Article:
Clarke, EL, Revie, C, Brettle, D et al. (6 more authors) (2018) Development of a novel tissue-mimicking color calibration slide for digital microscopy. Color Research and Application, 43 (2). pp. 184-197. ISSN 0361-2317

https://doi.org/10.1002/col.22187

This is the accepted version of the following article: Clarke EL, Revie C, Brettle D, Shires M, Jackson P, Cochrane R, Wilson R, Mello-Thoms C, Treanor D. Development of a novel tissue mimicking color calibration slide for digital microscopy. Color Research and Application (in press).

Development of a novel tissue-mimicking color calibration slide for digital microscopy

Emily L Clarke[1,2] MBBS PGCert(HR), Craig Revie[3] MSc, David Brettle[2] BSc MSc PhD FIPEM, Mike Shires[1], Pete Jackson[2] MPhil CSci FIBMS, Ravinder Cochrane[3] BSc PhD, Robert Wilson[3] BSc, Claudia Mello-Thoms[4] MS PhD, Darren Treanor*[1,2,5] MB BSc(Computing) PhD FRCPath

1. University of Leeds, Leeds, UK
2. Leeds Teaching Hospitals NHS Trust, Leeds, UK
3. FFEI Limited, Hemel Hempstead, UK
4. Faculty of Health Sciences, University of Sydney, Australia
5. Centre for Medical Image Visualisation, Linkoping University, Sweden

Corresponding Author:

Dr Darren Treanor

Leeds Institute of Cancer and Pathology

University of Leeds

Beckett Street

Leeds

LS9 7TF

UK

Tel +44(0)113 20 67546

Email: darrentreanor@nhs.net

Conflict of Interest Statement: Authors CR, RW and RC are employees of FFEI Limited. Author DT is on the advisory board of Sectra and Leica/ Aperio. He receives no personal remuneration for these boards. DT has had a collaborative research project with FFEI, where technical staff were funded by them. He received no personal remuneration. DT is a co-inventor on a digital pathology patent which has been assigned to Roche-Ventana on behalf of his employer. He will receive no personal remuneration.

Abstract

Digital microscopy produces high resolution digital images of pathology slides. Since no acceptable and effective control of color reproduction exists in this domain, there is significant variability in color reproduction of whole slide images. Guidance from international bodies and regulators highlights the need for color standardization.

To address this issue, we systematically measured and analyzed the spectra of histopathological stains. This information was used to design a unique color calibration slide utilizing real stains and a tissue-like substrate which can be stained to produce the same spectral response as tissue. By closely mimicking the colors in stained tissue, our target can provide more accurate color representation than film-based targets, whilst avoiding the known limitations of using actual tissue. The application of the color calibration slide in the clinical setting was assessed by conducting a pilot user-evaluation experiment with promising results.

With the imminent integration of digital pathology into the routine work of the diagnostic pathologist, it is hoped that this color calibration slide will help provide a universal color standard for digital microscopy thereby ensuring better and safer healthcare delivery.

Introduction

Background

Histochemical and immunohistochemical staining of tissue sections on glass slides are the foundation on which diagnoses in pathology is based - an estimated 300 million slides are stained in the US every year [1]. These techniques rely on the selective staining of tissue components which enables pathologists to be able to identify structures to form interpretations and make diagnoses. The most commonly used stain for morphological assessment is hematoxylin and eosin (H&E); a pink and purple stain.

Color consistency is not a major concern with the analogue optical microscope as it uses light to view the tissue directly and the known variation in slide staining has been accepted for many decades [2,3]. Despite these known inconsistencies, there is significant empirical evidence that pathologists can make successful diagnoses from slides with a wide variety of staining variation and using a range of optical microscopes, yet, the common laboratory practice of re-cutting and re-staining referred or 'foreign' slides suggests that pathologists prefer working with familiar colors when possible.

However, unlike light microscopy, digital microscopes and whole slide imaging (WSI) systems as used in digital pathology present issues with regards to color standardization since the pathologist does not observe the tissue directly. Moreover, the digitization process itself involves many stages (Figure 1), each of which substantially affects slide color [4].

[Figure 1 should go approximately here]

The big effect of varying the scanner and viewer independently on the color of H&E slides can be seen in Figure 2. Previous work by our group[2] has shown that there is up to 8% variation in H&E ratio when scanning the same slide on the same scanner on a different day, consistent with a significant visually noticeable difference in color.

[Figure 2 should go approximately here]

Such a wide variation in image appearance has caused concern regarding the ability of the digital image to facilitate accurate interpretation. A review of the issues surrounding color consistency in digital pathology was discussed at the International Color Consortium (ICC) Conference on Color in Medical Imaging [5], which highlighted that many of the issues were based around the absence of an established 'gold standard' for slide color. To try and resolve this issue, an ICC group on color in medical imaging [5,6] was formed. Recent guidance has also been published by the US Food and Drug Administration (FDA) [7], stipulating requirements to ensure that digital microscope images are displayed in a consistent and reliable fashion.

Color Calibration

Color calibration is a technique which seeks to match an array of perceived colors between two or more devices. End-to-end calibration describes the process of controlling color from source to output through each step of the imaging pathway.

In the print industry, end-to-end color calibration has been established for many

decades [8]. It is routinely used, for example, to ensure that the printed pages of a magazine match the original image. Color calibration works by imaging a set of colored patches of a known (measured) color. These colored patches are carefully selected to be representative of the colors encountered by the device being calibrated.

Calibrating a digital camera is challenging since the set of colors encountered when taking photographs varies widely and the importance of color regulation varies between colors. The Macbeth color checker [9,10] (Figure 3) addresses this issue through the inclusion of "memory colors" to ensure accurate reproduction of those critical colors. In a similar method to the print industry, photographers take a photo of the Macbeth color chart and the colors derived from each of the patches in the photo are compared to the 'known' values for those patches. These differences can then be used to derive a color profile which compensates for the color differences of that particular device.

[Figure 3 should go approximately here]

In radiology, where most images are greyscale rather than color, the need for end-to-end calibration of grey levels is recognized as essential to ensure diagnostic consistency of a system. Therefore, image calibration has been incorporated into Digital Imaging and Communications in Medicine (DICOM) Grayscale Standard Display Function (GSDF) requirements for radiological images [11] and has become a standard part of all radiology imaging workflows. Digitization of radiology, nearly 20 years ago, faced many of the same challenges, so it is important to ensure that lessons are learnt from these experiences [12] as digital pathology is integrated into clinical

practice.

Color calibration in pathology

Color calibration in digital pathology is still in its infancy, with a widespread lack of awareness of the importance of consistent color reproduction, or even an assumption that all digital imaging devices reproduce color in essentially the same way. This is evident from the fact that there are very few studies in the literature evaluating the importance of color on diagnosis in digital microscopy [13].

Recently, the FDA have released further recommendations [7] stating that whole slide imaging devices have their color consistency tested with a color calibration slide which would perform the same function as the Macbeth color chart in digital photography. This would involve scanning the color calibration slide on the scanner and comparing the measured color values of the patches with the known values to create a color profile which can be applied to WSI. The guidance also recommends that ideally, the color calibration device should have "similar spectral characteristics to stained tissue"[7].

Recognizing the importance of color reproduction and in an effort to comply with international guidance, several digital pathology vendors and research groups have attempted to develop color calibration systems for pathology images. Silverstein et al developed ICC color calibration profiles for three displays focusing on the display end of the imaging workflow [14] using an on-screen color target. The resulting profiles can be used to compare display performance, but as they do not include measurement of the color variation introduced by the imaging device (WSI instrument or digital

microscope), they cannot provide end-to-end color calibration. Others have used color calibration targets which involves imaging the target with the imaging device, then using the resulting image to develop a color correction profile specific to that device [15–18]. Most of these targets are film-based and although they will almost certainly produce improvements in color consistency between scanners, it is likely that the colors involved (cyan, magenta and yellow film or gel dyes) have substantially different spectra from that of histopathology stains, exacerbated by film also being denser than tissue. The resulting targets are likely to suffer from 'metameric failure' as they will not replicate tissue and stain spectra under all conditions.

To address these limitations, tissue-based color targets have been proposed, such as a section of mouse embryo with standard staining, or cell cultures. Although such targets have the advantage of accurately representing the target material to be imaged and employ spectra appropriate for histochemical staining, producing useful color phantoms based on stained tissue is difficult due to the significant variation between samples.

This paper describes the development of a novel color calibration assessment slide for digital pathology which overcomes the limitations outlined above. The primary objective is to create a target slide that includes a reproducible set of colors that are spectrally similar to the colors produced in stained tissue samples. This will necessitate investigation into the range of colors in tissue slides from representative stained sections, as well as histochemical stain interaction within stained tissue samples.

Materials and methods

Slide preparation and staining protocols

At the start of the project a list of staining protocols was created to represent the gamut of colors used in histopathology and cytopathology practice. The list was generated by a consultant histopathologist (DT) within our institution.

A calendar year of histopathology laboratory staining activity was extracted from the laboratory information system of our institution and the list of staining protocols used was sorted by frequency. This provided information on a total of 258,793 stained slides, of which 79% were H&E, 13% were Hematoxylin-Diaminobenzidine stain (H-DAB), and the remaining 8% comprised 66 different histochemical stains. This practice is likely reflective of similar academic institutions within the UK. The stains were sorted by frequency and those accounting for the top 96% of all stains were included in the list to account for the majority. This was supplemented by the commonly used cytological stains, as well as some additional stains added to include less well-represented parts of the color spectrum (e.g. green and yellow).

Staining strategies were developed in order to capture the color spectrum of each constituent stain without interference from other colors. This involved staining mouse embryo tissue slides using a standard staining protocol (usually with multiple stains) as well as slides stained with each of the stains independently. For example, the H&E staining protocol involved three slides that were sectioned and stained with (i) hematoxylin and eosin, full protocol, (ii) hematoxylin only and (iii) eosin only. For the

Martius Scarlet Blue staining protocol, four slides were sectioned and stained with (i) Martius scarlet blue, full protocol, (ii) Martius yellow only, (iii) Crystal scarlet only and (iv) Aniline blue only. In total, 200 reference slides were stained to represent all the staining protocols. The specific protocols used for tissue processing, sectioning and staining are included in Appendix 1.

Stained sections were recut at varying thicknesses (2 to 12 microns) to investigate the impact of tissue thickness on color.

Slides were visually assessed for quality of staining. For protocols involving multiple stains the overall result was assessed to ensure that all components were correctly stained and well differentiated with regards to intensity of color, and differentiation of staining in different tissue components by the constituent dyes. Where single stains were applied to tissue the intensity of color was the only assessment available.

Identification and measurement of colors

Each of the tissue slides was imaged at 40X magnification and approximately 300-400 spectral measurements were collected for each slide. To understand the range of independent colors in sets of measurements, we grouped measurements using a k-means clustering algorithm. This allowed us to identify measurements of contaminants such as dust in measurements of slides stained with a single stain and to identify measurements of regions with different proportions of stains. The location of each measurement on the slide was recorded so that we could later identify the region measured to determine whether the color was likely to be significant for a pathologist.

Figure 4 shows a typical tissue slide with the regions measured shown.

[Figure 4 should go approximately here]

Our measurement system comprised a digital microscope based on a commercial scanner which uses a custom Leica lens and white light LED illumination source [19] that was adapted to include a Hamamatsu C10082MD spectrophotometer [20]. The optics were adjusted so that the Hamamatsu aperture was aligned with the image plane of the digital microscope. The system was calibrated to align the measurement aperture to the high resolution (40X) image. We developed a software program in MATLAB R2015a [21] which allowed us to (a) scan a high resolution image, (b) analyze the high resolution image to select suitable measurement points and (c) make measurements and show the position and size of the measurement aperture in relation to the high resolution image.

The relative sensitivity across the measurement aperture was calculated by measuring a high-contrast edge in small intervals. This was repeated for horizontal and vertical edges. The high contrast edge was from a phantom slide which is used for geometric scanner calibration.

We calibrated our spectrophotometer using a set of neutral density filters and a Holmium slide for which measurements traceable to the National Physical Laboratory [22] have been provided. The result is shown in Supplementary Figures 1 and 2. For the range 400 – 700 nm, absorbance measurement error overall was less than 0.5%, with

repeatability error less than 0.2 %. The error in wavelength is less than 2nm for this spectral range.

Ethical approval for this work was obtained from Leeds West LREC 05-Q1205-22.

Creation of a color calibration slide

After obtaining results regarding histochemical stain interaction (see results section), we investigated of the ability of a number of materials to retain pathological stains. After many unsuccessful attempts, an appropriate biopolymer was identified.

Sheets of the biopolymer were stained using three different intensities of hematoxylin and three different intensities of eosin, spanning the range of absorbance measured in stained tissue. This was determined by including the maximum intensity for each stain in the tissue slides and selecting two intermediate stain intensity patches with approximately even visual spread. These sets of sheets were cut into 3 strips and then laid on top of each other to produce different color combinations in each overlapping region. This produced 16 H&E color patches. Further sheets of biopolymer were stained with other histopathological stains, including: Silver stain, Picric Acid, Neutral Red, Crystal Scarlet, Tartrazine, Aniline Blue, Fuchsin, Periodic Acid Schiff, Orange G and Light Green.

A prototype slide was manufactured to include a range of 2mm color patches of stained biopolymer, with an extended H&E region and a number of patches including other

stains (Figure 5).

[Figure 5 should go approximately here]

Pilot user-evaluation experiment

This participant blinded, user-evaluation experiment took place in May and June 2015.

Six glass pathology slides with different specimens and tissue stains were selected for evaluation by author DT (Figure 6). The cases were chosen to represent the spectrum of stains, including cases where color is important in diagnosis, as well as those known from previous experience to represent diagnostic difficulty when viewed as a digital slide [23,24]. The slides were scanned using the same commercial scanner at x40 magnification on the same day to produce 6 WSIs.

[Figure 6 should go approximately here]

An ICC color profile was created using our color calibration slide to calibrate each WSI. This involved measuring the colors (CIELAB) on the color calibration slide using a spectrophotometer and then scanning it to generate RGB values. Intermediate color values for the ICC Profile were interpolated from the 'known' values from the color calibration slide. Flare (where light from surrounding areas reduces image contrast) was estimated at 5% and a compensation applied.

Each color calibrated WSI was presented alongside its uncalibrated (i.e. original) counterpart on the experiment monitor. To blind the participants to the intervention, each image pair was assigned labels of 'Image A' and 'Image B'. 'Image A' remained on the left-hand side of the screen and 'Image B' on the right (Figure 7). The positioning of the calibrated or uncalibrated WSI (i.e. whether it was assigned to the 'Image A' or 'Image B' position) was alternated per participant.

[Figure 7 should go approximately here]

All experiments took place in the same windowless room, using the same equipment (Figure 7). The WSIs were viewed using bespoke FFEI software on a Barco Coronis Fusion MDCC6130 monitor [25]. The monitor luminance was set to 400 cd/m^2 and was color calibrated using the X-Rite® i1 Display Pro colorimeter [26] which incorporated the ambient lighting of the room. The settings used for monitor calibration were: CIE Standard Luminant D65, native contrast ratio, gamma 2.2. The lighting was standardized at 7 lux, and placed behind the participant, facing the floor to avoid glare. The display calibration was checked and recorded prior to each experimental session using the X-Rite® i1 Display Pro software, which permitted a maximum error of 15 DeltaE.

A Leica DMR microscope [19] was used during the experiment. This was fitted with a 10 x 25mm eyepiece and 2.5, 5, 10, 20, 40x dry objectives. Polarizer and filter settings were standardized, but participants could adjust the focus, interpupillary distance and substage condenser position. The daylight filter was applied and the rheostat was standardized at 7 where possible, as the light yellowed at lower light levels. However,

participants could adjust this if they found it uncomfortable. A script was used throughout each experimental session.

We invited all (approximately 40) consultant histopathologists employed at Leeds Teaching Hospitals NHS Trust at the time of the experiment, to take part. Informed, written consent was obtained. Participants compared glass slides using the optical microscope with the corresponding WSI, calibrated and uncalibrated, of the same tissue specimen. Details of the questions asked are outlined in Figure 8.

[Figure 8 should go approximately here]

When the glass slides were not being used, they were kept in an opaque envelope to help prevent fading.

Data were analyzed using SPSS v21 (IBM). Significance level was set at 0.05. Wilcoxon Signed Rank test was used to compare data and Kendall's Tau-b correlates are presented.

Results

Stain interactions

We examined the spectral measurements of the slides involved in each staining protocol, as described above. An example of the data obtained from the H&E staining protocol can be seen in Figure 9. By measuring the spectral absorbance of samples

with different staining levels separately and then in combination we conclude that the total spectral absorbance of hematoxylin, eosin and H&E is dependent only on the amount of stain present and therefore follows Beer-Lambert's law. As such, we are able to assume that the spectral absorbance is linearly proportional to the amount of staining. It is also apparent from the spectral data in Figure 9 that there is no interaction between hematoxylin and eosin stains and that they combine in a simple linear way. This knowledge enabled the layering of biopolymer, independently stained with eosin and hematoxylin, on top of one another, to create a representative gamut for H&E stained slides.

[Figure 9 should go approximately here]

All other stain combinations acted similarly to H&E in that they obey Beer-Lambert's Law and so were combined through simple addition. The only exception was HDAB, which has different spectral properties due to selective light scatter and therefore does not follow Beer-Lambert Law. However, in reality, our samples demonstrated that HDAB can be simulated to a similar accuracy achieved by other stains by including a number of patches stained with different intensities of DAB. Figure 10 shows the result of spectrophotometric analysis of HDAB absorbance and its modelling using combinations of H and DAB.

[Figure 10 should go approximately here]

Comparison of different thickness sections from 2 to 12 microns, showed that the shape of the spectrum is the same for the different thicknesses of material (see

Supplementary Figure 3).

Theoretical performance assessment

It is well known that when calibrating devices, the calibration target should include colors with similar spectra to those that the device has been designed to capture. Since our calibration slide uses the same set of stains as are used to stain tissue we hypothesize that their spectra are likely to be very similar. To confirm this assumption, we compared spectral measurements of hematoxylin- and eosin-stained biopolymer patches with measurements from mouse embryo slides stained with hematoxylin and eosin single stains (Figure 11a). Typical spectral data for hematoxylin and eosin (normalized by peak absorbance) are shown in Figure 11b. These and similar comparisons with the other stains used on the slide demonstrate that the spectra of stained tissue and the stained biopolymer patches used on our color calibration slide are very similar.

[Figure 11a&b should go approximately here]

For comparison, we examined the spectra from a photographic film color calibration slide (HutchColor HCT Ektachrome™ 35mm) which is one method that has been proposed for digital microscope calibration. These spectra, which arise from the set of three dyes used in the photographic process, are substantially different in shape from the spectra of pathology stains. This can be seen from the comparison in Figure 12 which shows the spectrum of a photographic film patch and a tissue color measurement. In our scanner, these two spectra produce the same scanner RGB

values but are visibly different, with a color difference of 5 CIEDE2000. This color difference is essentially a systematic error which cannot be removed by the calibration process.

[Figure 12 should go approximately here]

A comparison between color measurements of the photographic film calibration slide and color measurements of H&E-stained pathology slides is shown in Figure 13. This shows that the entire gamut for eosin and much of that for hematoxylin lies outside that of the photographic film color gamut and consequently photographic film color calibration slides are unable to predict any of the eosin colors. As there is no way to estimate colors outside the gamut of a calibration target, some form of extrapolation is necessary; this is likely to be very large as it essentially involves guessing these colors and values greater than 10 CIEDE2000 units are not uncommon in these situations.

[Figure 13 should go approximately here]

Although it is not possible to estimate precisely the difference in calibration accuracy between these two methods, our experience shows that the errors for a photographic film calibration target are likely to be several times larger than for our calibration slide.

Pilot user-evaluation experiment

Twelve participants took part in the pilot user-evaluation experiment who performed 72 trials in total (6 slides per participant). The mean working experience in pathology for the cohort was 21.5 years. Most pathologists self-reported to have seen 100-1000 WSIs prior to the experiment. Participants' opinion of digital pathology varied; when asked to rate how positively they felt about digital pathology on a scale of 0-10 (0 = very negative; 10 = very positive), the median score was 8 (IQR 3.25).

The monitor color calibration check using the X-Rite® i1 Display Pro had an average of 2.6 DeltaE across all experimental sessions.

The calibrated WSIs were reported to be closer in appearance to the optical microscope in 56% of trials (40/72) (Figure 14). On average over the 6 slides, seven pathologists (58%) thought that the color calibrated WSIs was more similar to the microscope.

[Figure 14 should go approximately here]

Regarding pathologist preference, 64% of trials (46/72) were in favor of the calibrated slide (Figure 14). On average over the six slides, nine pathologists (75%) favored the color calibrated WSIs.

The calibrated slide had no effect on diagnostic confidence in 53% of trials (38/72), but 38% of trials (27/72) found the calibrated slide improved diagnostic confidence. On average over the 6 slides, all pathologists reported that the calibrated slide increased or had no effect on diagnostic confidence. Overall, calibrated WSIs had a

higher median diagnostic confidence score as compared to the median diagnostic confidence score for the uncalibrated WSIs (Figure 15). Increased diagnostic confidence with the calibrated slides was correlated with preference for color calibration (r=0.499).

[Figure 15 should go approximately here]

There was no statistical difference between each of the six cases for all three main outcomes; similarity to microscope, pathologist preference or diagnostic confidence. Similarly, there was no difference in scores for each main outcome when the six cases were categorized into H&E slides and non-H&E slides (Figure 16).

[Figure 16 should go approximately here]

Discussion and Conclusion

Accurate color reproduction is an essential next step in digital pathology. It will enable the adoption of image analysis/ computer aided diagnosis, increased confidence in the suitability of whole slide imaging for clinical use (both by clinicians and regulators), and possibly improve quality or speed of diagnosis. To date, no practical and robust method for color calibration in digital pathology exists.

There are several repositories for stain measurement data such as Color Index International (http://www.color-index.com/) and Stains File (http://www.stainsfile.info/) which provide details of the stains as measured in isolation. However, we believe we

have produced the first systematic evaluation of the way in which stains interact with tissue samples and with each other.

We can confirm that the spectral absorbance of histochemical stains is linearly proportional to the amount of staining and therefore we can use a few differing concentrations of H&E stain in our calibration device to create a full gamut that is very like the gamut of hematoxylin and eosin from stained tissue in clinical practice. Moreover, since stains interact through simple addition, sheets of stained biopolymer can be layered without impacting on the spectral properties. This detailed knowledge of stain spectra has enabled creation of a color calibration slide that has similar spectral characteristics to stained tissue.

We have also been able to demonstrate that film-based color calibration slides have differing spectral characteristics to stained tissue with typical metameric error of 5 CIEDE2000 units. Additionally, there is color error of approximately 10 CIEDE2000 when correcting for colors that lie outside the film color calibration slide color gamut, which includes the entire gamut for eosin. Therefore, the use of film-based color calibration targets for digital microscopy would result in substantial color error despite correction.

We observed that an increase in the tissue thickness did not change the shape of the spectrum. This is an important finding as the accuracy of the color calibration derived from our color calibration slide is maintained despite variation in tissue section thickness.

Our findings from our clinical pilot indicate that calibration with the color calibration slide made the WSI closer in appearance to the microscope in over 50% of trials, which was a weaker effect than hypothesized. A possible reason for this may be that since pathologists are used to seeing uncalibrated WSIs (and several participants had seen large numbers of uncalibrated slides previously), they are 'self-calibrating' to this standard and thus simply recalling the colors that they are used to seeing [27]. This was illustrated very well by one participant: *'I am used to the (uncalibrated image), so to me it looks closer to the microscope'*. This reconfirms the need for color standardization to prevent pathologist adjustment to varying color profiles.

Color calibration did not affect similarity to the microscope, as strongly as preference. These results question the underlying assumption that 'the ideal' color for WSIs is the glass slide as viewed under the optical microscope. Perhaps there are more essential key factors when optimizing WSI color that have not yet been determined, for instance flare, illumination or color temperature.

Although the findings suggest that color calibration may improve diagnostic confidence, it should be emphasized that this study was not designed as a diagnostic accuracy study; the cases did not represent diagnostic subtleties and moreover, the diagnoses were provided to participants. This will inevitably have biased the results in favor of calibration having no effect on diagnostic confidence, so the true effect of calibration on diagnosis may be more pronounced than observed in this small study. It is also important to note, that diagnostic confidence may not be an accurate surrogate marker of diagnostic accuracy [28,29] and so follow on work should focus on a bigger study formally evaluating the effect of color calibration on diagnosis.

The link between pathologist preference and diagnostic confidence is a further area of interest. This study demonstrated a positive correlation between preference for, and diagnostic confidence with, the calibrated WSI. Whether this reflects a true benefit in the fitness-for-purpose of the image is not known. There may of course be other benefits with increased preference for an image aside from diagnostic accuracy, extending to user fatigue or speed of diagnosis. This is clearly a question for another follow-on study.

As previously mentioned, there is little published work in this area, with only a couple of studies having investigated the clinical effect of color calibration [28,30]. Krupsinki et al [28] concluded that monitor color calibration did not affect diagnostic accuracy, but did statistically increase average interpretation speed by 22 percent when viewing breast biopsy regions of interest . Hanna et al 2014 [30] found that pathologists preferred using an uncalibrated monitor and the group did not demonstrate an effect of color calibration on diagnosis. Our results did not agree with the findings of Hanna et al and this discrepancy may emphasize that (i) the effect of memory on color preference and (ii) the need for end-to-end color calibration using a color calibration slide, rather than monitor calibration alone, which does not correct for color changes introduced through the scanning process.

The main strengths of the research we present here is the high degree of methodological rigor requiring expertise in pathology, color science, imaging and cognitive psychology, as well as the large number of samples used to create the color calibration target.

Despite this, there are inevitable limitations which require discussion. Firstly, the choice of stains included reflects UK histopathology and cytopathology practice, which we have assumed to be broadly representative of worldwide pathology practice. However, we recognize that there may be significant international variability in both the stains used and the exact chemical composition of the dyes. We have made an attempted to account for this through analysis of Meyer's and Harris hematoxylin, but it is likely that there will be further, unaccounted variation. Additionally, whilst the color calibration target can be used to calibrate any digital slide scanner, further development will likely be needed to manufacture this device on a large scale. Also, although slides stained with HDAB showed relatively low color error, ideally further work should be done and additional patches should be included to eliminate any errors due to metamerism. Finally, the limited exploratory nature of the clinical pilot limits the applicability of the results, both relating to the small number of slides assessed, as well as the small sample size – a larger study focused on the diagnostic impact of color calibration is required to draw firm conclusions.

A prototype color calibration slide was tested by members of the International Color Consortium Medical Imaging Working Group[6] and a production version is expected to available commercially within the next year from FFEI Limited and industrial partners.

We are confident that, when refined, this approach will:

1. Provide a solid basis for end-to-end color calibration assessment for digital pathology,
2. Facilitate interoperability between digital microscope systems and between

pathology labs image transfer,

3. Increase reproducibility of automated image analysis in whole slide imaging,
4. Improve clinical efficiency and accuracy of digital microscopy and whole slide imaging, and
5. Facilitate work towards truly quantitative immunohistochemistry.

Acknowledgements

The authors thank Mrs. Doreen Crellin for her advice and assistance with this work.

Funding support for this work was provided from: 1. UK Technology Strategy Board, 2. Leeds Teaching Hospitals Charitable Trustees.

Author contributions:

- Study design: EC, CR, MS, DB, RC, DT
- Conducted the experiments: EC, CR, MS, PJ, RC, RW, DT
- Analyzed the data: EC, CR, DB, RC, DT, CM-T
- Wrote the manuscript: EC, CR, MS, PJ, DB, RC, RW, CM-T, DT

References

1. Little E. What difference does a difference make? In: Pathology Visions. San Diego; 2010.
2. Gray A, Wright A, Jackson P, Hale M, Treanor D. Quality control of histochemical stains: repeatable and objective quantification using virtual slides [Oral presentation]. 2013.
3. Gray A, Wright A, Jackson P, Hale M, Treanor D, James S. Quantification of histochemical stains using whole slide imaging: development of a method and demonstration of its usefulness in laboratory quality control. J Clin Pathol. 2015;68:192-199. doi:10.1136/jclinpath-2014-202526.
4. Badano A, Revie WC, Casertano A, et al. Consistency and standardization of color in medical imaging: a consensus report. J Digit Imaging. 2015;28:41-52.
5. Revie WC, Jackson P, Cheng WC, et al. International Color Consortium. ICC Draft White Paper: Digital Microscope Test Materials and Test Methods, Document Version 3. International Colour Consortium
6. International Colour Consortium Medical Imaging Working Group. http://www.color.org/groups/medical/medical_imaging_wg.xalter.
7. FDA. Technical Performance Assessment of Digital Pathology Whole Slide Imaging Devices; Guidance for Industry and Food and Drug Administration Staff. 2016.
8. Carrein G. White Paper Characteristics of CRT and LCD displays. White Pap.
9. McCamy C, Marcus H, Davidson J. A color-rendition chart. J App Photog Eng. 1976;2:95-99.
10. X-Rite. ColourChecker Classic for Image Reproduction.; 2014.
11. Digital Imaging and Communications in Medicine (DICOM) Part 14 : Grayscale Standard Display Function. Natl Electr Manuf Assoc.

2004;10(S1):3-4. doi:10.1007/BF03168637.

12. Krupsinski E. Virtual slide telepathology workstation of the future: lessons learned from teleradiology. Hum Pathol. 2009;40(8):1100-1111.

13. Clarke E, Treanor D. Colour in Digital Pathology: A Review. Histopathology. 2017;70(2):153-163.

14. Silverstein L, Hashmi S, Lang K, Krupinski E. Paradigm for achieving colour-reproduction accuracy in LCDs for medical imaging. J Soc Inf Disp. 2012;20:53-62.

15. Cheng W-C, Keay T, O'Flaherty N, et al. Assessing color reproducibility of whole-slide imaging scanners. 2013;8676:86760O. doi:10.1117/12.2007215.

16. Shrestha P, Hulsken B. Colour accuracy and reproducibility in whole slide imaging scanners. SPIE Digit Libr. 2014;1. doi:doi:10.1117/12.2048769.

17. Bautista P, Hashimoto N, Yagi Y. Color standardization in whole slide imaging using a color calibration slide. J Pathol Inform. 2014;5:4. doi:10.4103/2153-3539.126153.

18. Wei H, Brill MH, Park T. Evaluation of Targets for Color Calibrating Digital Images from an Optical Bright-Field Transmission Microscope. Colour Res Appl. 2014:1-8. doi:10.1002/col.21932.

19. Leica Microsystems (UK) Limited.

20. Hamamatsu Photonics UK. 2014.

21. Incorporated TM. MATLAB.

22. National Physical Laboratory. http://www.npl.co.uk/about/what-is-npl/.

23. Randell R, Ruddle RA, Thomas RG, Mello-Thoms C, Treanor D. Diagnosis of major cancer resection specimens with virtual slides: impact of a novel digital pathology workstation. Hum Pathol. 2014;45(10):2101-2106.

doi:10.1016/j.humpath.2014.06.017.

24. Randell R, Ruddle RA, Mello-Thoms C, Thomas RG, Quirke P, Treanor D. Virtual reality microscope versus conventional microscope regarding time to diagnosis: an experimental study. Histopathology. 2013;62(2):351-358. doi:10.1111/j.1365-2559.2012.04323.x.

25. Barco Limited.

26. Xrite Limited.

27. Bartels A, Zeki S. The architecture of the human color center. Eur J Neurosci. 2000;12(1):172-193.

28. Krupinski EA, Silverstein LD, Hashimi SF, Graham AR, Weinstein RS, Roehrig H. Observer Performance Using Virtual Pathology Slides: Impact of LCD Color Reproduction Accuracy. J Digit Imaging. 2012;25(6):738-743. doi:10.1007s/10278-012-9479-1.

29. Ng CS, Palmer CR. Analysis of diagnostic confidence and diagnostic accuracy: a unified framework. Br J Radiol. 2007;80(951):1550-1562.

30. Hanna MG, Monaco SE, Ahmed I, Parwani A V, Pantanowitz L. Impact of Monitor Color Calibration on Digital Pathology Interpretation. Pathol Informatics. 2015.

Figures and Figure Legends

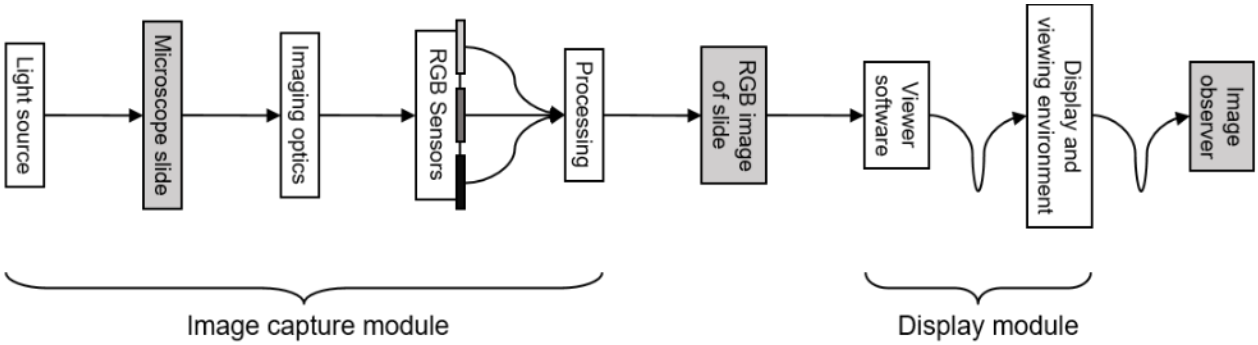
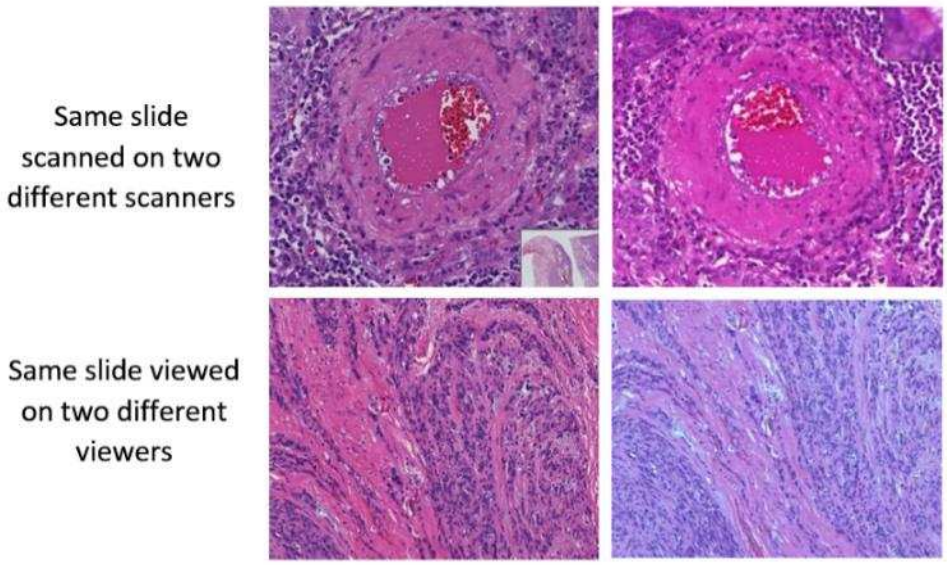


Fig 1: Imaging workflow from sample to display. A microscope slide sample is imaged to produce a digital image. The workflow illustrates the interaction of the image capture and display components to produce the final image observed by the user. Alterations at any point in this workflow can affect the final image observed. This figure was originally published in [13].



Reproduced with permission from Dr Yukako Yagi, MGH

Fig 1 [COLOR]: Examples of color variation

The top row shows a WSI scanned on one scanner but viewed in two different software viewers. The image on the bottom row shows the results of the same glass slide scanned on scanners from two different manufacturers. A marked difference in the color of both eosin and hematoxylin is seen in both cases. These differences illustrate the effect of uncontrolled color variation in the image capture and image display components of the workflow. Fig 2 is reproduced with permission from Dr Yukako Yagi, MGH.

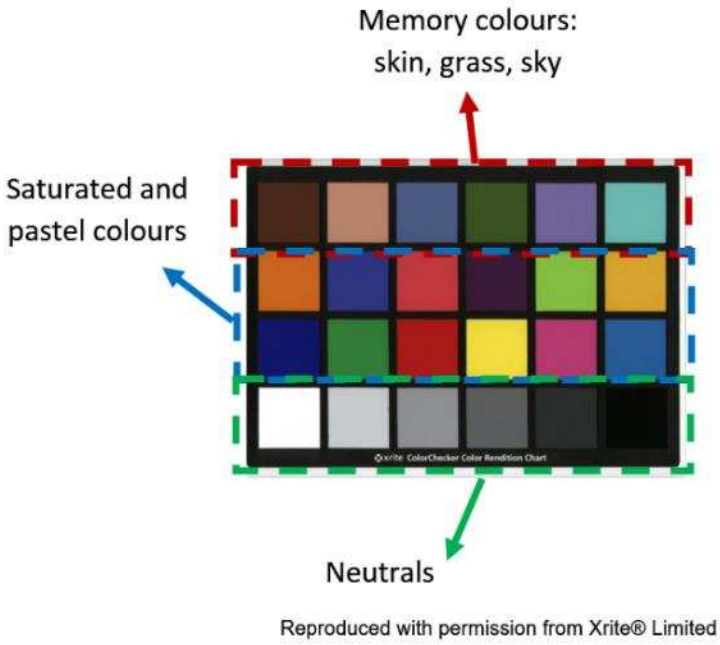


Fig 3 [COLOR]: Standard methods to calibrate color in digital imaging.

This standard Macbeth ColorChecker is used in color calibration of devices in other domains. It includes a variety of colors to illustrate those in the domain which are most important (skin, sky, grass), a range of different color hues (saturated and pastel) and a range of neutral colors from white through grey to black. Fig 3 is reproduced with permission from Xrite® Limited, Cheshire, UK.

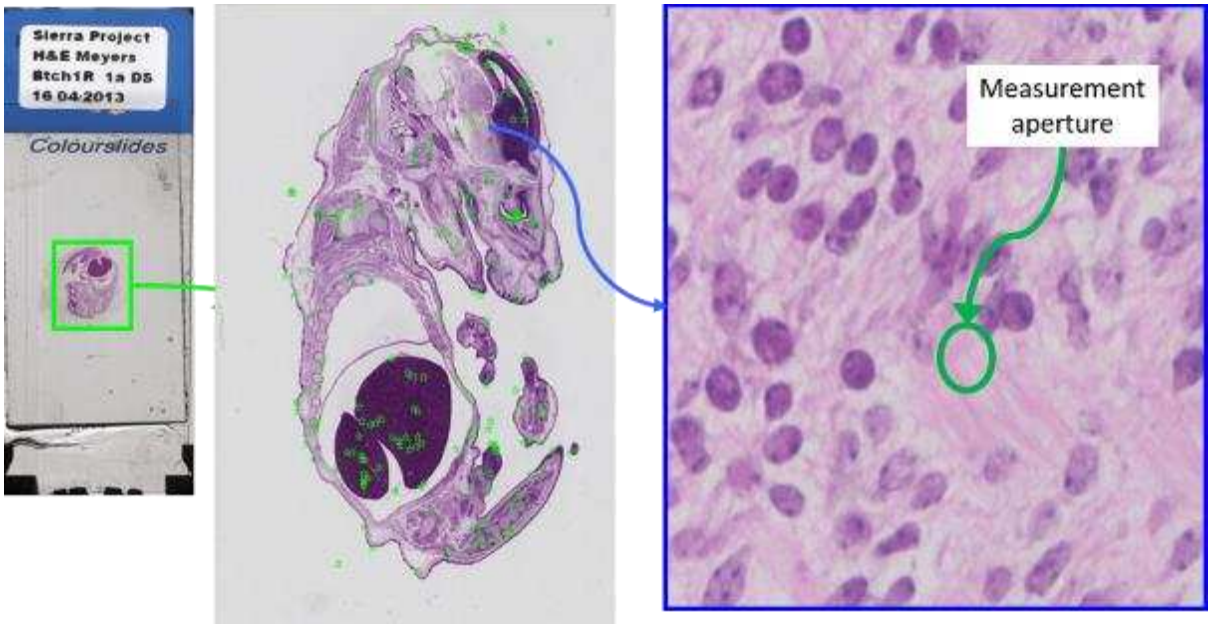


Fig 4 [COLOR]: Illustration of measurement method. The glass slide (left) was loaded onto the measurement apparatus. In the example shown, the glass slide contains an H&E stained section of mouse embryo. A low resolution whole slide image ("overview image") was taken (middle). This image was used to guide measurements, indicated by the green stars on the tissue. The measurement aperture, illustrated on the right, was small enough to allow measurement of microscopic areas of tissue. Measurements were taken across the entire sample to generate measurements which are representative of all the colors and intensities present on the slide.

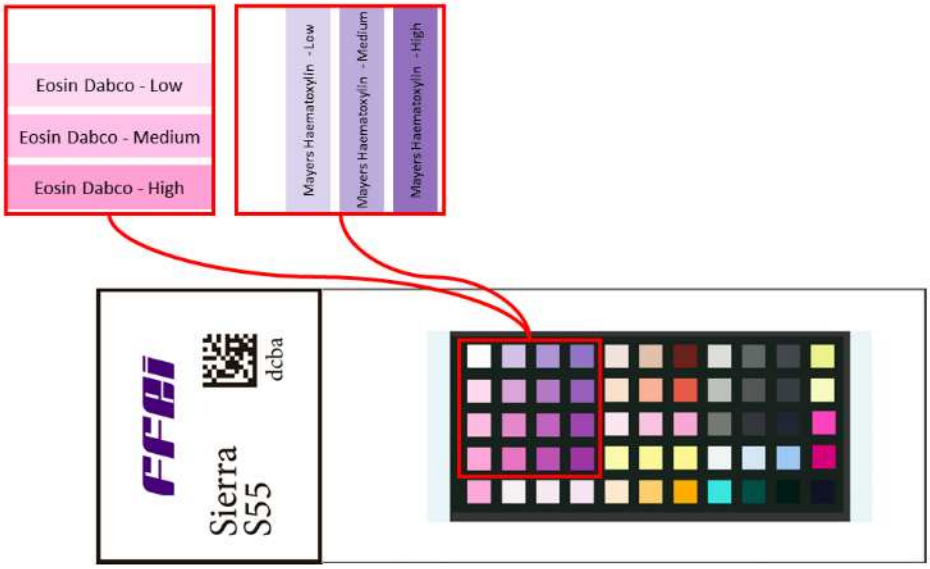


Fig 5 [COLOR]: Color calibration slide design. The H&E stain area is created using 3 strips of increasing Eosin intensity and 3 strips of increasing hematoxylin intensity. These strips are overlaid to provide 16 patches of controlled H&E intensity. The resulting H&E stain area is incorporated onto a glass slide, along with color patches stained with other histochemical stains.

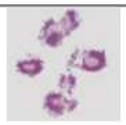
Slide Thumbnail	Tissue	Stain	Diagnosis/ Task
	Duodenal biopsy	H&E	In this normal biopsy, assess the villi and intra-epithelial lymphocytes
	Axillary lymph node	H&E	Assess the micrometastasis
	Breast tissue	ER nuclear immunostain	Look at 3 levels of expression of ER
	Cervical spin	PAP stain	Cervical spin with high grade dyskaryosis - assess the normal squamous cells and dyskaryotic cells
	Lung	ZN stain	Identify the ZN positive organisms
	Bowel	Gram stain	Look at the gram positive and gram negative bacteria

Fig 6. Details of the six tissue cases and recommended tasks included within the pilot user evaluation.

Fig 7 [COLOR]. Pilot user-evaluation system setup.

Question	Scale
Pertaining to each of the 6 tissue slides in turn: *"Which digital image is closer in appearance to the microscope slide?"*	1-7 Likert (1 = Image A is closer; 4 = They are as close as each other; 7 = Image B is closer)
Pertaining to each of the 6 tissue slides in turn: *"Which digital image do you prefer for diagnosis?"*	1-7 Likert (1 = I prefer Image A; 4 = I find them to same; 7 = I prefer Image B)
Pertaining to each of the 6 tissue slides in turn: *"How confident would you feel in your diagnosis from Image A, as compared to the microscope?"*	1-7 Likert (1 = I am definitely not as confident as the microscope; 4 = I am not as confident as the microscope; 7 = I am as confident as the microscope)
Pertaining to each of the 6 tissue slides in turn: *"How confident would you feel in your diagnosis from Image B, as compared to the microscope?"*	1-7 Likert (1 = I am definitely not as confident as the microscope; 4 = I am not as confident as the microscope; 7 = I am as confident as the microscope)

Fig 8: Pilot user-evaluation experiment questions

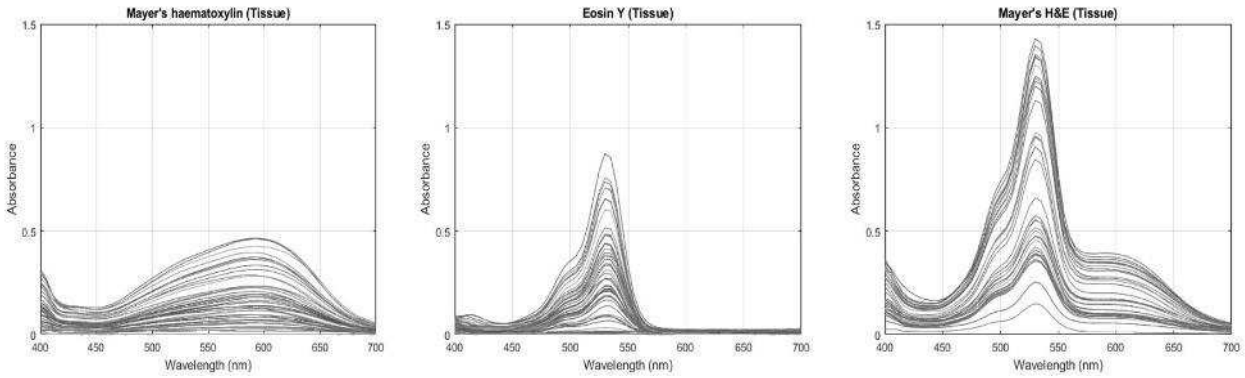


Fig 9 [COLOR]: Spectral results of tissue stained with hematoxylin only, eosin only and H&E

The left graph shows the spectra for tissue stained with hematoxylin only. The middle graph shows the spectra for tissue stained with eosin only. The right graph shows spectra for tissue stained with H&E. The total spectral absorbance of hematoxylin, eosin and H&E is dependent only on the amount of stain present and therefore H&E follows Beer-Lambert's law. Hence we can assume that the spectral absorbance is linearly proportional to the amount of staining. It is also apparent that there is no interaction between hematoxylin and eosin stains and that they combine in a simple linear way; the hematoxylin and eosin peaks are discernible from the H&E spectra.

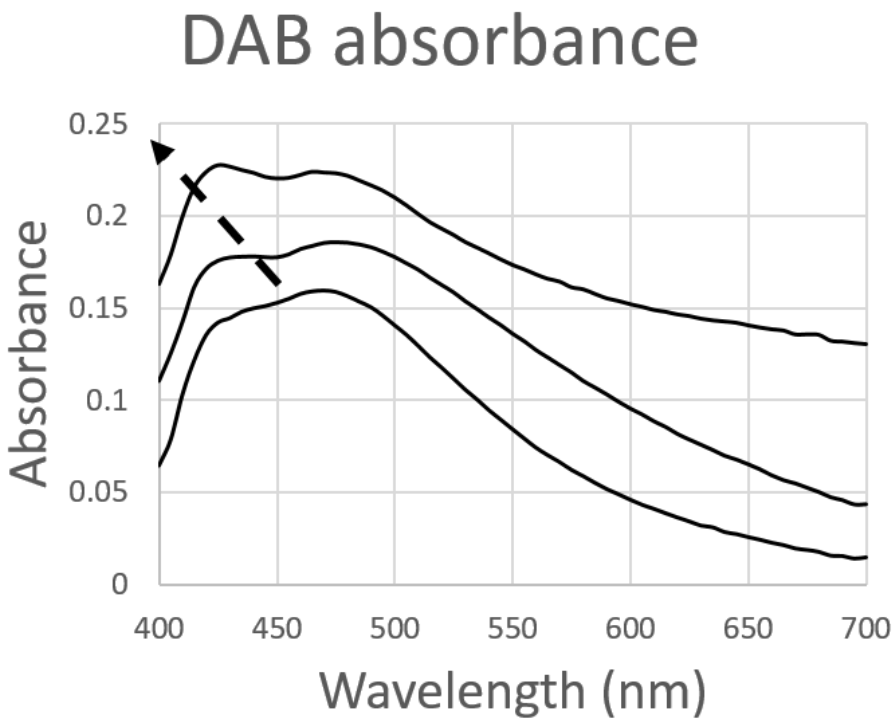


Fig 10: DAB measurement and modelling. This figure shows spectra obtained by taking multiple measurements of a slide stained with hematoxylin and DAB (only 3 measurements are shown, for illustration purposes). The absorbance peak shifts to the left as the amount of DAB increases (illustrated with a dashed arrow).

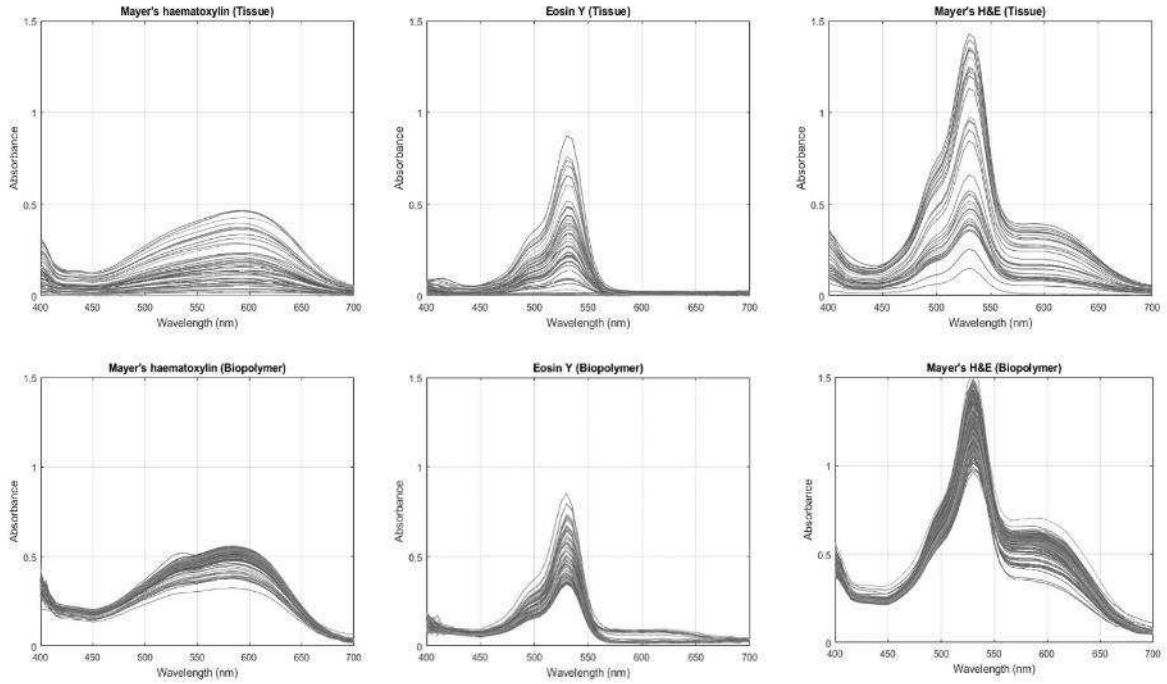


Fig 11a: H&E spectral results of assessment of the color calibration slide as compared to stained tissue.

The top row of graphs show the spectra obtained for stained tissue as shown in Figure 9 (top left = hematoxylin only, top middle = eosin only, top right = H&E). The bottom row of graphs show the spectra obtained from measuring the biopolymer (bottom left = hematoxylin only, bottom middle = eosin only, bottom right = H&E).

The range of spectra for hematoxylin only, eosin only and H&E obtained by measuring the color calibration slide are very similar to those obtained from stained tissue.

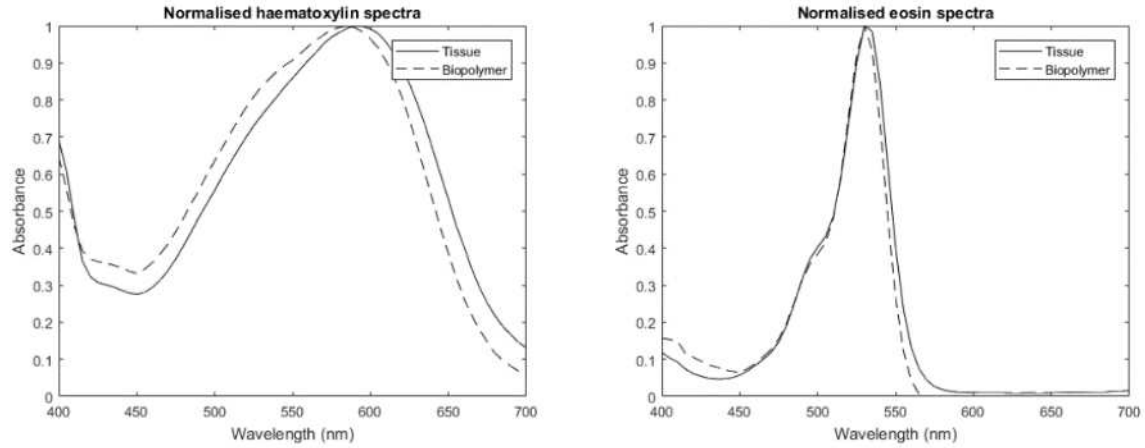


Fig 11b: A direct comparison of typical absorbance spectra, in biopolymer and in tissue. The graph on the left is for tissue and biopolymer stained with hematoxylin only and the graph on the right is for tissue and biopolymer stained with eosin only, with striking similarity in each. These spectra are normalized by peak absorbance.

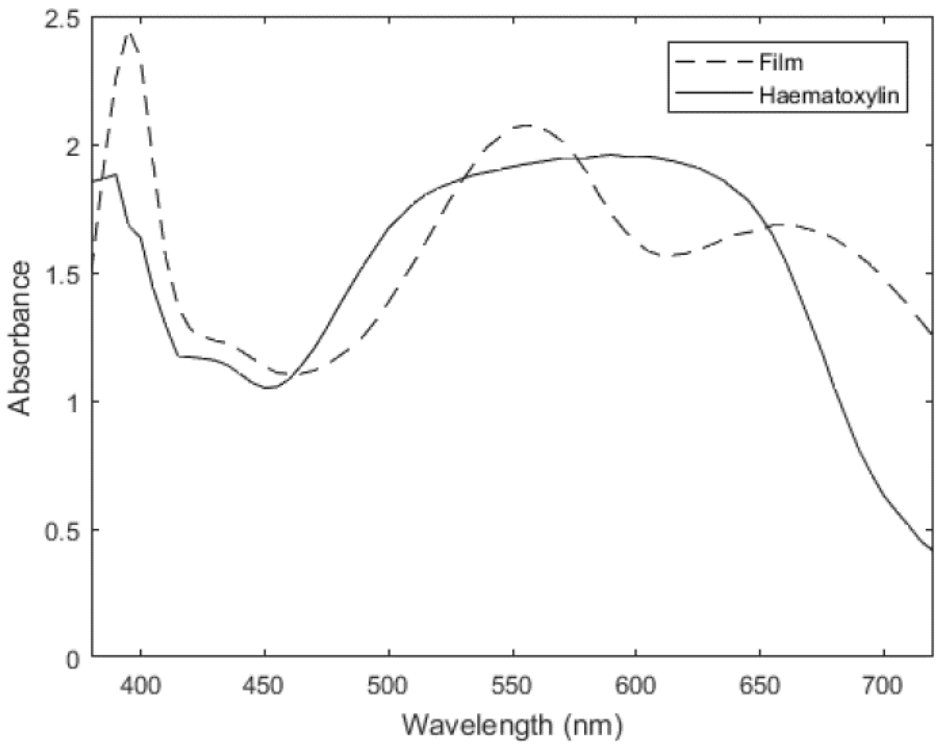


Fig 12. A comparison of film spectra and stained tissue. These patches have the same RGB values (48,43,69) but have a CIEDE2000 color difference of 5.3. This is systematic error and cannot be removed through the calibration process.

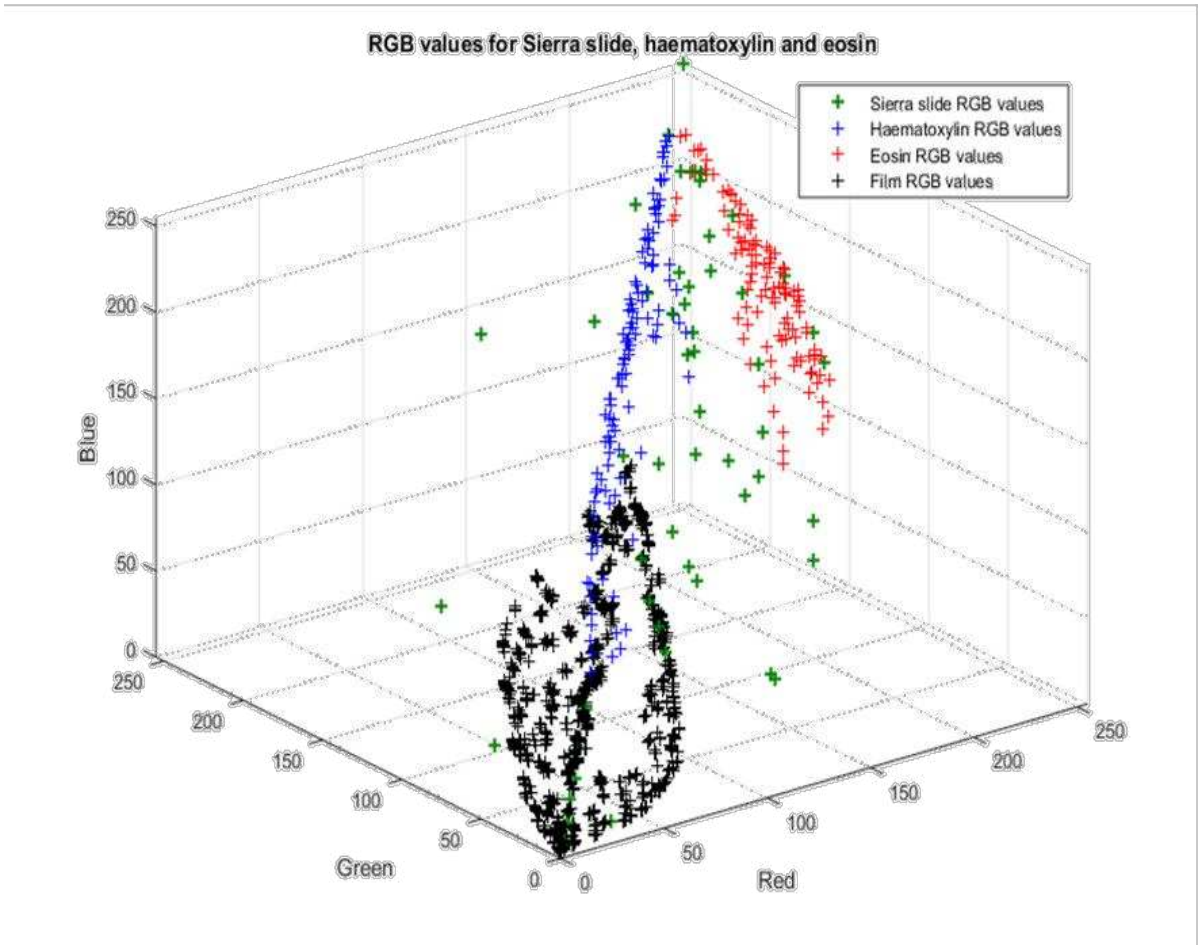


Fig 13. [COLOR] A comparison of RGB values for film (black), hematoxylin-stained tissue (blue), eosin-stained tissue (red) and the color calibration slide (green). The eosin RGB values do not overlap with the film RGB values and therefore film-based targets cannot accurately capture the color of eosin. Therefore, the use of film-based targets would result in color variation between images despite calibration. The spectra of the novel color calibration slide overlaps with the eosin and hematoxylin RGB values and therefore color calibration using the novel color calibration slide enables more accurate color calibration.

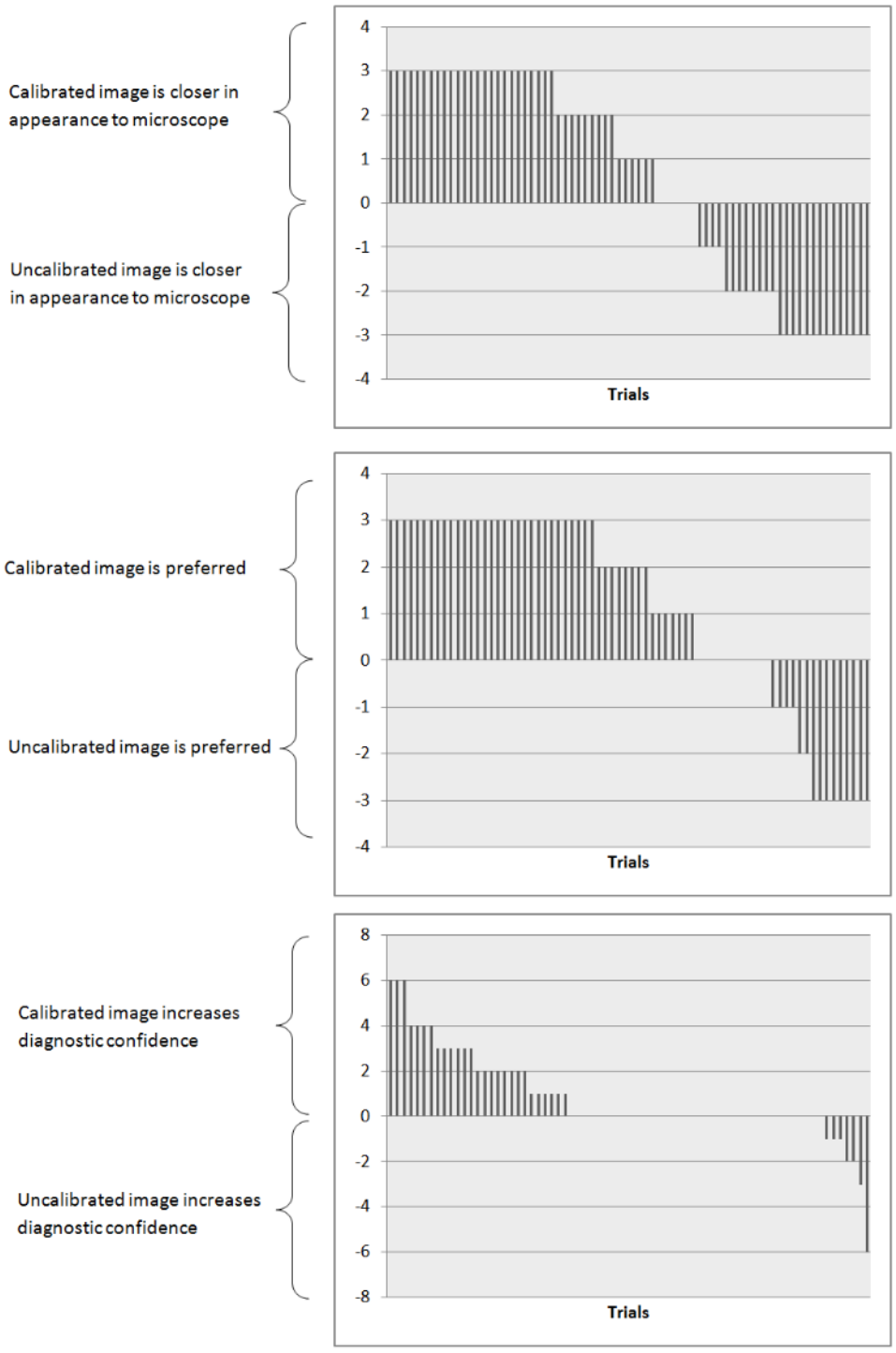


Fig 14. Scores by trial for each of the three main outcomes. Each bar represents one trial, ordered by score.

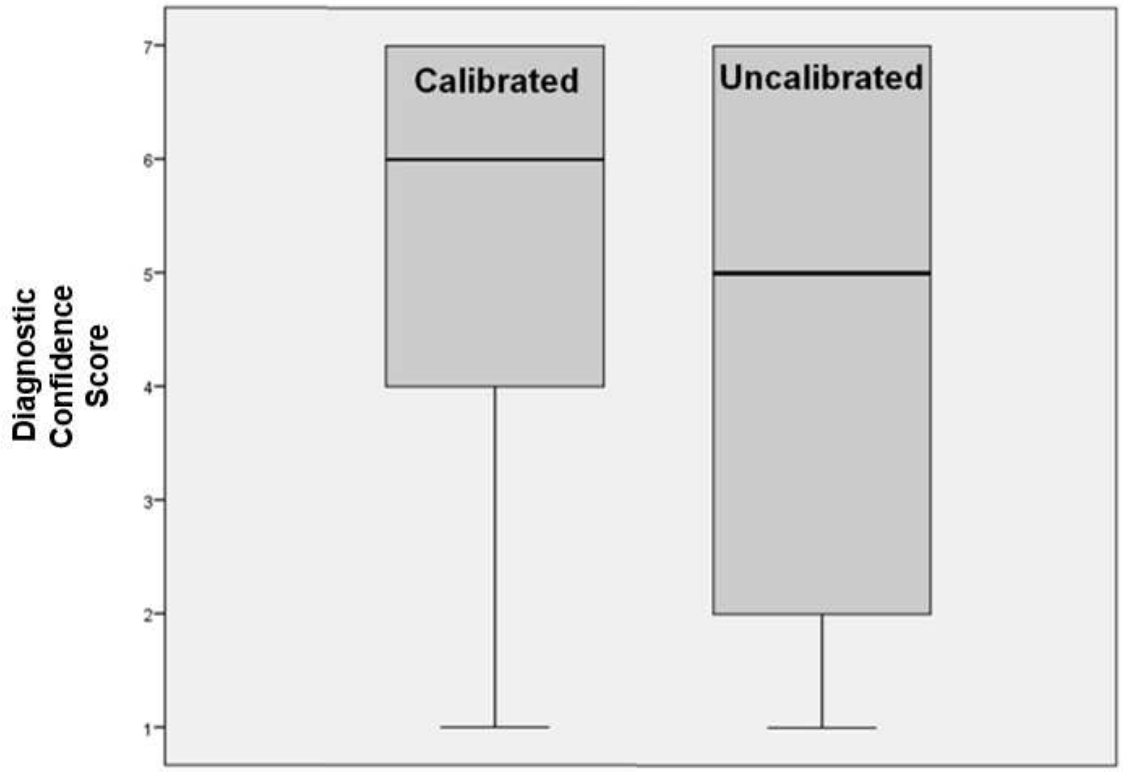


Fig 15. Median and IQR diagnostic confidence scores for calibrated slides versus uncalibrated slides. A score of 7 denotes 'as confident as the microscope'. A score of 1 denotes 'definitely not as confident as the microscope'.

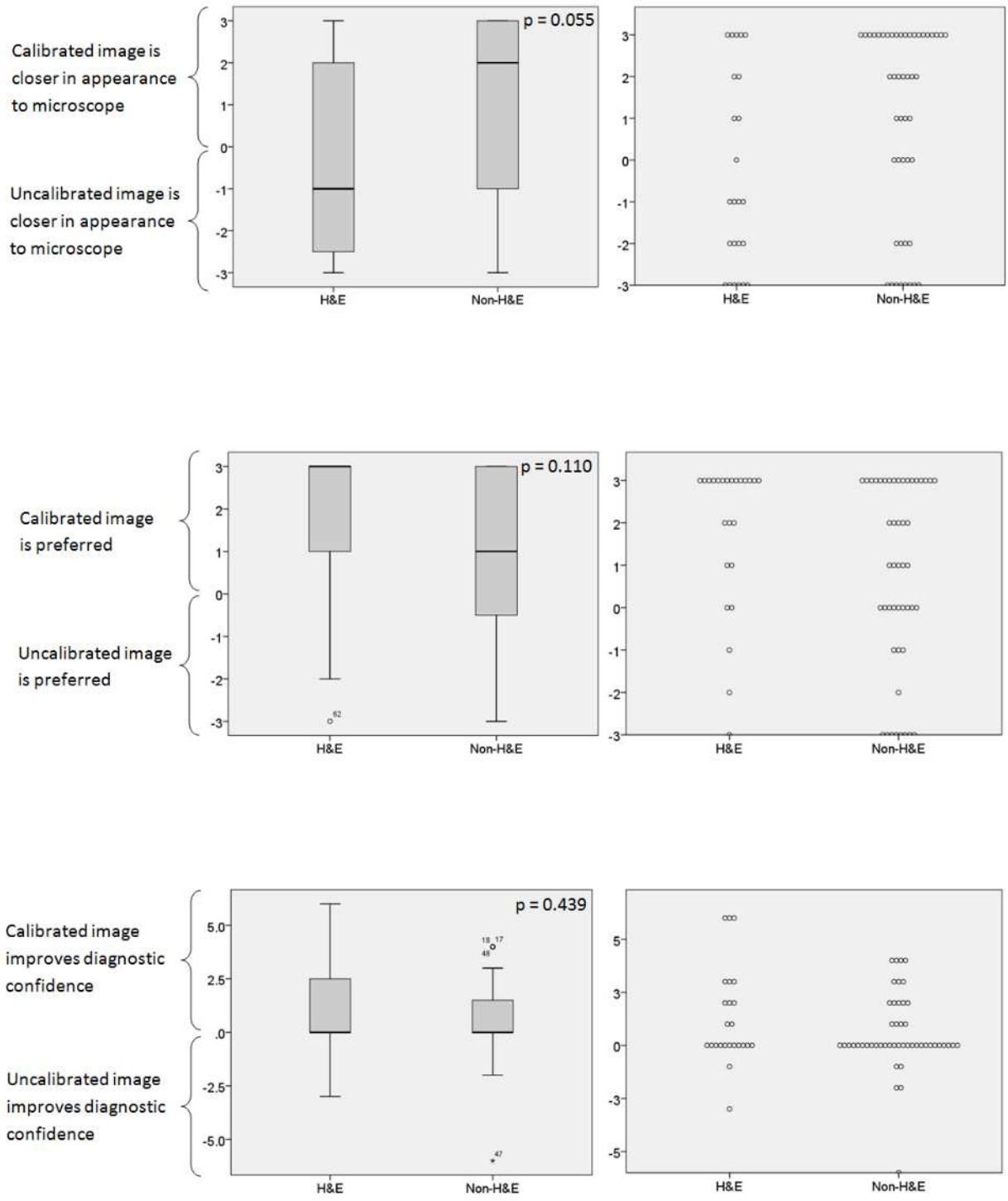


Fig 16. Scores by H&E stain vs. Non-H&E stain for each of the three main outcomes.

Author Biographical Information

Dr Emily Clarke (EC) is a Specialty Registrar in Histopathology and pre-doctoral National Institute of Health Research funded Academic Clinical Fellow in digital pathology at Leeds Teaching Hospitals NHS Trust and the University of Leeds.

W Craig Revie (CR), is employed by FFEI Limited, a manufacturer of imaging technology including digital microscopes. Craig has been involved in standards development and at present chairs the ICC Medical Imaging Working Group.

Dr Darren Treanor (DT) is a Consultant Histopathologist and leads the Digital Pathology Research Group at the University of Leeds. He holds degrees in both computing and medicine.