Image Analysis Using Machine Learning for Automated Detection of Hemoglobin H Inclusions in Blood Smears – A Method for Morphologic Detection of Rare Cells

Shir Ying Lee[1,2], Crystal M. E. Chen[1], Elaine Y. P. Lim[1], Liang Shen[3], Aneesh Sathe[4], Aahan Singh[4], Jan Sauer[4], Kaveh Taghipour[4], Christina Y. C. Yip[1]

[1]Department of Laboratory Medicine, Division of Haematology, National University Hospital, Singapore, [2]Department of Haematology-Oncology, National University Cancer Institute, Singapore, [3]Unit of Biostatistics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, [4]Qritive Pte Ltd, Singapore

Abstract

Background: Morphologic rare cell detection is a laborious, operator-dependent process which has the potential to be improved by the use of image analysis using artificial intelligence. Detection of rare hemoglobin H (HbH) inclusions in red cells in the peripheral blood is a common screening method for alpha-thalassemia. This study aims to develop a convolutional neural network-based algorithm for the detection of HbH inclusions. **Methods:** Digital images of HbH-positive and HbH-negative blood smears were used to train and test the software. The software performance was tested on images obtained at various magnifications and on different scanning platforms. Another model was developed for total red cell counting and was used to confirm HbH cell frequency in alpha-thalassemia trait. The threshold minimum red cells to image for analysis was determined by Poisson modeling and validated on image sets. **Results:** The sensitivity and specificity of the software for HbH+ cells on images obtained at ×100, ×60, and ×40 objectives were close to 91% and 99%, respectively. When an AI-aided diagnostic model was tested on a pilot of 40 whole slide images (WSIs), good inter-rater reliability and high sensitivity and specificity of slide-level classification were obtained. Using the lowest frequency of HbH+ cells (1 in 100,000) observed in our study, we estimated that a minimum of 2.4×10^6 red cells would need to be analyzed to reduce misclassification at the slide level. The minimum required smear size was validated on 78 image sets which confirmed its validity. **Conclusions:** WSI image analysis can be utilized effectively for morphologic rare cell detection. The software can be further developed on WISs and evaluated in future clinical validation studies comparing AI-aided diagnosis with the routine diagnostic method.

Keywords: Blood smear, convolutional neural network, hemoglobin H, machine learning, rare event detection

INTRODUCTION

Artificial intelligence (AI) using artificial neural network computational image analysis can be applied to many aspects of morphology-based laboratory analytics in hematology and cytopathology.[1-3] Convolutional neural network (CNN) algorithms can be trained to analyze images and subsequently classify them based on characteristic features. Successful applications of image analysis in the hematology laboratory include identification of malaria species, leukocyte differential counting, and classification and detection of acute leukemia and lymphoproliferative disease.[4-9]

Alpha-thalassemia, a genetic disorder of hemoglobin, is one of the most common genetic conditions worldwide. In high prevalence areas of East Asia and the Mediterranean,

How to cite this article: Lee SY, Chen CM, Lim EY, Shen L, Sathe A, Singh A, et al. Image analysis using machine learning for automated detection of hemoglobin H inclusions in blood smears – A method for morphologic detection of rare cells. J Pathol Inform 2021;12:18.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2021/12/1/18/313144

Access this article online

Quick Response Code:

Website:
www.jpathinformatics.org

DOI:
10.4103/jpi.jpi_110_20

an estimated 5%–15% of the population are carriers and 0.1%–0.5% have hemoglobin H (HbH) disease.[10-12] Detection of HbH inclusions within red blood cells is an established and specific method of screening for alpha-thalassemia carriers and HbH disease.[13] Inclusions are rare in carriers, with a quoted frequency from the literature of 1 in 1000–10,000 red cells, but are present in abundance or between 5% and 50% of red cells, in HbH disease.[13] HbH inclusion testing is widely performed in low resource countries, is inexpensive compared to genetic testing, and yet able to detect a large proportion of the clinically important alpha^0-deletion carriers.[14-16]

HbH inclusion detection relies heavily on the manual search for inclusions under light microscopy at high magnifications. Pathognomonic features are dark blue rounded inclusions conferring a pitted golf-ball like appearance to the red cell. The process is labor intensive and time-consuming given that the entire blood smear may contain few inclusions and is subject to interoperator variability similar to other operator-dependent tests such as screening for parasites, detection of rare leukemia cells, and quantification of fetomaternal hemorrhage.[17-19] Application of AI to assist in the detection of rare events carries the potential of improving detection rate, efficiency, and the quality of testing.

Detection of rare cells in peripheral blood is potentially challenging, whereas the analysis of normal or abnormal blood cells present in abundance in the smear would only require digitizing small sections of slides or a limited number of cells, the same approach for rare cells could potentially lead to false-negative results due to inadequate imaging. Hence large areas of the slide would need to be analyzed for rare cell detection. Consequently, the speed of image acquisition becomes relatively important when large areas of slides need to be imaged. Whole slide scanners are devices which currently provide some of the most rapid scans, though scanning is typically performed at lower magnifications such as ×20 and ×40,[20] well below the traditional magnification used for HbH inclusion detection.

Our primary aim was to develop an AI algorithm to detect HbH inclusions in blood smear images and to evaluate its utility as a diagnostic aid. To achieve the primary aim, a stepwise approach was taken to develop various steps of the process. This included evaluation of AI performance on images obtained at lower magnifications and on different image scanning platforms, the effect of storage on the quality of HbH blood smears, quantification of the total red cells within an image, estimation of the true frequency of HbH-positive red cells in alpha-thalassemia trait and validation of adequate image sampling as prerequisites before clinical evaluation. The following describes the intermediary steps to achieve the primary aim.

MATERIALS AND METHODS

The study was conducted prospectively between December 2017 and September 2020 at the Department of Laboratory Medicine, National University Hospital, Singapore. Ethics approval was granted by the Domain Specific Review Board of the National Healthcare Group (number 2017/01170). Anonymized blood smears were obtained from adults whose samples were submitted for thalassemia screening.

Blood smear preparation and hemoglobin H inclusion identification

HbH inclusion stain was performed using 1% Brilliant Cresyl Blue (BCB) staining solution on fresh K3-EDTA anticoagulated peripheral blood using a standardized protocol as previously described,[13] following which blood smears were made on glass slides. BCB is a supravital redox dye which causes precipitation of unstable HbH as rounded bluish inclusions and also stains the ribosomes of reticulocytes which appear filamentous, while normal mature red cells have a pale grey appearance. BCB differs from Romanowsky stains such as Wright-Giemsa used for leukocyte identification. In the routine diagnostic method, smears were observed for HbH inclusions by light microscopy using ×100 oil immersion lens by an experienced laboratory technologist, and HbH inclusion positive cells were verified by a second technologist. Two smears were inspected for cases with normocytic red cell indices and up to 6 smears for cases with microcytic red cell indices. The routine diagnostic method was used to classify smears into three slide-level categories as per usual practice: HbH-positive smear (rare HbH inclusions), HbH disease (abundant HbH inclusions), and HbH-negative smears. HbH-negative cases used in the study were additionally selected for normal hemoglobin and red cell indices.[21] As the aim of the study was to develop a software that could serve as an aid for morphological diagnosis rather than to attain the accuracy of genetic diagnosis, the results of the routine diagnostic method were used as the comparator for software performance.

Digital image capturing

Images of HbH-positive and HbH-negative blood smears were obtained at ×100 oil immersion objective (NA 1.25, 0.05 μm/pixel), ×60 objective (NA 0.8, 0.09 μm/pixel) and ×40 objective (NA 0.75, 0.13 μm/pixel) on the Olympus™ DP27 digital camera system attached to a BX53 microscope. Whole slide images (WSI) of the entire blood smear were obtained on the Hamamatsu NanoZoomer S60™ whole slide scanner at ×40 objective (numerical aperture 0.75, 0.23 μm/pixel) as shown in Figure 1A. Partial slide images (PSI) each with an area of 25 mm^2 were obtained on the Precipoint M8™ slide scanner at ×40 objective (numerical aperture 0.75, 0.16 μm/pixel) at regions where red cells were just overlapping. Digital camera and image acquisition settings were fixed during the study period.

Artificial intelligence identification of hemoglobin H inclusions

The two AI software applications used in the study were developed by the authors of this study. The first is an image analysis software which performs unbiased, automated analysis of digital image objects using deep learning techniques. The underlying machine learning model is a deep convolutional
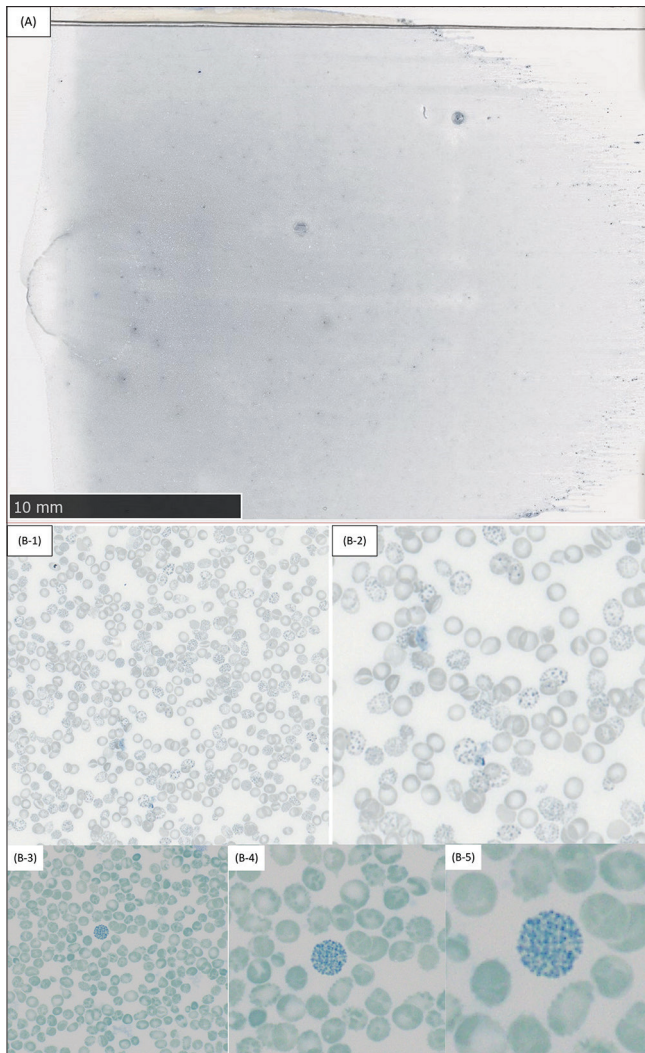
Figure 1: (A) Low-power view of a whole slide image of an entire blood smear stained by Brilliant Cresyl Blue obtained using ×40 objective on the Hamamatsu NanoZoomer S60™. (B1) Image of a case of HbH disease obtained using ×40 objective on the Precipoint™ slide scanner. (B2) The same image as in b1, as observed using digital magnification to ×80 showing preservation of cellular details and HbH inclusion bodies within numerous red cells. (B3-6) Images of a representative HbH inclusion positive red cell typically seen in alpha thalassemia trait obtained using ×40 objective on the Olympus™ imaging system at three different digital magnifications, showing the intracellular inclusions in detail. HbH: Hemoglobin H

neural network equipped with residual connections (ResNet). The neural network is based on a Region-Based Convolutional Network (RCNN) architecture with a ResNet-50 feature extraction backbone. This particular configuration was chosen for this study because neural networks with RCNN architectures have become the state of the art for accurate object detection, especially for small objects such as individual cells.[22] In order to reduce the training time, the weights of the feature extraction backbone were initialized to those of a model pretrained on the ImageNet dataset. The model was trained to identify HbH+ cells and predicts bounding boxes around them

on the basis of the features extracted by the backbone. Model weights were tuned throughout training to minimize two loss functions, the L1 loss on the bounding box coordinates and the cross-entropy loss on the prediction probability of bounding boxes. Both loss functions were given the same weight. The second software, Qritive Pantheon™, is a whole slide image viewer that enables users to inspect and annotate digital images as well as view predictions made by the AI software.

In the first phase, red cell images obtained at ×100 were individually segmented by software and annotated by two experienced technologists (C. ME. C, E. YP. L) into HbH inclusion positive (HbH+) and HbH inclusion negative (HbH-), i.e., the cell-level classification. The single-cell annotation formed the ground truth. The images were assigned into a training set and a test set. An independent development set was used to determine the model parameters with the best performance to select the final model. The test set was then used to evaluate the final model. The results of ground truth and software classification of cells in the test set were compared. Cells which were HbH+ by ground truth and software were true positive (TP) and cells HbH-by ground truth and AI were true negative (TN) or concordant events. Cells which were HbH+ by ground truth but HbH-by software were false negative (FN) and cells which were HbH-by ground truth but HbH+ by software were false positive (FP) or discordant events. A prediction confidence score (PCT), a numerical output in the range of 0–1 which serves as an indicator of the similarity between a given detection and objects in the training data, was generated by the software for each detection. Software performance values such as sensitivity (TP rate), specificity (TN rate), FP rate, false-negative rate, accuracy, and positive predictive value were calculated using standard formulae:[23]

$$\text{Sensitivity (TP rate)} = \frac{TP}{TP + FN}$$

$$\text{Specificity (TN rate)} = \frac{TN}{TN + FP}$$

$$\text{FP rate} = 1 - TNR = \frac{FP}{FP + TN}$$

$$\text{FN rate} = 1 - TPR = \frac{FN}{FN + TP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Positive predictive value} = \frac{TP}{TP + FP}$$

In the second phase, four objectives were studied.
1. Evaluation of the stability of HbH inclusions on storage: The stability of stained smears preserved by distyrene plasticizer and xylene (DPX) mounting media (CellPath, UK) when stored in the dark at room temperature of 22°C–24°C over 7 days was evaluated. Two technologists examined the smears daily for HbH inclusions and results were graded as acceptable if at least 20 HbH+ cells were

detected per smear. The proportion of observed positive cells out of total expected positives was determined

2. Evaluation of software on images at ×40 and ×60 and two imaging platforms. The AI software was tested on images obtained at ×40 and ×60 on the Olympus™ imaging system to evaluate its performance at the single-cell level on images captured at lower magnifications. Each single-cell image boundary was annotated into HbH+ or HbH−. The software performance at single-cell level was next tested on PSI obtained at ×40 on the Precipoint™ and WSI obtained at ×40 on the Hamamatsu™ slide scanners to evaluate its performance on images obtained by different imaging systems

   In addition, software performance at the per slide level was evaluated on WSI by comparing the result of AI-aided diagnosis of the slide with results of the routine diagnostic method. The AI-aided diagnosis was conducted by three assessors blinded to the actual result (C. ME. C, E. YP. L, SY. L). Each assessor independently appraised the AI-identified single cells and classified the slide into one of the three slide-level categories, i.e., HbH-positive slide, HbH-negative slide, and HbH disease. In case of disagreement between raters, a consensus review was adopted for the final slide classification.

3. Development of a model for total red cell estimation and determining the frequency of HbH+ cells in alpha-thalassemia trait: In order to estimate the number of red cells in an image, an intensity-based proxy metric was defined as follows: (1) The image was converted to grayscale so that cells appear lighter on a dark foreground. (2) The foreground was then segmented from the background using an Otsu-based thresholding algorithm. (3) The proxy metric was then defined as the sum of grayscale intensities of foreground pixels subtracted by the median of the background pixel grayscale intensities. The relationship between the metric and the actual cell counts was established using a linear model trained on a set of 75 manually annotated image patches. These patches were generated from cutouts of 11 different WSI and had dimensions ranging from 300 px × 300 px to 600 px × 600 px (at a resolution of 5.66 px/μm). The images were separated into a training set consisting of 40 images and a test set consisting of 35 images. The former was used to optimize the parameters of the model while the latter was used to assess its performance. In order to not bias, the parameter search toward larger images, the model was trained not on the absolute cell count but on the cell density per image.

   The frequency of HbH+ cells was determined by dividing the total number of AI-identified TP cells by the total red cell count generated by the model.

4. Determining the minimum number of red cells and size of smear to image for adequate sampling. We proceeded to model the probability of slide-level misdiagnosis according to the number of TP cells in the smear. On the basis that if two random variables are independent and identically distributed, the joint probability of multiple cells can be written as the multiplication of marginal probabilities.[24] The probability of a positive cell labeled as negative is equal to false-negative rate and the probability of a negative cell labeled as positive is equal to FP rate (FPR). Assuming that software predictions at the single-cell level are independent and identically distributed, the case misdiagnosis probability at a certain value of K, where K is the number of positive cells in the slide, can be calculated as:

Probability of labeling a positive slide as negative, Pslide $(N \mid P)$ $= Pcell \, (N \mid P)^K = FNR^K$

Probability of labeling a negative slide as positive, Pslide $(P \mid N)$ $= Pcell \, (P \mid N)^K = FPR^K$

We then applied Poisson modeling to determine the minimum number of red cells to analyze in order to achieve a high probability that at least a threshold number of positive cells would indeed be captured in the image. This step was undertaken so as to determine the minimum area of smear to be imaged for analysis, in order to reduce the likelihood of falsely labeling a slide as negative due to imaging of insufficient area containing no positive cells. Assuming random distribution of positive cells in the slide, by Poisson distribution, the probability of observing at least k abnormal cells in X can be written as:[25]

Probability$\left(\text{at least k abnormal cells in X}\right)$

$$= \frac{y^k e^{-y}}{k!} + \frac{y^{k+1} e^{-y}}{(k+1)!} + \frac{y^{k+2} e^{-y}}{(k+2)!} + \frac{y^{k+3} e^{-y}}{(k+3)!} + \dots$$

$$= 1 - e^{-y} - y e^{-y} - \frac{y^2 e^{-y}}{2!} - \dots - \frac{y^{k-1} e^{-y}}{(k-1)!}$$

where X denotes an area with x number of red cells, the frequency of occurrence of an abnormal cell is 1 in every N cells, and y is the expected number of abnormal cells in X and.

$y = \dfrac{x}{N}$ (details in Supporting information)

The number of red cells in each WSI was then estimated by multiplying the red cell count/μL obtained from the hematology analyzer by the volume of blood per smear (5 μL). The area of PSI to image and analyze was determined by dividing the desired minimum red cells to the image by the red cell density obtained using the model in section (3) above. These calculations were validated by evaluating the detection rate on WSI and PSI containing the required size.

Statistical analysis was performed on SPSS Statistics Version 26 (IBM Corporation, Somers, NY, USA) and GraphPad Prism Software Version 8 (La Jolla, CA, USA). Exact Clopper–Pearson binomial confidence interval (CI) for CI of rates, Fleiss' kappa for inter-rater reliability for more than 2 raters of categorical data, Kolmogorov–Smirnov normality test, and Kruskal–Wallis test for comparison of continuous data between groups were used. Parametric data were expressed as

mean and standard deviation (SD), and nonparametric data were expressed as median and range. The correlation was performed using Spearman's rank correlation coefficient for nonparametric data.

RESULTS

Blood smears from 110 individual cases, 78 rare HbH inclusion positive, 17 HbH disease, and 15 HbH inclusion negative, were used for the study. Table 1 summarizes the number of cases and images used for the entire study. In the first phase, 515 images each containing an average of 100 red cells were obtained using ×100 oil objective. 412 images formed the training set, 51 images formed the development set, and 52 images formed the test set. The CNN was trained on images in the training set and its performance tested on images in the test set. At a PCT >0.2, the sensitivity of the algorithm was 90.9%, specificity 99.0%, false-negative rate 9.1%, FP rate 1.0%, and overall accuracy was 97.6% [Table 2a]. False-negative cells occurred mainly in HbH disease due to finer inclusions and the large number of positive cells.

Evaluation of stability of hemoglobin H inclusions on storage

Twenty-seven samples comprising 22 alpha-thalassemia trait and 5 HbH disease were assessed on storage, with day 0

being the day of smear preparation. HbH inclusions remained visible after 7 days. Figure 2 shows representative images from day 0 to 7. After 7 days of storage, at least 20 HbH+ cells remained detected in all cases, giving 540 positive cells out of 540 expected positives (95% CI 99.3%–100%). The longer storage duration allowed for greater workflow flexibility as image acquisition could be performed up to 7 days after slide preparation without compromising image quality.

Evaluation of software on images at ×40 and ×60 and two imaging platforms

In high-resolution images obtained at ×40 on the different imaging platforms, cellular details were sufficiently preserved and recognizable visually [Figure 1B]. The software performance was tested on 140 annotated ×40 images and 200 annotated ×60 images from 14 alpha-thalassemia trait, 1 HbH disease, and 10 HbH-negative cases. At a PCT of 0.1 and above, the sensitivity was 91.64% and specificity was 99.94% on the ×40 images, while sensitivity was 93.07% and specificity was 99.99% on the ×60 images.

The software performance, evaluated on 51 PSI obtained on the Precipoint™ slide scanner as shown in Figure 3, showed a sensitivity of 90.0%, specificity of 99.9995%, FNR of 10%, FPR of 0.0005% or 1 FP in 200,000 cells, PPV of 82.8%
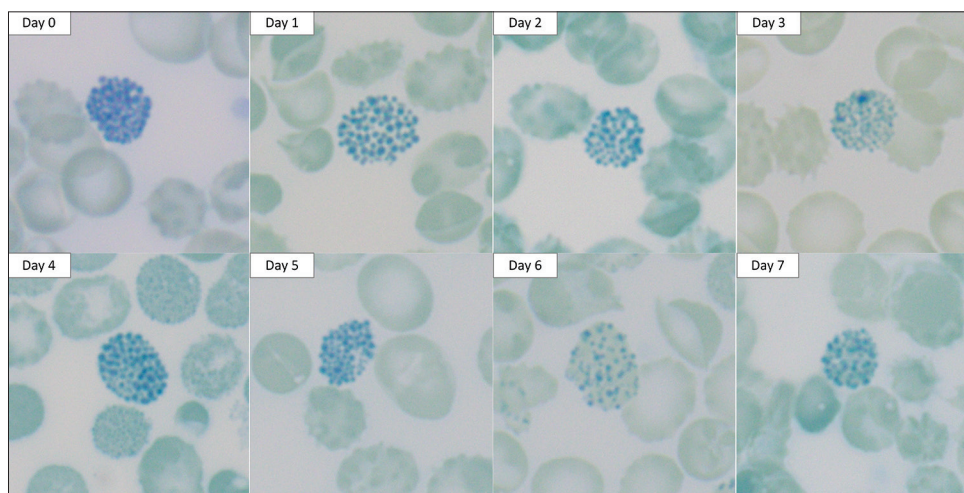

Figure 2: Representative HbH inclusion positive red cells observed over 7 days of storage. HbH inclusions remained visible under light microscopy when stored under DPX-mounting media in the dark at room temperature for up to 7 days. Each smear was considered stable on storage if 20 or more individual HbH inclusion positive cells remained recognizable over the duration of storage. HbH: Hemoglobin H

Table 1: Summary of number of cases and number of images used in the study

	Number of individual cases				Number of images/ smears for WSI
	Rare HbH inclusion positive	**HbH disease**	**HbH inclusion negative**	**Total**	
×100 images on Olympus™	33	10	0	43	515
×60 images on Olympus™	5	0	5	10	200
×40 images on Olympus™	9	1	5	15	250
×40 images on Precipoint™	17	3	0	20	177
×40 WSI on Hamamatsu™	14	3	5	22	118
Total	78	17	15	110	

HbH: Hemoglobin H, WSI: Whole slide images

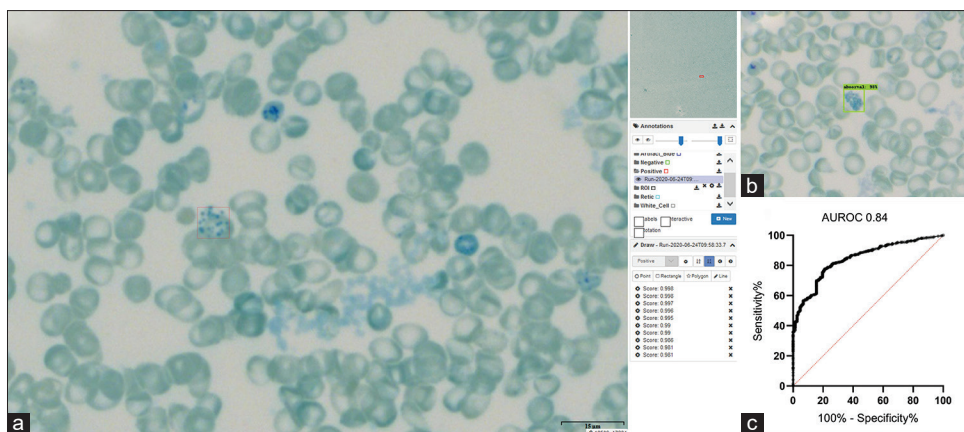**Figure 3:** Results of applying the software analysis on images of HbH blood smears obtained at ×40 objective. (a) Screenshot of the Qritive Pantheon™ user interface depicting the results of software analysis on an image obtained on the Precipoint™ slide scanner. AI identifications above 0.2 prediction confidence threshold are shown in the right-hand column. In this image, 10 confirmed HbH-positive identifications were detected by the software, with all having prediction confidence score of more than 0.98. (b) A HbH-positive cell with prediction confidence score of 0.98 is identified by the software on an image obtained on the Olympus™ imaging system. (c) ROC curve generated by comparing prediction confidence scores of true-positive versus true-negative cells on images obtained on the Precipoint™ slide scanner when identifications with prediction confidence score above 0.1 were considered. HbH: Hemoglobin H, ROC: Receiver operating characteristic

Table 2: Software performance at the single-cell level[a] on images obtained at (a) ×100 oil immersion objective on Olympus™ image system. (b) ×40 objective on the Precipoint™ imaging system

	Confirmed HbH+	Confirmed HbH−
(a)		
AI identified HbH+	828	40
AI identified HbH−	83	4149
(b)		
AI identified HbH+	576	120
AI identified HbH−	64	21,000,000[b]

[a]When positive identifications are defined by prediction confidence threshold >0.2, [b]Estimated from the cell count model referred to in methods, Section (3) development of a model for total red cell estimation. HbH: Hemoglobin H

and overall accuracy of 99.99% for identifications with PCT >0.2 [Table 2b]. The corresponding receiver operating characteristic curve showed an area under the receiver operating characteristic curve (AUROC) of 0.84 (95% CI 0.81–0.88, $P < 0.0001$).

When evaluated on WSI, the software performance at the cell level was lower. The software detected a total of 8230 identifications above a PCT of 0.2, of which 3679 were TP identifications, giving a positive predictive value of 44.7%. Sensitivity and specificity on WSI were not computed as the large number of identifications with PCT <0.2 were not individually recorded. The low positive predictive value was a result of far larger number of cells in WSI revealing further morphological categories such as reticulocytes and artifacts which were under-represented in the Olympus images. Figure 4 shows an example of WSI after software analysis in which red squares indicate the locations of HbH+ detections made by the software.

AI-aided diagnosis method was conducted on a pilot set of 30 HbH-positive and 10 HbH-negative WSI slides. Two HbH-negative smears had discordant results among the 3 assessors, i.e., 1 of 3 assessors misclassified the smears as HbH positive. The interrater kappa coefficient among the 3 assessors was 0.907, indicating good overall agreement. The consensus results were concordant with the results of the routine diagnostic method in all 40 slides, providing a slide-level sensitivity of 100% (95% CI 88.4%–100%) and specificity of 100% (95% CI 69.2%–100%).

Development of a model for total red cell estimation and determining the frequency of HbH+ cells in alpha-thalassemia trait

The model developed for cell count estimation was evaluated by comparing the cell density prediction with the ground truth and the model showed a good overall correlation ($R = 0.811$). The average ground truth cell density (number of cells/mm²) on the training data was 17761/mm² with a mean absolute error of 2461 (13.86%) and on test data was 17,935/mm² with a mean absolute error of 2865 (15.98%). Applying the model to 110 independent PSI of 25 mm² size at regions where red cells were just overlapping, the density of red cells was found to average 17,296/mm² (range 13,536–20,607; SD 2006; 2SD range 13,284–21,308/mm²).

Using the total red cell estimation derived from the model and the number of TP cells identified by software multiplied by a correction factor of 1.1 (correction factor = 1/sensitivity of the software), the true frequency of HbH+ cells was estimated in 11 cases of alpha-thalassemia trait. The frequency ranged from 1 in 13,619 to 1 in 91,890 (0.001%– 0.007%), with a median of 1 in 35070 (0.003%) and interquartile range of 1 in 19,057-1 in 57,781 (0.002%–0.005%). The frequency in smears from the same individual was comparable, but the frequency varied
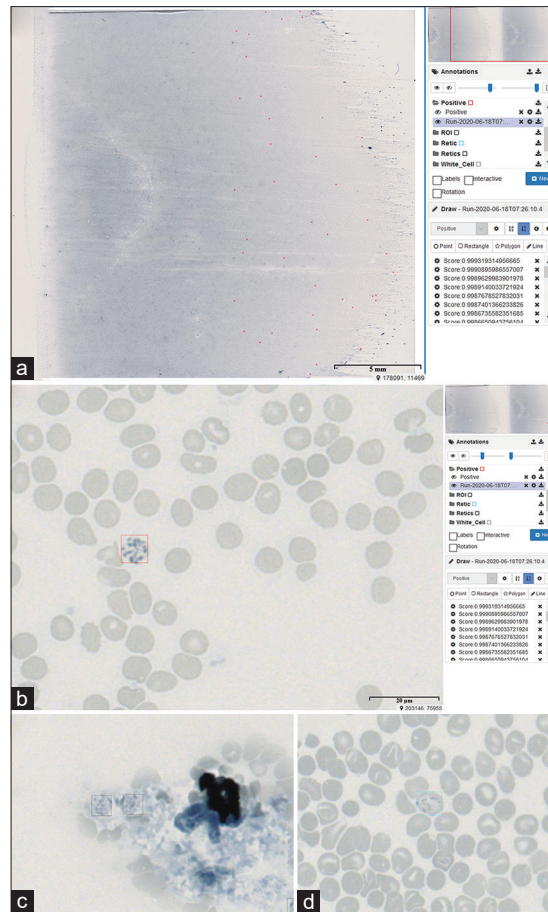
Figure 4: Results of applying the software analysis on whole slide images of HbH blood smears obtained at ×40 objective on the Hamamatsu NanoZoomer S60™ slide scanner. (a) Screenshot of the Qritive Pantheon™ user interface depicting the results of software analysis. The red dots on the whole slide image are the software identifications of HbH-positive cells detected above a prediction confidence threshold of 0.2. The right-hand column shows the list of identifications. (b) A higher magnification view of the same slide showing details of a confirmed HbH positive identification (red box). In this case, the identified cell had a prediction confidence score of 0.999. (c) False-positive identification of artifacts (black boxes) occurring particularly at the edges of the slide and likely representing stain precipitates. (d) False-positive identification of reticulocytes (light green box) occurred sporadically throughout the slide. HbH: Hemoglobin H

between different individuals with the alpha-thalassemia trait ($P < 0.0001$) [Figure 5]. Hence, to maximize the detection rate, we used the lowest observed frequency of approximately 1 in 100,000 for calculating the Poisson model in section 4.

Determining the minimum number of red cells and size of smear to image for adequate sampling

Using the software sensitivity of 91% and specificity of 99% at the single-cell level, the case misdiagnosis probability at different values of K, which is the number of positive cells in the image, was computed and shown in Table 3. As shown in Table 3, the larger the number of positive cells in the image, the lower the probability of misdiagnosis at the slide level.

By Poisson modeling, when the frequency of positive cells is 1 in 10,000, 214,700 red cells would need to be imaged to give a 99.99% confidence that at least 5 positive cells will be present in the image. At one-tenth, the frequency of positive cells, i.e., 1 in 100,000, the number of red cells to the image would be 10 times that for the frequency of 1 in 10,000.

Table 3: Probability of misdiagnosis at different values of K, where K is the number of positive cells in the slide. P_{slide} $(N|P)$ is the probability of labeling a positive slide as negative and P_{slide} $(P|N)$ is the probability of labeling a negative slide as positive, with C the corresponding chance of misdiagnosis expressed as 1 in $1/P$

K	P_{slide} $(N\|P)=0.091^K$	C_{slide} $(N\|P)$	P_{slide} $(P\|N)=0.01^K$	C_{slide} $(P\|N)$
1	0.091	1 in 11	0.01	1 in 100
2	0.008	1 in 125	0.0001	1 in 10^5
3	0.0007	1 in 1428	0.000001	1 in 10^6
4	0.00006	1 in 16,667	0.00000001	1 in 10^8
5	0.000006	1 in 166,667	0.0000000001	1 in 10^{10}

For higher levels of confidence, the number of red cells to the image would be progressively higher [full data table in Supplementary Information]. In this way, the Poisson model estimated that 2.4 million red cells would need to be imaged to provide a 99.999% confidence that at least 5 positive cells
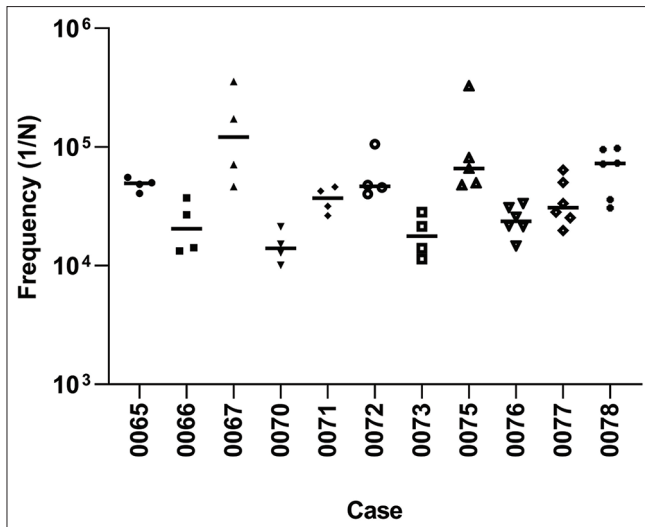
Figure 5: Frequency of HbH inclusion positive cells in 11 cases of alpha-thalassemia trait. Each dot represents the frequency in one smear area and horizontal lines represent the median frequency for the case. Different smear areas of each case contained HbH inclusion positive cells at comparable frequency, but the frequency varied between individuals with alpha-thalassemia trait (Kruskal–Wallis test with Dunn's multiple comparison, $P < 0.0001$). HbH: Hemoglobin H

would be present in the image when the frequency of positive cells is 1 in 100,000.

We estimated that for a case with red cell count of $4.5 \times 10^6/\mu L$, one smear would contain approximately 22.5×10^6 red cells, more than the minimum required of 2.4 million red cells, and would be sufficient for analysis. For calculation of the size of PSI required for analysis, assuming the lower limit of 2SD of cell density obtained in section 3, i.e., 13,284/mm^2, we estimated that 180 mm^2 would provide sufficient red cells for analysis. We validated these calculations on 78 independent image-sets comprising 60 WSI and 18 PSI of 180 mm^2 size. For the WSI, the number of confirmed HbH+ cells above PCT >0.99 ranged from 25 to 105 with a median of 63, and for the 180 mm^2-size PSI, the number of confirmed HbH+ cells ranged from 8 to 96 with a median of 53, demonstrating that all image-sets contained 5 or more HbH+ cells.

Discussion

In our study, we developed a machine-learning algorithm based on CNN which could identify HbH+ red cells on blood smears with good overall sensitivity of 91% and specificity of 99% on ×100 images. The software was applicable to images obtained at ×40 and ×60, although sensitivity was slightly lower than at ×100 as the PCT for identification had to be lowered to 0.1 in order to achieve an equivalent level of sensitivity. During the application of the software to two different slide scanner platforms, the software retained a high degree of specificity and sensitivity on the Precipoint platform, providing an AUROC of 0.84. On the other hand, software performance

on WSI at the single-cell level was suboptimal due to additional morphological classes under-represented in the original training set. This will necessitate future training of the algorithm with a larger number of morphological classes on WSI. Interestingly, when WSI was assessed in terms of the overall diagnostic accuracy at the slide level using an AI-aided diagnostic process, our pilot evaluation showed promising results as all cases were correctly classified at the slide level with good interrater reliability, suggesting high sensitivity of the software.

During the assessment of software performance, we placed greater value on high sensitivity for the following reasons. First, HbH inclusion detection is primarily a screening test for alpha-thalassemia trait and HbH disease, and a screening test would need to have a high level of sensitivity. Second, the AI identification can be designed as the first step to enhance the detection rate of rare cells, while the second step of operator verification of identified cells can be used to eliminate FP identifications.

Our study is the first to describe the application of AI-aided image analysis to the morphological detection of HbH inclusions. Despite the availability of other diagnostic modalities such as genetic testing, the morphologic review remains an indispensable and inexpensive technique available to most laboratories. Moreover, morphological rare cell detection remains commonplace despite its tedious nature. When employing image analysis for HbH inclusion testing, traditional procedures such as use of high magnification for cellular diagnosis and screening of multiple blood smears to detect rare cells had to be transformed to practical solutions by the use of high-resolution imaging at lower magnification, sensitive AI algorithms, and application of mathematical modeling. From mathematical modeling, it can be appreciated that the sensitivity at the case level is potentially higher than the sensitivity at the cell level because each smear would contain many positive cells. However, before image analysis, there needs to be sufficient image sampling in order to have a high degree of confidence that positive cells are indeed captured in the image in the first place. Therefore, we validated the Poisson model derived minimum red cell acquisition on several image sets.

Previous methods of image analysis for rare cell detection such as for the detection of cancer micro-metastases, utilized fluorescence or immunocytochemical staining to highlight pathological cells while obtaining images in a two-step process, with the initial screening scan at low magnification and the second scan of pathological cells at high magnification.[26-29] In contrast, our current study utilizes a one-stage scanning process at ×40 on a whole slide scanner which simplifies the automation process.[30] The need for sufficient sampling or acquisition of a sufficient number of background cells for rare event detection has been a recognized pre-requisite in other rare event methodologies such as flow cytometry.[31,32] To achieve the necessary scan area in a short time, whole slide scanners which enable rapid high-resolution image acquisition

at lower magnification are the most practical for translating this technology into clinical use. Here, we demonstrate a proof of concept that this simplified automation method is feasible.

There were several limitations to our current study. The first was the variable density of red cells in the blood smear, in particular the thick edge where several cell layers overlap, potentially obscuring positive cells. This did not appear to pose an impediment to the software as we observed positive detections in these areas, but it is plausible that some cells remained undetected. The use of hydrophilic-treated plastic plates which create monolayer blood smears may be able to overcome this inherent limitation of glass slides.[33] Differences in color saturation due to differences in microscope and imaging parameters caused slight differences between slides, with the potential for image misclassification both by human observer and software, hence image acquisition was conducted using standardized settings. Despite that, our results show that image acquisition on different platforms does impact software performance. The software algorithm should ideally compensate for these differences, otherwise, software development would need to be specific to slide scanners and settings. Guidelines for validation of whole slide scanners are available and standardization of digital imaging in hematology and pathology is currently in progress.[34-39] A frequent problem encountered on imaged blood smears was the presence of small areas of suboptimal focus which had gone unnoticed during the vetting process, and which have the potential to cause misclassification.[40] One of the pertinent issues, therefore, pertains to the algorithm being able to identify and analyze slightly off-focussed images. Additionally, inter-rater precision between the two annotators was not assessed before ground truth generation and could potentially have introduced confounders to software performance. Finally, although good specificity was achieved at the single-cell level, the absolute number of FP cells may seem significant to the observer due to the millions of cells being processed, and this context needs to be taken into consideration during the AI-aided diagnostic process.

Our experience using WSI on blood smears parallels some of the lessons learnt from cytopathology. As noted from the experience from cytopathology, uneven thickness of material requires multiple Z-plane scanning and Z-stacking, increasing the scanning times and file size which may limit the widespread adoption of WSI in high-throughput settings.[41] Suboptimal image quality also negatively impacts the subjective acceptance of WSI by the assessor, and subjective acceptance has been correlated with diagnostic accuracy.[42] In a systematic review of WSI in cytopathology, it was noted that there was good diagnostic concordance between WSI and light microscopy but these appeared lower than those reported in surgical pathology.[43] These technical challenges should be solved with future studies, as it is expected that the use of AI in WSI would increase in the future.[3]

CONCLUSION

The AI software developed presents a promising tool for AI-aided image analysis for automated detection of HbH inclusions in blood smears. Before clinical validation of such software, a prerequisite minimum area of the slide should be imaged for analysis. Future work would need to be conducted on platform-specific software training and multiclass classification of other cell types within WSI. Our study serves as groundwork for future clinical studies comparing the sensitivity, specificity, and relative efficiency of AI-aided diagnosis against the routine method. Collectively, the development process described could potentially be applied to other types of image-based rare cell detection to improve the efficiency of morphologic review.

Conflicts of interest

Authors: Lee SY, Chen CME, Lim EYP, Shen L, Yip CYC, have no conflicts of interest. Authors: Sathe A, Singh A, Sauer J, and Taghipour K, are employees of Qritive Pte. Ltd.

REFERENCES

1. Rodellar J, Alférez S, Acevedo A, Molina A, Merino A. Image processing and machine learning in the morphological analysis of blood cells. Int J Lab Hematol 2018;40 Suppl 1:46-53.
2. Mohammed EA, Mohamed MM, Far BH, Naugler C. Peripheral blood smear image analysis: A comprehensive review. J Pathol Inform 2014;5:9.
3. Landau MS, Pantanowitz L. Artificial intelligence in cytopathology: A review of the literature and overview of commercial landscape. J Am Soc Cytopathol 2019;8:230-41.
4. Zhao J, Zhang M, Zhou Z, Chu J, Cao F. Automatic detection and classification of leukocytes using convolutional neural networks. Med Biol Eng Comput 2017;55:1287-301.
5. Wang Q, Bi S, Sun M, Wang Y, Wang D, Yang S. Deep learning approach to peripheral leukocyte recognition. PLoS One 2019;14:e0218808.
6. Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G. Image analysis and machine learning for detecting malaria. Transl Res 2018;194:36-55.
7. Das DK, Ghosh M, Pal M, Maiti AK, Chakraborty C. Machine learning approach for automated screening of malaria parasite using light microscopic images. Micron 2013;45:97-106.
8. Eshel Y, Houri-Yafin A, Benkuzari H, Lezmy N, Soni M, Charles M, *et al*. Evaluation of the parasight platform for malaria diagnosis. J Clin Microbiol 2017;55:768-75.

9. Boldú L, Merino A, Alférez S, Molina A, Acevedo A, Rodellar J. Automatic recognition of different types of acute leukaemia in peripheral blood by image analysis. J Clin Pathol 2019;72:755-61.

10. Angastiniotis M, Modell B. Global epidemiology of hemoglobin disorders. Ann N Y Acad Sci 1998;850:251-69.

11. Modell B, Darlison M. Global epidemiology of haemoglobin disorders and derived service indicators. Bull World Health Organ 2008;86:480-7.

12. Modell B, World Health Organization. In: Modell B, editor. Guidelines for the Control of Haemoglobin Disorders. WHO/HDP/HB/GL/94.1. Geneva, Switzerland: World Health Organization; 1994.

13. Old J, Harteveld CL, Traeger-Synodinos J, Petrou M, Angastiniotis M, Galanello R. Prevention of Thalassaemias and Other Haemoglobin Disorders: Volume 2: Laboratory Protocols. Nicosia, Cyprus: Thalassaemia International Federation(c); 2012.

14. Chan AY, So CK, Chan LC. Comparison of the HbH inclusion test and a PCR test in routine screening for alpha thalassaemia in Hong Kong. J Clin Pathol 1996;49:411-3.

15. Langlois S, Ford JC, Chitayat D, CCMG Prenatal Diagnosis Committee, SOGC Genetics Committee. Carrier screening for thalassemia and hemoglobinopathies in Canada. J Obstet Gynaecol Can 2008;30:950-9.

16. Pan LL, Eng HL, Kuo CY, Chen WJ, Huang HY. Usefulness of brilliant cresyl blue staining as an auxiliary method of screening for α-thalassemia. J Lab Clin Med 2005;145:94-7.

17. Kim YA, Makar RS. Detection of fetomaternal hemorrhage. Am J Hematol 2012;87:417-23.

18. Dini L, Frean J. Quality assessment of malaria laboratory diagnosis in South Africa. Trans R Soc Trop Med Hyg 2003;97:675-7.

19. Thomson S, Lohmann RC, Crawford L, Dubash R, Richardson H. External quality assessment in the examination of blood films for malarial parasites within Ontario, Canada. Arch Pathol Lab Med 2000;124:57-60.

20. Rojo MG, García GB, Mateos CP, García JG, Vicente MC. Critical comparison of 31 commercially available digital slide systems in pathology. Int J Surg Pathol 2006;14:285-305.

21. Bain BJ. Haemoglobinopathy diagnosis: Algorithms, lessons and pitfalls. Blood Rev 2011;25:205-13.

22. Nguyen ND, Do T, Ngo TD, Le DD. An evaluation of deep learning methods for small object detection. J Electr Comput Eng 2020;2020:3189691. https://doi.org/10.1155/2020/3189691.

23. Tharwat A. Classification assessment methods. Appl Comput Inform 2018. https://doi.org/10.1016/j.aci.2018.08.003.

24. Florescu I. Probability and Stochastic Processes. New Jersey, United States: John Wiley & Sons Inc; 2014.

25. Ross SM. Introduction to Probability Models. Oxford, UK: Academic Press, 2003.

26. Kraeft SK, Ladanyi A, Galiger K, Herlitz A, Sher AC, Bergsrud DE, et al. Reliable and sensitive identification of occult tumor cells using the improved rare event imaging system. Clin Cancer Res 2004;10:3020-8.

27. Bauer KD, de la Torre-Bueno J, Diel IJ, Hawes D, Decker WJ, Priddy C, et al. Reliable and sensitive analysis of occult bone marrow metastases using automated cellular imaging. Clin Cancer Res 2000;6:3552-9.

28. Markovic S, Li B, Pera V, Sznaier M, Camps O, Niedre M. A computer vision approach to rare cell in vivo fluorescence flow cytometry. Cytometry A 2013;83:1113-23.

29. Mesker WE, Burg MJ, Oud PS, Knepfle CF, Ouwerkerk v- Velzen MC, Schipper NW, et al. Detection of immunocytochemically stained rare events using image analysis. Cytometry 1994;17:209-15.

30. Chandradevan R, Aljudi AA, Drumheller BR, Kunananthaseelan N, Amgad M, Gutman DA, et al. Machine-based detection and classification for bone marrow aspirate differential counts: Initial development focusing on nonneoplastic cells. Lab Invest 2020;100:98-109.

31. Hedley BD, Keeney M. Technical issues: Flow cytometry and rare event analysis. Int J Lab Hematol 2013;35:344-50.

32. Donnenberg AD, Donnenberg VS. Rare-event analysis in flow cytometry. Clin Lab Med 2007;27:627-52, viii.

33. Hashimoto M, Yatsushiro S, Yamamura S, Tanaka M, Sakamoto H, Ido Y, et al. Hydrophilic-treated plastic plates for wide-range analysis of Giemsa-stained red blood cells and automated Plasmodium infection rate counting. Malar J 2017;16:321.

34. Digital Imaging and Communications in Medicine (DICOM). Supplement 145: Whole Slide Microscopic Image IOD and SOP Classes; September, 2010.

35. Kratz A, Lee SH, Zini G, Riedl JA, Hur M, Machin S, et al. Digital morphology analyzers in hematology: ICSH review and recommendations. Int J Lab Hematol 2019;41:437-47.

36. U.S. Food and Drug Administration. Technical Performance Assessment of Digital Pathology Whole Slide Imaging Devices. Guidance for Industry and Food and Drug Administration Staff. Maryland, United States: Center for Devices and Radiological Health; 2016 Apr 20th. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/technical-performance-assessment-digital-pathology-whole-slide-imaging-devices. [Last accessed on 2020 Nov 11].

37. García-Rojo M, De Mena D, Muriel-Cueto P, Atienza-Cuevas L, Domínguez-Gómez M, Bueno G. New European union regulations related to whole slide image scanners and image analysis software. J Pathol Inform 2019;10:2.

38. Cross SF, Igali L, Snead D, Treanor D. Best practice recommendations for implementing digital pathology. Document No. G162. The Royal college of Pathologists; 2018. Available from: https://www.rcpath.org/document-library-search.html. [Last accessed on 2020 Nov 11].

39. Pantanowitz L, Sinard JH, Henricks WH, Fatheree LA, Carter AB, Contis L, et al. Validating whole slide imaging for diagnostic purposes in pathology: Guideline from the College of American Pathologists Pathology and Laboratory Quality Center. Arch Pathol Lab Med 2013;137:1710-22.

40. Kohlberger T, Liu Y, Moran M, Chen PC, Brown T, Hipp JD, et al. Whole-slide image focus quality: Automatic assessment and impact on AI cancer detection. J Pathol Inform 2019;10:39.

41. Eccher A, Girolami I. Current state of whole slide imaging use in cytopathology: Pros and pitfalls. Cytopathology 2020;31:372-8.

42. Capitanio A, Dina RE, Treanor D. Digital cytology: A short review of technical and methodological approaches and applications. Cytopathology 2018;29:317-25.

43. Girolami I, Pantanowitz L, Marletta S, Brunelli M, Mescoli C, Parisi A, et al. Diagnostic concordance between whole slide imaging and conventional light microscopy in cytopathology: A systematic review. Cancer Cytopathol 2020;128:17-28.