Mod Pathol 38 (2025) 100715

MODERN PATHOLOGY

USCAP UNITED STATES AND CANADIAN ACADEMY OF PATHOLOGY
Creating a Better Pathologist

Journal homepage: https://modernpathology.org/

Research Article

Physical Color Calibration of Digital Pathology Scanners for Robust Artificial Intelligence–Assisted Cancer Diagnosis

Xiaoyi Ji[a], Richard Salmon[b], Nita Mulliqi[a], Umair Khan[c], Yinxi Wang[a], Anders Blilie[d,e], Henrik Olsson[a], Bodil Ginnerup Pedersen[f,g], Karina Dalsgaard Sørensen[g,h], Benedicte Parm Ulhøi[i], Svein R. Kjosavik[j,k], Emilius A.M. Janssen[d,l,m], Mattias Rantalainen[a], Lars Egevad[n], Pekka Ruusuvuori[c,o,p], Martin Eklund[a], Kimmo Kartasalo[q,*]

[a] *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden;* [b] *PathQA Ltd, United Kingdom;* [c] *Institute of Biomedicine, University of Turku, Turku, Finland;* [d] *Department of Pathology, Stavanger University Hospital, Stavanger, Norway;* [e] *Faculty of Health Sciences, University of Stavanger, Stavanger, Norway;* [f] *Department of Radiology, Aarhus University Hospital, Aarhus, Denmark;* [g] *Department of Clinical Medicine, Aarhus University, Aarhus, Denmark;* [h] *Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark;* [i] *Department of Pathology, Aarhus University Hospital, Aarhus, Denmark;* [j] *The General Practice and Care Coordination Research Group, Stavanger University Hospital, Stavanger, Norway;* [k] *Department of Global Public Health and Primary Care, Faculty of Medicine, University of Bergen, Bergen, Norway;* [l] *Department of Chemistry, Bioscience and Environmental Engineering, University of Stavanger, Stavanger, Norway;* [m] *Institute for Biomedicine and Glycomics, Griffith University, Queensland, Australia;* [n] *Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden;* [o] *InFLAMES Research Flagship, University of Turku, Turku, Finland;* [p] *Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland;* [q] *Department of Medical Epidemiology and Biostatistics, SciLifeLab, Karolinska Institutet, Stockholm, Sweden*

ARTICLE INFO

Article history:
Received 15 August 2024
Revised 19 December 2024
Accepted 8 January 2025
Available online 16 January 2025

Keywords:
artificial intelligence
color calibration
computational pathology
foundation model
prostate cancer
whole slide scanning

ABSTRACT

The potential of artificial intelligence (AI) in digital pathology is limited by technical inconsistencies in the production of whole slide images (WSIs). This causes degraded AI performance and poses a challenge for widespread clinical application, as fine-tuning algorithms for each site is impractical. Changes in the imaging workflow can also compromise diagnostic accuracy and patient safety. Physical color calibration of scanners, relying on a biomaterial-based calibrant slide and a spectrophotometric reference measurement, has been proposed for standardizing WSI appearance, but its impact on AI performance has not been investigated. We evaluated whether physical color calibration can enable robust AI performance. We trained fully supervised and foundation model–based AI systems for detecting and Gleason grading prostate cancer using WSIs of prostate biopsies from the STHLM3 clinical trial (n = 3651) and evaluated their performance in 3 external cohorts (n = 1161) with and without calibration. With physical color calibration, the fully supervised system's concordance with pathologists' grading (Cohen linearly weighted κ) improved from 0.439 to 0.619 in the Stavanger University Hospital cohort (n = 860), from 0.354 to 0.738 in the Karolinska University Hospital cohort (n = 229), and from 0.423 to 0.452 in the Aarhus University Hospital cohort (n = 72). The foundation model's concordance improved as follows: from 0.739 to 0.760 (Karolinska), from 0.424 to 0.459 (Aarhus), and from 0.547 to 0.670 (Stavanger). This study demonstrated that physical color calibration provides a potential solution to the variation introduced by different scanners, making AI-based cancer diagnostics more reliable and applicable in diverse clinical settings.

Introduction

Application of artificial intelligence (AI) to digital pathology shows promise for tasks such as cancer detection and grading,[1]

* Corresponding author.
E-mail address: kimmo.kartasalo@ki.se (K. Kartasalo).

https://doi.org/10.1016/j.modpat.2025.100715

prognostication,[2] and prediction of molecular biomarkers.[3] However, inconsistencies in whole slide image (WSI) acquisition and postprocessing due to scanner instruments[4-6] can degrade AI performance.[7,8] Limited generalizability of AI models across scanners and sites is a considerable hurdle to their widespread clinical adoption as model tweaking and validation at each new deployment site is not feasible. Site-specific fine-tuning is also problematic in terms of medical device regulations and clinical validation of AI algorithms.

Even after an AI model has been fine-tuned to a given laboratory, unexpected changes can arise in the imaging workflow, such as damage or wear of scanner components or software updates. This can lead to incorrect diagnoses and poses a threat to patient safety. Statistical methods like conformal prediction[9] have been proposed for detecting changes in the data generation process, indicating that the AI outputs may no longer be reliable. Although detecting such issues is important, avoiding them altogether through robust AI algorithms and normalization or standardization techniques would be preferred. With normalization, we refer to reducing variation across WSIs in relation to an arbitrary reference input data set, whereas standardization produces image characteristics conforming to a universal reference standard.

The brute-force approach for improved AI generalization is to use training data from an increased number of sources. However, it will be difficult to account for all possible sources of variation that new scanner instruments may produce in the future. A widely applied approach is data augmentation, that is, perturbation of the colors, contrast, or other image characteristics randomly during training.[10] It is, however, challenging to design universally applicable augmentations that cover all real-world variance while not impacting the correlations between image features and output labels.[11] Data augmentation designed to improve performance on a given data set may not generalize to other data sources, where different types or magnitudes of variation may be present.

Computational stain normalization is an alternative for minimizing discrepancies in image characteristics. Classical methods often rely on estimating the hematoxylin and eosin stain components from a WSI and transforming them to match a reference image.[12] These types of methods tend to produce mixed results.[10] More recent learning-based methods use cycle-consistent generative adversarial networks (cycle-GANs), which automate image-to-image style translation without paired data[13] and have proven well suited for stain normalization.[14] Challenges include artifact generation, need for model retraining for new data sources and changing conditions, and sensitivity to the choice of reference WSIs. In addition to requiring large data sets, which is problematic for small research groups or local clinics, these methods do not provide simple, quantitative readouts for quality assurance (QA).

As an alternative, a physical calibrant has been proposed for standardizing WSI appearance.[15] According to the FDA, a calibration slide should contain a set of measurable and representative color patches, which possess spectral characteristics similar to stained tissue.[16] By scanning and analyzing the calibration slide, it is possible to estimate the International Color Consortium (ICC) profile of a scanner[17] (https://www.color.org/specification/ICC.1-2022-05.pdf), and WSIs produced by this scanner can then be calibrated to conform to a target ICC profile. As an alternative, scanner manufacturers also have the opportunity to embed ICC profiles into WSIs, but currently, only few scanners do this, and their color profiling processes and accuracy vary. Calibration has been shown to improve the visual concordance of WSIs with microscopy and the diagnostic confidence of pathologists.[4] Although standardization of WSIs to a specific physical standard as such is not currently crucial for AI-based analysis, application of color calibration also provides a normalization effect, removing problematic variation in WSI characteristics via a traceable and explainable process.

We hypothesized that physical color calibration of scanners can provide a consistent and lightweight means of standardizing WSIs, resulting in improved AI generalization across different clinical sites. The computational pathology community has focused considerable efforts on the generalization problem but physical calibration is, to the best of our knowledge, unexplored as a potential solution. To test our hypothesis, we applied a commercial color calibration slide to standardize WSIs of prostate biopsies collected at 4 different sites each using a different scanner model and assessed the effect of calibration on the diagnostic performance of a fully supervised AI system we have developed earlier[1] for the intensively studied task of detecting and Gleason grading prostate cancer.[7,18-21] In addition, to benchmark the performance of the calibration slide against computational normalization methods, we applied the Macenko and cycle-GAN algorithms to the data sets and assessed their effect on AI performance.

Foundation models have recently been introduced for computational pathology.[22-24] These large-scale models, typically pretrained in a task-agnostic manner using vast amounts of unlabeled heterogeneous pathological data, are hypothesized to have the capacity to generalize well across various sites and scanners and to be adaptable to a wide range of downstream tasks with minimal fine-tuning.[24,25] To investigate this hypothesis and the potential role of foundation models in solving the generalization problem in computational pathology, we additionally implemented another AI system relying on the recently released visual transformer–based foundation model UNI[24] and analyzed the effect of physical color calibration on its cross-site generalization performance.

Materials and Methods

Sample Collection

For AI training, we used prostate core needle biopsies collected in the STHLM3 clinical trial (ISRCTN84445406)[26] of 2012-2014 in Stockholm, Sweden. In total, 3651 biopsy cores from 957 participants were digitized with 20× magnification using an Aperio AT2 scanner (Leica Biosystems) at SciLifeLab. For AI tuning and evaluation, we collected data from 3 external sites. From Karolinska University Hospital, Stockholm, Sweden, we obtained 329 biopsy cores (73 patients) scanned by a Hamamatsu NanoZoomer S360 C13220 scanner (Hamamatsu Photonics) with 20× magnification. From Aarhus University Hospital, Aarhus, Denmark, we obtained 102 biopsy cores (42 patients) scanned by a Hamamatsu NanoZoomer 2.0-HT C9600-12 scanner with 20× magnification. From Stavanger University Hospital, Stavanger, Norway, 1228 biopsy cores (200 patients) were scanned with a Hamamatsu NanoZoomer S60 C13210 with 40× magnification. All cohorts were collected as part of clinical trials or using an otherwise controlled sampling scheme to ensure clinically representative patient characteristics (for details, see Supplementary Methods).

All STHLM3 and Karolinska University Hospital cores were graded following the International Society of Urological Pathology (ISUP) grading classification by an experienced urologic pathologist (L.E.), who also delineated the cancerous areas with a marker pen and measured the linear cancer extent. Slides from Aarhus

and Stavanger cohorts were similarly graded by pathologists from the corresponding hospitals. All core needle biopsy slides in this study were routinely stained with hematoxylin and eosin in the respective laboratory at each site.

For training the cycle-GAN normalization model, a tuning set specific to each site is required. From Karolinska University Hospital, 100 (30.4%) slides were assigned for tuning the hyperparameters of the AI model and for training the GAN. From Aarhus University Hospital and Stavanger University Hospital, 30 (29.4%) and 368 (30.0%) slides, respectively, were assigned to a tuning set for training the GAN. The remaining slides from each site were used as test sets. Slides were assigned to tuning and test sets randomly, stratified by ISUP grade.

The study is conducted in agreement with the tenets of the Helsinki Declaration. The collection of patient samples was approved by the Stockholm regional ethics committee (permits 2012/572-31/1, 2012/438-31/3, and 2018/845-32), the Swedish Ethical Review Authority (permit 2019-05220), and the Regional Committee for Medical and Health Research Ethics (REC) in Western Norway (permits REC/Vest 80924, REK 2017/71). Informed consent was provided by the participants in the Swedish data set. For the other data sets, informed consent was waived by the institutional review board due to the usage of deidentified prostate specimens in a retrospective setting. Table 1 presents the detailed composition of training, tuning, and test data.

Artificial Intelligence System

We based the fully supervised AI system on a previously published model,[1] where the tissue in each WSI is divided into patches (~540 × 540 μm). The individual patches are classified into benign or malignant and further into Gleason pattern 3, 4, or 5, using InceptionV3[27] deep neural networks. For aggregating patch-level predictions into slide-level predictions of cancer presence, cancer length, and ISUP grade, we used gradient boosted trees.

In addition, we implemented a weakly supervised AI model relying on the transformer–based foundation model UNI.[24] To ensure maximal image quality and generalization performance, we adjusted our patch size to match the UNI pretraining parameters (224 × 224 μm). Briefly, the features of every patch are extracted by a ViT-Large architecture,[28] which was pretrained using the DINOv2 self-supervised learning framework,[29] and patch-level features are aggregated into a WSI-level feature vector using a gated attention mechanism.[30] Finally, a linear classifier uses these aggregated features to produce slide-level predictions of the ISUP grade in the form of a continuous regression output (rounded to the nearest integer to provide the final ISUP grade, with 0 indicating a benign sample).

Both models were trained using approximately 2.9 million patches from the STHLM3 data set, scanned on the Aperio AT2 scanner. For comparing color normalization or calibration methods, the models were retrained using patches processed with each method as input (see Color Calibration and Normalization methods). Hyperparameters and decision thresholds for cancer detection were optimized based on the Karolinska University Hospital tuning set to maximize area under the curve for cancer detection, Cohen κ for ISUP grading, and linear correlation for cancer length estimation (see Statistical analysis for details), and applied to all test sets without modifications. This cohort was chosen for tuning to ensure consistency of the reference standard (grading by L.E.) between training and tuning data, which allowed hyperparameter selection to optimize the technical generalizability of the models to data from a different laboratory and scanner, without simultaneously introducing a change of reference standard. This provided a competitive baseline in the absence of color management and ensured a fair comparison between different color management methods.

For the fully supervised model, deep neural networks were implemented in Python (3.6.9) using TensorFlow (2.3.0)[31] and boosted trees using XGBoost (1.2.1).[32] UNI was implemented in

Table 1
Baseline characteristics of included prostate biopsy slides

Variable	Digitized biopsy slides						
	STHLM3 (number of patients = 957)	Karolinska University Hospital (number of patients = 73)		Aarhus University Hospital (number of patients = 42)		Stavanger University Hospital (number of patients = 200)	
	Training	Tuning	Test	Tuning	Test	Tuning	Test
No. of slides	3651	100	229	30	72	368	860
Scanner	Aperio AT2	Hamamatsu NanoZoomer S360 C13220		Hamamatsu NanoZoomer 2.0-HT C9600-12		Hamamatsu NanoZoomer S60 C13210	
Cancer length, number of slides (%)							
No cancer	739 (20.2)	33 (33.0)	75 (32.8)	13 (43.3)	30 (41.7)	261 (70.9)	609 (70.8)
>0-1 mm	752 (20.6)	8 (8.0)	24 (10.5)	3 (10.0)	2 (2.8)	24 (6.5)	55 (6.4)
>1-5 mm	1105 (30.3)	25 (25.0)	52 (22.7)	2 (6.7)	16 (22.2)	32 (8.7)	86 (10.0)
>5-10 mm	691 (18.9)	25 (25.0)	50 (21.8)	7 (23.3)	17 (23.6)	16 (4.3)	46 (5.4)
>10 mm	364 (10.0)	9 (9.0)	28 (12.2)	5 (16.7)	7 (9.7)	35 (9.5)	64 (7.4)
Cancer grade, number of slides (%)							
Benign	739 (20.2)	33 (33.0)	75 (32.8)	13 (43.3)	30 (41.7)	261 (70.9)	609 (70.8)
ISUP 1 (3+3)	1156 (31.6)	19 (19.0)	45 (19.6)	8 (26.7)	18 (25.0)	62 (16.8)	145 (16.9)
ISUP 2 (3+4)	459 (12.6)	19 (19.0)	44 (19.2)	7 (23.3)	18 (25.0)	18 (4.9)	43 (5.0)
ISUP 3 (4+3)	309 (8.5)	15 (15.0)	34 (14.8)	0	1 (1.4)	13 (3.5)	32 (3.7)
ISUP 4 (4+4, 3+5, and 5+3)	576 (15.8)	6 (6.0)	18 (7.9)	2 (6.7)	5 (6.9)	7 (1.9)	15 (1.7)
ISUP 5 (4+5, 5+4, and 5+5)	412 (11.3)	8 (8.0)	13 (5.7)	0	0	7 (1.9)	16 (1.9)

ISUP, International Society of Urological Pathology.

Python (3.8.10) using PyTorch (2.1.1) and the required dependencies,[24] and the pretrained weights were obtained from https://huggingface.co/MahmoodLab/UNI. Computing was performed on GPU clusters Alvis, part of the National Academic Infrastructure for Supercomputing in Sweden, and Berzelius, at the National Supercomputer Centre. For details, see Appendix.

Color Calibration and Normalization

For physical color calibration, we used Sierra (PathQA Ltd).[33] The spectrophotometrically characterized slide contains patches of biopolymer treated with varying combinations of stain intensities representative of typical histopathology stains (Fig. 1 and Supplementary Fig. S1) and is made to exactly replicate the FDA guidelines for WSI precision and accuracy testing.[16,34] The manufacturer has registered as a creator of ICC profiles that meet the ICC standard (ICC creator tag PAQA), and slides are measured before distribution using a calibrated, highly accurate spectrophotometer traceable to UK National Physical Laboratory Standards. The ΔE calculations are conducted according to ISO 11664-6:2022. The ICC profiles estimated based on Sierra are DICOM and ICCv4 compliant and adhere to ISO 15076-1:2010. We scanned Sierra on each of the 4 scanners using settings identical to the biopsy specimens, and ICC profiles for each scanner were generated at PathQA. We also evaluated applying the embedded ICC profile by the scanner vendor available for the Aperio AT2 scanner. To calibrate images produced by a given scanner, we mapped the image patches from the input space defined by the scanner's ICC profile to a standard RGB color space with the *ImageCms* module (1.0.0 pil) from the *Pillow* library (8.0.0) in Python.

As a baseline computational color normalization approach, we used the method of Macenko et al,[12] modified to estimate slide-level instead of patch-level stain vectors. Initially, luminosity was slide-level corrected based on 100 random patches from the WSI under study. A reference stain vector, representing the normalized stain target, was then estimated from 2000 luminosity-corrected patches randomly sampled from all training WSIs. After establishing the reference stain vector, slide-level stain vectors were estimated from 100 randomly sampled tiles per WSI, and all WSIs were normalized to the reference target. The code was adapted from the StainTools (https://github.com/Peter554/StainTools) and Staining Unmixing and Normalization in Python packages (https://github.com/schaugf/HEnorm_python).

As a state-of-the-art computational color normalization method, we implemented the unsupervised GAN-based method, Cycle-GAN.[13] A representative target WSI was selected from the STHLM3 data set (training) by L.E., and the rest of the WSIs were used as source samples. A separate GAN model was trained for each of the 4 sites based on a site-specific tuning set (see Sample Collection) and applied on the remaining slides from each site to transform them to match the appearance of the target WSI. For details, see Appendix.

Statistical Analysis

We evaluated the concordance of AI with pathologists in ISUP grading using linearly weighted Cohen κ, in cancer detection using sensitivity and specificity, and in cancer length estimation using the linear correlation coefficient. For the fully supervised models, which produce classwise output probabilities, the effects of physical color calibration and computational normalization methods on the discrimination and calibration properties of the models were additionally studied with receiver operating characteristics (ROC) analysis and calibration curves. The calibration curves indicate the true frequency of the positive label against its predicted probability for binned predictions. All evaluations were performed on slide-level predictions. All CIs were 2-sided with 95% confidence level and calculated from 1000 bootstrap samples. Statistically significant differences were assessed using the McNemar test for sensitivity and specificity and the *z*-test for ISUP κ and cancer length correlation.

Results

We first assessed the ability of the fully supervised AI system to detect and Gleason grade prostate cancer, first without any color management and then with each of the calibration or normalization techniques applied to both the training and test data. All test data were external, representing samples prepared in a different laboratory and scanned on a different scanner than the training data. In the Karolinska University Hospital set, Cohen linearly weighted κ for Gleason grading improved from 0.354 (95% CI, 0.288-0.417) to 0.738 (95% CI, 0.692-0.784) using Sierra color calibration. For cancer detection without calibration, we observed sensitivity and specificity of 0.955 (95% CI, 0.917-0.987) and 0.987 (95% CI, 0.955-1.000) and, with Sierra calibration applied, values of 0.987 (95% CI, 0.966-1.000) and 0.813 (95% CI, 0.722-0.897), respectively. The correlation coefficients between cancer lengths reported by the AI system and the pathologists exhibited a modest enhancement from 0.824 (95% CI, 0.781-0.863) to 0.834 (95% CI, 0.790-0.876) when Sierra calibration was applied. On the Aarhus data set without calibration vs with calibration, the Cohen κ coefficient was 0.423 (95% CI, 0.319-0.541) vs 0.452 (95% CI, 0.346-0.557), the sensitivity was in both cases 0.952 (95% CI, 0.875-1.000), the specificity was 0.967 (95% CI, 0.889-1.000) vs 0.933 (95% CI, 0.826-1.000), and the linear correlation was 0.811 (95% CI, 0.706-0.895) vs 0.820 (95% CI, 0.691-0.914). In the case of the Stavanger data set without calibration vs with calibration, we observed Cohen κ coefficients of 0.439 (95% CI, 0.383-0.491) vs 0.619 (95% CI, 0.578-0.657), sensitivity of 0.518 (95% CI, 0.456-0.580) vs 0.904 (95% CI, 0.867-0.938), specificity of 0.997 (95% CI, 0.992-1.000) vs 0.959 (95% CI, 0.941-0.974), and linear correlation coefficients of 0.817 (95% CI, 0.758-0.863) vs 0.844 (95% CI, 0.788-0.892). For context, typical interpathologist κ values are in the approximate range of 0.5 to 0.8.[35]

To compare Sierra with computational color normalization, we conducted otherwise identical experiments using Macenko and cycle-GAN methods for processing the training, tuning, and test data. Compared with the baseline of no normalization, the Macenko method provided a clear improvement on the Karolinska University Hospital data with κ for Gleason grading of 0.650 (95% CI, 0.588-0.710), sensitivity and specificity for cancer detection of 0.981 (95% CI, 0.955-1.000) and 0.880 (95% CI, 0.803-0.949), respectively, and linear correlation for cancer length of 0.845 (95% CI, 0.806-0.883). However, it led to slight deterioration of performance on Aarhus data with κ of 0.408 (95% CI, 0.295-0.513), sensitivity and specificity of 0.952 (95% CI, 0.881-1.00) and 0.867 (95% CI, 0.731-0.971), respectively, and linear correlation of 0.749 (95% CI, 0.557-0.901), and considerably degraded performance on Stavanger data with κ of 0.281 (95% CI, 0.221-0.340), sensitivity and specificity of 0.378 (95% CI, 0.316-0.443) and 0.954 (95% CI, 0.936-0.971), respectively, and linear correlation of 0.578 (95% CI, 0.502-0.645). Cycle-GAN considerably improved performance on

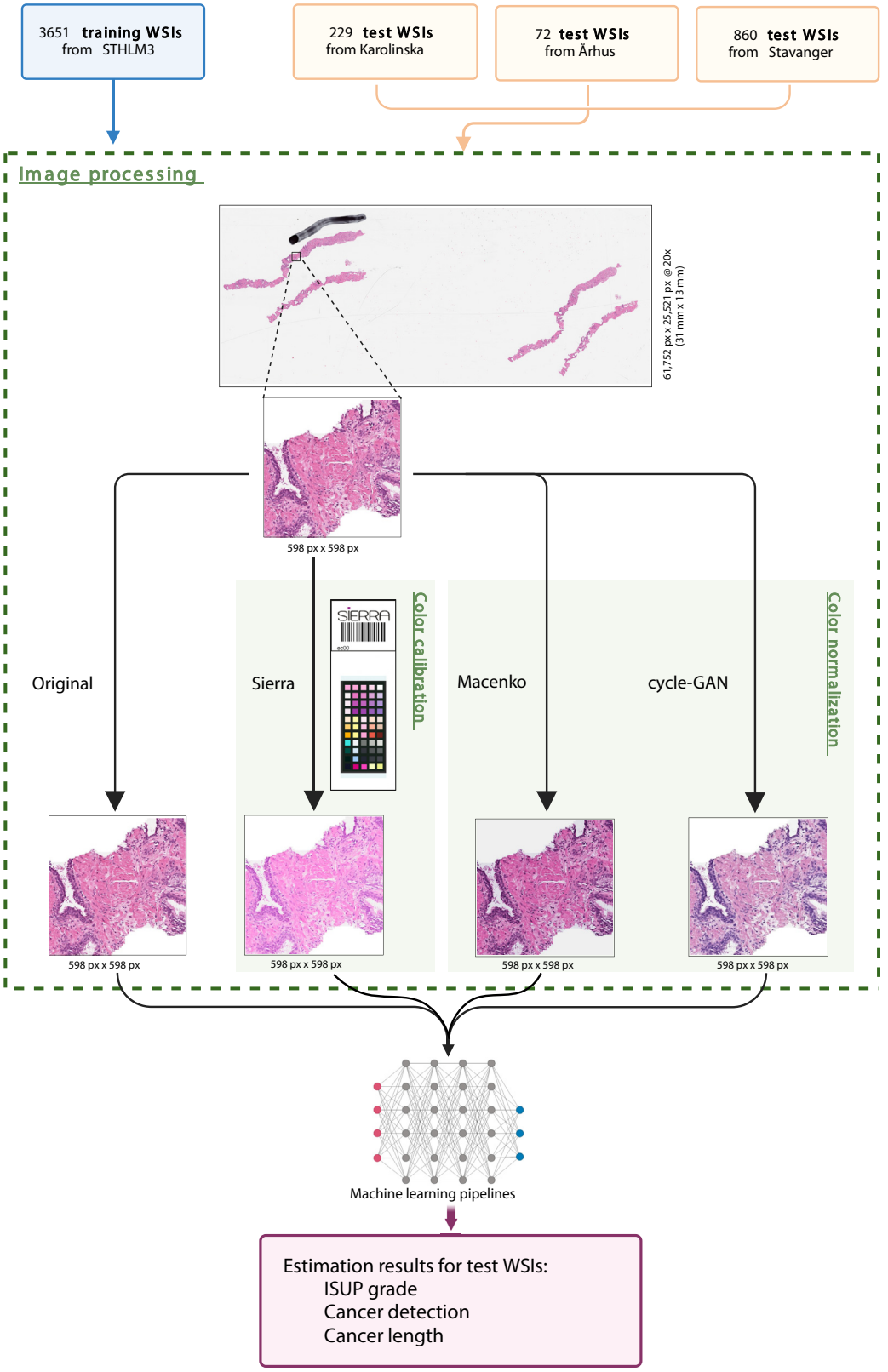

Figure 1.
Overview of the study pipeline. The tissue region in the input whole slide image (WSI) is first split into patches. Here, one patch is taken as an example for demonstrating the result of color calibration and normalization. Models were trained independently on the original data, Sierra-calibrated data, and data normalized with the Macenko or cycle-GAN algorithms. The details of the AI systems used ("Machine learning pipelines") are presented in Supplementary Figure S1. This figure was created using BioRender (https://BioRender.com/v05u633). GAN, generative adversarial networks.

the Karolinska University Hospital data set with κ of 0.655 (95% CI, 0.597-0.711), sensitivity and specificity of 0.961 (95% CI, 0.929-0.987) and 0.96 (95% CI, 0.909-1.000), respectively, and linear correlation of 0.889 (95% CI, 0.856-0.919). Its performance on the Aarhus data set was mediocre, with κ of 0.368 (95% CI, 0.249-0.479), sensitivity of 0.952 (95% CI, 0.880-1.000), specificity of 0.833 (0.692-0.960), and linear correlation of 0.720 (0.505-0.889). For the Stavanger slides, cycle-GAN performed comparably with the Sierra color calibration, with κ of 0.623 (95% CI, 0.578-0.663), sensitivity and specificity of 0.968 (95% CI, 0.944-0.988) and 0.882 (95% CI, 0.856-0.906), respectively, and linear correlation of 0.862 (95% CI, 0.812-0.907).

We performed ROC and calibration curve analyses to investigate whether the improvements in model performance achieved with Sierra calibration are due to improved discrimination capacity between benign vs malignant samples or improved model calibration (Fig. 2). We observed minimal differences between models relying on noncalibrated vs calibrated data in terms of discriminative capacity measured with area under the ROC curve: 0.971 (95% CI, 0.942-0.992) vs 0.989 (95% CI, 0.973-0.999) for Karolinska, 0.964 (95% CI, 0.898-1.000) vs 0.963 (95% CI, 0.913-1.000) for Aarhus, and 0.982 (95% CI, 0.972-0.990) vs 0.976 (95% CI, 0.963-0.986) for Stavanger. In contrast, the calibration curves showed marked differences indicative of improved model calibration.

Finally, to assess whether physical color calibration could even be beneficial in the context of foundation models, we evaluated the performance of the AI system based on the UNI model using the same training and testing data set structure. The κ for ISUP grading was improved with Sierra calibration from 0.739 (95% CI, 0.689-0.786) to 0.760 (95% CI, 0.710-0.804) for Karolinska University Hospital data, from 0.424 (95% CI, 0.302-0.546) to 0.459 (95% CI, 0.330-0.587) for Aarhus data, and from 0.547 (95% CI, 0.504-0.582) to 0.670 (95% CI, 0.631-0.711) for Stavanger data. The sensitivity and specificity values without color calibration were 0.994 (95% CI, 0.980-1.000) and 0.813 (95% CI, 0.718-0.900) for Karolinska; 0.952 (95% CI, 0.881-1.000) and 0.533 (95% CI, 0.360-0.720) for Aarhus; and 0.968 (95% CI, 0.948-0.988) and 0.669 (95% CI, 0.632-0.707) for Stavanger, respectively. With Sierra color calibration, sensitivity and specificity values were 0.987 (95% CI, 0.966-1.000) and 0.787 (95% CI, 0.691-0.877) for Karolinska; 0.952 (95% CI, 0.881-1.000) and 0.667 (95% CI, 0.500-0.846) for Aarhus; and 0.912 (95% CI, 0.875-0.945) and 0.847 (95% CI, 0.819-0.876) for Stavanger, respectively.

The performance metrics of the fully supervised and UNI-based models are summarized in Tables 2 and 3, respectively. Corresponding confusion matrices for the 2 models are presented for cancer detection in Supplementary Tables S1 and S2 and for ISUP grading in Supplementary Figures S2 and S3. Scatter plots of predicted and pathologist-reported cancer lengths are presented in Supplementary Figure S4. As an alternative baseline to uncorrected WSIs, we evaluated using the scanner's embedded color profile for calibration (available only for the Aperio AT2), but the resulting performance was inferior to using uncorrected data (Supplementary Tables S3 and S4).

Discussion

Our experiments demonstrated that a physical calibrant slide serves as a robust color calibration method for AI-assisted computational pathology. Physical calibration exhibited a consistently positive or neutral effect across different data sources on AI performance in detecting prostate cancer, Gleason grading, and estimating cancer extent in biopsies. Although the effect on discriminatory capacity was rather modest, the improvement in model calibration was striking (Fig. 2). This was reflected in considerably improved Gleason grading performance measured at a classifier operating point that was specified on the tuning set and then applied to the other cohorts without further adjustments (Table 2). For example, the more than 100% improvement in Cohen κ on the Karolinska University Hospital cohort represents a difference between an AI model that could be considered a risk to patient safety and one that would perform comparably with pathologists. Similarly, in the Stavanger University Hospital cohort, without any calibration the sensitivity of the AI model dropped to 51.8%. When calibration was applied, the sensitivity remained at 90.4%, much closer to the intended tuning set value. Ultimately, a discrete classification decision, for example, a cancer diagnosis or grade is typically needed and although often neglected in AI studies, model calibration is of crucial importance for practical applications.[36] In the Aarhus cohort, physical calibration had a neutral effect on AI performance. We hypothesized that this may be due to a lower degree of initial miscalibration between the training data and the Aarhus data or due to the minimal proportion of high-grade cases in this cohort. High-grade cases (ISUP 3-5) tend to be the most challenging in terms of AI generalization because of their lower prevalence in training data compared with benign and ISUP 1-2 cases, and cohorts enriched for low-grade tumors could thus be expected to benefit less from calibration approaches.

Computational normalization techniques have the advantage of correcting for variations arising from other sources than the scanner and are therefore likely to outperform physical calibration on some data cohorts in terms of absolute performance but require considerable training or reference data sets. This is exemplified by the slight advantage of cycle-GAN normalization compared with Sierra calibration on the largest cohort of the study, Stavanger University Hospital. In contrast, for the Aarhus data set, relying on only 30 slides to train the cycle-GAN, relatively weak performance was observed, particularly in terms of cancer grading. Moreover, complex models like cycle-GAN risk generating image artifacts commonly referred to as hallucinations,[14,37] where the tissue morphology is unintentionally altered. Although physical calibration was not the top-performing method in all data sets when compared with computational color normalization, the stable performance of the method is a crucial advantage and contrasts to the unpredictable behavior of the computational approaches. Physical calibration does not require training or reference samples apart from the calibrant itself, which may explain the consistent performance observed also for the smaller cohorts in this study. Indeed, this may provide a route to scaledown AI to sites producing less data such as research groups and local clinics, which in reality represent more numerous use cases than big data from regional hospitals.

Besides the issues with scaling to new sites, computational normalization methods are problematic in terms of medical device regulations with respect to model validation and transparency of the processes being used. To account for changes over time in the sample preparation and scanning, approaches like the cycle-GAN would require retraining and revalidation of the model periodically at each site, which would be highly challenging in practice. Physical calibration provides independent quantification against a universal standard for real-time QA or good laboratory practice, using an established methodology that can be integrated anywhere in the postscanning workflow, including into the scanners themselves. The few scanner manufacturers that already

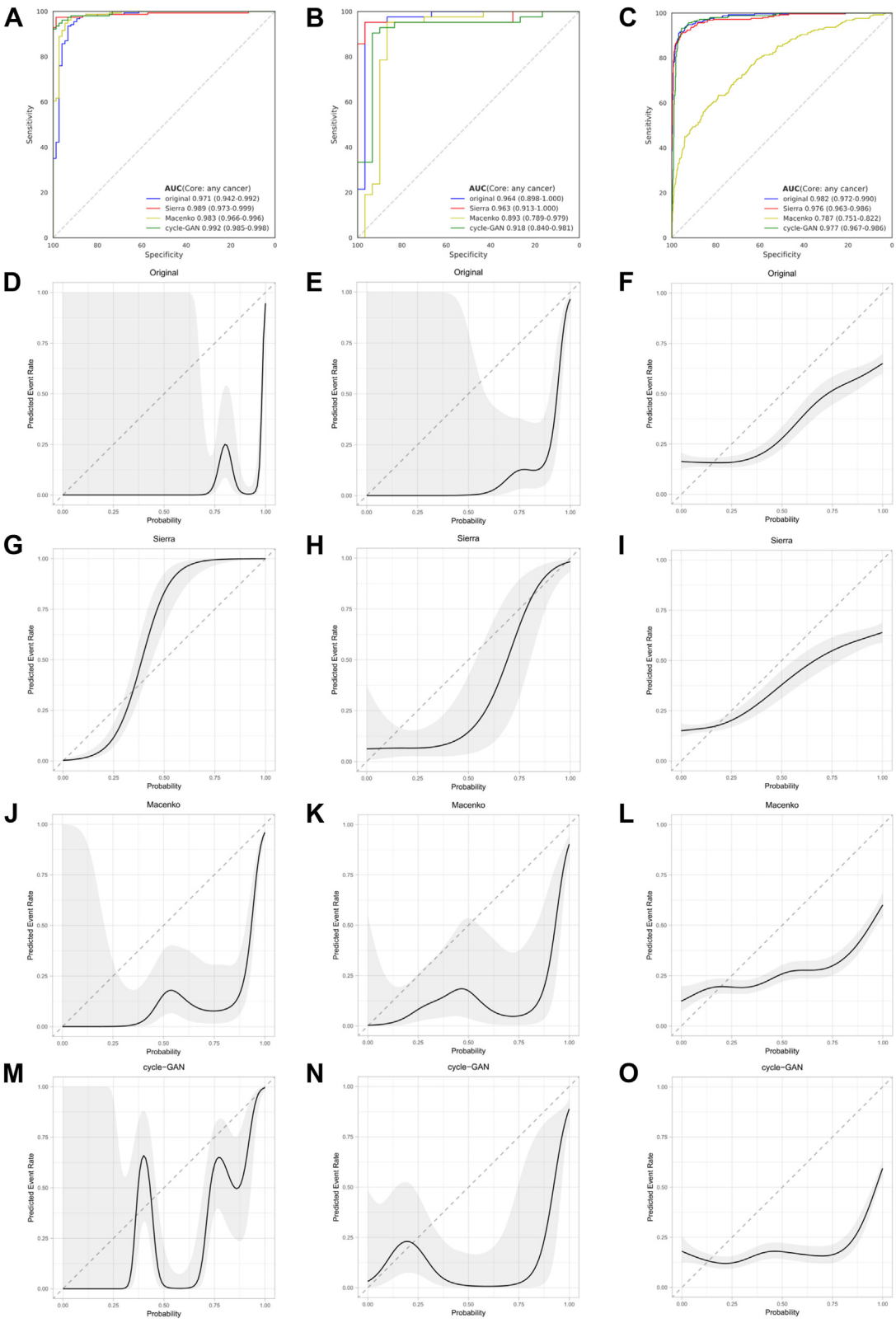

Figure 2.
ROC curves with AUC (A-C) and calibration curves (D-O) for cancer detection in individual cores, with original, Sierra-calibrated, Macenko-normalized or cycle-GAN normalized WSIs. Columns from left to right: data from Karolinska University Hospital, Aarhus University Hospital, and Stavanger University Hospital. Dashed gray lines in ROC curves represent the baseline curve corresponding to random guessing, whereas the ones in calibration curves represent an ideally calibrated model. The 90% CIs for calibration curves are visualized using the gray ribbon. AUC, area under the curve; GAN, generative adversarial networks; ROC, receiver operating characteristic; WSIs, whole slide images.

Table 2
Evaluation metrics for slide-level cancer diagnosis performance using the fully supervised models

Cohort & metric	Original	Sierra	Macenko	cycle-GAN
Karolinska University Hospital				
Sensitivity	0.987 (0.966-1.000)	0.955 (0.917-0.987)	0.981 (0.955-1.000)	0.961 (0.929-0.987)
Specificity	0.813 (0.722-0.897)	0.987 (0.955-1.000)↑	0.880 (0.803-0.949)	0.960 (0.909-1.000)↑
ISUP κ	0.354 (0.288-0.417)	0.738 (0.692-0.784)↑[a]	0.650 (0.588-0.710)↑	0.655 (0.597-0.711)↑
Length correlation	0.824 (0.781-0.863)	0.834 (0.790-0.876)	0.845 (0.806-0.883)↑[a]	0.889 (0.856-0.919)↑[a]
Aarhus University Hospital				
Sensitivity	0.952 (0.875-1.000)	0.952 (0.875-1.000)	0.952 (0.881-1.000)	0.952 (0.880-1.000)
Specificity	0.967 (0.889-1.000)	0.933 (0.826-1.000)	0.867 (0.731-0.971)	0.833 (0.692-0.960)
ISUP κ	0.423 (0.319-0.541)	0.452 (0.346-0.557)	0.408 (0.295-0.513)	0.368 (0.249-0.479)
Length correlation	0.811 (0.706-0.895)	0.820 (0.691-0.914)[a]	0.749 (0.557-0.901)↓	0.720 (0.505-0.889)↓
Stavanger University Hospital				
Sensitivity	0.518 (0.456-0.580)	0.904 (0.867-0.938)↑	0.378 (0.316-0.443)↓	0.968 (0.944-0.988)↑[a]
Specificity	0.997 (0.992-1.000)	0.959 (0.941-0.974)↓	0.954 (0.936-0.971)↓	0.882 (0.856-0.906)↓
ISUP κ	0.439 (0.383-0.491)	0.619 (0.578-0.657)↑	0.281 (0.221-0.340)↓	0.623 (0.578-0.663)↑
Length correlation	0.817 (0.758-0.863)	0.844 (0.788-0.892)↑	0.578 (0.502-0.645)↓	0.862 (0.812-0.907)↑[a]

Measurements in this table include sensitivity and specificity for cancer detection, linearly weighted Cohen κ for ISUP grade (ISUP κ), and the linear correlations between cancer lengths estimated by the AI system and the pathologist (length correlation). Values in parentheses indicate 95% CIs. For each color correction method (Sierra, Macenko, and cycle-GAN), upward and downward arrows indicate a statistically significant ($P < .05$) improvement or decline, respectively, relative to the model without color correction (original).

[a] The color correction method provided a statistically significant ($P < .05$) improvement compared to both of the other 2 color correction methods.

embed ICC profiles into WSIs use their own generic profiles to a variety of accuracies, with the aim of masking hardware variability and rendering desired output colors. However, not all scanners use ICC profiles, and instead attempt to create color fidelity purely by the choice of optical and lightpath components, which inherently creates large variability in accuracy and supplier batch effects. Owing to these reasons, a universal and vendor-agnostic method is still currently needed to standardize the variable or absent use of ICC profiles by scanner manufacturers.

Foundation models have recently emerged as powerful new tools for computational pathology[22-24,38-40] and represent a shift away from the traditional approach of training highly specialized "narrow" models for each particular application, instead providing a step in the direction of multipurpose models applicable to diverse tasks. Using foundation models pretrained on massive data sets representing different tissue types, clinical sites, and scanners as stepping stones for building models for a specific task not only accelerates AI development and decreases the need for large amounts of task-specific training data but may also lead to more robust models due to the exposure to a large degree of variations in data characteristics during the task-agnostic pretraining process. Our experiments on the state-of-the-art foundation model UNI[24] confirmed its superior generalization capacity compared with a task-specific, narrow AI model. However, although the effect was less pronounced than with the task-specific model and only statistically significant in the Stavanger cohort, physical color calibration still led to improved grading performance (Table 3) even when applied to the UNI-based model. This is in line with emerging evidence suggesting that even foundation models are not fully immune to batch effects.[41,42] Moreover, even the generalization performance of a foundation model is ultimately dependent on the fraction of all plausible variations covered by its pretraining data and unfortunately, the space of laboratory and scanner dependent variation is not fixed but will change over time. This is likely to result in model drift, or "AI aging",[43-45] which even foundation models are not immune to. With this in mind, characterization and control of input data may still have an important role in computational pathology even in the foundation model era.

Table 3
Evaluation metrics for slide-level cancer diagnosis performance using the models with UNI

Cohort & metric	Original	Sierra
Karolinska University Hospital		
Sensitivity	0.994 (0.980-1.000)	0.987 (0.966-1.000)
Specificity	0.813 (0.718-0.900)	0.787 (0.691-0.877)
ISUP κ	0.739 (0.689-0.786)	0.760 (0.710-0.804)
Aarhus University Hospital		
Sensitivity	0.952 (0.881-1.000)	0.952 (0.881-1.000)
Specificity	0.533 (0.360-0.720)	0.667 (0.500-0.846)
ISUP κ	0.424 (0.302-0.546)	0.459 (0.330-0.587)
Stavanger University Hospital		
Sensitivity	0.968 (0.948-0.988)	0.912 (0.875-0.945)↓
Specificity	0.670 (0.632-0.708)	0.847 (0.819-0.876)↑
ISUP κ	0.547 (0.504-0.582)	0.670 (0.631-0.711)↑

Measurements in this table include sensitivity and specificity for cancer detection and linearly weighted Cohen κ for ISUP grade (ISUP κ). Values in parentheses indicate 95% CIs. For the model using color correction (Sierra), upward and downward arrows indicate a statistically significant ($P < .05$) improvement or decline, respectively, relative to the model without color correction (original).

The study has a number of limitations. First, Sierra addresses the digital color fidelity differences between scanners and calibrates them to a standard based on spectral ground truth, but it does not account for variations due to tissue processing and staining chemistry that occur before a glass slide is imaged. Although physical calibration still matched or outperformed the computational normalization methods in this study, calibration could likely be further improved by developing physical techniques similar to Sierra, applicable to correcting for variation in tissue staining. Moreover, techniques such as conformal prediction can be used as the last line of defense to detect data quality issues not captured by physical calibrants.[9] Second, we included 4 scanner models from 2 vendors, and it is likely that the impact of Sierra would be greater with increased variability introduced by inclusion of more vendors and new scanner models released over time. Third, there were time delays between the scanning of Sierra and the actual slides used for evaluation (over 1 year for the STHLM3 and Karolinska University Hospital cohorts). The observed positive impact of calibration provides some evidence of

the stability of both the Sierra calibration method and the included scanners. Still, improved calibration is to be expected with prospective scanner profiling. Each Sierra slide is stable for 100 uses or 12 months, allowing repeated calibrations. Further investigations into the potential variability due to scanners drifting from factory settings over time may have implications for the recommended frequency of Sierra profiling.

Gleason grading is crucial for treatment decisions in prostate cancer and its subjectivity[35] presents both a challenge and an opportunity for AI-assisted pathology. Consequently, AI-based Gleason grading has attracted considerable attention in recent years[20] and provided a relatively well-defined and algorithmically mature example application for the current study. Importantly, neither the problems with generalization of AI algorithms nor the color calibration technology evaluated in this study is specific to prostate cancer grading, and the methodology presented addresses a fundamental and universal issue in all WSI. To the best of our knowledge, this is the first study to evaluate the efficacy of physical color calibration as a potential solution to the problem of cross-site generalization of AI algorithms, and we expect subsequent studies to apply the same approach to other tasks, tissue types, and disease states.

There are fundamental task-dependent differences in the quality requirements for AI algorithms. For example, the occasional errors committed by chatbots like Chat-GPT typically do not bear similarly catastrophic consequences as those of AI algorithms that clinicians rely on for medical diagnostics. If AI is to be widely applied in digital pathology, all aspects of the data processing chain need to be scrutinized and quality controlled rigorously, similarly to the processes applied to other medical measurement instruments[46] or even simple laboratory tools like pipettes. Ensuring consistent image quality is important also when the actual diagnostic task is performed by pathologists,[47] but even more so if parts of the decision-making process are delegated to AI algorithms, which currently have limited capabilities to report issues in their input data or to adapt to changes dynamically in the manner a human expert does. Errors conducted by machines are also often tolerated to a lesser extent than errors by human experts, which sets the bar high for medical AI.[48]

As a step in this direction, our results suggest that physical color calibration can be a reliable approach for ensuring safe, quality-assured, and robust deployment of computational pathology AI algorithms across different clinical sites, using a standardized and simple method that does not require extra data-handling skills or site-specific tuning. A further positive impact is that the methodology allows for interlaboratory digital QA, which would be beneficial for data sharing and industry regulation if this method gains wider adoption. We believe that such improved techniques and processes for QA should not be seen merely as incremental technical tweaks to existing AI methodology but as the next essential step in translating AI algorithms from emerging technology into ubiquitous, routine clinical tools.

Acknowledgments

The authors thank Carin Cavalli-Björkman, Astrid Björklund, and Britt-Marie Hune for assistance with scanning and database support; Simone Weiss for assistance with scanning in Aarhus; and Silja Kavlie Fykse and Desmond Mfua Abono for scanning in Stavanger. ICC profile guidance was contributed by Craig Revie of the International Color Consortium, Louise Collins of PathQA, Ltd, and Jacqui Deane and John Stevenson-Hoare of FFEI, Ltd. The authors acknowledge the participants of the STHLM3 study and the NordCaP project who contributed with the clinical material that made this study possible.

Author contributions

X.J., R.S., N.M., P.R., M.E., and K.K. conceived and designed the study. A.B., B.G.P., K.D.S., B.P.U., S.R.K., E.A.M.J., L.E., and M.E. collected the data. X.J., R.S., N.M., U.K., Y.W., M.R., P.R., and K.K. analyzed and interpreted the data. X.J., R.S., U.K., Y.W., M.E., and K.K. drafted the paper. X.J., R.S., N.M., U.K., Y.W., A.B., H.O., B.G.P., K.D.S., B.P.U., S.R.K., E.A.M.J., M.R., L.E., P.R., M.E., and K.K. critically reviewed the paper for important intellectual content. R.S., B.G.P., K.D.S., S.R.K., E.A.M.J., M.R., L.E., P.R., M.E., and K.K. obtained the funding. K.K. had full access to all data in the study, taking responsibility for the integrity of the data and the accuracy of the data analysis. All authors read and approved the final version of the paper.

Data Availability

The deidentified data used for model training are available for noncommercial research purposes subject to a CC BY-SA-NC 4.0 license as part of the PANDA challenge data set and are freely downloadable after registration at https://www.kaggle.com/c/prostate-cancer-grade-assessment. Data representing the testing cohorts can be made available through contact with M.E. (Karolinska University Hospital), S.R.K. (Stavanger University Hospital), and B.G.P. (Aarhus University Hospital) under research collaboration and data-sharing agreements. The fully supervised AI models and associated code are released under the CC BY-SA-NC 4.0 license and available at https://github.com/prcimage/supplementary-physical-color-calibration. The weakly supervised UNI models are implemented with requested authentication and following the instruction at https://github.com/mahmoodlab/uni, which also follows the CC-BY-NC-ND 4.0 license.

Declaration of Competing Interest

R.S. is the CEO and founder of PathQA, Ltd. Y.W. is an employee of Stratipath AB. M.R. is a cofounder and shareholder of Stratipath AB. N.M, L.E., M.E., and K.K. are shareholders of Clinsight AB.

Funding

R.S. received funding from Innovate UK (Future Leaders Fellowship MR/V023314/1). U.K. received funding from University of Turku (graduate school), Finland. A.B. received a grant from the Health Faculty at the University of Stavanger, Norway. B.G.P and K.D.S received funding from Innovation Fund Denmark (8114-00014B) for the Danish branch of the NordCaP project. M.R. received funding from Swedish Research Council and Swedish Cancer Society. P.R. received funding from the Research Council of Finland (341967) and Cancer Foundation Finland. M.E. received funding from Swedish Research Council, Swedish Cancer Society, Swedish Prostate Cancer Society, Nordic Cancer Union, Karolinska Institutet, and Region Stockholm. K.K. received funding from the SciLifeLab & Wallenberg Data Driven Life Science Program (KAW 2024.0159), David and Astrid Hägelen Foundation, Instrumentarium Science Foundation, KAUTE Foundation, Karolinska Institute Research Foundation, Orion Research Foundation, and Oskar Huttunen Foundation. Computations were made possible by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at C3SE, partially funded by the Swedish Research Council

(2022-06725 and 2018-05973), by the supercomputing resource Berzelius provided by National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg foundation, and by CSC—IT Center for Science, Finland. The funders of the study and the providers of computing infrastructure had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Ethics Approval and Consent to Participate

This is a retrospective study that does not require patient consent.

Supplementary Material

The online version contains supplementary material available at https://doi.org/10.1016/j.modpat.2025.100715.

References

1. Ström P, Kartasalo K, Olsson H, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol*. 2020;21(2):222–232.
2. Amgad M, Hodge JM, Elsebaie MAT, et al. A population-level digital histologic biomarker for enhanced prognosis of invasive breast cancer. *Nat Med*. 2024;30(1):85–97. https://doi.org/10.1038/s41591-023-02643-7
3. Fu Y, Jung AW, Torne RV, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer*. 2020;1(8):800–810.
4. Clarke EL, Treanor D. Colour in digital pathology: a review. *Histopathology*. 2017;70(2):153–163.
5. Rojo MG, García GB, Mateos CP, García JG, Vicente MC. Critical comparison of 31 commercially available digital slide systems in pathology. *Int J Surg Pathol*. 2006;14(4):285–305.
6. Patel A, Balis UGJ, Cheng J, et al. Contemporary whole slide imaging devices and their applications within the modern pathology department: a selected hardware review. *J Pathol Inform*. 2021;12:50.
7. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301–1309.
8. Swiderska-Chadaj Z, de Bel T, Blanchet L, et al. Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. *Sci Rep*. 2020;10(1):14398.
9. Olsson H, Kartasalo K, Mulliqi N, et al. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nat Commun*. 2022;13(1):7761.
10. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med Image Anal*. 2019;58:101544.
11. Marini N, Otalora S, Wodzinski M, et al. Data-driven color augmentation for H&E stained images in computational pathology. *J Pathol Inform*. 2023;14:100183.
12. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2009:1107–1110. https://doi.org/10.1109/ISBI.2009.5193250
13. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017:2242–2251. https://doi.org/10.1109/ICCV.2017.244
14. de Bel T, Bokhorst JM, van der Laak J, Litjens G. Residual cyclegan for robust domain transformation of histopathological tissue slides. *Med Image Anal*. 2021;70:102004.
15. Bautista PA, Hashimoto N, Yagi Y. Color standardization in whole slide imaging using a color calibration slide. *J Pathol Inform*. 2014;5(1):4.
16. Center for Devices and Radiological Health. Technical performance assessment of digital pathology whole slide imaging devices. U.S. Food and Drug Administration. Accessed May 25, 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/technical-performance-assessment-digital-pathology-whole-slide-imaging-devices
17. Version 4 ICC specification. Accessed May 25, 2023. https://www.color.org/v4spec.xalter
18. Nagpal K, Foote D, Tan F, et al. Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. *JAMA Oncol*. 2020;6(9):1372–1380.
19. Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol*. 2020;21(2):233–241.
20. Bulten W, Kartasalo K, Chen PC, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med*. 2022;28(1):154–163.
21. Marletta S, Eccher A, Martelli FM, et al. Artificial intelligence-based algorithms for the diagnosis of prostate cancer: a systematic review. *Am J Clin Pathol*. 2024;161(6):526–534.
22. Wang X, Yang S, Zhang J, et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med Image Anal*. 2022;81:102559.
23. Vorontsov E, Bozkurt A, Casson A, et al. Virchow: a million-slide digital pathology foundation model. Preprint. Posted online September 14, 2023. arXiv 2309.07778. https://doi.org/10.48550/ARXIV.2309.07778
24. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. *Nat Med*. 2024;30(3):850–862.
25. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. Preprint. Posted online August 16, 2021. arXiv 2108.07258. https://doi.org/10.48550/ARXIV.2108.07258
26. Grönberg H, Adolfsson J, Aly M, et al. Prostate cancer screening in men aged 50-69 years (STHLM3): a prospective population-based diagnostic study. *Lancet Oncol*. 2015;16(16):1667–1676.
27. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016:2818–2826. https://doi.org/10.1109/CVPR.2016.308
28. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. *International Conference on Learning Representations*. 2020. Accessed June 13, 2024. https://openreview.net/pdf?id=YicbFdNTTy
29. Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning robust visual features without supervision. *Preprint*. Posted online April 14, 2023. arXiv 2304.07193. https://doi.org/10.48550/ARXIV.2304.07193
30. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. *International Conference on Machine Learning*. PMLR; 2018:2127–2136.
31. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *Preprint*. Posted online March 14, 2016. arXiv 1603.04467. https://doi.org/10.48550/ARXIV.1603.04467
32. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785–794. https://doi.org/10.1145/2939672.2939785
33. Clarke EL, Revie C, Brettle D, et al. Development of a novel tissue-mimicking color calibration slide for digital microscopy. *Color Res Appl*. 2018;43(2):184–197.
34. Badano A, Revie C, Casertano A, et al. Consistency and standardization of color in medical imaging: a consensus report. *J Digit Imaging*. 2015;28(1):41–52.
35. Egevad L, Swanberg D, Delahunt B, et al. Identification of areas of grading difficulties in prostate cancer and comparison with artificial intelligence assisted grading. *Virchows Arch*. 2020;477(6):777–786.
36. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):1–7.
37. Cohen JP, Luck M, Honari S. Distribution matching losses can hallucinate features in medical image translation. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, eds. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer; 2018:529–536.
38. Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*. 2024;630(8015):181–188.
39. Lu MY, Chen B, Williamson DFK, et al. A visual-language foundation model for computational pathology. *Nat Med*. 2024;30(3):863–874.
40. Filiot A, Ghermi R, Olivier A, et al. Scaling self-supervised learning for histopathology with masked image modeling. *Preprint*. Posted online September 14, 2023. medRxiv 2023.07.21.23292757. https://doi.org/10.1101/2023.07.21.23292757
41. Kömen J, Marienwald H, Dippel J, Hense J. Do histopathological foundation models eliminate batch effects? A comparative study. Preprint. Posted online November 8, 2024, 05489. https://doi.org/10.48550/ARXIV.2411.05489
42. Yun J, Hu Y, Kim J, Jang J, Lee S. EXAONEPath 1.0 patch-level foundation model for pathology. *Preprint*. Posted online August 1, 2024. arXiv 2408.00380. https://doi.org/10.48550/ARXIV.2408.00380
43. Vela D, Sharp A, Zhang R, Nguyen T, Hoang A, Pianykh OS. Temporal quality degradation in AI models. *Sci Rep*. 2022;12(1):11654.
44. Nelson K, Corbin G, Anania M, Kovacs M, Tobias J, Blowers M. Evaluating model drift in machine learning algorithms. In: *2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*. 2015:1–8. https://doi.org/10.1109/CISDA.2015.7208643

45. Kore A, Abbasi Bavil E, Subasri V, et al. Empirical data drift detection experiments on real-world medical imaging data. *Nat Commun*. 2024;15(1):1887.
46. Romanchikova M, Thomas SA, Dexter A, et al. The need for measurement science in digital pathology. *J Pathol Inform*. 2022;13:100157.
47. Wright AI, Clarke EL, Dunn CM, Williams BJ, Treanor DE, Brettle DS. A point-of-use quality assurance tool for digital pathology remote working. *J Pathol Inform*. 2020;11:17.
48. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. *J Consum Res*. 2019;46(4):629–650.