Journal of Pathology Informatics 17 (2025) 100423

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

An international study of stain variability in histopathology using qualitative and quantitative analysis

Catriona Dunn [a,*], David Brettle [a], Chantell Hodgson [b], Robert Hughes [b], Darren Treanor [a,c,d,e,f]

[a] National Pathology Imaging Co-operative, Leeds Teaching Hospitals NHS Trust, Beckett Street, Leeds, UK
[b] UK NEQAS Cellular Pathology Technique, Haylofts, St Thomas Street, Haymarket, Newcastle, UK
[c] Department of Histopathology, Leeds Teaching Hospitals NHS Trust, Beckett Street, Leeds, UK
[d] Department of Pathology and Data Analytics, University of Leeds, Beckett Street, Leeds, UK
[e] Department of Clinical Pathology and Clinical and Experimental Medicine, Linköping University, Linköping, Sweden
[f] Centre for Medical Image Science and Visualisation, Linköping University, Linköping, Sweden

ARTICLE INFO

Keywords:
H&E
Stain variation
Quality assurance
Digital pathology

ABSTRACT

Hematoxylin and eosin (H&E) staining accounts for over 80% of slides stained worldwide. Although routinely used, there are high levels of variation between labs due to different staining methods. Staining is a pivotal part of slide preparation, but quality control is largely subjective, with overall clinical assurance provided by external quality assessment (EQA) services, underpinned by expert assessment. Digital pathology offers the potential to provide objective quantification of stain, through color analysis, to augment EQA assessment.

This large-scale study evaluated H&E staining in 247 international labs participating in the UK NEQAS CPT EQA programme. Tissue sections were circulated to each lab to stain using their routine H&E staining protocol. The slides were reviewed by independent expert UK NEQAS CPT assessors, and quantitative digital analysis was conducted, comprising of H&E color deconvolution and color difference determination (ΔE).

Most labs (69%) achieved an EQA score indicating good or excellent staining, with high inter-observer concordance to support this (92.5% within one mark of each other). H&E color difference, ΔE, showed 60% of labs were within 2 ΔE of the mean, which is considered as only perceptible through close observation. There was little correlation found between H&E intensity and assessor score, however, the H&E intensity ratio indicated a trend with assessor score suggesting there may be an optimal stain relationship that should be investigated further.

The presented hybrid analysis combines expert analysis with objective data. This has the potential to inform upon optimal tissue staining and allows us to consider quantitative standards of H&E staining in pathology practice.

Introduction

Hematoxylin and eosin (H&E) staining accounts for over 80% of all slides stained worldwide.[1] Although this is a 'standard' stain combination, there are still high levels of variation between labs due to different stain manufacturers and staining regimes.[2] The stained color of tissue is important for diagnostic evaluation by pathologists, and for artificial intelligence (AI) analysis in digital pathology, which is also increasingly being used for molecular pathology applications.[3]

In traditional pathology, using microscopy, variation between slides can be induced by physical parameters including fixation, tissue sectioning, and the staining process. This variability can lead to re-sectioning and staining if the slide produced is not considered to be adequately stained by the reviewing pathologist. In digital pathology, these variations can be further compounded by variations introduced during the digitization process, where color reproduction may not be accurate or consistent within or between scanners.[4,5] Gray et al.[2] found that there was 0.5% variation between slides scanned in the same scanner on the same day, but this rose to 8% when scanned on different days. There are more significant variations between scanners, by design, where different manufacturers use different technical specifications and signal processing, often resulting in various image color presentation between scanners.[6]

Anecdotally humans have a good tolerance for a range of stain variability in histopathology, but stain color variation is more of an issue when in digital pathology. Arguably, one of the greatest benefits of digital progression in pathology is the development of image analysis and AI to enhance

Abbreviations: Artificial intelligence, AI; cellular pathology technique, CPT; delta E, ΔE; directional delta E, dΔE; External Quality Assessment, EQA; hematoxylin and eosin, H&E; lightness (L*a*b* space), L*; whole slide image, WSI.

* Corresponding author at: National Pathology Imaging Co-operative, Sir Robert Ogden Centre, St James University Hospital, Beckett Street, Leeds, West Yorkshire LS9 7TF, UK.
E-mail address: catriona.dunn@nhs.net (C. Dunn).

http://dx.doi.org/10.1016/j.jpi.2025.100423
Received 25 October 2024; Received in revised form 6 January 2025; Accepted 7 February 2025
Available online 12 February 2025

diagnostic processes. But these techniques can be faltered by variations in the color of whole slide images (WSIs), particularly between datasets stained in different institutions. This is because many algorithms are sensitive to domain shift and depend upon measurements of color intensities to establish thresholds and pattern detection.[7,8] AI training data cannot easily cover the full breadth of stain variation found in clinical practice, therefore AI lacks portability between institutions and can fail when applied to datasets that vary in staining presentations compared to the training dataset. Computational pre-processing is often adopted to mitigate for stain variation. For example, in the literature, several investigations into the accuracy of AI using original datasets compared to stain 'normalized' datasets have found significant improvements when using normalized data.[9–12] Stain normalization may reduce the effect of stain variation on AI outputs, however, it has been evidenced that even when using stain normalization, lab-specific characteristics can still be easily interpreted by deep learning algorithms and may have the potential to bias the results.[13] Furthermore, pre-processing has the potential to introduce extra variation and may be less effective at the extreme ends of the staining spectrum, e.g., weakly stained or over-saturated WSIs.[14]

Stain variation can also impact upon pathologist examination meaning that the assessment of H&E staining quality is important in conventional microscopy as well as in the use of digital pathology. The role of external quality assessment (EQA) and proficiency testing is therefore essential to provide assurance that a high standard of testing and clinical quality is achieved in a lab for the correct patient results. EQA schemes are an essential part of patient diagnostics, to ensure early detection of any lab errors and additionally, to provide education, help, and support to organizations in correcting issues.[15] In the UK, H&E pathology slide EQAs are undertaken by a neutral, external body, UK NEQAS CPT (National External Quality Assessment Service Cellular Pathology Technique; Newcastle, UK), where experienced assessors qualitatively examine H&E-stained slides, according to pre-established criteria, to assign a semi-quantitative score.[16] For more than 3 decades UK NEQAS CPT has evolved, developed, and adapted cellular pathology EQA and proficiency testing, to match lab needs, and are a key element of meeting lab accreditation standards and UKAS assessment.[17,18] UK NEQAS CPT schemes and services are also accredited to the complementary ISO/IEC 17043:2023 Proficiency Testing standards.[19]

The digitization of slides introduces the opportunity for additional image quality assessment using computational methods. One example of this is quantitative measurements of WSI color.[2,20,21] Digital color can be measured using mathematical models describing color, often in a three-dimensional color space covering color and intensity, such as RGB, HSV, and CIE L*a*b* color spaces.[22,23] Color is represented in pixels on digital displays as various intensities of red, green, and blue (RGB), therefore an often-used color space for digital color measurement is RGB color space. In digital pathology research, however, color is often converted from non-linear RGB into different spaces, including optical density (OD) space and CIE L*a*b* color space.[20,24] OD space is linear and directly corresponds with the concentration of stain absorbing light. Due to this, it is particularly useful in digital pathology for H&E 'deconvolution' of WSIs into H&E color space, to estimate values for hematoxylin-only, and eosin-only, based upon approximations of the two stains within the WSI.[25] H&E deconvolution is typically undertaken in OD space and is adopted into stain normalization algorithms for normalization of the individual stain amounts.[26] CIE (International Commission on Illumination) L*a*b* color space is also commonly used, and has three channels: L* (representing perceptual Lightness), a* (representing color from red to green), and b* (representing color from yellow to blue). CIE L*a*b* color space is especially useful for color comparison because it provides the ability to calculate Delta E 2000 (ΔE).[27] ΔE is a pre-established metric that provides a measure of human perceptible color difference between two colors in CIE L*a*b* color space, where a ΔE of 0 means the colors are the exact same, and of 100 means the colors are the exact opposite.[28]

This article presents an international H&E staining audit using a hybrid evaluation of existing qualitative analysis of analogue slides, by UK NEQAS CPT, and quantitative color analysis of the digitized slides. The objectives were to gain a snapshot into the current landscape of stain variability, to develop techniques to enhance routine digital pathology EQA processes through utilization of digital quantitative measurements, and to understand if there is any correlation between expert assessor score and H&E color intensity.

Methods

Optimally processed, formalin-fixed, and paraffin-embedded bovine intestine was obtained in collaboration with an approved tissue supplier as part of a service level agreement between UK NEQAS CPT, an accredited ISO 15189 or equivalent laboratory. Blocks were sectioned to 4 μm thickness and mounted onto glass slides following lab standard operating procedures. Single slides, containing the unstained tissue preparations, were supplied to 247 participating labs to be stained using their routine H&E staining protocol. All participating labs were members of the UK NEQAS CPT EQA service, a UKAS accredited proficiency testing provider (no. 8268, ISO/IEC 17043:2023). The stained slides were returned to UK NEQAS CPT, within a 4-week window following staining, for review at the subsequent assessment.

Expert analysis

There were a total of 10 UK NEQAS CPT assessors (five pairs). The assessors were experienced Biomedical Scientists, Clinical Scientists, Advanced Practitioners, and Pathologists. They were specialists in their field of expertise, and underwent rigorous training, both prior to and during assessment sessions. Prior to the slide assessment, all the assessors passed a competency test that checked the concordance of their scoring compared with a working group. During the slide assessment, witness audits were undertaken where the assessors were observed to ensure that the SOP was followed, and the slide assessment was carried out correctly. Inter-observer statistics were also collected to ensure consistency in scoring between the pairs, with an acceptance threshold of 95% scoring within two marks of each other, and a threshold of 85% scoring within one mark of each other to be considered as 'excellent'.

Each slide was assessed on a microscope, using the published assessment criteria for H&E staining.[29] The microscopical assessment procedure was performed in pairs, any borderline or failed slides underwent secondary assessment by assessors. To remove opportunity for bias, all slides were given unique reference numbers and assessors did not have access to participant details. As the slides had been supplied with pre-mounted unstained sections, only the staining quality was assessed. Guidelines for an assessors score out of 10 are detailed in Table 1, where a score of 10 indicated optimal H&E staining, and a score of 4 or less showed sub-optimal staining:

The slide presentation was fully considered for visualization of nuclear detail and tissue discrimination. Nuclei must be stained purple-blue with hematoxylin and the intensity must be strong enough to allow clear demonstration of nuclear detail at a medium power, but not too intense to cause a loss of the chromatin granularity or excessive cytoplasmic or connective

Table 1
UK NEQAS CPT assessor scoring.

Score out of 10	Description
<5	A score of less than 5 shows poor demonstration, where the participant has failed to clearly display the expected results.
5 or 6	A score of 5 or 6 is a pass. While demonstration is appropriate the expected results are sub-optimal, and improvements are required overall.
7 or 8	A score of 7 or 8 shows good, appropriate demonstration of the expected results and an acceptable level of quality.
9 or 10	A score of 9 or 10 shows excellent demonstration of the expected results and a high level of quality. This assessment represents the current best practice for stain assessment.

tissue staining. Where the hematoxylin has been over differentiated, minimal cytoplasmic or connective tissue background staining with hematoxylin must remain. This background, if present, must not reduce the effectiveness of the nuclear demonstration or affect the color and selectiveness of the eosin. The eosin must be selective enough to demonstrate tissue morphology and different components such as collagen, cytoplasm, red blood cells, cellular granules, amyloid, etc. The intensity must be appropriate to the section thickness and the hematoxylin intensity. Where the eosin is too weak, it will fail to allow selective demonstration of different components at low power. If the eosin intensity is too strong, the color and detail of the nuclear stain will be obscured, and selectivity will be reduced.

Digital analysis

After review by UK NEQAS CPT, the H&E-stained control slides were provided to the National Pathology Imaging Co-operative (NPIC; Leeds, UK) for quantitative digital analysis. All slides were digitized in a Leica GT450 DX (Leica Biosystems; Nußloch, Germany) whole slide imaging scanner, at 40× magnification (0.263 μm per pixel), with JPEG compression at 91 quality (using libjpeg), according to the standard scanning protocol and saved as 'svs' files. All images were internally checked for quality using the NPIC standard quality control (QC) process, with repeat scans for any WSIs with digital QC issues such as stitching or focus artifact. To reduce the likelihood of scanner variation between the WSIs, all slides were scanned in one scanner on 1 day, with the small number of QC re-scans undertaken within 5 days.

The WSIs were opened in QuPath (https://qupath.github.io/),[30] with no color profile applied, and extracted as lower quality PNG images (down-sample value = 10). Down-sampling of the images reduced the resolution by combining and averaging the color of neighboring pixels, resulting in a reduced image size with fewer pixels. It effectively resulted in averaging of the signal which had no impact on the subsequent net color measurement of the whole tissue region that was required in this work. Using MATLAB (Natick, Massachusetts, USA), the PNG images were cropped to include as little white background as possible. These processes were undertaken to reduce the images' size for increased processing speed. The cropped images were color deconvolved into H&E-separated images, using the adapted implementation by Landini et al.[31] of the Ruifrok and Johnson method.[32] This was not an absolute deconvolution, but a mathematical approximation of the separated stains.

Using MATLAB, the tissue areas on the original H&E images and the deconvolved hematoxylin-only and eosin-only images, were detected. The average CIE L*a*b* values were collected from this area on each image. The L* (Lightness) values were utilized as an intensity measurement from each WSI, with values plotted onto histograms to visualize the spread of the L* intensity values measured.

Delta E 2000 (ΔE), a measurement of perceptual color difference, was calculated by comparing the mean L*a*b* values of each H&E slide with the mean value of all the submitted H&E slides (equating to a mean of 247 laboratories). ΔE has values between 0 and 100; see Table 2A for an overview of the ΔE values.[28]

To determine if each slide was lighter or darker than the average of all submitted slides, the ΔE measurement was adjusted for intensity change by multiplying ΔE values, where the intensity was darker, by −1. This resulted in positive ΔE values being lighter than the average, and negative ΔE values being darker than the mean of participants. We refer to this as directional ΔE (dΔE). For digital pathology, we propose that dΔE values up to ±2 are considered to be equivalent; between ±2 and ±5, there is a perceptible difference from the mean, but it is acceptable; between ±5 and ±10, the differences are easily perceptible and ideally should be moved closer to the mean; we suggest that values above ±10 are outliers (Table 2B). It is important to note that as dΔE gets further from 0, in either direction, color variation becomes increasingly perceptible.

Relationship of digital data to assessor score

The intensity values (L*) measured from the H&E and deconvolved hematoxylin-only and eosin-only images were plotted against each slide's corresponding UK NEQAS CPT assessor score to indicate any association between intensity and assessor score. The hematoxylin:eosin (H:E) ratio was also plotted against assessor score to indicate association. The H:E ratio was calculated for each slide by dividing the deconvolved hematoxylin image intensity by the corresponding deconvolved eosin image intensity, for each slide. To measure the strength of any relationship between the continuous (color intensity and ratio) and ordinal (NEQAS assessor score) variables, Kendall's Tau b (τb) was calculated.[33]

Results

Expert analysis

The UK NEQAS CPT assessment results can be seen in Fig. 1 and show 69% of the laboratories scored 8 or higher, meaning their staining was good or excellent. 17% of laboratories obtained the highest score of 10. The lowest score obtained was 5, meaning poor demonstration of staining, but only 1% of laboratories achieved this score. No labs failed the assessment.

The results of the inter-observer concordance tests found 100% achieved scores within two marks of each other and 92.59% were within one mark of each other. This passed the acceptance threshold and was considered as 'excellent'.

Table 2
(A) Delta E 2000 (ΔE) values for human color perception and, (B) our proposed directional ΔE (dΔE) ranges applicable for a digital pathology H&E standard.

(A)	ΔE	PERCEPTION
	≤1	Not perceptible by a human
	1–2	Perceptible through close observation
	2–10	Perceptible at a glance
	11–49	Colors are more similar than different
	100	Colors are exact opposite
(B)	dΔE	DIGITAL PATHOLOGY CONSIDERATION
	≤2	Staining is considered equivalent and is acceptable
	2 < 5	There is a perceptible difference in stain, but it is acceptable
	5 < 10	Stain differences are easily perceptible staining should be adjusted to be closer to the target stain levels.
	10+	Staining levels are outliers from the target levels and remedial action is required
	+/−	Negative ΔE values are darker than the mean, positive ΔE values are lighter

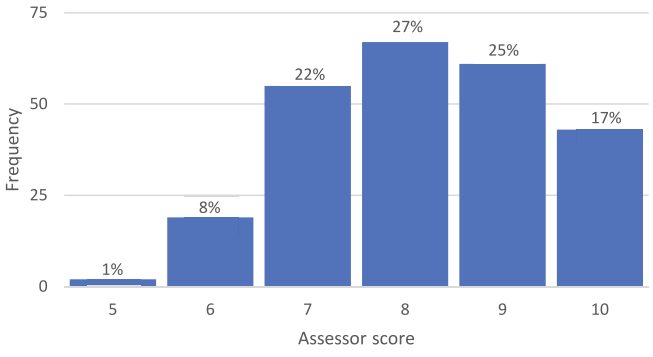


Fig. 1. Expert assessor score results. A histogram depicting the results of the UK NEQAS CPT expert assessment of control slides (bovine intestine) stained with H&E in 247 participating laboratories. The bins represent the frequency of laboratories that obtained each assessor score (between 5 and 10, no laboratory scored under 5), where a score of 5 showed poor demonstration of staining, and a score of 10 showed excellent demonstration. The percentage of labs that achieved each score is shown above each bin on the graph.

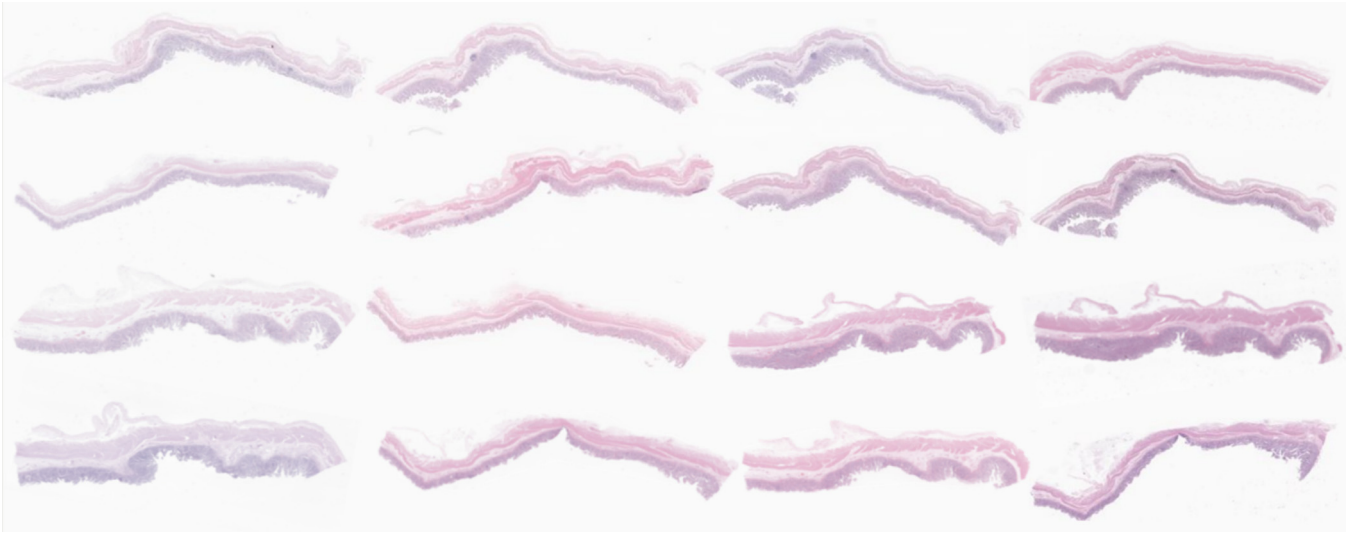


Fig. 2. Submitted slide images. Selection of thumbnail images of WSIs of H&E-stained bovine intestine sections, submitted for assessment, showing a range of staining presentations present across labs.

Digital analysis

The 247 slides were digitized to produce WSIs, a selection of WSI thumbnails can be seen in Fig. 2. Mean intensity values were measured from the whole tissue area in the H&E images, and the deconvolved hematoxylin-only and eosin-only images. The intensity measurement used was perceptual lightness, L*, from L*a*b* color space, where the L* value can range from 0 (black) to 100 (white). The lower the intensity value, the darker the average stain intensity, and conversely, the higher the intensity value means the lighter the average stain intensity. The histograms in Fig. 3 show the spread of intensity measured.

Fig. 4 shows the color difference (ΔE) calculated between the average L*a*b* values from each laboratory's H&E-stained slide and the mean L*a*b* values of all submitted H&E-stained slides, using the proposed directional ΔE metric (dΔE), where positive ΔE values indicate the image is lighter than the mean, and negative values indicate that the image is darker than the mean. For visualization, Fig. 5 contains examples of images with a range of dΔE values. 60% of labs were within ±2 dΔE from the mean, 94% within ±5 dΔE, and 100% were within ±10 dΔE.

Relationship of digital data to assessor score

Fig. 6 shows boxplots for (A) H&E, (B) deconvolved hematoxylin, and (C) deconvolved eosin stain intensity plotted against assessor score. No correlation was found between assessor score and hematoxylin intensity ($\tau b = 0.02$), and weak negative correlations were found for H&E ($\tau b = -0.22$), and eosin ($\tau b = -0.34$) intensities.

A boxplot displaying the H:E ratio plotted against assessor score can be seen in Fig. 6D. As the assessor score increased, the median H:E ratio increased, but there was only a weak positive correlation found ($\tau b = 0.21$). Excluding the results for assessor score 5, because it only had two datapoints, Fig. 6D did show that the H:E ratio distribution narrowed as the assessor score increased. There was a 60% reduction in interquartile range (IQR) from an assessor score of 6 (IQR = 0.040) to a score of 10 (IQR = 0.017).

Discussion

This study was a valuable opportunity to understand stain variation across a wide and diverse range of UK NEQAS CPT accredited histopathology labs. The data generated by NPIC were independent of the UK NEQAS CPT slide assessment and provided complimentary quantitative data for the assessment of the color variation of the digitized WSIs.

The expert analysis by UK NEQAS CPT found that 69% of labs scored 8 or above which was considered good or excellent, and the digital analysis showed good perceptual agreement (within ±2 dΔE) of color intensity for approximately 60% of the labs. An additional 34% of labs were within what could be argued to be an acceptable range of up to ±5 dΔE. However, these values should be taken with caution as there is a visible difference between images with 2 dΔE compared to a 5 dΔE.

When comparing the assessor scores with measured stain intensity, no strong correlation was found. There was a large distribution of H&E intensities for most assessor scores, but despite the noise, there was an indication that an H:E ratio of between 0.94 and 0.99 may be a factor in a higher assessment score. If this was found to be the case, it may help labs optimize their staining processes to standardize clinical agreement. There was, however, a broad distribution of H:E ratios for most scores which suggests there are other, perhaps more proportionally influential, factors that impact upon obtaining a high assessor score.

Overall, one interpretation of these results is that there is generally a core of good inter-laboratory agreement. However, this study was limited to UK NEQAS CPT accredited labs only, and therefore is not fully representative. It is likely that there would be much larger variation present if this study were to be extended to other labs. Additionally, this work compared the average, overall intensity of the tissue in each image, and it is predicted that a more detailed analysis, comparing, e.g., nuclei color, may have highlighted more significant variation between labs; future work will address this. It is important to note that the calculation of dΔE purely provides an indication of color difference between each lab and an 'arbitrary' value of the average L*a*b* of all the labs. The mean of all the labs was chosen as the base color because, to date, there is no 'gold-standard' H&E stain intensity or stain duration recommendation for labs to achieve, and therefore the dΔE is indicative only to color difference from the mean. All of labs were found to be performing at a diagnostically acceptable level and therefore using the mean may allow us to set a realistic and achievable initial level for standards based on color intensity. dΔE did not have correlation with expert assessor score which was to be expected due to the H&E intensity measurements not having a correlation with assessor score.

We believe the dΔE metric is a new concept that allows the impact of the color variation on the resultant clinical image to be quantified with respect to stain intensity. This is of importance, for example, if you used the conventional ΔE as a stain control metric where you would need to know whether to increase or decrease stain intensity to bring the stain intensity to within set parameters. We believe this metric is valid as we make the assumption that, nominally, we are aiming for the same, or very similar, color each time we stain, so any variation is along the same color curve in color

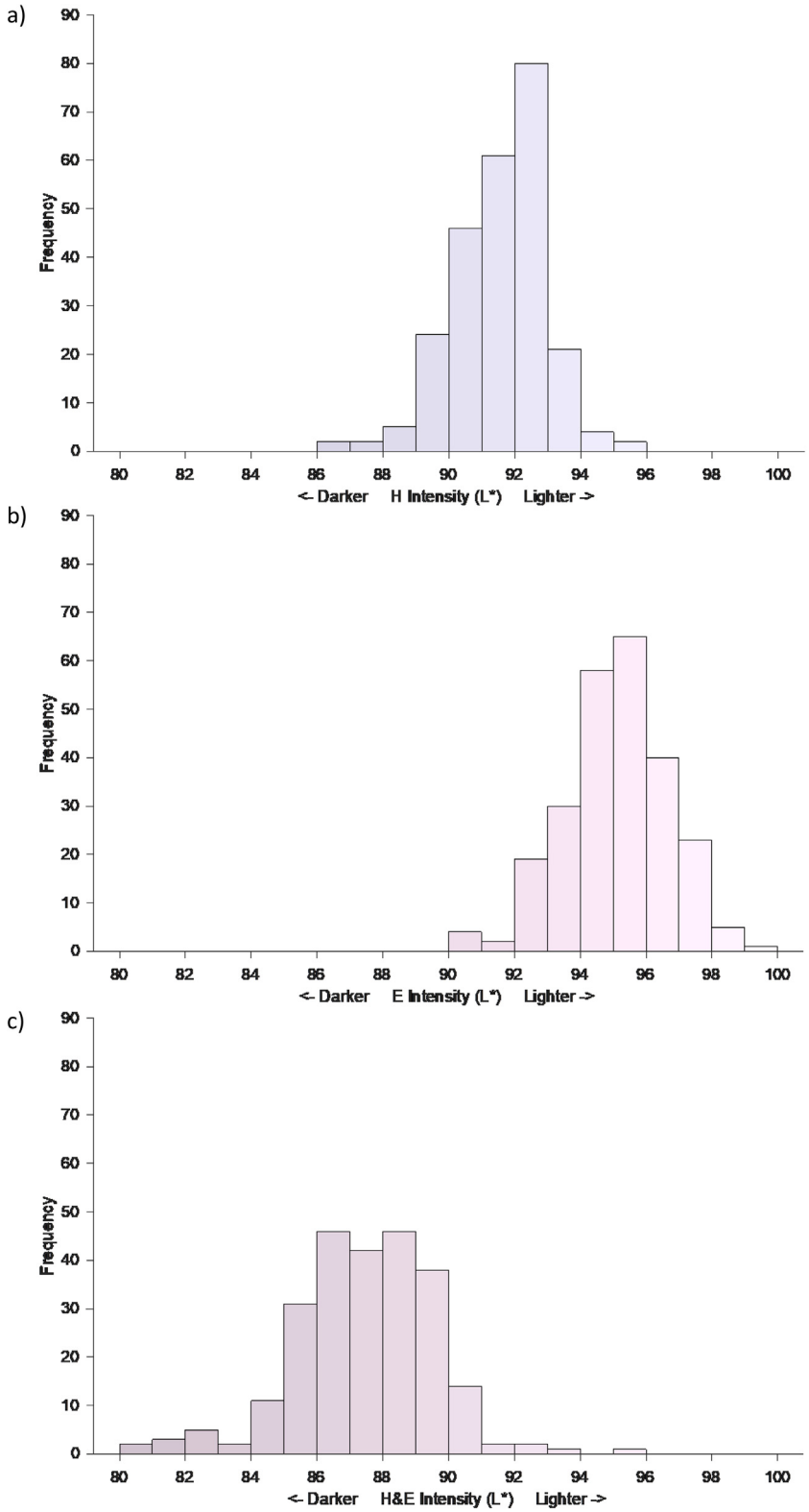
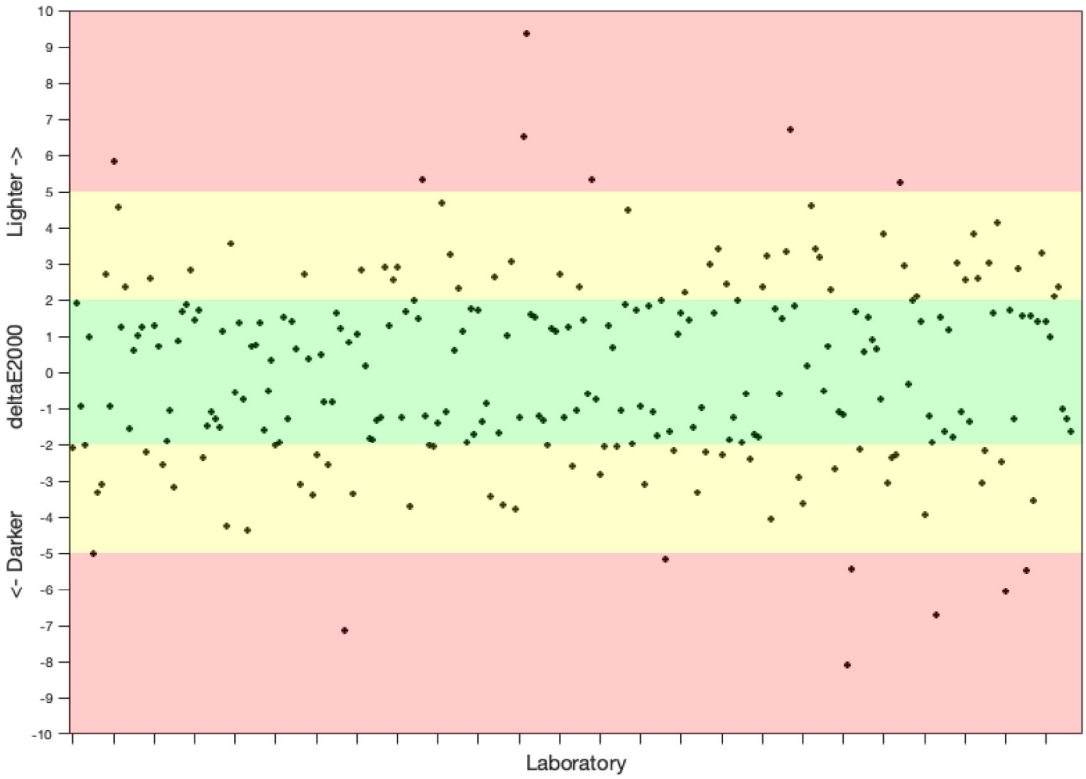


Fig. 3. H&E intensity. Histograms showing frequency of the average intensity (L*) measured from the 247 slides for the deconvolved hematoxylin (a) and eosin-only (b), and the H&E (c) images. The lower the intensity value, the darker the staining intensity, with the bin colors representing the median color measured from each histogram bin of slides.



± 0 to 2	Considered to be equivalent to the mean.
± 2 to 5	Perceptible difference from the mean, but acceptable
± 5 to 10	Easily perceptible difference from the mean and should be closer to the

Fig. 4. Directional Delta E. Directional Delta E 2000 (dΔE) results comparing the color difference of each slide compared with the average for all the submitted slides. The colored sections of the graph represent suggested approximate boundaries for stain variation limits in pathology images.

space. Caution should be used if the two-color points being compared vary widely as this assumption would not then be valid. This is a limitation of this work as we calculated dΔE on H&E stains, which is effectively two compounded color curves which may have more variability than a single stain. In addition, another limitation with the color work, especially the color deconvolution, is that different H&E stains can yield different color values for a* and b* (in L*a*b* color space) which will have affected ΔE values. For color deconvolution, a 'typical' color reference was used, based on Ruifrok and Johnson,[32] which may not have been the same as the actual stain used across all 247 labs. We believe this may have resulted in more variable results, as seen in Fig. 6, although weak trends were still observed, despite the noise, suggesting that they may be a real effect.

Another significant limitation of this study is that four tissue blocks were used to generate the assessment slides. Although all blocks were taken from the same tissue sample, there was a wide range of cellularity and stroma between the samples (see Fig. 5), which will have impacted upon the mean intensities used in this study. Although this is a limitation, it is also a real-world example of how many labs will stain different samples for control tissue for internal and EQA. Future work could look at using more homogeneous tissue, although this can be difficult to achieve. Another possibility would be to stratify the digital analysis, for example, by cell count and tissue type/area analysis.

A limitation of the qualitative assessment of the slides is that human observations are subjective measurements that can be susceptible to bias (e.g., local preference of staining) and both intra- and inter-observer variation.[34,35] To limit this, and ensure high consistency between assessors, UK NEQAS CPT undertake stringent testing and, for this study, found 92.5% of assessors scored within one mark of each other during testing, meaning low inter-observer variability was found. A limitation of the quantitative assessment of the WSIs is that digital images are susceptible to scanner-induced image acquisition variation, including variation in colour.[2,4] Future work could utilize a scanner color calibration slide alongside the sample scanning.

The results highlight that there is no accepted quantitative intensity threshold for an acceptable range of staining. Based on the dΔE, we have proposed new definitions of ΔE ranges that maps stain variability more closely to practical clinical stain QC. When used in a large-scale stain assessment study using nominally 'similar' tissue, such as the one presented here, it allows these values to be used as stain thresholds or standards. However, this does not allow routine stain QC unless consistent control tissue were used, but control tissue has its own variations and may be not consistent between samples; this is a known limitation within cellular pathology. Cell lines could be used as a stain control but controlling the reproducibility of the sample thickness would be challenging and any thickness variation would be an additional confounder for stain intensity measurements. We have previously proposed that a stain assessment biopolymer[36] can be used in this way which, when combined with the data presented here, presents the possibility of a quantitative tool for routine lab QC of stain

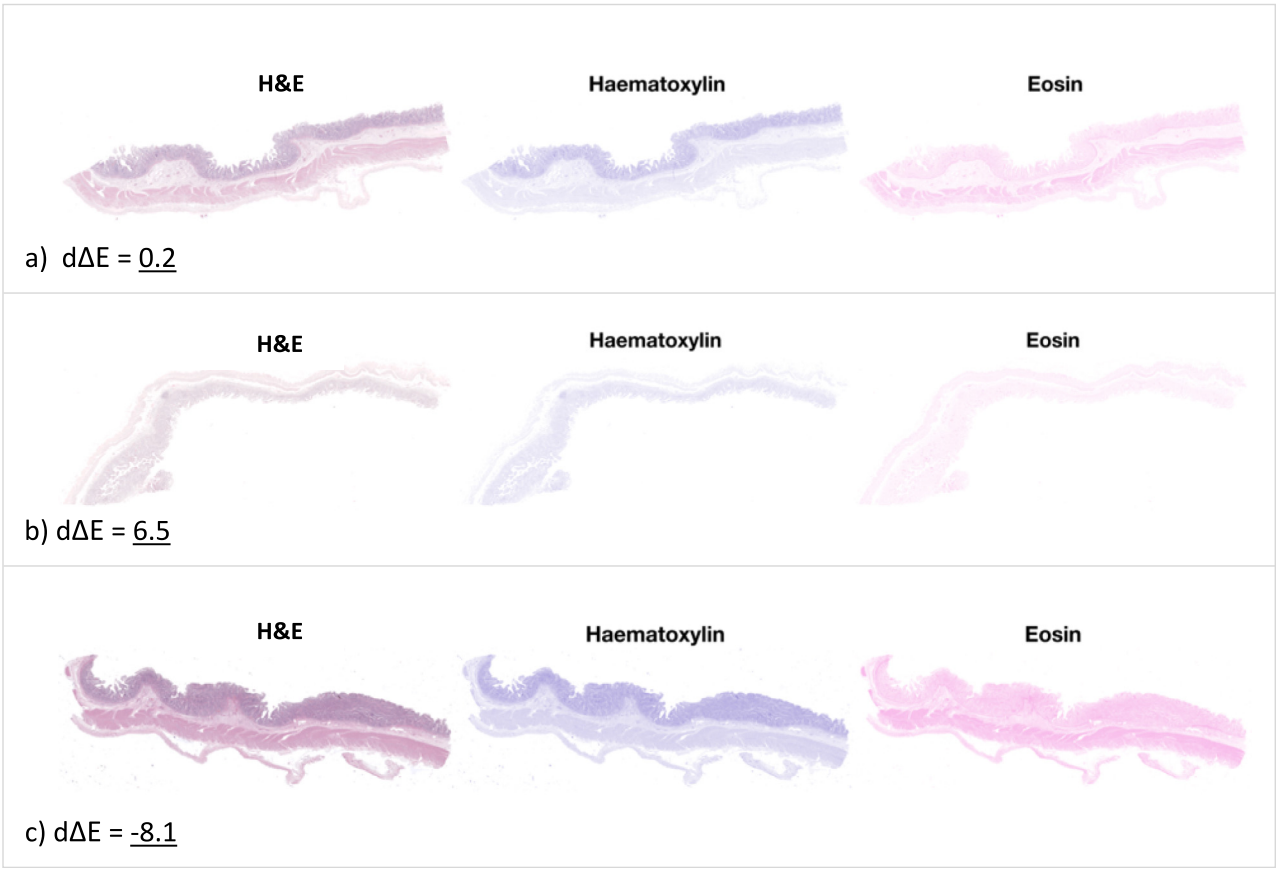


Fig. 5. Slides that showed the largest color difference. Example thumbnail images showing a range of directional Delta E 2000 (dΔE) values. Fig. 5a shows a H&E image (left) with a dΔE of 0.2, i.e., equivalent to the mean of all the submitted slides. It also shows the H&E image deconvolved into estimated hematoxylin-only and eosin-only images. Fig. 5b shows a H&E image with a dΔE of +6.5, a lower intensity than the average, and the associated deconvolved images. Fig. 5c shows an image with a dΔE of −8.1, a higher intensity than the average, and the associated deconvolved images.

consistency alongside current methods of qualitative QC and to augment EQA programmes.

Extremes of staining may materially impact upon the discernible features of clinical interest due to a compromise in signal, i.e., under- or over-staining of tissue structure for visualization.[37] It is accounting for this variation that we believe causes AI algorithms to underperform and therefore removing the outliers should provide a more consistent dataset optimizing AI performance. Future work needs to be undertaken to evidence this.

This study was an opportunity to gain 'point in time' quantitative knowledge of the landscape of stain variability across a large number of diverse labs. The results of this study will be valuable to inform future quality assurance/QC strategy in pathology. As the evolution of pathology, from analog to digital, increases, so too will the need for quality metrics and standards to ensure the value of the digital pathology dataset and the wide scale adoption of computational analysis in histopathology. Certainly, in this study, we saw a wide difference between the WSIs at the extremes of the range and there is an increasing understanding from the literature that variation does materially affect AI performance, so understanding and addressing these variations may be important practical steps towards realizing real-world AI in histopathology.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Catriona Dunn, David Brettle, Darren Treanor report financial support was provided by UK Research and Innovation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

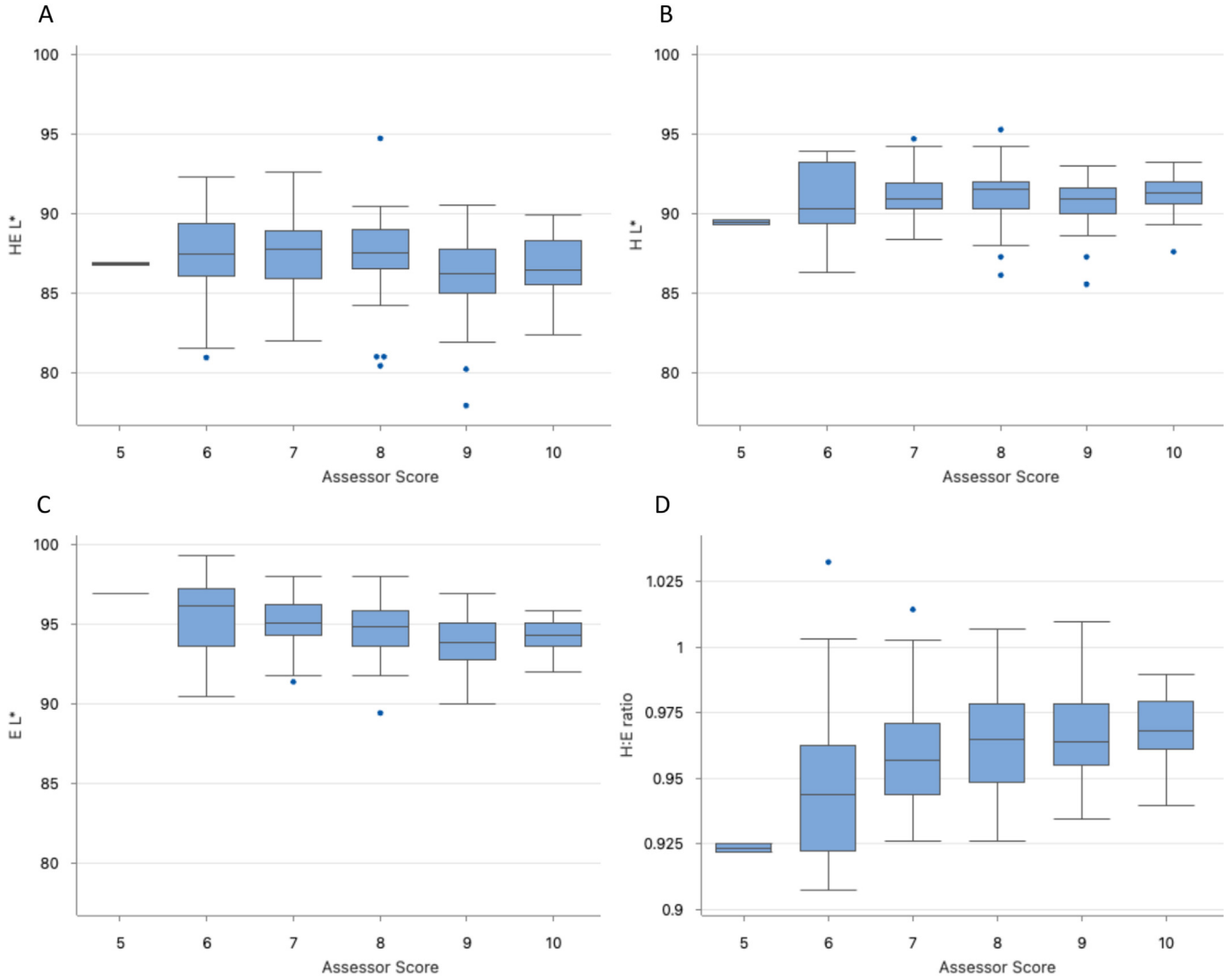


Fig. 6. H&E intensity and assessor score. Box plots comparing the intensity (L* from L*a*b* color space) measured from (A) H&E, (B) deconvolved hematoxylin (H) and (C) deconvolved eosin (E) images, plotted against UK NEQAS CPT assessor score. The box plot in (D) compares hematoxylin to eosin ratio (H:E) with UK NEQAS CPT assessor score. The number of samples (*n*) for assessor scores 5–10 were 2, 19, 55, 67, 61, and 43, respectively.

References

1. Frost & Sullivan. Global Tissue Diagnostics Market, Forecast to 2022. Accessed 16/01/2024: https://store.frost.com/global-tissue-diagnostics-market-forecast-to-2022.html 2018.
2. Gray A, Wright A, Jackson P, Hale M, Treanor D. Quantification of histochemical stains using whole slide imaging: development of a method and demonstration of its usefulness in laboratory quality control. J Clin Pathol Mar 2014;68(3):192–199. https://doi.org/10.1136/jclinpath-2014-202526.
3. Couture HD. Deep learning-based prediction of molecular tumor biomarkers from H&E: a practical review. J Personal Med 2022;12(12):2022. https://doi.org/10.3390/jpm12122022.
4. Pye H, Brettle D, Kaye D, Dunn C, Humphries M, Treanor D. Physical Metrics of Scanner Introduced Variation across 5 Different Makes and Models of WSI Scanner. Accessed 01/06/2024, Available here: https://npic.ac.uk/wp-content/uploads/sites/71/2023/01/HPye-Poster.pdf.
5. Humphries MP, Kaye D, Stankeviciute G, et al. Development of a multi-scanner facility for data acquisition for digital pathology artificial intelligence. medRxiv 2023. https://doi.org/10.1002/path.6326.2023.11. 07.23297408.
6. Duenweg SR, Bobholz SA, Lowman AK, et al. Whole slide imaging (WSI) scanner differences influence optical and computed properties of digitized prostate cancer histology. J Pathol Informatics 2023;14, 100321. https://doi.org/10.1016/j.jpi.2023.100321.
7. Vasiljević J, Feuerhake F, Wemmert C, Lampert T. Towards histopathological stain invariance by unsupervised domain augmentation using generative adversarial networks. Neurocomputing 2021;460:277–291. https://doi.org/10.48550/arXiv.2012.12413.
8. Csurka G. A comprehensive survey on domain adaptation for visual applications. In: *Csurka G, ed. Domain Adaptation in Computer Vision Applications. Springer International Publishing; 2017. p. 1-35.*
9. Ciompi F, Geessink O, Bejnordi BE, et al. The importance of stain normalization in colorectal tissue classification with convolutional networks. IEEE 2017:160–163.
10. Salvi M, Molinari F, Acharya UR, Molinaro L, Meiburger KM. Impact of stain normalization and patch selection on the performance of convolutional neural networks in histological breast and prostate cancer classification. Comput Methods Prog Biomed Update 2021;1, 100004. https://doi.org/10.1016/j.cmpbup.2021.100004.
11. Madusanka N, Jayalath P, Fernando D, Yasakethu L, Lee B-I. Impact of H&E stain normalization on deep learning models in cancer image classification: performance, complexity, and trade-offs. Cancers 2023;15(16):4144.
12. 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.
13. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun 2021;12(1):4423. https://doi.org/10.1038/s41467-021-24698-1.
14. Roy S, Kumar Jain A, Lal S, Kini J. A study about color normalization methods for histopathology images. Micron Nov 2018;114:42–61. https://doi.org/10.1016/j.micron.2018.07.005.
15. James D, Ames D, Lopez B, Still R, Simpson W, Twomey P. External quality assessment: best practice. J Clin Pathol 2014;67(8):651–655. https://doi.org/10.1136/jclinpath-2013-201621.
16. UK NEQAS Cellular Pathology Technique. Staining Criteria Handbook. https://cdn.website-editor.net/8fbe2120dfac419988bb54de58b21798/files/uploaded/NEQMANMA005%20Special%20stain%20criteria%20handbook%20muscle%20edition%20004%20june%2017.pdf.

17. International Organization for Standardization. *ISO 15189: 2022 Medical Laboratories Requirements for Quality and Competence*. ISO. 2022.
18. The UK Accreditation Body (UKAS). Assessment of a Medical Laboratory's Approach to the Assurance of Clinical Staff Competence. Accessed 01/07/2024: https://www.ukas.com/resources/technical-bulletins/technical-bulletin-ukas-position-paper-assessment-of-a-medical-laboratorys-approach-to-the-assurance-of-clinical-staff-competence-and-use-of-eqa/ 2024.
19. International Organization for Standardization. *ISO 17043: Conformity assessment — General Requirements for the Competence of Proficiency Testing Providers*. ISO. 2023.
20. Chlipala E, Bendzinski CM, Chu K, et al. Optical density-based image analysis method for the evaluation of hematoxylin and eosin staining precision. J Histotechnol 2020;43(1):29–37. https://doi.org/10.1080/01478885.2019.1708611.
21. Chlipala EA, Butters M, Brous M, et al. Impact of preanalytical factors during histology processing on section suitability for digital image analysis. Toxicol Pathol 2021;49(4):755–772. https://doi.org/10.1177/0192623320970534.
22. Palus H. Representations of colour images in different colour spaces. In: *Sangwine SJ, Horne REN, eds. The Colour Image Processing Handbook. Springer: US; 1998. p. 67–90.*
23. Clarke EL, Treanor D. Colour in digital pathology: a review. Histopathology 2017;70(2):153–163. https://doi.org/10.1111/his.13079.
24. Bautista PA, Hashimoto N, Yagi Y. Color standardization in whole slide imaging using a color calibration slide. J Pathol Informatics 2014;5(1):4.
25. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol 2001;23(4):291–299.
26. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. IEEE 2009:1107–1110.
27. Sharma G, Wu W, Dalal EN. The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations. Color Res Appl 2005;30(1):21–30. https://doi.org/10.1002/col.20070.
28. Minaker SA, Mason RH, Chow DR. Optimizing color performance of the ngenuity 3-dimensional visualization system. Ophthalmol Sci 2021;1(3), 100054. https://doi.org/10.1016/j.xops.2021.100054.
29. UK NEQAS Cellular Pathology Technique. NEQMANMA034 UK NEQAS CPT Specialist Techniques Assessment Criteria Handbook. Accessed 01/07/2024: www.ukneqascpt.org 2024.
30. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. Sci Rep 2017;7(1), 16878. https://doi.org/10.1038/s41598-017-17204-5.
31. Landini G, Martinelli G, Piccinini F. Colour deconvolution: stain unmixing in histological imaging. Bioinformatics 2020;37(10):1485–1487. https://doi.org/10.1093/bioinformatics/btaa847.
32. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol Aug 2001;23(4):291–299.
33. Corder GW, Foreman DI. *Nonparametric Statistics for Non-statisticians*. John Wiley & Sons, Inc.. 2011
34. Acs B, Fredriksson I, Rönnlund C, et al. Variability in breast cancer biomarker assessment and the effect on oncological treatment decisions: a nationwide 5-year population-based study. Cancers 2021;13(5):1166.
35. Wright KC, Melia J, Moss S, Berney DM, Coleman D, Harnden P. Measuring interobserver variation in a pathology EQA scheme using weighted κ for multiple readers. J Clin Pathol 2011;64(12):1128–1131.
36. Dunn C, Brettle D, Cockroft M, Keating E, Revie C, Treanor D. Quantitative assessment of H&E staining for pathology: development and clinical evaluation of a novel system. Diagn Pathol 2024;19(1):42. https://doi.org/10.1186/s13000-024-01461-w.
37. Wright AI, Dunn CM, Hale M, Hutchins GGA, Treanor DE. The effect of quality control on accuracy of digital pathology image analysis. IEEE J Biomed Health Inform Feb 2021;25(2):307–314. https://doi.org/10.1109/JBHI.2020.3046094.