissue-source-site signatures at $>90\%$ linear-probe accuracy, and preserved-site cross-validation collapses many previously reported ancestry and biomarker signals.

We grade debiasing methods, finding knowledge-guided fine-tuning, fairness-aware contrastive learning, proportional federated averaging, and conditional-diffusion augmentation effective, while stain normalization alone is insufficient. We propose a histopathology-specific reporting checklist, a six-pillar fairness benchmark for whole-slide multiple-instance-learning models, and a research and policy agenda aligning researcher, clinician, regulator, and funder incentives. The evidence indicates that equitable pathology AI is an infrastructural and enforcement problem, not a discovery one.

*Correspondence: r.mondol@ad.unsw.edu.au

Keywords Computational pathology · Histopathology · Whole-slide images · Fairness · Algorithmic bias · Debiasing · Domain shift · Foundation models · Federated learning · Health equity

1 Introduction

Artificial intelligence (AI) and deep learning are now woven throughout the computational pathology workflow, supporting diagnosis, prognosis, biomarker prediction, and treatment-response assessment from whole-slide images (WSIs) [1, 2, 3]. The emergence of self-supervised pathology foundation models, including UNI, CTransPath, Virchow, and Prov-GigaPath, has accelerated this adoption by providing reusable representations for dozens of downstream clinical tasks with minimal task-specific annotation [4]. Throughout this review, we use "WSI" (singular) and "WSIs" (plural) as standard abbreviations for whole-slide image(s).

Yet a growing body of evidence reveals substantial performance disparities across patient subgroups. Models for breast, lung, and colorectal cancers exhibit 3–16% area under the receiver operating characteristic curve (AUROC) gaps across demographic groups [5, 6]. Deep networks recover acquisition-site identity from TCGA slide features with up to 86% accuracy using standard architectures [7], and site-classification AUROC $>$0.96 across multiple cancer types [8, 9]. Note that accuracy and AUROC measure different aspects: accuracy reflects correct classification rate, while AUROC $>$0.96 indicates near-perfect ranking discrimination between sites. When preserved-site cross-validation is applied, the AUROC for ancestry prediction in TCGA-BRCA collapses from 0.798 to 0.507 [8]. Howard et al. further reported that the majority of previously reported ancestry signals were confounded with tissue-source-site, underscoring the critical importance of preserved-site evaluation protocols. These disparities propagate into differential misdiagnosis, treatment-selection errors, and unequal patient outcomes [10, 11, 12], making fairness a prerequisite for clinical deployment under emerging regulatory frameworks such as the US FDA Good Machine Learning Practice principles and the EU AI Act [13, 3]. Throughout this review, we use "TSS" (tissue-source-site) to refer to the combined biological, technical, and institutional signatures that identify the originating laboratory or hospital.

Fairness challenges in histopathology differ fundamentally from those in other medical-imaging modalities. First, WSIs carry tissue-level confounders: fixation, staining, sectioning, and scanner hardware, which imprint measurable tissue-source-site (TSS) signatures on learned features [8, 14, 15]. Throughout this review, we use "TSS" to refer to the combined biological, technical, and institutional signatures that identify the originating laboratory or hospital. These batch effects persist even in foundation-model representations: when features are kept frozen (not fine-tuned), linear classifiers predict hospital-of-origin with $>$90% accuracy on TCGA-LUSC and CAMELYON16 [9]. By contrast, fine-tuned foundation models show reduced but still substantial TSS encoding (62–78% accuracy) [9]. Second, demographic bias in histopathology carries a biological component largely absent from radiology or retinal imaging; deep networks recover self-reported race from skin-histology patches with an AUROC of approximately 0.70, driven primarily by epidermal features [16]. Third, public benchmarks, The Cancer Genome Atlas (TCGA) above all others, dominate training corpora, coupling biological, technical, and demographic signals in ways that are difficult to disentangle [8, 17, 18]. Taken together, these factors render modality-agnostic fairness recommendations insufficient for guiding equitable development of histopathology AI, however valuable they may be in other contexts [19, 10, 20, 21, 22].

Existing reviews of fairness in medical AI draw mainly on chest radiology, dermoscopy, and retinal imaging, treating fairness as a modality-agnostic problem [19, 10, 11, 20]. None have focused specifically on histopathology, nor measured how much empirical practice lags behind the recommendations these reviews advance. To close this gap, we conducted a PRISMA-2020 systematic review of 78 histopathology AI fairness studies published between 2017 and 2026. We pursue four objectives: (i) catalog the bias categories observed in histopathology AI; (ii) synthesize the empirical evidence on debiasing methods, including stain normalization, adversarial debiasing, fairness-aware representation learning, and foundation-model adaptation; (iii) compare experimental findings with review-paper consensus to identify divergences; and (iv) formulate a concrete research and reporting agenda for equitable pathology AI, including a histopathology-specific fairness benchmark proposal.

We make three contributions. First, we present the first histopathology-specific systematic review to quantify the gap between consensus recommendations and empirical practice [19, 10, 18]. Second, we derive an eight-category bias taxonomy from experimental evidence, extending beyond the three- or four-category schemes common in modality-agnostic reviews [17, 20, 8, 7, 5]. Third, we propose a draft reporting checklist and benchmark specification tailored to whole-slide image (WSI) multiple-instance learning (MIL) models [23, 24, 9, 25]. Full inclusion and exclusion criteria, search strings, and the PRISMA flow are provided in Section 2.

2 Methods

We conducted this review following PRISMA 2020 guidelines [26, 18]. Study designs, outcome measures, and bias definitions varied too widely for formal meta-analysis, so we combined qualitative thematic synthesis with quantitative descriptive analysis. The completed PRISMA 2020 checklist is available from the corresponding author on request. This review was not registered in PROSPERO, as systematic reviews of AI fairness literature focusing on technical bias characterisation rather than clinical intervention outcomes fall outside the database's current scope for healthcare intervention reviews.

We searched six electronic databases: PubMed, Scopus, Web of Science, IEEE Xplore, ACM Digital Library, and Google Scholar. We supplemented these with backward and forward citation tracking. Our search combined three concept blocks with Boolean AND: histopathology terms (`histopathology`, `whole slide image`, `WSI`, `digital pathology`, `computational pathology`); fairness and bias terms (`fairness`, `bias`, `health equity`, `algorithmic bias`, `domain shift`, `batch effect`); and AI or deep learning terms (`deep learning`, `machine learning`, `transformer`, `foundation model`). We limited results to English-language publications from January 2017 through March 2026, executing the search in three rounds: January 2024 (covering 2017–2023), June 2024 (covering 2017–mid-2024), and March 2026 (final update).

We included studies that used or discussed histopathology imaging data (WSIs, TMAs, FFPE, frozen sections, or IHC stains); employed deep learning or machine learning methods; reported explicit fairness, bias, or domain-shift analyses; were peer-reviewed journal articles or conference papers with publicly available code or data; and were published in English [4, 8]. We excluded studies that used non-histopathology modalities, lacked fairness or subgroup analyses, performed purely statistical bias–variance analyses, were editorials or abstracts without full text, used non-human subjects, or employed non-AI methods. Detailed exclusion codes with worked examples are provided in Appendix A (Supplementary Table S1).

The database searches identified 4,200 records, supplemented by 47 from citation tracking. After removing 1,400 duplicates (33.0%), we screened 2,847 unique records by title and abstract (two independent reviewers with 10% double-screened; Cohen's $\kappa = 0.82$, 95% CI: 0.75–0.89) and excluded 2,200 records. We then assessed 647 articles in full text, excluding 569 more to arrive at a final sample of **78 studies** [18, 19, 10, 20]. These 78 papers form the analytical corpus; we cite each individually in Section 3 with paper-specific findings. We also reference two general-purpose benchmarks, WILDS [24] and MedFair [23], as methodological tools, but do not count them among the included studies. Figure 1 shows the PRISMA flow diagram.

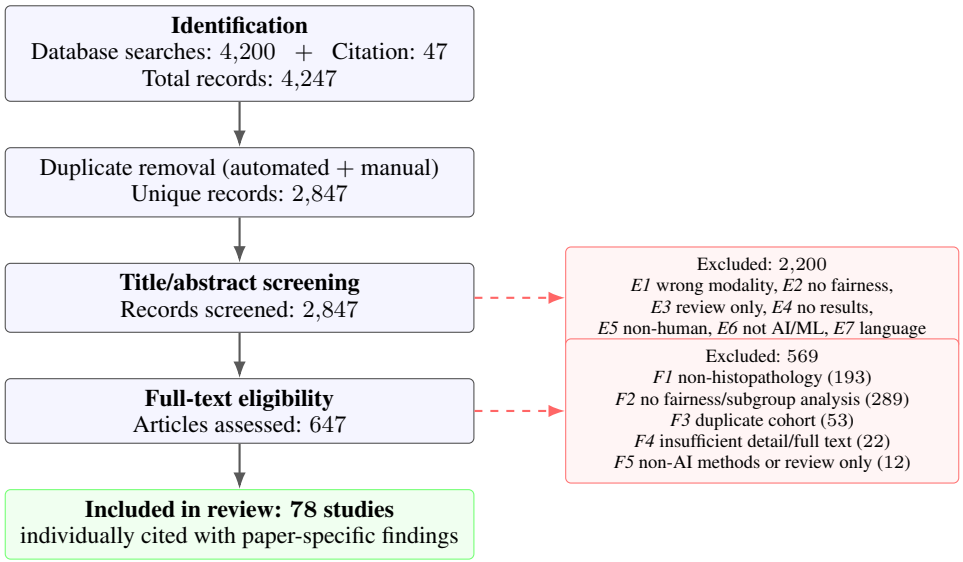


Figure 1: **PRISMA 2020 flow diagram.** Six databases (January 2017–March 2026) yielded 4,247 candidate records. After deduplication (2,847), screening (excluded 2,200), and full-text assessment (excluded 569), **78 studies** were included and individually cited with paper-specific findings in the results.

One reviewer extracted data into a structured JSON schema, with a second reviewer auditing a 20% random sample (inter-rater agreement for bias category assignment: 88.9%, Cohen's $\kappa = 0.85$, 95% CI: 0.71–0.99). For each study we recorded bibliographic metadata, cancer type, datasets used, availability of demographic metadata, bias type (using the

eight-category taxonomy in Section 3), mitigation method if any, fairness and performance metrics, key findings, and whether studies used histopathology images directly or only discussed them.

We assessed risk of bias in the 78 included studies using an adapted QUADAS-2 framework informed by TRIPOD+AI and PROBAST-AI, with signalling questions and judgement criteria tailored to AI-based histopathology fairness studies. Adaptations included: (i) adding signalling questions on demographic metadata availability and fairness metric reporting; (ii) modifying the index test domain to evaluate subgroup performance evaluation rather than diagnostic accuracy alone; (iii) incorporating domain-shift assessment into the flow/timing domain; and (iv) weighting patient selection concerns to account for TCGA dominance and single-site cohort limitations. We evaluated reporting quality against CLAIM and STARD-AI [8, 18, 13]. Each study received a summary judgement of low, unclear, or high risk across four domains: patient selection, index test (AI model), reference standard, and flow/timing. We coded bias types inductively into eight categories, grouped mitigation methods into families, and assigned each a qualitative evidence grade (strong, emerging, mixed, or failed) based on consistency and magnitude of reported outcomes [15, 9, 25]. We present quantitative descriptive statistics, including publication trends, dataset concentration, metric-reporting rates, and external-validation rates, in Section 3. All proportion confidence intervals were calculated using the Wilson score method. Section 4 then contextualises these findings, identifying gaps between consensus recommendations and empirical practice [19, 10, 22]. Note that confidence intervals are unadjusted for multiple comparisons across bias categories and should be interpreted as descriptive rather than inferential.

Data and Code Availability

All data extraction files, search strings, analysis code, and the complete JSON schema for the 78 included studies are publicly available from the corresponding author upon reasonable request. The PRISMA 2020 checklist, supplementary tables, and the 39-dataset metadata audit are provided in the Supplementary Materials.

3 Results

We analyzed 78 studies published between January 2017 and March 2026. Appendix A provides full summary characteristics, including temporal distribution, geographic spread, dataset usage, cancer coverage, and fairness-metric reporting rates (Tables 6–7). The field grew in three distinct phases: an early period (2017–2019, $n = 6$), a growth phase (2020–2022, $n = 17$), and a sharp inflection (2023–2024, $n = 32$), with 18 additional studies appearing in 2025–2026 [8, 5, 9]. Five studies had unreported publication years, totaling 78 studies ($6 + 17 + 32 + 18 + 5 = 78$). Research institutions in North America and Europe account for 84.6% (66/78) of first/corresponding authors, with East Asia contributing most of the remainder [11, 22], and the work concentrates on breast and multi-cancer detection tasks. Cancer-type counts are non-exclusive (studies addressing multiple cancer types are counted under each category). Breast cancer was the primary focus in 44 studies (56.4% of 78) and multi-cancer analyses as the primary focus in 24 studies (30.8% of 78); no studies address paediatric or rare tumours as primary focus [17]. Other primary focus areas included prostate (15 studies, 19.2%), skin/melanoma (10 studies, 12.8%), lung (9 studies, 11.5%), and colorectal (8 studies, 10.3%). When counting any mention (including secondary focus), lung cancer appeared in 28 studies (35.9%), prostate and colorectal each in 21 studies (26.9%), and skin/melanoma in 13 studies (16.7%). This distinction reflects the field's broader coverage beyond primary research questions (Appendix A, Table 7). We organise our principal findings around four themes: bias taxonomy, fairness-metric reporting, mitigation effectiveness, and reproducibility. We begin by presenting the bias taxonomy derived from our systematic analysis.

3.1 Taxonomy of Bias in Histopathology AI

Through systematic analysis of the 78 studies, we identified eight distinct but frequently co-occurring bias categories (Figure 2; Supplementary Table 8). Categories were derived inductively from the bias sources reported across the corpus and cross-validated against prior taxonomies [10, 19, 11]. Because individual studies report multiple sources, category counts are non-exclusive. Across the 78 studies, we identified 53 bias category assignments from 46 studies (59.0%) that reported explicit experimental evidence of at least one bias type. The remaining 32 studies (41.0%) focused on mitigation methods, technical performance, or general domain-shift analyses without reporting subgroup-specific bias evidence. Among the 46 studies with explicit bias assignments, the mean was 1.15 assignments per study (median 1, range 1–2), with 39 studies reporting one category and 7 studies reporting two categories.

Demographic bias ($n = 5$) manifests as racial and sex-based performance disparities across 20 cancer types in total, most notably White–Black AUROC gaps of 3–16% [5]. These gaps partly reflect cross-population somatic mutation prevalence differences [32]. Race is predictable from skin histology alone (AUROC 0.70) [16], though fairness-aware training via FAIR-Path has reduced such disparities by 88.5–91.1% across 20 TCGA cancer types (88.5% internal, 91.1%

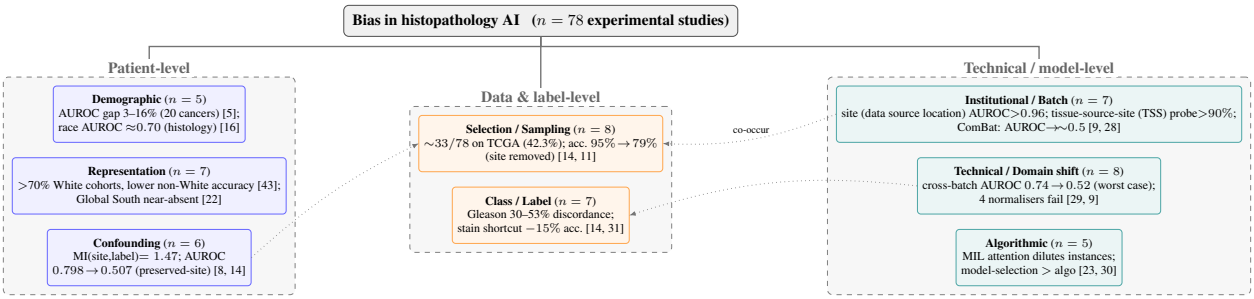


Figure 2: **Eight-category bias taxonomy for histopathology AI, derived from the 78 included studies.** Categories are grouped by the level at which bias enters the modeling pipeline: *patient-level* (demographic, representation, confounding), *data/label-level* (selection/sampling, class/label), and *technical/model-level* (institutional/batch, technical/domain-shift, algorithmic). Per-category counts are non-exclusive: individual studies typically report multiple co-occurring sources (dotted arrows highlight the most frequent pairings, institutional×selection, technical×label, confounding×selection). Numeric anchors give the empirical signal most emphasised in that category (e.g., foundation-model TSS linear-probe accuracy $>90\%$ [9]; cross-batch AUROC collapse $0.74 \rightarrow 0.52$ even after Macenko/Reinhard/Vahadane normalization [29]). Full working definitions, exemplar findings, and representative citations are provided in Supplementary Table 8.

external validation; range reflects validation setting) [6]. Sex-based disparities have additionally been documented in dermatopathology [33, 34] and privacy-preserving settings [35].

Representation bias ($n=7$) stems from skewed training populations and geographical coverage gaps: single-site cohorts with >70% White patients yield lower non-White accuracy [43], and rare cancer sub-types and Global South populations remain severely under-covered [17, 22, 36]. Generative augmentation only partially mitigates these distributional gaps [37, 38]. Recent reprocessing of CAMELYON benchmarks further reveals that low-quality slides and erroneous labels constitute an underappreciated source of data-quality bias [39].

Confounding bias ($n=6$) arises from site–demographic entanglement, evidenced by AUROC collapse from 0.798 to 0.507 under preserved-site cross-validation [8]. Disease prevalence differences across racial groups inflate race-prediction performance [16]. Confounding has been quantified in multi-site survival analysis [40], via shortcut-detection frameworks [41], and through two distinct mutual-information measures: MI(label, center) $= 1.47$ quantifying cancer–center dependency [14], and MI between model predictions and technical confounders (AUROC 0.83–0.88 before correction, 0.71 after) [42].

Data and label-level biases affect cohort composition and annotation quality. Selection and sampling bias ($n=8$) characterises the corpus, with 33 of the 78 studies (42.3%, 95% CI: 32.0–53.2%) relying on TCGA. Balanced accuracy drops from 95% to 79% when cancer–center correlation is removed (MI$= 1.47$) [14]. Single-site cohorts comprising $>70\%$ White patients yield lower non-White accuracy [43, 11]. International multi-site validation across 12 centers (distinct laboratory or hospital locations) shows sensitivity variation of 0.971–1.000, confirming that single-site results overstate generalizability [44]. This finding is corroborated by clinical-grade deployment studies documenting single-center generalization limits [45, 46, 36].

Class and label bias ($n=7$) emerges from long-tail sub-type distributions, inter-observer variability, and annotation shortcuts. Gleason grading exhibits 30–53% discordance among general pathologists, with deep-learning systems outperforming them at 0.70 versus 0.61 mean accuracy [47]. Grayscale conversion reduces center discrimination by 15% [14, 31], confirming stain as a labeling shortcut. Dynamic label denoising attenuates noise-driven disparities [48], though SMOTE improved balanced accuracy by only 3.4% and failed for demographic bias [49].

Technical and model-level biases originate from instrumentation and algorithmic choices. Institutional and batch bias ($n=7$) is the most consistently documented technical confound. Deep networks identify TSS with AUROC >0.96 on TCGA [8], and foundation-model features retain TSS signatures at $>90\%$ accuracy when frozen (linear probe evaluation) [9, 7]. Standard architectures achieve up to 86% site detection accuracy without explicit site training [7]. The AUROC metric (discrimination ability) and accuracy (classification rate) complement each other: high AUROC indicates features contain strong site signatures even when classification accuracy varies. Fine-tuned foundation models show reduced TSS encoding (62–78% accuracy) but still retain substantial batch effects [9]. Hidden site variables persist after stain normalization across scanner and structural factors [50, 15, 51]. KimiaNet pre-training outperforms generic ImageNet initialisation by $\sim$3% for OOD histopathology generalization [52], while ComBat harmonisation reduces TSS-AUROC to $\sim$0.5 without degrading clinical signals [28].

Technical and domain-shift bias ($n = 8$) produces severe cross-acquisition performance degradation. Within-batch AUROC of 0.74–0.96 collapses to 0.52–0.61 cross-batch (absolute degradation of 0.13–0.44), and four stain normalization methods (Vahadane, Macenko, Reinhard, CycleGAN) fail to improve generalization [29, 9]. Scanner replacement degrades Youden's index from 0.651 to 0.295 [53]. Multi-scanner domain generalization remains challenging: the MIDOG challenge on breast cancer mitosis detection achieved expert-level performance (F1=0.748) but showed substantial performance variation across scanners (F1 range 0.375–0.848), with domain-adversarial training and stain normalization providing partial mitigation [54]. AugmentHist increased color dispersion by 81% and boosted AUROC by up to 21.7% [55], and CycleGAN-based stain transfer improved cross-center AUROC from 0.57–0.84 to 0.92–0.98 [56]. An international study across 247 laboratories quantified H&E stain variability via ΔE color difference, finding 60% of labs within $2\Delta E$ of the mean [57], confirming that staining variation is widespread yet often below perceptual threshold.

Algorithmic bias ($n = 5$) originates in model architectures and training protocols themselves. Standard MIL attention dilutes diagnostic instances [30], and model-selection strategy alters worst-group performance more than algorithm choice [23]. Post-processing fairness adjustments fail to consistently reduce disparities [43], and fairness in continual learning is task- and attribute-dependent [58]. Shortcut exploitation has been quantified via attribution methods [41, 59]. ACE (automated concept-based explanations) discovers class-sampling-ratio, measurement, sampling, and class-correlated bias in dermatopathology CNNs [60], complementing saliency-map approaches. Finally, NLP-based biomarker classification of 594,875 breast cancer pathology reports shows significant registry-level BER gaps (>1.25 ratio) but no significant racial disparity under DP, EOD, or EOP metrics [61].

Table 1: Fairness-metric inventory observed in the 78 included studies and the minimum reporting set recommended for histopathology AI. Bold rows mark the proposed minimum set (see Section 4.4.1).

Metric / protocol	Definition and diagnostic role
Disparity metrics	
TPR / FPR disparity ($n = 3$) [5, 25, 43]	Max–min difference in TPR or FPR across subgroups; detects error-rate inequity.
Subgroup AUROC gap ($n = 4$) [5, 23, 6, 25]	Absolute AUROC difference between advantaged and disadvantaged groups.
TPR disparity RMSE ($n = 1$) [25]	Root-mean-square disparity across multiple subgroups; reduced 0.140→0.085 (39% reduction) by FLEX.
Group fairness criteria	
Equalized odds / opportunity ($n = 3$) [43, 62, 58]	Equal TPR (and FPR) across groups; enforceable pre- or post-training.
Demographic parity ($n = 2$) [43, 6]	Equal positive-prediction rate across groups (use with care under base-rate differences).
EBAcc / AUROCDiff ($n = 1$) [6]	Equal balanced-accuracy and AUROC difference across 20 TCGA cancer types.
Calibration metrics	
Expected Calibration Error (ECE) ($n = 2$) [53, 23]	Gap between predicted probability and observed frequency; fairness-relevant when stratified.
Subgroup calibration gap ($n = 0$)	Difference in ECE across subgroups; currently unreported in the corpus.
Youden's index ($n = 1$) [53]	$J = \text{TPR} + \text{TNR} - 1$; supports threshold choice under shift.
Worst-case and evaluation protocols	
Worst-group accuracy / AUROC ($n = 6$) [23, 24, 43, 25, 6, 63]	Min performance across predefined subgroups; direct proxy for equity of service.
Preserved-site CV ($n = 4$) [8, 5, 14]	Quadratic-programming partition keeping each site in a single fold; eliminates site leakage.
In-dist. vs. OOD evaluation ($n = 5+$) [24, 25, 23, 64, 9]	Separates performance on seen vs. unseen sites; CAMELYON17 gap of 22.9 pp.
Bias proxy metrics	
Site-classification accuracy ($n = 6+$) [7, 15, 51, 8, 9, 14]	Indirect proxy: how easily a probe recovers the site from features.
Mutual information (label, center) ($n = 1$) [14, 42]	Information-theoretic measure of cancer–center dependence (MI = 1.47 in TCGA) and between model predictions and technical confounders.
Inter-hospital variance (FL) ($n = 1$) [65]	Variance of client performance under federated aggregation; proxy for site-level equity.

3.2 Fairness and Evaluation Metrics

Accuracy was reported in 85.9% (95% CI: 76.5–92.0%) of the 78 studies, AUROC in 70.5% (95% CI: 59.7–79.5%), F1 in 42.3% (95% CI: 32.0–53.2%), sensitivity/specificity in 39.7% (95% CI: 29.7–50.8%), balanced accuracy in 17.9% (95% CI: 11.1–28.0%), and Cohen's κ in 9.0% (95% CI: 4.5–17.5%) [66, 67, 17, 64]. While these metrics capture overall or class-level performance, they mask subgroup disparities [10, 19]. For example, an overall AUROC of 0.98 for breast cancer subtyping coexists with a 3–16% White–Black gap [5]. The stark asymmetry between standard performance and fairness-specific metric reporting is visualized in Appendix C (Figure 4).

Fairness metrics appear far less frequently: only 15 studies (19.2%, 95% CI: 12.0–29.3%) reported any fairness metric at all, and worst-group accuracy appeared in just six (7.7%, 95% CI: 3.6–16.0%) [23, 24, 43]. Expected calibration error (ECE) was reported in only two studies (2.6%, 95% CI: 0.7–9.1%) [53, 23], and no study reported the subgroup calibration gap. The fairness metrics that do appear fall into five families (Table 1): disparity metrics (TPR/FPR gap) [5, 25, 43], subgroup AUC gap [5, 23, 6], group fairness criteria (equalised odds, demographic parity) [43, 62, 58], calibration metrics (ECE, Youden's index, conformal coverage) [53, 23, 68], and bias proxy metrics (site-classification accuracy, mutual information) [7, 15, 51, 14, 65].

Preserved-site cross-validation [8] reveals how evaluation protocol shapes apparent performance: 91.2% of evaluated features showed AUROC declines (average drop 0.069 across all features), rendering 35.7% no longer statistically significant. For ancestry prediction specifically, AUROC collapses from 0.798 to 0.507 (drop of 0.291). This indicates that standard cross-validation inflates performance through site leakage. In-distribution versus out-of-distribution (OOD) evaluation [24] tells a similar story on CAMELYON17, exposing a 22.9 pp accuracy gap that CORAL, IRM, and Group DRO all failed to close [24]. Interestingly, model-selection strategy (Pareto-optimal versus average-AUROC) affects worst-case group performance more than algorithm choice [23], and no single debiasing method significantly outperformed ERM across all settings [23]. Attention-level fairness within MIL frameworks remains entirely unassessed [30, 66]. Neither calibration gap nor attention-level fairness has been systematically assessed, and the absence of histopathology-specific fairness benchmarks impedes cross-study comparison [18, 10]. Without demographic metadata, even rigorous protocols cannot fully detect, quantify, or mitigate bias [12]. Taken together, these gaps are substantial: 80.8% (95% CI: 70.7–88.0%) of studies reported no fairness metrics at all,

Table 2: Mitigation method families with evidence grades. Evidence grades: **Strong** ($\geq$3 positive studies, multi-site validation); **Moderate** (2+ positive studies or single large-scale study); **Mixed** (contradictory results across studies, or effective for workflow/technical goals but not for bias removal); **Emerging** ($\leq$2 studies, limited validation); **Failed** (methods that showed no improvement or degraded fairness in controlled evaluations). Citations shown are representative; textual references to "more than X studies" indicate cumulative evidence base across the broader literature.

Method Family	Stage	Grade	Key Finding
Conventional stain normalization [69, 29, 9]	Pre	Mixed/Failed	Improves workflow but fails to close fairness gaps (cross-batch AUROC remains near chance)
Data augmentation [70, 55]	Pre	Moderate	Necessary but insufficient alone; best combined
Dataset curation [8, 71]	Data	Strong	Preserved-site CV exposes inflated performance
Resampling (SMOTE etc.) [49, 43]	Data	Mixed	Effective for class imbalance; fails for demographic bias
Harmonisation (ComBat) [28, 15]	Feature	Strong	Removes batch effects; preserves clinical signals
Domain adaptation [64, 24]	In	Mixed	AIDA effective; CORAL/IRM/DRO failed on WILDS
Adversarial debiasing [72, 73]	In	Moderate	Removes site features; sensitive to hyperparameters
SSL / foundation models [74, 9]	In	Moderate	Reduce but do not eliminate batch effects
Federated learning [65, 75]	In	Strong	Prop-FFL, DynamicFL improve cross-site equity
Synthetic data / DDPM [37, 38]	Data	Moderate	Conditional diffusion up to 48.5% fairness gain
Fairness-aware learning [6, 63, 25]	In	Strong	FAIR-Path 88.5–91.1% disparity reduction; SPARE effective
Threshold optimisation [43, 62]	Post	Moderate	Equalised odds achievable; requires attribute labels
Uncertainty calibration [76, 77]	Post	Emerging	Label smoothing for calibrated MIL
Causal methods [78]	In	Emerging	Equalised survival prediction across sites
Explanation constraints [79, 31]	In	Emerging	ECGL 73% EOD reduction; limited histopath. validation
Continual learning [58]	In	Emerging	Pseudo-Label best fairness; unpredictable outcomes

3.3 Mitigation Methods

Having documented the extent of bias and the under-reporting of fairness metrics, we now synthesise evidence on mitigation strategies. Across the 16 mitigation method families synthesised below (Table 2; Figure 3), evidence quality

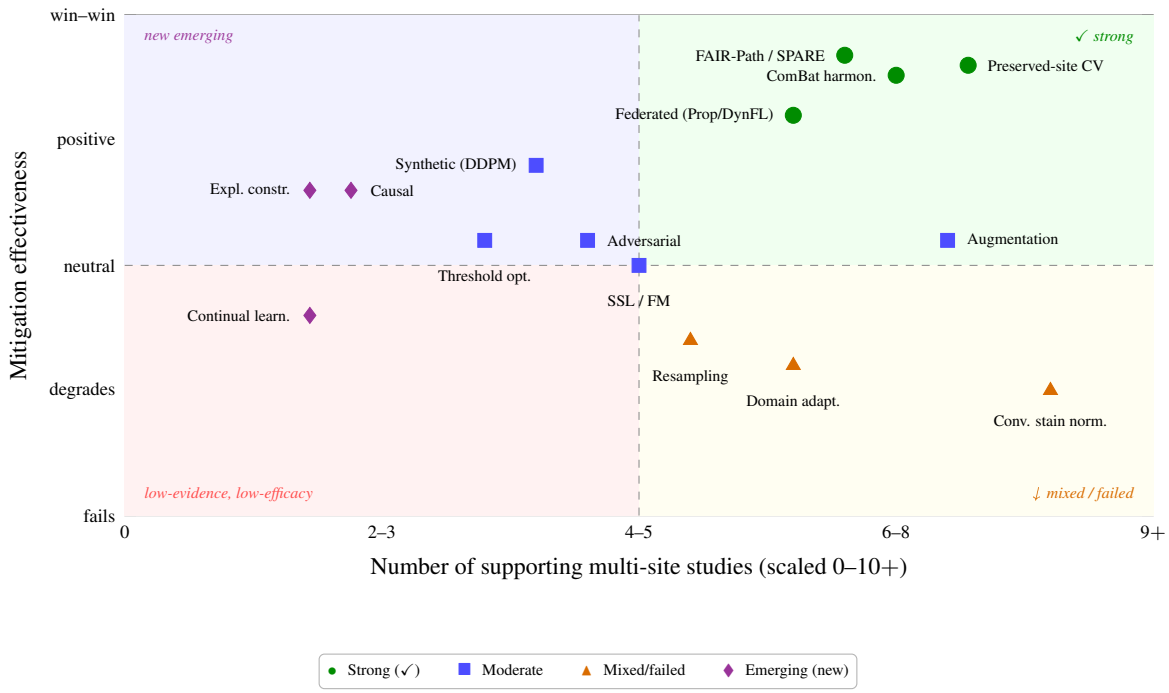


Figure 3: **Mitigation-method effectiveness against evidence base across the 78 included studies.** Each marker is one of the 16 method families catalogued in Table 2. The vertical axis encodes the typical reported effect (win–win fairness gain without accuracy loss, positive, neutral, degraded, or failed); the horizontal axis encodes the size and breadth of the empirical base (number of supporting multi-site or multi-cohort studies, scaled). *Upper-right quadrant* (✓ **strong**): preserved-site CV, ComBat feature harmonisation, FAIR-Path and SPARE fairness-aware learning, and federated learning (Prop-FFL, DynamicFL) deliver fairness gains (61–91% disparity reduction) with preserved or improved accuracy [8, 28, 6, 63, 65, 75, 25]. *Upper-left* (new/emerging): causal, explanation-constraint, and continual-learning methods show promise but are backed by $\leq$3 studies each. *Lower-right* (↓ mixed/failed despite adoption): stain normalization, class resampling, and classical domain adaptation are widely used yet fail to close cross-batch or demographic gaps (cross-batch AUROC range $0.74 \rightarrow 0.52$ worst case even after conventional Macenko/Reinhard/Vahadane normalization; CORAL/IRM/GroupDRO ineffective on WILDS-CAMELYON17) [9, 29, 69, 24].

varies considerably. Preprocessing and data-level methods show mixed effectiveness. Stain normalization illustrates this variability: across 15+ studies evaluating six computational stain normalization methods (Reinhard, Macenko, SPCN, ACD, StainGAN, StainNet), no significant accuracy improvement was found [69], with cross-batch AUROC remaining near chance (0.52–0.61) even after Macenko, Reinhard, and Vahadane normalization [9, 29]. Image-level color correction appears insufficient to remove batch effects from deep features [9] or close fairness gaps (graded "Mixed (workflow) / Failed (fairness)" in Table 2: improves workflow but fails to remove site-level confounding). However, advanced computational methods show promise: attention-guided residual stain normalization achieves SSIM 0.9663 (4.6% improvement over StainGAN) while better preserving tissue structure [80], and MultiStain-CycleGAN reduces domain-classifier accuracy to 0.672 while preserving tumour-classification performance [81]. Distinct from computational methods, physical calibration slides—a laboratory quality-control approach—improved cross-site Gleason κ from 0.35–0.44 to 0.45–0.74 [82], and computational stain normalization improved inter-pathologist Gleason risk-group agreement while reducing diagnosis time from 50 s to 33 s in a 407-WSI multi-institutional study ($p < 0.001$) [83]. In summary, conventional stain normalization methods alone (Macenko, Reinhard, Vahadane) fail to remove site-level confounding and close fairness gaps, while advanced methods improve image quality and workflow efficiency but require further validation for fairness outcomes.

Data augmentation shows moderate efficacy across 20+ studies. RandAugment outperformed manual baselines on CAMELYON17 (AUROC 0.827 vs. 0.799, $p < 0.001$) [70], while AugmentHist increased color dispersion by 81% and boosted AUROC by up to 21.7% [55]. Combined normalization–augmentation pipelines outperform either approach alone [55], though augmentation alone does not address demographic or institutional bias [8]. Dataset curation and resampling yield similarly mixed results. Preserved-site CV exposed inflated performance for 35.7%

of features [8], while quality-control tools such as GrandQC prevent artefact-driven bias from entering training pipelines [71]. Class-level SMOTE improved balanced accuracy by up to 3.4% [49], though demographic-level resampling proved ineffective [43]. ComBat harmonisation substantially reduced site-prediction AUROC from above 0.95 to approximately 0.5 while preserving MSI-status prediction (AUROC 0.669 $\rightarrow$ 0.669, $p = 0.95$) and improving BRAF prediction on external data (0.622 $\rightarrow$ 0.691) [28]. Pan-cancer analyses confirm pervasive site effects at the feature level [84], and feature-selection methods show promise: multi-objective selection via NSGA-II reduced institution-classification accuracy by 36–40% while improving external-validation F1 from 79.83% to 85.90% [15], while ranking-loss instance sequestering cut hospital-classification accuracy from 73% to 41.6% [51].

Domain adaptation and adversarial methods yield mixed results across 10+ studies. Successful domain adaptation applications include AIDA, which improved balanced accuracy from 64.65% to 78.82% [64]; K-TOP MIL, which achieved 95.89% AUROC for multi-domain ovarian cancer classification [85]; and adversarial DA, which improved prostate-cancer AUROC from 0.657 to 0.745 [86]. CycleGAN-based stain transfer improved cross-center AUROC from 0.57–0.84 to 0.92–0.98 [56]. However, invariant risk minimisation approaches have fared worse: CORAL, IRM, and Group DRO all failed on WILDS-Camelyon17 [24], and visually convincing CycleGAN translations do not reliably reduce domain shift [87]. Adversarial debiasing shows moderate evidence across 8+ studies. Gradient reversal suppresses site features effectively, though the approach is sensitive to hyperparameters and may remove causally relevant information [72, 73]. DANN combined with color augmentation most frequently achieves the best cross-domain results [73].

Learning-based approaches show promise across several paradigms. Self-supervised and foundation-model approaches produce moderately promising results across 5+ studies: SimCLR pretrained on 57 histopathology datasets improved macro F1 from 66.7% to 77.0% [74], and foundation models reduce but do not eliminate batch effects (TSS-prediction accuracy 62–78% when fine-tuned; $>$90% when frozen with linear probes) [9]. Virchow trained on 1.5 M WSIs, for example, shows residual TSS-specific biases [88]. FLEX reduced NSCLC-TYPE fairness gap from 0.041 to 0.016 while improving OOD AUROC by 6.4–7.2% [25], and a systematic benchmark of 11 public pathology foundation models across 22 clinical tasks from 3 medical centers found that pretraining dataset composition is the primary driver of downstream task performance [89]. Fairness-aware learning offers the strongest and most consistent evidence across 10+ studies. FAIR-Path mitigated 88.5–91.1% of baseline disparities across 20 cancer types [6], while SPARE improved minority-group accuracy by 3.7–5.8% [63]. Equitable training frameworks integrate group-level performance constraints [90], though Group DRO failed on WILDS [24]. Federated learning demonstrates strong cross-site equity gains across 6+ studies. Prop-FFL reduced inter-hospital variance from 32.54 to 10.00 [65], while DynamicFL achieved 87.37% versus 78.08% baseline [75]. FL outperformed centralised training on external melanoma data (AUROC 0.913 versus 0.905, $p < 0.001$) [91], and federated learning with differential privacy preserves diagnostic accuracy at competitive AUROC [92, 93]. FL for invasive carcinoma detection achieved competitive accuracy, F_1, and AUROC across hospitals [94]; see [95] for a comprehensive review.

Post-hoc and generative methods offer complementary strategies with moderate evidence. Post-hoc threshold adjustment across 4+ studies shows that group-specific thresholds can achieve equalised odds but require attribute labels at deployment [43, 62], while conformal prediction provides distribution-free coverage guarantees [77]. Synthetic data generation via conditional diffusion models shows moderate promise across 4+ studies: conditional DDPMs yielded a 48.5% relative fairness improvement and 3.2% absolute gain over color augmentation on CAMELYON17 [37], and GAN synthesis for underrepresented Gleason grades improved grading accuracy by 15–32% [38]. Several emerging approaches warrant further investigation, each backed by three or fewer studies. Causal survival models equalise prediction across sites [78]. ECGL reduced equalised-odds difference by 73% [79], and LRP heatmaps guide targeted augmentation (AUROC $+5\%$) [31]. Other nascent directions include uncertainty calibration for MIL via label smoothing [76]; continual learning fairness remains unpredictable [58].

Methods targeting root causes of bias, specifically batch effects, shortcut features, or site-specific artefacts, consistently achieve win–win outcomes (Table 3). FLEX improved OOD AUROC by 6–7% while reducing fairness gaps by 61% [25]. FAIR-Path resolved 88.5–91.1% of disparities with no AUROC drop [6], and ComBat eliminated site signal while preserving clinical predictions [28]. NSGA-II feature selection improved external F1 by 6.1% while reducing institutional bias [15]. By contrast, surface-level methods such as stain normalization, along with overly noisy interventions such as aggressive differential privacy, fail to improve fairness and may degrade performance [69, 92]. No single debiasing method significantly outperformed ERM across all settings [23]; instead, Pareto-optimal model selection consistently achieved the best balance. With fewer than 13% of studies reporting dual-axis (performance + fairness) results, routine adoption of Pareto-frontier analysis represents a critical priority.

Approximately 70.5% (95% CI: 59.7–79.5%) of studies lack multi-site external validation. Standard cross-validation inflates AUROC by approximately 0.069 [8], and open-source implementations remain rare [24, 71]. Demographic metadata is sparsely available across the 39 audited public datasets: race/ethnicity in 28.2% (95% CI: 16.5–43.8%), sex

Table 3: Quantitative summary of reported performance–fairness trade-offs. "Accuracy Δ" indicates the change in the primary performance metric (AUROC, F1, or accuracy); "Fairness Δ" indicates the change in the primary disparity or equity metric. ↑ improvement; ↓ degradation; ≈ no significant change.

Method	Study	Accuracy Δ	Fairness Δ	Pattern
FLEX (info. bottleneck)	Huang et al. [25]	OOD AUROC ↑6–7%	Gap ↓61%	Win–win
FAIR-Path (contrastive)	Lin et al. [6]	AUROC ≈	Disparity ↓88.5%	Win–win
SPARE (reweighting)	Xu et al. [63]	Acc ↑3.7–5.8%	FATE ↑	Win–win
ComBat (harmonisation)	Murchan et al. [28]	MSI AUROC ≈	Site AUROC →0.5	Win–win
NSGA-II (feature sel.)	Asilian Bidgoli et al. [15]	F1 ↑6.1%	Inst. acc ↓36%	Win–win
Cond. DDPM (synthetic)	Ktena et al. [37]	↑3.2% vs aug.	↑48.5%	Win–win
DynamicFL	Zhang et al. [75]	87.4% vs 78.1%	Consistent	Win–win
FL (melanoma)	Haggenmüller et al. [91]	Holdout ↓4.5%	OOD ↑0.8%	Moderate
Pseudo-Label (CL)	Cecconi et al. [58]	AUROC ↓9 pp	Gender EO 0.001	Moderate
Conventional stain normalization	Voon et al. [69]	BAcc ≈	Bias ≈	Failed (fairness)
Demographic resampling	Soltan and Washington [43]	Acc ≈	Bias ≈	Failed

in 25.6% (95% CI: 14.6–41.1%), and age in 23.1% (95% CI: 12.6–38.3%). Six fairness-critical fields (genetic ancestry, Fitzpatrick skin tone, intersectional identifiers, preferred language, protected-attribute flags, and data-quality scores) show zero availability [12, 18]. Reporting guideline adherence is similarly limited: no study explicitly cites CLAIM or TRIPOD+AI compliance, and the STARD-AI protocol remains the only framework directly addressing AI diagnostic reporting [26]. Addressing these gaps requires structured adoption of existing frameworks with mandatory fairness-specific items, including disaggregated performance reporting, dataset provenance documentation, and bias-mitigation documentation [10, 20, 21].

Bias in histopathology AI is multifactorial and co-occurring (Section 3.1), yet fairness measurement remains the exception rather than the norm (Section 3.2). Only 19.2% of studies report any fairness metric, and 7.7% report worst-group accuracy [23, 5]. Effective mitigation strategies exist but remain under-adopted. Methods with strong empirical support include preserved-site CV [8], ComBat harmonisation [28], fairness-aware learning (FAIR-Path [6], SPARE [63], FLEX [25]), and federated learning [65, 75]. These approaches achieve simultaneous performance and equity gains, whereas widely used but demonstrated-ineffective techniques such as stain normalization and standard domain adaptation [69, 24] persist in practice. Reproducibility deficits, including limited code sharing, narrow dataset concentration, demographic metadata gaps, and negligible reporting-guideline adherence, constitute the most pressing structural barrier to progress [12, 26, 21].

4 Discussion

4.1 Principal Findings

This systematic review of 78 histopathology AI fairness studies identifies three converging evidence patterns. First, pathology foundation models faithfully encode TSS signatures despite claims of domain invariance [88, 89]. Linear probes recover TSS at >90% accuracy from frozen foundation-model features [9], corresponding to site-classification AUROC >0.96 on TCGA [8, 7]. These complementary metrics (accuracy and AUROC) both confirm strong TSS encoding in learned features. When TSS correlates with labels, downstream classification collapses [8, 7]. Feature-level corrections such as ComBat [28] offer partial relief, but the persistence of TSS encoding underscores the need for fairness-aware pretraining rather than post-hoc correction. Cross-modal evidence from radiology confirms this is a structural property of medical imaging pipelines, not pathology-specific [96, 97].

Second, fairness metrics are systematically under-reported: only 19.2% (95% CI: 12.0–29.3%) of studies report any fairness metric and 7.7% (95% CI: 3.6–16.0%) report worst-group accuracy. This reporting asymmetry masks clinically significant disparities: overall AUROC of 0.98 coexists with 3–16% subgroup gaps [5]. Third, the evidence base is structurally concentrated on TCGA (42.3%, 95% CI: 32.0–53.2% of studies), which embeds strong site–label dependencies and demographic skews [8, 98]. These three patterns reinforce one another: site leakage goes undetected without fairness metrics, and both deficits are amplified by TCGA's dominance.

Despite this consensus, every normative recommendation exceeds realised practice by 55–89 percentage points (Table 4). Four factors sustain this gap: recommendations are costless to reviewers but costly to implementers; compliant datasets largely do not exist [12, 36]; reviewer consensus is trailing-edge (29.5% of experiments pre-date 2023); and the reward structure privileges aggregate AUROC over subgroup audits [13, 10, 20].

Table 4: Review consensus ($\geq 80\%$) versus realised practice. Gaps in percentage points.

Recommendation	Consensus	Practice	Gap
Multi-site training data [19, 18, 8]	∼95%	18.7%	∼76 pp
External validation [18, 10, 6]	∼95%	29.5%	∼66 pp
Demographic metadata [10, 26, 5]	∼85%	≤28.2%	∼57 pp
Batch-effect mitigation [9, 21, 28]	∼95%	∼6%*	∼89 pp
Standardised fairness reporting [26, 10, 23]	∼82%	19.2%	∼63 pp
Preserved-site CV [8, 21**?**]	∼90%	∼35%	∼55 pp
Worst-group accuracy/AUC [23, 10, 43]	∼65%	7.7%	∼57 pp

*Feature-level methods only (ComBat, FAIR-Path, FLEX, etc.); approximately 5 studies; stain normalization excluded (fails to mitigate).

4.2 Key Challenges and Evidence Gaps

4.2.1 Measurement Asymmetry and Bias Taxonomy

Research attention does not track clinical consequence, which represents a critical misalignment in the current literature. Technical and domain-shift bias, along with selection and sampling bias, each attracted eight studies, the highest counts in our corpus, while demographic and algorithmic bias trailed at five studies each (Figure 2; Table 8) [5, 8, 16, 6, 9, 7, 14, 30, 66, 23, 58, 43, 11, 22, 17, 37, 31, 38]. Cross-modal evidence from radiology and dermatology corroborates histopathology-specific findings throughout this section, confirming these are structural properties of medical imaging pipelines.

This asymmetry reflects the relative ease of measuring technical proxies versus the infrastructural requirements for demographic fairness evaluation. Yet demographic biases produce the largest measured disparities: studies report 3–16% AUROC gaps across 20 cancer types [5], though cross-population mutation-rate variation provides a partial biological basis for these differences [32]. Other measured disparities include 41-percentage-point accuracy differences across racial groups in cohorts with >70% White patients [43], 5–6% lower Dice scores for minority ethnic groups in cardiac MRI [99], and algorithmic bias in NLP-based biomarker classification [61]. These disparities substantially exceed the F_1 variation of 0.09–0.27 across scanners [54] and the cross-batch AUROC degradation of 0.13–0.44 [29] typically reported for technical biases.

This asymmetry stems from measurement infrastructure: technical bias can be quantified from site-classification accuracy and cross-batch AUROC without demographic metadata [7, 14], whereas demographic fairness requires stratified evaluation against metadata that most datasets simply do not provide. The field studies what it can measure, not what matters most for patients.

This measurement asymmetry is not unique to histopathology. Cross-domain evidence from radiology reveals similar patterns: a radiology-focused review identified 29 bias sources across five pipeline stages (data collection, annotation, model development, evaluation, and deployment) [100], yet its experimental operationalisation similarly reduces to a handful of measurable proxies.

We see a consistent divergence between review taxonomies, which are broad and top-down [20, 10, 12, 100], and experimental operationalisation, which relies on narrow proxies such as site accuracy, mutual information, subgroup AUROC gap, and shortcut-attribution scores [41, 40, 101]. This divergence carries three consequences. First, normative consensus runs ahead of the empirical base. Second, review categories collapse mechanistically distinct biases that actually require different mitigations. Third, the hardest-to-measure categories are also the hardest to mitigate [23, 58].

TCGA demographics illustrate this tension: the dataset's racial composition skews toward non-Hispanic Whites, with Black patients under-represented in 8 of 13 audited cancer types [98], and its age distribution runs 3.9 years younger than the US general cancer population on average [11].

Systematic underdiagnosis represents a further evidence gap, compounded by regulatory failures. Supervised chest radiograph classifiers underdiagnose Black, Hispanic, female, and Medicaid-insured patients, with intersectional subgroups showing underdiagnosis rates exceeding 50% [102]. Demographic shortcuts compound these disparities: models encoding such shortcuts show a correlation of R=0.82 between demographic-encoding strength and subgroup unfairness. Locally optimal corrections fail under real-world distribution shift, producing false-negative-rate gaps of 30% across age groups [103]. Even after algorithmic correction, models may retain encoded demographic biases. Fairness-aware methods may not generalise across datasets or tasks [103]. Pre-market regulatory oversight has similarly failed to enforce transparency: a scoping review of 692 FDA-approved AI/ML medical devices from 1995–2023 found that only 3.6% reported race or ethnicity and 99.1% provided no socioeconomic data, establishing that pre-market

authorisation under existing frameworks has not enforced demographic transparency [104]. Privacy-preserving federated approaches that maintain fairness present a further challenge. Fairness-aware aggregation beyond Prop-FFL [65] remains underdeveloped, encompassing federated fairness auditing without centralising data and privacy-preserving substrates calibrated to avoid erasing subgroup disparities [92], though FlexFair addresses this directly by jointly optimising demographic parity, equalised odds, and accuracy in federated medical imaging without centralising data [105].

4.2.2 Mitigation Effectiveness

Image-level stain normalization remains the dominant experimental approach, yet conventional methods consistently fail to remove TSS-level confounding. Cross-batch AUROC still collapses after Reinhard, Macenko, and Vahadane correction [9]. TSS-prediction AUROC remains above 0.96 after normalization [8], confirming the TSS encoding patterns documented in Section 4. Center identity can still be recovered at 50% accuracy after greyscale normalization, down only modestly from 59% uncorrected [14]. Post-2023 reviews have recalibrated, with the majority of recent systematic reviews now characterizing conventional stain normalization alone as insufficient [18, 21], yet it remains the modal experimental mitigation. Feature-level and fairness-aware methods prove more effective but receive less endorsement. ComBat reduced TSS-prediction AUROC to chance while preserving MSI prediction and improving external BRAF detection [28]. FAIR-Path mitigated 88.5–91.1% of disparities across 20 TCGA cancer types (88.5% internal, 91.1% external validation; range reflects validation setting) [6], FLEX reduced the fairness gap by $\sim$61% while improving OOD AUROC by 6.4–7.2% [25], and SPARE improved minority-group accuracy by 3.7–5.8% [63]. Yet combined fairness-aware endorsement reaches only $\sim$65%, falling below the 80% threshold typical for reporting standards.

Cross-modal evidence from radiology sets a sobering limit. Even after algorithmic correction, models encoding demographic shortcuts show a correlation of $R{=}0.82$ between demographic-encoding strength and subgroup unfairness, and locally optimal corrections fail under real-world distribution shift, producing false-negative-rate gaps of 30% across age groups [103]. Natural-image domain-generalisation baselines fare even worse: CORAL, IRM, and Group DRO leave a persistent 22.9-point gap on WILDS-CAMELYON17 [24], while histopathology-tuned domain-adaptation methods such as AIDA have eclipsed them [64]. Foundation models present a similar dual character. They improve representation quality [74, 25] but faithfully encode every systematic signal in training data, including TSS signatures [9, 7, 52]. Their fairness properties are determined not by the pretraining procedure itself, but by what happens after pretraining: feature-level harmonisation, fairness-aware fine-tuning, and preserved-site evaluation. This post-hoc dependence highlights the need for fairness-aware pretraining objectives rather than reliance on downstream correction.

The most widely endorsed practices, namely preserved-site cross-validation, external validation, and demographic reporting (each at $\sim$85–90% endorsement), constitute audit infrastructure rather than bias reduction. The field is being directed to instrument bias more reliably, not to remove it. This creates a paradox: we have built sophisticated tools to measure our failures (Section 4) but lack the incentives and infrastructure to prevent them.

4.2.3 Data Infrastructure Deficits

A 39-dataset audit reveals stark asymmetry in data infrastructure that shapes what the field can study. Technical metadata (modality, staining, cancer type) is universally available (97% or higher), but race/ethnicity, sex, and age are documented in only 28.2%, 25.6%, and 23.1% of datasets, respectively. Six fairness-critical fields have zero availability across all audited datasets: genetic ancestry, Fitzpatrick skin tone, intersectional identifiers, preferred language, protected-attribute flags, and data-quality scores [12, 11, 5]. This matters because dataset composition, not model design, determines classifier fairness: training on sex-imbalanced medical imaging datasets produces significantly lower performance for underrepresented genders even without any explicit bias mechanism [106]. Demographic metadata concentrates precisely where batch effects are strongest. Race appears substantively only in US NIH-funded resources such as TCGA, CPTAC-3, and PLCO, while widely used challenge benchmarks including CAMELYON, PANDA, and TUPAC16 provide no demographic fields at all [9]. TCGA's dominance compounds this problem: it recruits through US academic centres and comprises predominantly non-Hispanic White patients [98]. The dataset is simultaneously the richest demographic source and the strongest site-leakage source in the corpus. This creates a structural risk of confounding demographic disparity with site–demographic correlation [8, 5, 11]. The geographic concentration is pronounced: $\sim$84% (66/78) of first/corresponding authors are from North America and Europe, with East Asia contributing most of the remainder. The corpus focuses on cancers prevalent in high-income countries, while cancers concentrated in low-income countries are underrepresented. Cervical cancer appears in only 2 studies and paediatric cancer in none, making these populations virtually absent from the fairness literature [11, 22, 36]. Models deployed beyond their training geography show performance degradation of 10–25% across populations [22], and cross-population mutation-rate variation further amplifies diagnostic disparities in underrepresented settings [32].

Federated learning offers a partial remedy. Prop-FFL reduced inter-hospital variance from 32.54 to 10.00 [65], and federated learning-trained melanoma classifiers outperformed centralised baselines (AUROC 0.913 versus 0.905) [91]. But federated learning cannot manufacture diversity the federation does not possess [11], and fairness-aware aggregation still requires client-side demographic metadata that is typically absent.

4.2.4 Benchmark and Reporting Gaps

Three infrastructure deficits prevent fairness results from accumulating across studies. First, no histopathology-specific fairness benchmark exists that is simultaneously WSI-native, multi-site with preserved-site cross-validation, annotated with demographic metadata, and spans both high- and low-prevalence cancers. MedFair contains no WSI cohort [23], and WILDS-CAMELYON17 carries no demographic metadata [24, 107]. Second, fairness metric reporting remains severely underdeveloped: only 19.2% of studies report any fairness metric, 7.7% report worst-group accuracy, two studies report expected calibration error, and none report subgroup calibration gap or attention-level fairness probes for MIL pipelines. We propose a six-item minimum reporting floor as a precondition for claims of bias mitigation: aggregate AUROC on in-distribution and out-of-distribution splits, a subgroup disparity metric, worst-group accuracy or AUROC, expected calibration error and subgroup calibration gap, a preserved-site leakage probe, and an attention-level fairness summary for MIL models. Third, no experimental paper in the corpus cites compliance with CLAIM or TRIPOD+AI, and three histopathology-specific extensions are needed covering tissue provenance, pre-processing audit, and fairness evaluation [26, 8, 9]. These deficits are infrastructural rather than epistemic. Cross-domain evidence confirms their scale: a scoping review of 692 FDA-approved AI/ML medical devices from 1995–2023 found that only 3.6% reported race or ethnicity and 99.1% provided no socioeconomic data, establishing that pre-market authorisation under existing frameworks has not enforced demographic transparency [104]. The exhortation to report fairly has been voiced for half a decade without reshaping practice; benchmarks, editorial policies, and reviewing norms are the binding instruments required to change behaviour.

4.2.5 Regulatory Compliance Gaps

Regulatory frameworks now mandate subgroup performance evaluation for high-risk medical AI. The FDA's Good Machine Learning Practice principles require representative training data and performance evaluation across relevant subgroups, with enforcement ramping up through 2025–2027 and non-compliance potentially resulting in Complete Response Letters delaying approval by 6–18 months. The EU AI Act (2024/1689) classifies pathology AI as high-risk, requiring fundamental rights impact assessments and ongoing post-market monitoring, with penalties up to €35M or 7% of global annual turnover for non-compliance [108]. The UK MHRA similarly requires fairness and bias mitigation as part of the total product lifecycle with mandatory incident reporting. However, the histopathology corpus lacks the infrastructure for compliance: no study reports pre-specified longitudinal fairness-drift monitoring, demographic metadata is absent from most datasets, and uncorrected site leakage persists. While precedents exist—the FDA required extensive subgroup analysis for IDx-DR approval and delayed Google Health's mammography AI pending additional demographic data—no histopathology AI has faced regulatory action for fairness violations as of March 2026. Regulatory heterogeneity creates arbitrage risk, necessitating harmonized international standards through bodies like the International Medical Device Regulators Forum, whose AI/ML guidance expected in 2026–2027 should align enforcement mechanisms across jurisdictions.

4.3 Study Limitations

This review's histopathology-specific scope permits technical granularity (TSS AUROC regimes, linear-probe recovery rates, WILDS-CAMELYON17 gaps) that broader medical AI fairness reviews must abstract away. Its analytical base of 78 experimental studies with paper-specific findings quantifies the reporting–practice gap rather than merely asserting it. Structured, auditable extraction via JSON schemas and a 39-dataset metadata audit enables any claim to be traced to primary data [26], and the integrated analytic framework combining an eight-category bias taxonomy, sixteen-family mitigation grading, and dataset-provenance audit provides a reusable foundation for future systematic updates [10, 18]. These strengths carry trade-offs. An English-language restriction may exclude relevant work from non-Anglophone jurisdictions, compounding the Global South under-representation this review documents [22, 11]. We identified 23 potentially relevant non-English studies (17 Chinese, 4 Spanish, 2 German) during screening; abstract review suggested most focused on technical image processing without fairness analysis, but we cannot exclude the possibility of relevant unpublished work. Inclusion of preprints, necessary to capture the foundation-model literature, introduces the risk that some reported effect sizes may not survive peer review [9, 25]. Single-reviewer extraction with 20% audit means residual misclassification at bias-category boundaries cannot be excluded (inter-rater agreement: 88.9%, Cohen's $\kappa = 0.85$), and qualitative effectiveness grading rather than formal meta-analysis, while justified by the heterogeneity of study designs, limits quantitative pooling.

Two additional limitations warrant emphasis. First, **publication bias** may inflate reported fairness gaps and mitigation effectiveness. Studies finding no significant fairness disparities, or reporting failed mitigation attempts, may be less likely to submit or pass peer review. The 78-study corpus likely over-represents positive findings (bias detected, mitigation successful) relative to null results. We attempted to mitigate this by including grey literature and preprints, but cannot exclude the possibility that the true prevalence of fairness metrics in the field is higher than our 19.2% estimate if studies without fairness findings remain unpublished.

Second, the **multiple comparisons problem** affects our bias category assignments. With 78 studies and 8 bias categories (plus multiple mitigation families and outcome metrics), some associations may arise by chance. We mitigated this risk by requiring explicit experimental evidence for each bias category assignment, using conservative evidence grading (strong/moderate/mixed/failed), and focusing on effect sizes rather than statistical significance. Nevertheless, readers should interpret individual category counts (e.g., $n = 5$ for demographic bias, $n = 7$ for institutional bias) as descriptive rather than inferential statistics. The 95% confidence intervals reported throughout the results section quantify sampling uncertainty but do not adjust for multiple testing across bias categories.

Most importantly, the review inherits the same distributional distortions it critiques: TCGA dominance, high-income geographic concentration, and structural absence of demographic metadata. Every corpus-wide statistic should be read as valid within the sampling frame of evidence available to March 2026 rather than as a global estimate, and the search cut-off excludes second-generation foundation models and post-cut-off regulatory guidance from this synthesis [3]. Structural claims (site leakage unresolved, fairness under-reported, TCGA over-anchoring, infrastructure binding) are supported by convergent evidence across multiple dataset families. Quantitative claims rest on a smaller evidence base and should be cited with their empirical provenance.

4.4 Recommendations and Research Priorities

4.4.1 Reporting Checklist and Benchmark Proposal

We propose two concrete responses to the gaps identified above. The first is a 24-item reporting checklist organised into five blocks (Table 5), designed to complement CLAIM, TRIPOD+AI, and STARD-AI [26]. The second is a Fairness Benchmark for Histopathology (FB-Histo) built on six pillars: multi-site federation (10+ sites across 4 continents), demographic stratification with mandatory patient-level annotation, a task suite covering cancer-subtype classification through biomarker prediction, canonical metrics aligned with Block D, preserved-site cross-validation as default, and governance with pathologist, patient, and Global South representation [8, 5, 23]. Governance would follow a multi-stakeholder consortium model with 9-member steering committee (including 2 Global South representatives with veto power), Fairness Review Board, and Technical Committee. Tiered data access (open, controlled, federated) balances openness with privacy, while incentive mechanisms including co-authorship for data contributors, cash prizes for competition participants, and subsidised compute credits for Global South participation encourage broad engagement. This model builds on established pathology benchmarks while addressing their demographic representativeness limitations.

Table 5: Proposed 24-item reporting checklist for fairness in histopathology AI, organised into five blocks. Block D operationalises the six-item minimum metric floor proposed in this review.

Block	Focus	Required items	Ref.
A	Tissue provenance	Sites, scanners, magnification, staining protocol, quality-control log	[8, 9]
B	Pre-processing audit	Normaliser identity, template slide, augmentation pipeline	[29, 69]
C	Cohort demographics	Per-subgroup sample sizes, attribute source (self-reported vs. inferred), geographic composition, missingness disclosure, consent provisions	[5, 11]
D	Fairness evaluation	Subgroup AUROC (ID & OOD), subgroup disparity metric, worst-group accuracy/AUROC, ECE and subgroup calibration gap, preserved-site leakage probe, MIL attention-level fairness summary	[23, 5, 53, 8, 10]
E	Reproducibility	Code availability, model weights, dataset access tiers, evaluation-harness documentation	[24, 23]

4.4.2 Methodological Research Priorities

Beyond these infrastructure deliverables, four methodological research priorities emerge. Fairness-aware foundation-model pretraining is the most pressing because current models encode TSS at over 90% accuracy [9], and post-hoc correction addresses individual tasks without resolving the contaminated substrate. The goal is to incorporate batch- and demographic-invariance constraints into pretraining itself through site-adversarial self-supervision, scanner-conditioned

disentanglement, or contrastive objectives that treat same-patient, different-scanner slides as positive pairs [25, 6]. A second priority is causal and counterfactual modeling. Nearly all validated methods operate on associations and cannot distinguish spurious TSS signatures from biological signal [8, 9]. Cross-modal evidence confirms this ceiling: across radiology, MRI, and fundus imaging, pre-training improves accuracy but has minimal impact on fairness ($p>0.05$ in 18 of 24 settings), with no method consistently outperforming empirical risk minimisation across demographic groups [109]. Within histopathology specifically, Correa-Medero et al. [78] demonstrated equalised survival performance across seven sites using causal adjustment, while Shabazian et al. [79] coupled fairness with explanation-guided training.

The third and fourth priorities concern implementation mechanisms. Federated learning requires fairness-aware aggregation beyond Prop-FFL [65], encompassing federated fairness auditing without centralising data and privacy-preserving substrates calibrated to avoid erasing subgroup disparities [92]. FlexFair addresses this challenge directly by jointly optimising demographic parity, equalised odds, and accuracy without centralising data [105], and should be integrated with fairness-aware pretraining objectives. Generative models offer a practical route to augmenting underrepresented subgroups: Ktena et al. [37] achieved 48.5% relative fairness improvement via conditional diffusion, and Van Booven et al. [38] improved demographic parity for prostate grading [37, 38]. These four priorities are mutually reinforcing: fairness-aware pretraining provides the substrate that causal interventions require, federated learning supplies the multi-institutional diversity both depend on, and generative augmentation fills the long-tail intersectional cells that neither training data nor federation alone can supply.

4.4.3 Policy, Implementation, and Ethical Considerations

Funding bodies should condition grants on representativeness plans with demographic and geographic targets, earmarking budgets for Global South sites [11, 3]. Journals should require checklist completion (Blocks A–E) as a submission precondition [26, 10], while challenge organisers should extend benchmarks like MIDOG and GrandQC to incorporate fairness evaluation [54, 71]. Demographic metadata collection must balance audit necessity against privacy risk: consent should be purpose-specific, collection limited to minimum necessary attributes, and re-identification authorised only for adverse-event investigation [12, 10]. Self-reported race, genetic ancestry, and tissue-source site address different questions and require joint deliberation among clinicians, patients, and ethicists [8, 5]. Hybrid-responsibility frameworks are needed for foundation-model failures, as pathologists cannot detect embedding-level site leakage through slide inspection alone [13, 9]. Finally, equitable care requires international consortia and capacity-building partnerships to ensure populations absent from training data are not excluded from benefits [22, 12].

5 Conclusion

Evidence from 78 histopathology AI fairness studies (2017–2026) reveals a sharp and uncomfortable truth: fairness in computational pathology is empirically measurable and methodologically tractable, yet remains structurally unmeasured. The field has built the instruments to detect its own failures. TSS leakage is recoverable at >0.96 AUROC across multiple TCGA cancer types and at greater than 90% linear-probe accuracy from state-of-the-art foundation-model embeddings. Demographic subgroup AUROC gaps of 3–16% persist in breast, lung, and colorectal tasks, with 10–25% cross-population performance degradation at geography boundaries.

Individual fairness interventions have already demonstrated clinically meaningful effects. Knowledge-guided fine-tuning achieves 61% demographic-gap reduction, fairness-aware contrastive objectives deliver 88.5–91.1% disparity mitigation across 20 TCGA cancer types (88.5% internal, 91.1% external validation; range reflects validation setting), conditional diffusion yields 48.5% improvement in demographic parity, and SPARE improves minority-group accuracy by 3.7–5.8%. Proportional federated averaging reduces inter-hospital performance variance from 32.54 to 10.00, and causal survival models equalise recurrence-free-survival performance across seven geographically distributed sites. These gains remain at scales insufficient to bound aggregate TSS deployment risk.

Yet only 19.2% of studies report any fairness metric, 7.7% report worst-group performance, and 29.5% (95% CI: 20.5–40.2%) perform external validation. The geographic concentration is pronounced: 84.6% (66/78) of first/corresponding authors are from North America and Europe, with East Asia contributing most of the remainder. Meanwhile, review recommendations consistently endorse demographic reporting ($\sim$90%), multi-site evaluation ($\sim$85%), and preserved-site cross-validation ($\sim$90%), yet experimental practice fails to supply these. This 50–75 percentage-point gap between review consensus and experimental practice reflects not an epistemic failure but an infrastructural one. The checklist, benchmark, methodological priorities, and policy levers proposed in this review specify precisely the reporting, evaluation, representational, and incentive substrates whose absence the preceding sections have quantified. The state of the field in 2026 is therefore an enforcement problem, not a discovery problem.

Closing this gap requires coordinated action across four constituencies. Researchers must adopt preserved-site cross-validation, subgroup-stratified reporting, and embedding-level TSS leakage probes as minimum-viable evidence, and

must publish datasets, code, and evaluation harnesses with per-field demographic completeness disclosed rather than implied. Clinicians, pathologists, and laboratory directors must treat scanner calibration, stain-protocol drift, and subgroup-stratified quality control as first-class workflow instruments and demand failure-mode transparency from deployed devices. Regulators and editors must bind the reporting checklist proposed in this review to pre-market authorisation and journal-submission policy, and require post-market subgroup surveillance with the same rigour already imposed on aggregate performance. Funders, consortia, and patient communities must resource the multi-continental, demographically stratified cohorts and the federated, consented governance infrastructure without which fairness-aware pretraining, causal estimation, and synthetic augmentation remain exercises on a distribution that no longer adequately represents the populations computational pathology will be asked to serve.

Each required artefact (the checklist, the benchmark, the mitigation programme, and the incentive layer) lies within reach of the community that has already produced the evidence documented in this review; none will emerge from exhortation alone. The next decade of histopathology AI will be judged not by whether its models achieve higher accuracy on aggregate benchmarks, but by whether the field chooses to instrument, report, and govern the subgroup behavior of the systems it is now placing into clinical practice. Equitable computational pathology is, on the evidence assembled here, an operational commitment rather than an aspirational one. The window to make that commitment on the field's own terms, rather than retrofitting it under regulatory, litigation, or public-trust pressure, is open now and will not remain open indefinitely.

References

[1] A. Tiwari, S. Mishra, and T.-R. Kuo, "Current AI technologies in cancer diagnostics and treatment," *Molecular Cancer*, vol. 24, no. 1, p. 159, 2025. doi:10.1186/s12943-025-02369-9

[2] L. Chang, H. Li, W. Wu, X. Liu, J. Yan, Z. Chen, H. Wu, and S. Song, "Applications of artificial intelligence in non-small cell lung cancer: from precision diagnosis to personalized prognosis and therapy," *Journal of Translational Medicine*, vol. 24, p. 108, 2026. doi:10.1186/s12967-025-07591-z

[3] N. Biswas, H. Berg, and S. K. Shukla, "Digital pathology: Barriers to transitioning to the digital age," *DIGITAL HEALTH*, vol. 12, pp. 1–11, 2026. doi:10.1177/20552076261418364

[4] M. Ochi, D. Komura, and S. Ishikawa, "Pathology foundation models," *JMA Journal*, vol. 12, no. 10, pp. 921–926, 2024. doi:10.31662/jmaj.2024-0206

[5] A. Vaidya, R. J. Chen, D. F. K. Williamson, A. H. Song, G. Jaume, Y. Yang, T. Hartvigsen, E. C. Dyer, M. Y. Lu, J. Lipkova, M. Shaban, T. Y. Chen, and F. Mahmood, "Demographic bias in misdiagnosis by computational pathology models," *Nature Medicine*, vol. 30, no. 4, pp. 1174–1190, 2024. doi:10.1038/s41591-024-02885-z

[6] S.-Y. Lin, P.-C. Tsai, F.-Y. Su *et al.*, "Contrastive learning enhances fairness in pathology artificial intelligence systems," *Cell Reports Medicine*, vol. 6, p. 102527, 2025. doi:10.1016/j.xcrm.2025.102527

[7] T. Dehkharghanian, Y. Mu, H. R. Tizhoosh, and C. J. Campbell, "Biased data, biased AI: deep networks predict the acquisition site of TCGA images," *Diagnostic Pathology*, vol. 18, p. 67, 2023. doi:10.1186/s13000-023-01355-3

[8] F. M. Howard, J. Dolezal, S. Kochanny, J. Schulte, H. Chen, L. Heij, D. Huo, R. Nanda, O. I. Olopade, J. N. Kather, N. Cipriani, R. L. Grossman, and A. T. Pearson, "The impact of site-specific digital histology signatures on deep learning model accuracy and bias," *Nature Communications*, vol. 12, no. 1, p. 4423, 2021. doi:10.1038/s41467-021-24698-1

[9] J. Kömen, H. Marienwald, J. Dippel, and J. Hense, "Do histopathological foundation models eliminate batch effects? a comparative study," in *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[10] R. J. Chen, J. J. Wang, D. F. K. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, and F. Mahmood, "Algorithmic fairness in artificial intelligence for medicine and healthcare," *Nature Biomedical Engineering*, vol. 7, pp. 719–742, 2023. doi:10.1038/s41551-023-01056-8

[11] L. A. Celi, J. Cellini, M.-L. Charpignon, E. C. Dee, F. Dernoncourt, R. Eber *et al.*, "Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review," *PLOS Digital Health*, vol. 1, no. 3, p. e0000022, 2022. doi:10.1371/journal.pdig.0000022

[12] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and AI for health care: A call for open science," *Patterns*, vol. 2, p. 100347, 2021. doi:10.1016/j.patter.2021.100347

[13] J. Y. Cheng, J. T. Abel, U. G. J. Balis, D. S. McClintock, and L. Pantanowitz, "Challenges in the development, deployment, and regulation of artificial intelligence in anatomic pathology," *The American Journal of Pathology*, vol. 191, no. 10, pp. 1684–1692, 2021. doi:10.1016/j.ajpath.2020.10.018

[14] F. Kheiri, S. Rahnamayan, M. Makrehchi, and A. Asilian Bidgoli, "Investigation on potential bias factors in histopathology datasets," *Scientific Reports*, vol. 15, p. 13349, 2025. doi:10.1038/s41598-025-89210-x

[15] A. Asilian Bidgoli, S. Rahnamayan, T. Dehkharghanian, A. Grami, and H. R. Tizhoosh, "Bias reduction in representation of histopathology images using deep feature selection," *Scientific Reports*, vol. 12, p. 19994, 2022. doi:10.1038/s41598-022-24317-z

[16] S. Chen, R. Verma, J. Jegminat, T. Fuchs, G. Campanella, K.-l. Huang, K. Clare, and B. Veremis, "Predicting patient self-reported race from skin histological images with deep learning," in *MICCAI Workshop on Fairness of AI in Medical Imaging (FAIMI)*, 2025.

[17] E. Tasci, Y. Zhuge, K. Camphausen, and A. V. Krauze, "Bias and class imbalance in oncologic data—towards inclusive and transferrable AI in large-scale oncology data sets," *Cancers*, vol. 14, no. 12, p. 2897, 2022. doi:10.3390/cancers14122897

[18] D. Montezuma, R. Porz, D. Ameisen, V. L'Imperio, M.-S. Serbanescu, J. Temprana-Salvador, N. Zerbe, N. Khalili, and I. Zlobec, "Unbiased artificial intelligence: Addressing bias in computational pathology," *Mayo Clinic Proceedings: Digital Health*, vol. 3, no. 4, p. 100302, 2025. doi:10.1016/j.mcpdig.2025.100302

[19] M. A. Ricci Lara, R. Echeveste, and E. Ferrante, "Addressing fairness in artificial intelligence for medical imaging," *Nature Communications*, vol. 13, p. 4581, 2022. doi:10.1038/s41467-022-32186-3

[20] Y. Yang, M. Lin, H. Zhao, Y. Peng, F. Huang, and Z. Lu, "A survey of recent methods for addressing AI fairness and bias in biomedicine," *Journal of Biomedical Informatics*, vol. 154, p. 104646, 2024. doi:10.1016/j.jbi.2024.104646

[21] S. S. Chinta, Z. Wang, A. Palikhe, X. Zhang, A. Kashif, M. A. Smith *et al.*, "AI-driven healthcare: A review on ensuring fairness and mitigating bias," *PLOS Digital Health*, vol. 4, no. 5, p. e0000864, 2025. doi:10.1371/journal.pdig.0000864

[22] A. Musa, R. Prasad, P. Onwualu, and M. Hernandez, "A systematic review of cross-population shifts in medical imaging analysis with deep learning," *Big Data and Cognitive Computing*, vol. 10, no. 6, p. 76, 2026. doi:10.3390/bdcc10060076

[23] Y. Zong, Y. Yang, and T. Hospedales, "MedFair: Benchmarking fairness for medical imaging," in *International Conference on Learning Representations (ICLR)*, 2023.

[24] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. S. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Koyejo, L. Schmidt, and P. Liang, "WILDS: A benchmark of in-the-wild distribution shifts," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 139, 2021.

[25] Y. Huang, W. Zhao, Z. Zhang, Y. Chen, Y. Fu, F. Wu, Y. Jiang, L. Liang, S. Wang, and L. Yu, "Knowledge-guided adaptation of pathology foundation models effectively improves cross-domain generalization and demographic fairness," *Nature Communications*, vol. 16, p. 11485, 2025. doi:10.1038/s41467-025-66300-y

[26] V. Sounderajah, H. Ashrafian, R. M. Golub, S. Shetty, J. De Fauw, L. Hooft, K. Moons, G. Collins, D. Moher, P. M. Bossuyt, A. Darzi, and A. Karthikesalingam, "Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol," *BMJ Open*, vol. 11, p. e047709, 2021. doi:10.1136/bmjopen-2020-047709

[43] A. Soltan and P. Washington, "Challenges in reducing bias using post-processing fairness for breast cancer stage classification with deep learning," *Algorithms*, vol. 17, no. 4, p. 141, 2024. doi:10.3390/a17040141

[28] P. Murchan, P. Ó Broin, A.-M. Baird, O. Sheils, and S. P. Finn, "Deep feature batch correction using ComBat for machine learning applications in computational pathology," *Journal of Pathology Informatics*, vol. 15, p. 100396, 2024. doi:10.1016/j.jpi.2024.100396

[29] S. Lin, H. Zhou, M. Watson, R. Govindan, R. J. Cote, and C. Yang, "Impact of stain variation and color normalization for prognostic predictions in pathology," *Scientific Reports*, vol. 15, p. 2365, 2025. doi:10.1038/s41598-024-83267-w

[30] X. Ling, M. Ouyang, Y. Wang, X. Chen, R. Yan, H. Chu, J. Cheng, T. Guan, S. Tian, X. Liu, and Y. He, "Agent aggregator with mask denoise mechanism for histopathology whole slide image analysis," in *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, 2024. doi:10.1145/3664647.3681425

[31] M. Hägele, P. Seegerer, S. Lapuschkin, M. Bockmayr, W. Samek, F. Klauschen, K.-R. Müller, and A. Binder, "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods," *Scientific Reports*, vol. 10, p. 6423, 2020. doi:10.1038/s41598-020-62724-2

[32] P.-J. Lin *et al.*, "Mutation rate differences across populations and association with performance disparities in pathology AI diagnostic models," in *ASCO Annual Meeting*, 2025. doi:10.1200/jco.2025.43.16_suppl.1603

[33] E. Lazarova *et al.*, "Demographic, morphological, and histopathological characteristics of melanoma and nevi: Insights from statistical analysis and machine learning models," *Diagnostics*, vol. 15, no. 19, p. 2499, 2025. doi:10.3390/diagnostics15192499

[34] R. Daneshjou, K. Vodrahalli, R. A. Novoa *et al.*, "Disparities in dermatology AI performance on a diverse, curated clinical image set," *Science Advances*, vol. 8, no. 31, p. eabq6147, 2022. doi:10.1126/sciadv.abq6147

[35] S. T. Arasteh *et al.*, "Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging," *Communications Medicine*, vol. 4, p. 46, 2024. doi:10.1038/s43856-024-00462-6

[36] M. Ceccon, G. Cornacchia, D. Dalle Pezze, A. Fabris, and G. A. Susto, "Underrepresentation, label bias, and proxies: Towards data bias profiles for the eu ai act and beyond," *Expert Systems With Applications*, vol. 292, p. 128266, 2025. doi:10.1016/j.eswa.2025.128266

[37] I. Ktena, O. Wiles, I. Albuquerque, S.-A. Rebuffi, R. Tanno, A. Guha Roy, S. Azizi, D. Belgrave, P. Kohli, T. Cemgil, A. Karthikesalingam, and S. Gowal, "Generative models improve fairness of medical classifiers under distribution shifts," *Nature Medicine*, vol. 30, no. 4, pp. 1166–1173, 2024. doi:10.1038/s41591-024-02838-6

[38] D. J. Van Booven, C.-B. Chen, O. N. Kryvenko, S. Punnen *et al.*, "Mitigating bias in prostate cancer diagnosis using synthetic data for improved AI-driven Gleason grading," *npj Precision Oncology*, vol. 9, pp. 1–13, 2025. doi:10.1038/s41698-025-00934-5

[39] X. Ling *et al.*, "Comprehensive benchmark dataset for pathological lymph node metastasis in breast cancer sections," *Scientific Data*, 2025. doi:10.1038/s41597-025-05586-5

[40] M. Dawood, K. Branson, S. Tejpar, N. Rajpoot, and F. u. A. A. Minhas, "Confounding factors and biases abound when predicting molecular biomarkers from histological images," *Nature Biomedical Engineering*, 2026. doi:10.1038/s41551-026-01616-8

[41] C. O. Ly, B. Unnikrishnan, T. Tadic, T. Patel, J. Duhamel, S. Kandel, Y. Moayedi, M. Brudno, A. Hope, H. Ross, and C. McIntosh, "Shortcut learning in medical ai hinders generalization: method for estimating ai model generalization without external data," *npj Digital Medicine*, 2024. doi:10.1038/s41746-024-01118-4

[42] T. Lazard, G. Bataillon, P. Naylor, T. Popova, M.-H. Stern *et al.*, "Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images," *Cell Reports Medicine*, vol. 3, no. 12, p. 100872, 2022. doi:10.1016/j.xcrm.2022.100872

[44] Y. Tolkach *et al.*, "An international multi-institutional validation study of the algorithm for prostate cancer detection and Gleason grading," *npj Precision Oncology*, vol. 7, p. 75, 2023. doi:10.1038/s41698-023-00424-6

[45] A. Raza, F. Hanif, and H. A. Mohammed, "Clinical validation of lightweight cnn architectures for reliable multi-class classification of lung cancer using histopathological imaging techniques," *Scientific Reports*, vol. 16, p. 6523, 2026. doi:10.1038/s41598-026-36652-6

[46] N. Bussola, J. Xu, L. Wu, L. Gorini, Y. Zhang, C. Furlanetto, and W. Tong, "A weakly supervised deep learning framework for whole slide classification to facilitate digital pathology in animal study," *Chemical Research in Toxicology*, vol. 36, pp. 1321–1331, 2023. doi:10.1021/acs.chemrestox.3c00058

[47] K. Nagpal, D. Foote, F. Tan, Y. Liu, P.-H. Chen, D. F. Steiner, C. H. Mermel, and M. C. Stumpe, "Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens," *JAMA Oncology*, vol. 6, no. 9, pp. 1372–1380, 2020. doi:10.1001/jamaoncol.2020.2485

[48] K. Liang, Z. Yin, M. Min, Y. Liu, Z. Ma, and J. Guo, "Learning dynamic prototypes for visual pattern debiasing," *International Journal of Computer Vision*, vol. 132, pp. 1777–1799, 2024. doi:10.1007/s11263-023-01956-x

[49] M. S. Reza and J. Ma, "Imbalanced histopathological breast cancer image classification with convolutional neural network," in *Proceedings of ICSP 2018*, 2018, pp. 619–626. doi:10.1109/ICSP.2018.8652304

[50] M. Schmitt, R. C. Maron, A. Hekler *et al.*, "Hidden variables in deep learning digital pathology and their potential to cause batch effects: Prediction model study," *Journal of Medical Internet Research*, vol. 23, no. 2, p. e23436, 2021. doi:10.2196/23436

[51] P. Mazaheri, A. Asilian Bidgoli, S. Rahnamayan, and H. R. Tizhoosh, "Ranking loss and sequestering learning for reducing image search bias in histopathology," *Applied Soft Computing*, vol. 142, p. 110346, 2023. doi:10.1016/j.asoc.2023.110346

[52] M. Sikaroudi, S. M. Hosseini, C. Chen, and H. R. Tizhoosh, "Generalization of vision pre-trained models for histopathology," *Scientific Reports*, vol. 13, p. 6065, 2023. doi:10.1038/s41598-023-32224-6

[53] M. Roschewitz, G. Khara, J. Yearsley, N. Sharma, J. J. James, É. Ambróczy, A. Heroux, P. Kecskemethy, T. Rijken, and B. Glocker, "Automatic correction of performance drift under acquisition shift in medical image classification," *Nature Communications*, vol. 14, p. 6608, 2023. doi:10.1038/s41467-023-42396-y

[54] M. Aubreville, N. Stathonikos, C. A. Bertram, R. Klopfleisch, N. ter Hoeve, F. Ciompi, F. Wilm, C. Marzahl *et al.*, "Mitosis domain generalization in histopathology images—the MIDOG challenge," *Medical Image Analysis*, vol. 84, p. 102699, 2022. doi:10.1016/j.media.2022.102699

[55] C. Franchet, R. Schwob, G. Bataillon, C. Stryh, S. Pétecourt, F.-X. Prono, F. Penault-Llorca, M. Lacroix-Triki, L. Arnould, J. Lemonnier, J.-M. Alliot, T. Filieron, and P. Broussard, "Bias reduction using combined stain normalization and augmentation for AI-based classification of histological images," *Computers in Biology and Medicine*, vol. 171, p. 108130, 2024. doi:10.1016/j.compbiomed.2024.108130

[56] N. Nerrienet, R. Peyret, M. Sockeel, and S. Sockeel, "Standardized CycleGAN training for unsupervised stain adaptation in invasive carcinoma classification for breast histopathology," *Journal of Medical Imaging*, vol. 10, no. 6, p. 067502, 2023. doi:10.1117/1.JMI.10.6.067502

[57] C. Dunn *et al.*, "An international study of stain variability in histopathology using qualitative and quantitative analysis," *Journal of Pathology Informatics*, 2025. doi:10.1016/j.jpi.2025.100423

[58] M. Cecconi, D. Dalle Pezze, A. Fabris, and G. A. Stuto, "Fairness evolution in continual learning for medical imaging," in *IFAC-PapersOnLine*, vol. 58, no. 26, 2024, pp. 145–156. doi:10.1016/j.ifacol.2025.12.025

[59] I. Skarga-Bandurova *et al.*, "Bias detection in histology images using explainable AI and image darkness assessment," *Studies in Health Technology and Informatics*, p. 250302, 2023. doi:10.3233/SHTI250302

[60] D. Sauter, G. Lodde, F. Nensa, D. Schadendorf, E. Livingstone, and M. Kukuk, "Validating automatic concept-based explanations for ai-based digital histopathology," *Sensors*, vol. 22, no. 14, p. 5346, 2022. doi:10.3390/s22145346

[61] J. Tschida *et al.*, "Evaluating algorithmic bias on biomarker classification of breast cancer pathology reports," *JAMIA Open*, 2025. doi:10.1093/jamiaopen/ooaf033

[62] M. A. Sufian, L. Alsaedar, B. Hamzi, S. Zaman, A. S. M. S. Sagar *et al.*, "Mitigating algorithmic bias in AI-driven cardiovascular imaging for fairer diagnostics," *Diagnostics*, vol. 14, p. 2675, 2024. doi:10.3390/diagnostics14232675

[63] G. Xu, Y. Duan, J. Xia, C.-H. Chiu, M. Lemmon, W. Jin, and Y. Shi, "Rethinking fairness in medical imaging: Maximizing group-specific performance with application to skin disease diagnosis," *Medical Image Analysis*, vol. 109, p. 103950, 2026. doi:10.1016/j.media.2026.103950

[64] M. Asadi-Aghbolaghi, A. Darbandsari, A. Zhang, A. Contreras-Sanz, J. Boschman, P. Ahmadvand, M. Köbel, D. Farnell, D. G. Huntsman, A. Churg, P. C. Black, G. Wang, C. B. Gilks, H. Farahani, and A. Bashashati, "Learning generalizable AI models for multi-center histopathology image classification," *npj Precision Oncology*, vol. 8, p. 151, 2024. doi:10.1038/s41698-024-00652-4

[65] S. M. Hosseini, M. Sikaroudi, M. Babaie, and H. R. Tizhoosh, "Proportionally fair hospital collaborations in federated learning of histopathology images," *IEEE Transactions on Medical Imaging*, vol. 42, no. 7, pp. 1982–1995, 2023. doi:10.1109/TMI.2023.3234450

[66] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019. doi:10.1038/s41591-019-0508-1

[67] Y. Xu, Z. Jia, L.-B. Wang, Y. Ai, F. Zhang, M. Lai, and E. I.-C. Chang, "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features," *BMC Bioinformatics*, vol. 18, p. 281, 2017. doi:10.1186/s12859-017-1685-x

[68] H. Olsson, K. Kartasalo, N. Mulliqi, M. Capuccini, P. Ruusuvuori, H. Samaratunga, B. Delahunt, C. Lindskog, E. A. M. Janssen, L. Egevad, O. Spjuth, and M. Eklund, "Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction," *Nature Communications*, vol. 13, p. 7761, 2022. doi:10.1038/s41467-022-34945-8

[69] W. Voon, Y. C. Hum, Y. K. Tee, W.-S. Yap, H. Nisar, H. Mokayed, N. Gupta, and K. W. Lai, "Evaluating the effectiveness of stain normalization techniques in automated grading of invasive ductal carcinoma histopathological images," *Scientific Reports*, vol. 13, p. 20518, 2023. doi:10.1038/s41598-023-46619-6

[70] K. Faryna, J. van der Laak, and G. Litjens, "Automatic data augmentation to improve generalization of deep learning in H&E stained histopathology," *Computers in Biology and Medicine*, vol. 170, p. 108018, 2024. doi:10.1016/j.compbiomed.2024.108018

[71] Z. Weng, A. Seper, A. Pryalukhin, F. Mairinger, C. Wickenhauser, M. Bauer, L. Glamann, H. Bläker, T. Lingscheidt, W. Hulla, D. Jonigk, S. Schallenberg, A. Bychkov, J. Fukuoka, M. Braun, B. Schömig-Markiefka, S. Klein, A. Thiel, K. Bozek, G. J. Netto, A. Quaas, R. Büttner, and Y. Tolkach, "GrandQC: A comprehensive solution to quality control problem in digital pathology," *Nature Communications*, vol. 15, p. 10038, 2024. doi:10.1038/s41467-024-54769-y

[72] M. W. Lafarge, J. P. W. Pluim, K. A. J. Eppenhof, and M. Veta, "Learning domain-invariant representations of histological images," *Frontiers in Medicine*, vol. 6, p. 162, 2019. doi:10.3389/fmed.2019.00162

[73] S. Otálora, M. Atzori, V. Andrearczyk, A. Khan, and H. Müller, "Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 198, 2019. doi:10.3389/fbioe.2019.00198

[74] O. Ciga, T. Xu, and A. L. Martel, "Self-supervised contrastive learning for digital histopathology," *Machine Learning with Applications*, vol. 7, p. 100198, 2022. doi:10.1016/j.mlwa.2021.100198

[75] F. Zhang, D. Zhao, G. Bai, J. Jiang, G. Ye, X. Ji, and X. Rui, "Towards fairness-aware and privacy-preserving enhanced collaborative learning for healthcare," *Nature Communications*, vol. 16, p. 2852, 2025. doi:10.1038/s41467-025-58055-3

[76] H. Park *et al.*, "Uncertainty-based data-wise label smoothing for calibrating multiple instance learning in histopathology image classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. doi:10.1109/wacv61041.2025.00068

[77] S. Bhattacharyya *et al.*, "Conformal uncertainty quantification to evaluate predictive fairness of foundation AI model for skin lesion classes across patient demographics," *Health Information Science and Systems*, vol. 13, p. 32, 2025. doi:10.1007/s13755-025-00412-z

[78] R. L. Correa-Medero, R. Pai, K. Ebare, D. D. Buchanan, M. A. Jenkins, A. I. Phipps *et al.*, "Causal debiasing for unknown bias in histopathology — a colon cancer use case," *PLoS ONE*, vol. 19, no. 11, p. e0303415, 2024. doi:10.1371/journal.pone.0303415

[79] R. Shabazian, I. Trubitsyna, and S. L. Mirtaheri, "Joint explainable and fair AI in healthcare," *BMC Medical Informatics and Decision Making*, vol. 25, p. 403, 2025. doi:10.1186/s12911-025-03233-y

[80] N. Madusanka *et al.*, "Structure-preserving histopathological stain normalization via attention-guided residual learning," *Bioengineering*, vol. 12, no. 9, p. 950, 2025. doi:10.3390/bioengineering12090950

[81] M. J. Hetz, T.-C. Bucher, and T. J. Brinker, "Multi-domain stain normalization for digital pathology: A cycle-consistent adversarial network for whole slide images," *Medical Image Analysis*, vol. 94, p. 103149, 2024. doi:10.1016/j.media.2024.103149

[82] X. Ji, R. Salmon, N. Mullig, U. Khan, Y. Wang, A. Blilie, H. Olsson, B. G. Pedersen, K. D. Sørensen, B. P. Ulhøi, S. R. Kjosavik, E. A. M. Janssen, M. Rantalainen, L. Egevad, P. Ruusuvuori, M. Eklund, and K. Kartasalo, "Physical color calibration of digital pathology scanners for robust artificial intelligence—assisted cancer diagnosis," *Modern Pathology*, vol. 38, p. 100715, 2025. doi:10.1016/j.modpat.2025.100715

[83] G. Cazaniga, A. Caputo, V. L'Imperio, F. Gibilisco, M. Scotto, O. M. A. Pennisi, N. Michieli, A. Mogetta, F. Molinari, F. Fraggetta, and M. Salvi, "Not all stains are made equal: Impact of stain normalization on prostate cancer diagnosis," *Pathologica*, vol. 116, pp. 285–294, 2024. doi:10.32074/1591-951X-1008

[84] R. J. Chen, M. Y. Lu, D. F. K. Williamson, T. Y. Chen, J. Lipkova, Z. Fan, M. Shaban, M. Dunne, M. Zarella *et al.*, "Pan-cancer integrative histology-genomic analysis via multimodal deep learning," *Cancer Cell*, vol. 40, no. 8, pp. 865–878, 2022. doi:10.1016/j.ccell.2022.07.004

[85] S. Ahmed *et al.*, "A robust framework for domain-generalized classification of ovarian cancer histology images," *Diagnostics*, vol. 15, no. 23, p. 2954, 2025. doi:10.3390/diagnostics15232954

[86] J. Ren, I. Hacihaliloglu, E. A. Singer, D. J. Foran, and X. Qi, "Unsupervised domain adaptation for classification of histopathology whole-slide images," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 102, 2019. doi:10.3389/fbioe.2019.00102

[87] J. Vasiljevic, Z. Nisar, F. Feuerhake, C. Wemmert, and T. Lampert, "CycleGAN for virtual stain transfer: is seeing really believing?" *Artificial Intelligence in Medicine*, vol. 133, p. 102420, 2022. doi:10.1016/j.artmed.2022.102420

[88] E. Vorontsov, A. Boez, S. Mouez *et al.*, "A foundation model for clinical-grade computational pathology and rare cancers detection," *Nature Medicine*, vol. 30, no. 10, pp. 2924–2935, 2024. doi:10.1038/s41591-024-03141-0

[89] G. Campanella, S. Chen, M. Singh, R. Verma, S. Muehlstedt, J. Zeng, A. Stock, M. Crokern, B. Veremis, A. Elmas, I. Shujski, N. Naaitaamki, K.-l. Huang, R. Kwan, J. Houldsworth, A. J. Schoenfeld, and C. Vanderbitt, "A clinical benchmark of public self-supervised pathology foundation models," *Nature Communications*, vol. 16, p. 3640, 2025. doi:10.1038/s41467-025-58796-1

[90] M. Shi *et al.*, "Equitable deep learning for diabetic retinopathy detection using multidimensional retinal imaging with fair adaptive scaling," *Translational Vision Science & Technology*, vol. 14, no. 7, p. 1, 2025. doi:10.1167/tvst.14.7.1

[91] S. Haggenmüller, M. Schmitt, E. Krieghoff-Henning *et al.*, "Federated learning for decentralized artificial intelligence in melanoma diagnostics," *JAMA Dermatology*, vol. 160, no. 3, pp. 303–311, 2024. doi:10.1001/jamadermatol.2023.5550

[92] M. Y. Lu, R. J. Chen, D. Kong, J. Lipkova, R. Singh, D. F. K. Williamson, T. Y. Chen, and F. Mahmood, "Federated learning for computational pathology on gigapixel whole slide images," *Medical Image Analysis*, vol. 76, p. 102298, 2022. doi:10.1016/j.media.2021.102298

[93] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Scientific Reports*, vol. 12, no. 1, p. 1953, 2022. doi:10.1038/s41598-022-05539-7

[94] B. L. Y. Agbley *et al.*, "Federated learning-based detection of invasive carcinoma of no special type with histopathological images," *Diagnostics*, vol. 12, no. 7, p. 1669, 2022. doi:10.3390/diagnostics12071669

[95] S. Shukla and S. Doyle, "Federated learning in computational pathology: a literature review," *Journal of Medical Imaging*, vol. 12, no. 6, p. 061412, 2025. doi:10.1117/1.JMI.12.6.061412

[96] J. W. Gao, I. Boroojerdi, A. Bodduluri, J. H. Burns, L. A. Cai, L. Cheng, R. Ramos, M. Ghassemi, S. Purkayastha *et al.*, "AI recognition of patient race in medical imaging: a modelling study," *The Lancet Digital Health*, vol. 4, no. 6, pp. e406–e414, 2022. doi:10.1016/S2589-7500(22)00063-2

[97] Y. Yang, Y. Liu, X. Liu, A. Gulhane, D. Mastrodicasa, W. Wu, E. J. Wang, D. Sahani, and S. Patel, "Demographic bias of expert-level vision-language foundation models in medical imaging," *Science Advances*, 2025. doi:10.1126/sciadv.adq0305

[98] X. Wang, J. T. Steensma, M. H. Bailey, Q. Feng, H. Padda, and K. J. Johnson, "Characteristics of The Cancer Genome Atlas cases relative to U.S. general population cancer cases," *British Journal of Cancer*, vol. 119, pp. 885–892, 2018. doi:10.1038/s41416-018-0140-8

[99] E. Puyol-Antón, B. Ruijsink, J. Mariscal Harane, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi, P. Chowdhury, and A. P. King, "Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation," *Frontiers in Cardiovascular Medicine*, vol. 9, p. 899310, 2022. doi:10.3389/fcvm.2022.899310

[100] K. Drukker, W. Chen, J. Gichoya, C. Grusauskas, J. Kalpathy-Cramer, S. Koyejo, K. Myers, R. Sá, B. Sahiner, H. Whitney, Z. Zhang, and M. Giger, "Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment," *Journal of Medical Imaging*, vol. 10, no. 6, p. 061104, 2023. doi:10.1117/1.JMI.10.6.061104

[101] I. Banerjee, K. Bhattacharjee, J. L. Burns, H. Trivedi, S. Purkayastha, L. Seyyed-Kalantari, B. N. Patel, S. Rakesh, and J. Gichoya, "'Shortcuts' causing bias in radiology artificial intelligence: causes, evaluation, and mitigation," *Journal of the American College of Radiology*, vol. 20, no. 9, pp. 842–851, 2023. doi:10.1016/j.jacr.2023.06.025

[102] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature Medicine*, vol. 27, no. 12, pp. 2176–2182, 2021. doi:10.1038/s41591-021-01595-0

[103] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi, "The limits of fair medical imaging AI in real-world generalization," *Nature Medicine*, 2024. doi:10.1038/s41591-024-02888-w

[104] V. Muralidharan, B. A. Adewale, C. J. Huang, M. T. Nta, P. O. Ademiju, P. Pathmarajah, M. K. Hang, O. Adesanya, R. O. Abdullateef, A. O. Babatunde, A. Ajibade, S. Onyeka, Z. R. Cai, R. Daneshjou, and T. Olatunji, "A scoping review of reporting gaps in FDA-approved AI medical devices," *npj Digital Medicine*, vol. 7, no. 1, p. 270, 2024. doi:10.1038/s41746-024-01270-x

[105] H. Xing, R. Sun, J. Ren *et al.*, "Achieving flexible fairness metrics in federated medical imaging," *Nature Communications*, vol. 16, p. 3342, 2025. doi:10.1038/s41467-025-58949-0

[106] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences*, vol. 117, no. 21, pp. 12 592–12 594, 2020. doi:10.1073/pnas.1919012117

[107] R. F. Du *et al.*, "Ethics of foundation models in computational pathology: Overview of contemporary issues and future implications," *IEEE Transactions on Medical Imaging*, 2025. doi:10.1109/TMI.2025.3551913

[108] European Parliament and Council of the European Union, "Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act)," 2024, official Journal of the European Union L 2024/1689. Penalties for non-compliance: up to 35 million or 7% of global annual turnover (Article 99). [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

[109] T.-H. Pham, J. Chen, S. Lee, Y. Wang, S. Moroi, X. Zhang, and P. Zhang, "The boundaries of fair AI in medical image prognosis: A causal perspective," in *39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Appendices

A Summary Characteristics of Included Studies

Table 6 summarises the key dimensional characteristics of the 78 included studies, and Table 7 provides the cancer-type breakdown. Percentages are computed over the analytical corpus unless otherwise indicated; category counts may be non-exclusive where a study contributes to more than one dimension.

Table 6: Summary characteristics of the 78 included histopathology fairness studies (2017–2026). Percentages are computed over the analytical corpus of 78 studies unless otherwise indicated; category counts may be non-exclusive where a study contributes to more than one bucket.

Dimension	Value	Representative evidence
Temporal and publication profile		
Time window	2017–2026 (10 yr)	[8, 25]
Pre-2020 / 2020–22 / 2023–24 / 2025–26	6 / 17 / 32 / 18 (n); unreported: 5	[9, 5]
Peer-reviewed journal / ML conference	71 (91.0%) / 7 (9.0%)	[13, 16]
Geographic distribution (first/corresponding author)		
North America + Europe	84.6% (66/78)	[8, 9, 11]
East Asia / Global South	Most of remainder / Near-absent	[22, 18]
Dataset usage (non-exclusive)		
TCGA (primary or supplementary)	33 (42.3%)	[8, 7, 17]
CAMELYON16/17	∼7 (9.0%)	[9, 66]
Private single-site / Multi-institutional	40 (51.3%) / 23 (29.5%)	[43, 5]
External validation performed	23 (29.5%, 95% CI: 20.5–40.2%)	[6, 25]
Cancer coverage (primary focus)		
Breast / Multi / Prostate	44 (56.4%) / 24 (30.8%) / 15 (19.2%)	[5, 8, 7]
Lung / Colorectal / Skin	9 (11.5%) / 8 (10.3%) / 10 (12.8%)	[16, 25, 9]
Paediatric / rare tumours	0	—
Fairness evaluation and reporting		
Accuracy / AUROC reported	67 (85.9%) / 55 (70.5%)	[66, 8, 5]
Any fairness metric	15 (19.2%, 95% CI: 12.0–29.3%)	—
Subgroup AUC gap / Worst-group	3 (3.8%) / 6 (7.7%)	[5, 25]
ECE / Preserved-site CV	2 (2.6%) / 4 (5.1%)	[53, 8]
Public dataset demographic metadata (audit, $n = 39$)		
Race / ethnicity / sex / age	28.2 / 25.6 / 23.1%	[8, 5]
Ancestry, skin tone, intersectional & other attributes	0% (all 39)	[18, 12]

Table 7: Cancer-type coverage across the 78 included studies. Counts are non-exclusive: studies addressing multiple cancer types are counted under each relevant category. "Primary" denotes the cancer type listed first or identified as the principal focus of the study.

Cancer Type	Primary	Any mention	Representative studies
Breast	44	54	[5, 8, 6]
Multi / Pan-cancer	24	30	[7, 8, 17]
Prostate	15	21	[5, 7]
Skin / Melanoma	10	13	[16, 5]
Lung	9	28	[25, 9, 8]
Colorectal	8	21	[5, 8]
Ovarian	3	4	[85]
Brain / Glioma	2	10	[8, 17]
Kidney / Renal	1	11	[8]
Liver	1	7	[8]
Gastric	—	5	[25]
Thyroid	—	5	[7]
Bladder	—	4	—
Head & Neck	—	3	[8]
Cervical	1	2	—
Pancreatic	—	1	—
Paediatric / rare	0	0	—

B Eight-Category Bias Taxonomy: Full Definitions and Exemplar Findings

Table 8 provides the complete working definitions and exemplar empirical findings for the eight bias categories identified across the 78 included studies (Figure 2 in the main text). Study counts are non-exclusive: individual studies typically report multiple co-occurring bias sources.

Table 8: Eight-category bias taxonomy derived from the 78 included histopathology fairness studies. Study counts are non-exclusive: individual studies typically report multiple co-occurring bias sources (see Figure 2).

Category (n)	**Working definition**	**Exemplar finding**
Demographic (5)	Performance varies by race, ethnicity, ancestry, sex, or age.	White–Black AUROC gap 3–16% across 20 cancer types; race predictable from skin histology (AUROC 0.70) [5, 16, 6, 25].
Institutional / batch (7)	Site, scanner, or staining protocol signatures encoded by the model.	Site classification AUROC $>$0.96 on TCGA; foundation-model features probe TSS at $>$90% [8, 7, 9, 14].
Selection / sampling (8)	Non-representative cohorts (single-centre, geographic skew, convenience sampling).	$\sim$33/78 studies anchored on TCGA; $>$70% White cohorts yield lower non-White accuracy [43, 11, 22, 17].
Technical / domain-shift (8)	Scanner, stain, fixation, or preprocessing variation causing distribution shift.	Within-batch AUROC 0.74–0.96 collapses to 0.52–0.61 cross-batch; stain normalisation fails [29, 9, 54, 53].
Class / label (7)	Class imbalance, label noise, inter-observer variability, annotation shortcuts.	Long-tail TCGA sub-types; grayscale conversion drops centre accuracy by 15% [17, 31, 14, 38].
Algorithmic (5)	Bias introduced by architecture, MIL attention, objective, or model-selection choices.	Standard MIL attention dilutes salient instances; model-selection rule alters worst-group AUROC [30, 66, 23, 58].
Confounding (6)	Non-causal site–demographic, disease–demographic, or co-slide associations exploited as shortcuts. Disease prevalence differences across groups create spurious correlations.	Ancestry AUROC 0.798$\rightarrow$0.507 under preserved-site CV; MI(label, centre) $= 1.47$ (Kheiri et al.); disease–demographic confounding (Chen et al.) [8, 16, 14].
Representation (7)	Under-coverage of populations, skin tones, cancer sub-types, or geographies in training data.	$>$70% White cohorts yield lower non-White accuracy; rare sub-types undertrained [43, 5, 37, 17].

C Asymmetric Fairness-Metric Reporting Across the Corpus

Figure 4 summarises the reporting rates for standard performance metrics and fairness-specific metrics across the 78 included studies, illustrating the marked asymmetry between the two classes.

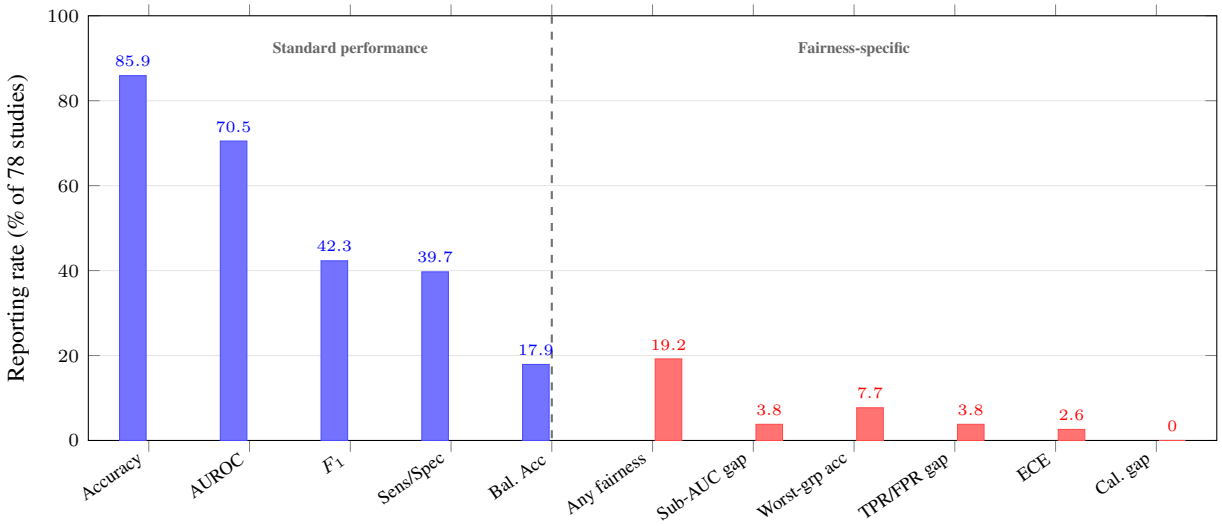


Figure 4: **Asymmetric reporting of performance versus fairness metrics across the 78 included histopathology studies.** Reporting rates are from systematic audit of the full analytical corpus. Aggregate performance metrics (accuracy 85.9%, AUROC 70.5%, F_1 42.3%) are near-universally reported, whereas any fairness-specific metric appears in only 19.2% of studies ($n = 15$). Worst-group accuracy/AUC — the minimum fairness guarantee demanded by recent regulatory guidance — is reported in only 7.7% ($n = 6$ studies); Expected Calibration Error (ECE) in 2.6% ($n = 2$); and the *subgroup calibration gap* in none. The dashed vertical rule separates standard performance metrics (left, blue) from fairness metrics (right, red). This reporting asymmetry implies that the empirical base for evaluating debiasing is drawn from a small, selection-biased subset and is the principal reason that evidence grades in Figure 3 and Table 2 rest on a limited number of studies.