Pathology (■ xxxx) **xxx(xxx)**, xxx

HAEMATOLOGY

Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation

Leander van Eekelen[1,2], Hans Pinckaers[2,3], Michiel van den Brand[3,4], Konnie M. Hebeda[3], Geert Litjens[2,3]

[1]*Faculty of Biomedical Engineering, Technical University Eindhoven, Eindhoven, the Netherlands;* [2]*Computational Pathology Group, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, the Netherlands;* [3]*Department of Pathology, Radboud University Medical Center, Nijmegen, the Netherlands;* [4]*Pathology-DNA, Rijnstate Hospital, Arnhem, the Netherlands*

Summary

Cellularity estimation forms an important aspect of the visual examination of bone marrow biopsies. In clinical practice, cellularity is estimated by eye under a microscope, which is rapid, but subjective and subject to inter- and intraobserver variability. In addition, there is little consensus in the literature on the normal variation of cellularity with age. Digital image analysis may be used for more objective quantification of cellularity. As such, we developed a deep neural network for the segmentation of six major cell and tissue types in digitized bone marrow trephine biopsies. Using this segmentation, we calculated the overall bone marrow cellularity in a series of biopsies from 130 patients across a wide age range. Using intraclass correlation coefficients (ICC), we measured the agreement between the quantification by the neural network and visual estimation by two pathologists and compared it to baseline human performance. We also examined the age-related changes of cellularity and cell lineages in bone marrow and compared our results to those found in the literature. The network was capable of accurate segmentation (average accuracy and dice score of 0.95 and 0.76, respectively). There was good neural network-pathologist agreement on cellularity measurements (ICC=0.78, 95% CI 0.58−0.85).

We found a statistically significant downward trend for cellularity, myelopoiesis and megakaryocytes with age in our cohort. The mean cellularity began at approximately 50% in the third decade of life and then decreased ±2% per decade to 40% in the seventh and eighth decade, but the normal range was very wide (30−70%).

Key words: Bone marrow; cellularity; segmentation; deep learning; digital pathology.

Received 26 August 2020, revised 7 July, accepted 14 July 2021
Available online: xxx

INTRODUCTION

Bone marrow trephine biopsies are routinely used in haematology to investigate myeloid diseases and for staging lymphomas. During the visual examination of a biopsy, various aspects of the tissue are evaluated, such as the marrow architecture and the cellularity of haematopoietic and stromal components.[1] The estimation of cellularity allows for a biopsy to be roughly categorised as hypo-, normo- or hypercellular, depending on the age and the clinical circumstances, which can give an indication of the activity of haematopoiesis. Many bone marrow diseases are characterised by either hypercellularity (e.g., myeloproliferative neoplasms) or hypocellularity (e.g., aplastic anaemia). Therefore, cellularity estimation within the clinical context helps to guide the differential diagnostic process.

Categorising biopsies as hypo-, normo- or hypercellular relies on the subjective internal reference frame of the pathologist and knowledge of the normal variations of bone marrow cellularity with age. There is a general consensus that cellularity decreases with age,[1] but there is little consensus in the published literature on the mean and variation of cellularity per age. Several studies published in the last 60 years found results that disagree on the rate of decrease and at which ages this decrease takes place.[2–5]

We attribute these differences in results across studies to differences in clinical features of the examined cohorts (patients with and without haematopoietic diseases, necropsies) and to the wide variety in methods used for determining cellularity. In histology, cellularity is typically expressed as a visual estimate of the percentage of surface area in the marrow cavity occupied by active haematopoietic marrow. This visual estimation is rapid but semiquantitative in nature. In contrast, point counting (histomorphometry)[6] by using a microscope eyepiece with a graticule is considered to be a more accurate method of quantifying cellularity, but is slow and labour-intensive.

Studies have also used automated (analog) image analysis systems to quantify cellularity.[7,8] A practical alternative is comparing the tissue to a range of photographic examples of marrows with different cellularity.[9]

Quantifying bone marrow cellularity for routine diagnostics using digital image analysis offers advantages over manual microscopy techniques. Areas are easily measured and compared and cells can be counted quickly and exhaustively, giving an objective quantification instead of an

Print ISSN 0031-3025/Online ISSN 1465-3931
DOI: https://doi.org/10.1016/j.pathol.2021.07.011

Please cite this article as: van Eekelen L et al., Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation, Pathology, https://doi.org/10.1016/j.pathol.2021.07.011

estimation of proportions. The latter is more comparable to the cell counts in the bone marrow smears. Several recent studies reported cellularity measurement using digital image analysis techniques on digitised slides that show good agreement with references of visual estimate or point counting.[4,10–12] These studies have used traditional machine learning techniques, while the field of medical image analysis has shown a shift towards generally better performing deep learning systems.[13] Deep learning has already been applied to the segmentation of erythropoiesis and myelopoiesis,[14] but no work has been published on the simultaneous segmentation of all major cell types in bone marrow, which would allow for a more detailed analysis of the tissue.

In this paper, we present the first neural network for the automated segmentation (i.e., pixelwise labelling) of six major cell and tissue types in diagnostic bone marrow trephine biopsies: erythropoiesis, myelopoiesis, megakaryocytes, lipocytes, trabecular bone, and erythrocytes. We then show how this segmentation can be used to quantify overall bone marrow cellularity in a series of biopsies from 130 patients across a wide age range. We also evaluate the agreement between the neural network and two pathologists and compare this to human-level performance. Lastly, we quantify trabecular bone surface and various other cell ratios and compare between age groups and gender.

MATERIALS AND METHODS

Materials

For this retrospective study, 157 bone marrow trephine biopsies performed for staging of lymphoid or solid cancers in the period 2017–2020 were selected from the archive of the Radboud University Medical Center. The biopsies were obtained from the posterior iliac crest and processed according to a previously published protocol.[15] All biopsies were negative for disease. Eleven biopsies were irretrievable. Biopsies were excluded if they were severely fragmented, consisted predominantly of cortical bone, showed extensive artifacts (crush/tears) or did not have sufficient identifiable marrow for visual estimation of cellularity by the pathologists. This excluded 16 biopsies, resulting in 130 biopsies from patients with an age range of 6–83 years (mean of 57 years, 69 male and 61 female patients; Table 1). The majority of patients was diagnosed with lymphoproliferative disease (n=113), while some were diagnosed with mastocytosis (n=9), solid tumour metastases (n=4), or Langerhans cell histiocytosis (n=3). A stem cell donor without disease was also included. The need for informed consent was waived by the institutional review board of the Radboud UMC (2020–6483).

For each patient, the routine periodic acid–Schiff (PAS) stained glass slide was used. The PAS stain allows for better differentiation between myelopoietic cells (PAS positive) and erythropoietic cells (PAS negative) than an H&E stain,[1] thus facilitating the annotation that was needed for training the neural network. All slides were scanned with a Panoramic Flash II 250 scanner (3DHistech, Hungary) using a 40× objective, resulting in digital whole-slide images (WSIs) with a pixel resolution of 0.25 µm.

The annotations used to train the neural network were made by an experienced pathologist (KMH) and two trained students. In the training and tuning set, erythropoiesis, myelopoiesis, lipocytes, and megakaryocytes were randomly selected and sparsely annotated with point annotations at their approximate cell centres. To fully cover cells, the point annotations were extended to circles with a diameter equal to the approximated dataset-wide average diameter for that particular cell type. Trabecular bone and fields of erythrocytes were annotated with polygon masks. In the test set, all cells in randomly spaced bounding boxes were exhaustively point annotated and extended to circles. Cells inside the bounding boxes that did not belong to one of the six classes (for example plasma cells) were not included in the evaluation.

For estimation of the cellularity by the pathologists and the neural network, coarsely marked regions of interest (ROIs) were annotated. The two trained students marked ROIs that were free of major haemorrhage, cortical bone, lymphoid aggregates and crush artifacts in all WSIs. Non-fat unstained spaces in the ROIs, such as large vessel lumens, sinuses or tissue tears were excluded by the use of a tissue segmentation network,[16] which treated white structures greater than approximately 200 µm as background. A typical biopsy could contain multiple ROIs, each multiple millimeters in length. An example of the annotation and ROIs is shown in Fig. 1.

Methods

Training and evaluation of neural network for segmentation

We developed a neural network for the segmentation of six major cell and tissue types in bone marrow biopsies: erythropoiesis, myelopoiesis, megakaryocytes, lipocytes, trabecular bone, and erythrocytes. For the selection of the network architecture, we chose a fully convolutional[17] neural network with a VGG16-like[18] architecture, as this was most suitable for the sparse point annotation available for training the neural network. The network consisted of 10 convolutional layers separated by ReLU non-linearities, and a softmax function at the end for classification. Details are given in the supplementary data (Appendix A). The network was trained using patches sampled from areas annotated in the training set. During training, extensive data augmentation was applied to the patches according to the HSV-Light method published by Tellez *et al.*[19] to improve the ability of the neural network to generalise to the stain variation found in the dataset. The neural network was developed using Python 3.6 with Keras (2.3.0) as a framework.[20]

Multiple pre- and postprocessing steps were taken before and after the application of the neural network. The tissue segmentation network was used to differentiate between background and bone marrow; the neural network was only applied to the bone marrow. During application, noise was reduced by test time augmentation (averaging over the segmentation output of 8 rotations/flips of the input patch) followed by median filtering (5×5 pixels). Lastly, connected component analysis was used to remove spuriously detected objects that were smaller than 50% of the dataset-wide average diameter of the predicted cell/tissue type.

Table 1 Demographic data and bone marrow cellularity per age category

Age, years	No. cases	Male/female	Bone marrow cellularity (%)[a]	Min	Max
0–9	1	1/0	56.5	56.5	56.5
10–19	2	1/1	53.9 ± 16.6	37.3	70.5
20–29	7	5/2	51.6 ± 4.6	32.3	65.7
30–39	10	5/5	48.0 ± 4.0	25.1	68.5
40–49	12	5/7	46.5 ± 3.5	20.4	63.9
50–59	33	18/15	42.5 ± 2.0	16.1	62.5
60–69	37	22/15	39.2 ± 1.9	15.3	62.8
70–79	23	9/14	39.5 ± 2.1	24.7	68.2
80–100	5	3/2	33.9 ± 6.8	6.8	52.3

[a] Reported values are mean ± standard error of mean (SEM), measured via neural network.

Please cite this article as: van Eekelen L et al., Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation, Pathology, https://doi.org/10.1016/j.pathol.2021.07.011

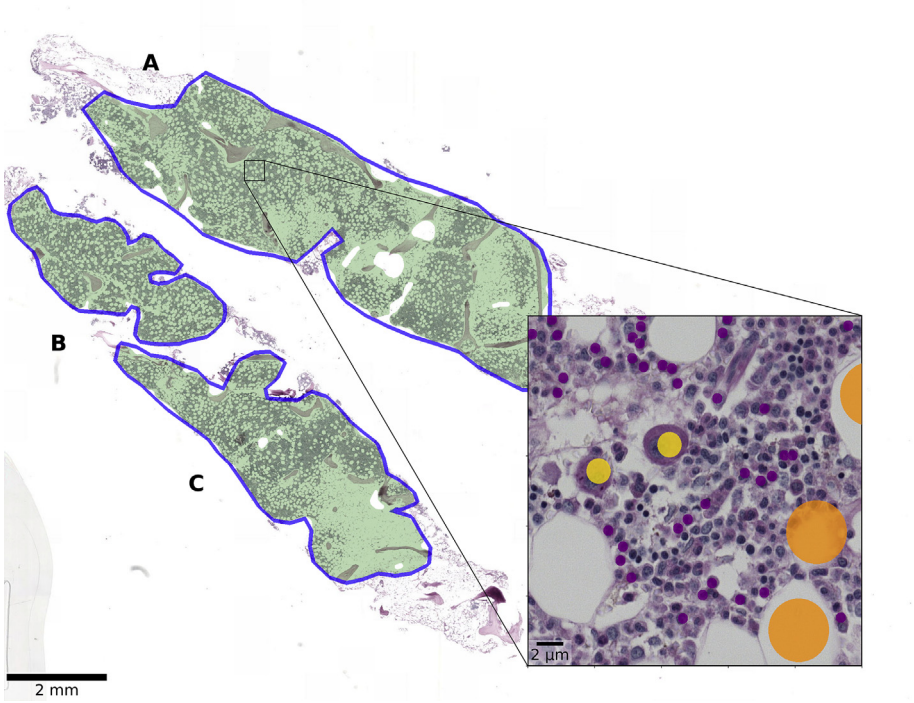

Fig. 1 An overview of the annotation used for training the neural network for estimating the cellularity in biopsies. Two trained students selected regions of interest (ROIs) of representative PAS-stained marrow (marked in blue), that were free of large areas of haemorrhage, cortical bone or crush artifacts. A tissue segmentation network was used to exclude white space larger than approximately 200 µm, such as tissue tears or large vessel lumen: all tissue included in the calculation of cellularity by the tissue segmentation network is shown in green. Note the heterogeneous distribution of cellularity within ROIs A and C. The inset shows the sparse annotation used to train the neural network. Cells were annotated at their approximate centre. This point annotation was extended to circles of an average diameter for that particular cell type. Myelopoiesis is shown in purple, megakaryocytes are shown in yellow, and lipocytes are shown in orange.

Experiments

In total, 130 bone marrow biopsies were used for this study. Fig. 2 shows the study design, detailing which biopsies were used for which experiments.

Twenty-one randomly selected WSIs were used for the development and evaluation of the neural network, split in a training, tuning, and test set of 14, 5, and 2 WSIs, respectively. The tuning set was used to tune the hyperparameters of the neural network and monitor for overfitting during training. In total, 7864 annotations were made across the training and tuning set. The network weights that performed best on the tuning set were applied as a final model on the test set. The full annotation for the test consisted of nine bounding boxes (on average 500×500 µm) across the two WSIs with a total number of 11,444 annotations. The test set was used after the finalisation of the neural network to evaluate its segmentation performance. This performance was evaluated per cell type on a pixelwise basis by measuring accuracy and Dice score.[21] We also evaluated each cell type on an objectwise basis by measuring the number of annotated cells that were segmented as one class for 50% or more of their area. If this class was correct, the prediction was counted as a true positive, else it was counted as false positive. We report this object detection rate in a normalised confusion matrix with values ranging from 0 (no detected objects) to 1 (all objects detected).

To quantify the cellularity of biopsies with the neural network, we applied the network to all ROIs in the WSIs, resulting in a prediction of the cell/tissue type of every pixel. Then, within the ROIs, all pixel occurrences of haematopoiesis (erythropoiesis, myelopoiesis, megakaryocytes) and fat were counted. For each WSI, the total cellularity of the bone marrow over all ROIs was calculated as:

$$C = \#\,\text{haematopoiesis}/(\#\,\text{haematopoiesis} + \#\,\text{fat}) * 100$$

One pathologist (KMH) performed visual estimation of bone marrow cellularity in all 130 WSIs. For WSIs with heterogeneous cellularity (determined visually), the median cellularity of all ROIs was taken. The agreement between the network and the pathologist was measured on 109 WSIs, kept separate from the 21 WSIs used for the development of the neural network to avoid bias. Overall agreement was measured for the 109 WSIs and in three subgroups, stratified according to cellularity: low (<40%), normal (40–60%) or high (>60%), as determined by the neural network. To establish a baseline for human performance, we also measured the agreement between two pathologists (KMH and MVDB) on five of the 109 WSIs (randomly selected). To increase the number of available data points for the statistical analysis, we measured agreement on a per-ROI basis (39 ROIs total, 8 ROIs per WSI on average). The pathologists each performed independent visual estimation twice with a washout period of 1 month, allowing us to measure both inter- and intra-rater agreement. We also measured the agreement between the network and the median of the pathologists. All visual estimation was done in increments of 5%, while the pathologists were blind to the age and sex of the patient.

The cellularity measurements by the neural network and the pathologist on all 130 WSIs were used to create separate age-related trendlines for cellularity, which were compared to trendlines in the literature. We also made an age-related trendline for trabecular bone surface (TBS). TBS was calculated as the percentage of bone pixels in the ROIs as predicted by the neural network.

Statistical analysis

For all agreement analyses, we calculated the intraclass correlation (ICC) using a two-way random effect model with absolute agreement [ICC(2,1)]. ICC values below 0.5 were considered poor, between 0.50 and 0.75 moderate, between 0.75 and 0.90 good and above 0.90 excellent, according to guidelines proposed by Koo and Li.[22] Data normality was tested using D'Agostino's K:[2] this suggested that cellularity (both visually estimated and algorithmically) was distributed normally, but the age of patients was not. To determine the correlation between age and cellularity in different age categories, a Kruskal-Wallis H-test was used (the non-parametric equivalent of a one-way ANOVA F-test). A significance level of 0.05 was chosen for all statistical tests. Because few patients at the extremes of the age spectrum were in the cohort, we excluded patients below 20 years and above 80 years from any statistical tests, regression or conclusions regarding age-cellularity correlation. All statistical tests were performed using Python 3.6 with the NumPy (1.17.2),[23] pandas (0.25.1),[24] scipy (1.3.1),[25] pingouin (0.3.5),[26] matplotlib (3.1.1)[27] and statsmodels (0.11.1)[28] packages.

Please cite this article as: van Eekelen L et al., Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation, Pathology, https://doi.org/10.1016/j.pathol.2021.07.011

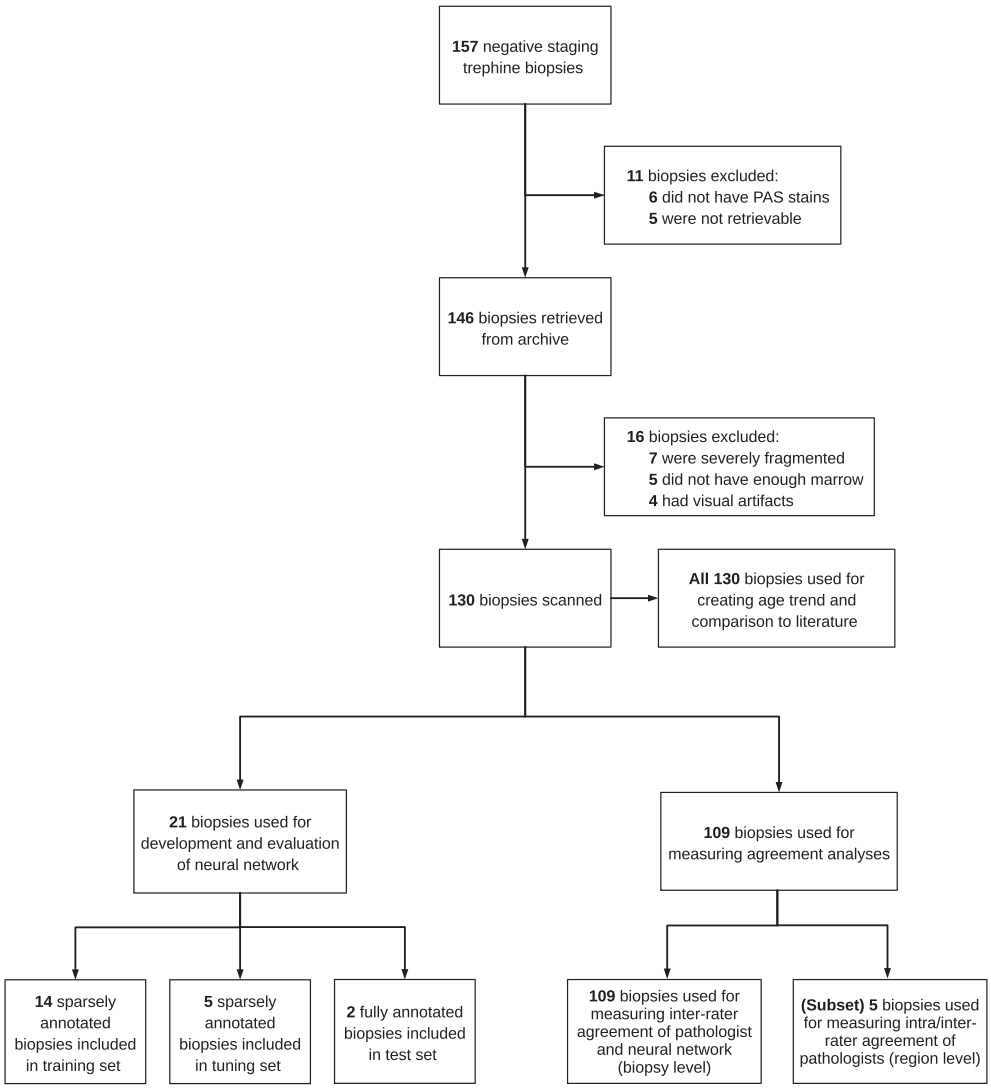

Fig. 2 Study design. (*) Cases were stratified according to cellularity as measured by the neural network.

RESULTS

Bone marrow segmentation

The pixelwise accuracy and Dice scores for each cell type/tissue are shown in Table 2. Overall, the average accuracy and Dice score were 0.95 and 0.76, respectively. Trabecular bone was segmented best with a pixelwise accuracy and Dice score of 0.9964 and 0.9896, respectively. The lowest pixelwise accuracy was achieved on lipocytes with 0.8796, while the lowest Dice score was obtained on megakaryocytes: 0.4769.

The average object detection rate was 0.83. The detection rates of objects (segmented for ≥50% of their area by a single class) are shown in Fig. 3. The detection rates for erythropoiesis, myelopoiesis, and megakaryocytes were good (0.86, 0.76 and 0.95, respectively). When incorrectly detected, erythropoiesis was mostly confused with myelopoiesis, and vice versa. All areas of trabecular bone were correctly detected. For erythrocyte accumulations (haemorrhages), which are barely PAS stained, the detection rate was low (0.43), often being confused for megakaryocytes, lipocytes, or myelopoiesis.

Table 2 Pixelwise performance metrics for the bone marrow segmentation per cell type

	Pixelwise accuracy	Pixelwise Dice score
Erythropoiesis	0.9626	0.7053
Myelopoiesis	0.9251	0.8106
Megakaryocytes	0.9862	0.4769
Trabecular bone	0.9964	0.9896
Lipocytes	0.8796	0.8800
Erythrocytes	0.9882	0.7104

Please cite this article as: van Eekelen L et al., Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation, Pathology, https://doi.org/10.1016/j.pathol.2021.07.011

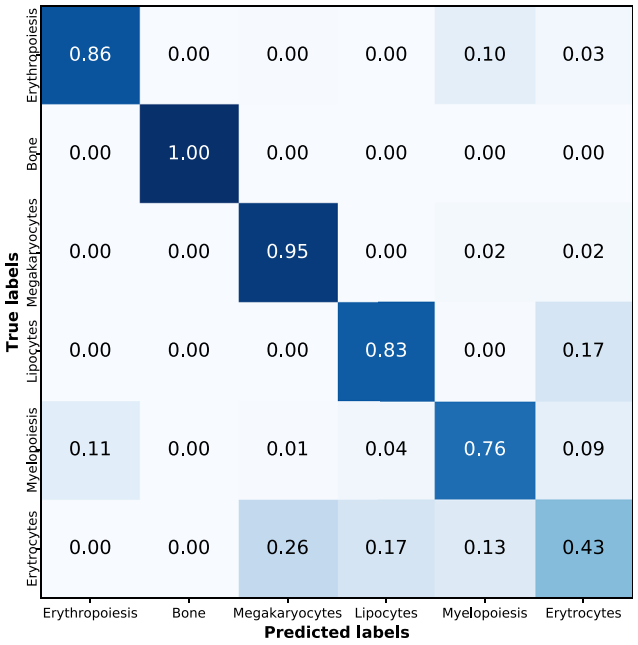

Fig. 3 The detection rate of objects (segmented for ≥50% of their area by a single class) in the test set is shown in a confusion matrix. True labels are the annotations by the pathologist, predicted labels are by the neural network. The average detection rate was 0.83. When incorrectly detected, erythropoiesis and myelopoiesis were most often misclassified as one another. Erythrocyte accumulations were often mistaken for megakaryocytes, lipocytes, or myelopoiesis, resulting in a low detection rate for erythrocytes.

In visual examples of the segmentation it can be seen that erythropoietic islands were segmented well. Bone, lipocytes, and megakaryocytes were generally segmented well, but were often oversegmented at their borders, meaning that smaller cells surrounding these structures may be incorrectly segmented. Visual examples of the segmentation are shown in Fig. 4, alongside the annotations and the original images. Fig. 4G–I shows that dark, uniform patches of plasma and the tightly clustered cells were sometimes mistaken for megakaryocytes.

Agreement of cellularity quantification

We plot the cellularity measurements by the network in all 109 WSIs against the estimations of the pathologist (KMH) in Fig. 5A. The associated agreement metrics are summarised in Table 3. Overall, there was a strong correlation (R^2=0.7) and moderate to good agreement (ICC=0.78, 95% CI 0.58–0.85). The mean difference between the two values was −4.68 (95% CI −21.09–11.73), indicating that the neural network systematically measured a lower cellularity than the pathologist's estimations.

Agreement was substantially lower when biopsies were stratified into low (<40%), normal (40–60%) and high (>60%) cellularity subgroups (as measured by the neural network). However, the ICC 95% confidence intervals of the three subgroups were comparable, as were their mean differences, indicating that the difference between the network and pathologist was not associated with the cellularity of the biopsy.

The agreement metrics for the two pathologists and the neural network on ROIs in five randomly selected WSIs are shown in Table 4. Individual measurements and estimations are shown in Fig. 5B,C. Overall, intra-rater agreement was good to excellent (ICC=0.98 for both pathologists) and the mean difference between rounds was approximately 1%. The inter-rater agreement of the pathologists was lower in both rounds (ICC=0.887, 95% CI 0.61–0.99; and ICC=0.934, 95% CI 0.67–1.0) and lower still for the inter-rater agreement of the neural network and the median of the pathologists' estimation (ICC=0.818 and 0.797). We show the agreement of the network with the individual pathologists in Table 4. As was the case for the whole-slide level measurements, the neural network systematically measured a lower cellularity than the pathologists (approximately −10%).

Age and gender-related changes of cellularity

For the analysis of age-related changes in cellularity, we divided the 130 patients into age categories according to decade (Table 1). Fig. 6A shows the age-related cellularity, both by the pathologist (KMH) and the neural network, and in Fig. 6B we compare this to age-related cellularity measured by Hartsock *et al.*,[2] Ogawa *et al.*[3] and Hagiya *et al.*[4] We found a steady downward trend of cellularity with age. An H-test confirms that the median cellularity differs between the age categories (H=13.1, p=0.02). Fig. 6C shows age-related cellularity stratified by gender. Ignoring the few cases at the extremes of the age spectrum, the male patients showed no age-related change in cellularity (H=3.96, p=0.6), while female patients showed higher levels below 50 years and a continuous downward trend throughout life (H=13.6, p=0.02). We show the age-related cellularity, proportions of cell types and other metrics of individual biopsies in scatterplots in Supplementary Fig. 2–4 (Appendix A).

Age and gender-related changes of trabecular bone surface

The age-related changes of TBS were examined using the same age categories as for cellularity. Results are shown in Fig. 6D, both for all patients and stratified per gender. There was a significant downward trend for all patients taken together (H=21.1, p<0.001) and for female patients (H=12.2, p=0.03), but not for male patients (H=10.3, p=0.07).

DISCUSSION

In this study we presented a neural network capable of segmenting clinically relevant cell lineages and tissue in digitised PAS-stained bone marrow trephine biopsy slides. We used the segmentation output to investigate the relationship between bone marrow cellularity and age and compared to sources in the literature that used different techniques of quantification. We also quantified agreement between the neural network and visual estimates of different pathologists and compared to human-level performance. In addition to the haematopoietic cellularity, we quantified trabecular bone surface and various cell ratios, and compared these between age groups and gender. The technique shown here is one small example of the potential of machine learning in future clinical reporting of bone marrow trephine biopsies.

The neural network performed well at segmenting cells and tissue, with an average pixelwise accuracy and Dice score of 0.95 and 0.76, respectively, and an average object detection rate of 0.83. Erythropoiesis, myelopoiesis, bone, and lipocytes were segmented well, but the segmentation performance on megakaryocytes and erythrocyte aggregates

Please cite this article as: van Eekelen L et al., Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation, Pathology, https://doi.org/10.1016/j.pathol.2021.07.011

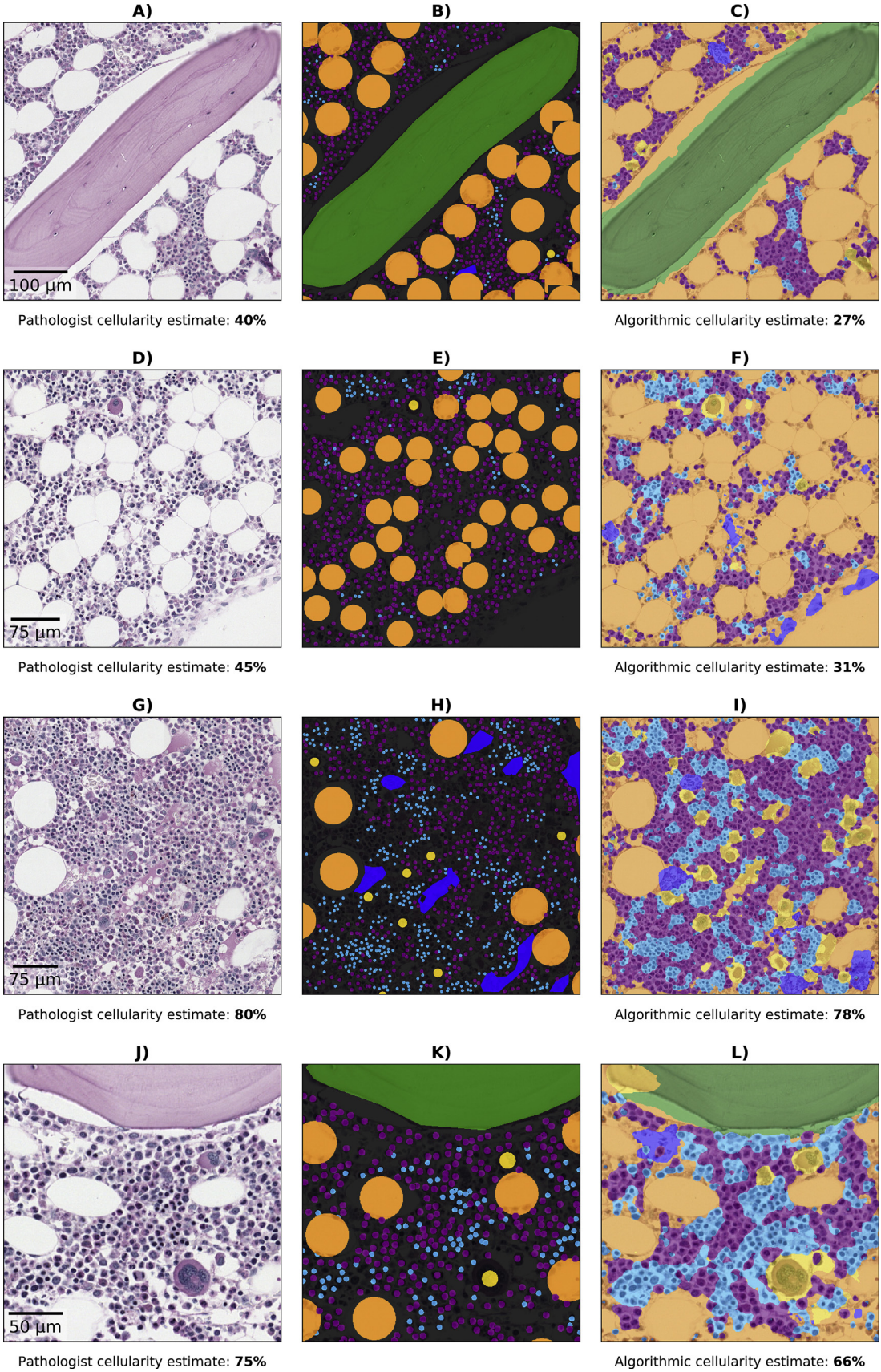

Fig. 4 An overview of segmentation and cellularity estimation results from the neural network compared to the annotations and cellularity estimation by the pathologist (KMH) From left to right, the original PAS-stained biopsies (A,D,G,J), annotation (B,E,H,K) and segmentation results (C,F,I,L) are shown. Erythropoiesis (light blue) was generally well segmented, as were lipocytes (orange), megakaryocytes (yellow) and trabecular bone (green). Performance on intercellular space (no annotation) was low, often segmented as megakaryocytes or lipocytes, as intercellular space was not included in annotations for the training of the network. Tightly clustered cells were regularly incorrectly segmented as megakaryocytes.

Please cite this article as: van Eekelen L et al., Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation, Pathology, https://doi.org/10.1016/j.pathol.2021.07.011

(haemorrhage) was lower than average: the Dice score for megakaryocytes was 0.4769 and the detection rate of erythrocyte aggregates was 0.43. For megakaryocytes, the network had the tendency to incorrectly segment tightly clustered, dark cells as megakaryocytes, possibly because the network mistook those cells as the nuclei of megakaryocytes. The low detection rate of erythrocytes reflected the fact that segmentation performance of intercellular space was low: light coloured tissue cavities were often predicted as lipocytes and darker coloured patches of plasma/erythrocytes were often predicted as megakaryocytes**.** We believe this is a consequence of the sparse point annotation used for training the neural network. This annotation method does not offer the neural network any information on what is between cells, causing it to be unreliable in these areas.

When using the segmentation output to measure the bone marrow cellularity of biopsies at a whole-slide level, there was moderate to good agreement with visual estimation by the pathologist (KMH) (ICC=0.78, 95% CI 0.58–0.85). At a ROI level, results were similar but with a larger confidence interval (ICC=0.818, 95% CI 0.63–1.0, in round 1; and ICC=0.797, 95% CI 0.61–0.98, in round 2). The intra-rater agreement of both pathologists was very high and their inter-rater agreement was slightly lower but still moderate to good (ICC=0.887, 95% CI 0.61–0.99, in round 1; and ICC=0.934, 95% CI 0.67–1.0, in round 2). It should be noted that the pathologists both specialise in haematopathology, were from the same medical centre, and regularly align their method of estimation with each other. The inter-rater agreement of the neural network was slightly lower than the inter-rater agreement of the pathologists.

The neural network systematically measured a lower cellularity than the estimation of the pathologists, as indicated by a mean difference of approximately −5% at a whole-slide level and −10% at a ROI level. We attribute this lower prediction to oversegmentation of lipocytes at their border, leading to a larger area measurement for fat tissue. This was caused by the fact that extending the point annotation to

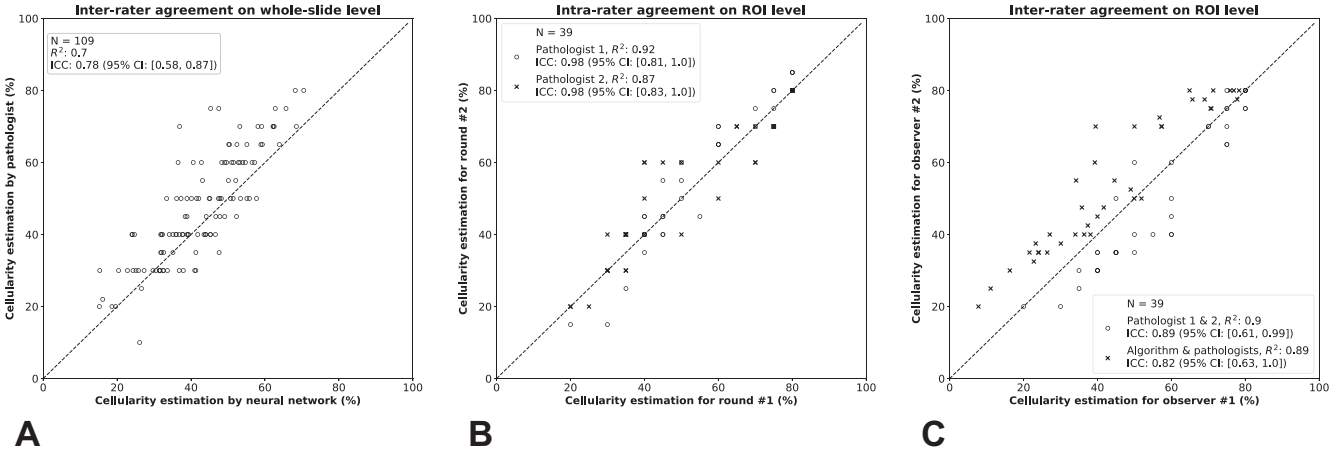

Fig. 5 Scatter plots depicting the (A) inter-rater agreement on a whole slide level of the neural network and the pathologist (KMH), (B) the intra-rater agreement on a region of interest (ROI) level of both pathologists, and (C) the inter-rater agreement on a ROI level of both pathologists over the first round, and the algorithm and the median of the pathologists' predictions.

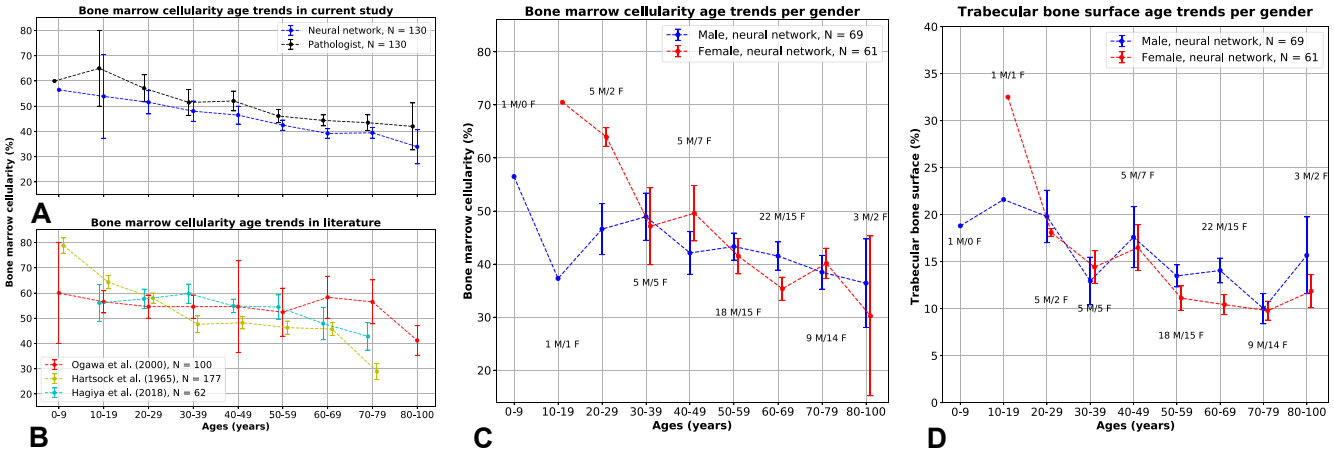

Fig. 6 (A) The relationship between age and mean cellularity per age category determined by (i) applying the neural network to 130 biopsies and measuring the percentage of haematopoiesis in the total marrow in the resulting segmentation, and (ii) visual estimation by a pathologist (KMH). A Kruskal–Wallis H-test found a significant difference between the median cellularity of the age categories (H=13.1, p=0.02). (B) Age trends found in literature. (C) For the present study, when stratified by gender, a downward trend in cellularity for females was seen, but not for males, which was confirmed by H-tests (male: H=3.96, p=0.6; female: H=13.6, p=0.02). (D) The relationship between age and mean trabecular bone surface (TBS) per age category calculated as the percentage of pixels predicted as bone by the neural network in the regions of interest. An H-test found a significant difference between the median TBS of the age categories for females (H=12.2, p=0.03), but not for males (H=10.3, p=0.07). For Fig. 6C,D, age categories are labelled with the number of males (M) and females (F). Error bars in all plots represent the standard error of the mean (SEM). Points in all plots are slightly offset around their corresponding age category to increase legibility.

Please cite this article as: van Eekelen L et al., Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation, Pathology, https://doi.org/10.1016/j.pathol.2021.07.011

Table 3 Intraclass correlation values (ICC) and mean differences between the cellularity estimation of the neural network and the pathologist (KMH) on a whole-slide level (n=109) and stratified according to cellularity

Selection	ICC (95% CI)	Mean difference, % (95% CI)
All (n=109)	0.775 (0.58–0.87)	−4.68 (−21.09–11.73)
Low cellularity[a] (<40%, n=47)	0.500 (0.19–0.70)	−4.86 (−20.67–10.94)
Normal cellularity[a] (40–60%, n=53)	0.401 (0.15–0.60)	−4.06 (−21.92–13.81)
High cellularity[a] (>60%, n=9)	0.220 (0.00–0.67)	−7.61 (−15.35–0.13)

CI, confidence interval.
[a] As measured by the neural network.

Table 4 Intraclass correlation values (ICC) and mean differences for cellularity estimation in regions of interest (n=39) across 5 WSIs by two pathologists and the neural network

Selection	ICC (95% CI)		Mean difference, % (95% CI)	
Intra-rater agreement				
Pathologist #1 (KMH)	0.978 (0.81–1.0)		−0.897 (−12.19, 10.39)	
Pathologist #2 (MVDB)	0.976 (0.83–1.0)		−1.154 (−15.27–12.96)	
Intra-rater agreement	Round 1	Round 2	Round 1	Round 2
Pathologist #1 and #2	0.887 (0.61–0.99)	0.934 (0.67–1.0)	6.026 (−7.14–19.19)	5.769 (−8.19–19.72)
Neural network and median of pathologists	0.818 (0.63–1.0)	0.797 (0.61–0.98)	−9.574 (−22.52–3.38)	−10.60 (−22.16–0.96)
Neural network and pathologist #1	0.713 (0.59–0.97)	0.729 (0.59–0.97)	−12.587 (−24.28–0.87)	−13.49 (−26.94–0.03)
Neural network and pathologist #2	0.884 (0.68–0.99)	0.852 (0.67–0.98)	−6.562 (−23.45–10.33)	−7.715 (−21.2–5.83)

Inter-rater values are shown for round 1 and 2.

circles resulted in varying degrees of over/undersized annotation. This may have also caused the network to sometimes fail to recognise haematopoiesis from lipocytes, especially in marrow regions of low cellularity.

The problem of unreliable performance in intercellular space and the oversegmentation of lipocytes could be solved by using more sophisticated ways to extend point annotations to full annotation (e.g., NuClick[29] or region-growing algorithms) or by training with fully annotated areas of bone marrow. Thereby, we offer the network more information on intercellular space and the border of cells. This is also one of many important prerequisites for applying this technique on biopsies where morphology and cell diameter can vary significantly due to the presence of a pathology. For this, purpose-made annotation on slides containing the pathology of interest could also be made.

Our results are comparable to previous studies that investigate agreement between digital image analysis and visual estimation. Hagiya *et al.*[4] compared automatic cellularity measurements by a proprietary HALO Imaging algorithm to visual estimates of three pathologists and found good agreement between both methods (ICC=0.81). Kim *et al.*[12] developed a technique to count the nucleated cells per unit area in a manually selected region of a digital scan of bone marrow and found a high correlation (R^2=0.816) with visual estimation. Both studies[4,12] also found inter-rater agreement of visual estimation by pathologists comparable to our study (ICC=0.88–0.91 and 0.870, respectively). Nielsen *et al.*[10] trained a support vector machine to segment biopsies in red (haematopoiesis) and yellow (lipocytes) marrow and used this in a similar way to the present paper to calculate cellularity, with good agreement with visual estimation (ICC=0.799). They reported a higher Dice score for lipocytes (0.9001 versus our 0.88), possibly because they used full segmentation over point annotation. For red marrow, they did not distinguish between erythropoiesis, myelopoiesis, and megakaryocytes, prohibiting direct comparison of the segmentation performance. Most recently, Tratwal *et al.*[11] developed a semi-automatic technique for determining cellularity via segmenting biopsies into bone, nucleated cells (haematopoiesis), lipocytes and interstitium and found good correlation (R^2=0.85 for human trephine biopsies). In contrast to our approach, theirs requires manual background/artifact delineation as a preprocessing step in addition to ROI selection.

Using the segmentation output of the neural network, we found a statistically significant downward trend for bone marrow cellularity and trabecular bone volume with age in a cohort of 130 (predominantly) lymphoma patients.

According to the observed trend, mean cellularity begins at approximately 50% in the third decade of life and then decreases ±2% per decade to 40% in the seventh and eighth decade. The trend found by the pathologist (KMH) in the same 130 cases is consistently 5% higher, but follows a similar rate of decrease. When stratifying by gender, females over 20 years old showed a similar (significant) age-related decrease in cellularity, while this was not apparent in male patients over 20 years old. Although it is tempting to speculate that this effect is caused by activation of the marrow by menstrual blood loss or hormonal influences, this difference between genders should be interpreted with caution, given the relatively small number of patients in some age categories. We are also hesitant to interpret the age-related trend of TBS. While the TBS trend was in the same order-of-magnitude as previous papers,[30–32] the used ROIs did not ensure that bone was proportionally sampled since the focus of the project was on the marrow and not on the bone, possibly giving skewed values.

Please cite this article as: van Eekelen L et al., Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation, Pathology, https://doi.org/10.1016/j.pathol.2021.07.011

A steady downward trend of age-related bone marrow cellularity was not observed in previous studies. Hartsock *et al.*[2] found that cellularity progressed in three distinct periods: cellularity steadily decreased from 80% to ±50% in the first three decades, then remained constant until the eighth decade, and then decreased to approximately 30%. Ogawa *et al.*[3] found no definite age-related change: their results indicate cellularity remains constant between 60% and 55% throughout the first eight decades of life and only significantly decreases to 40% in the ninth and tenth decade. Both Hagiya *et al.*[4] and Schnitzler *et al.*[5] found no decrease in cellularity with age over an adult population.

We attribute this lack of consensus in the literature on the relationship between bone marrow cellularity and age to the differences in clinical features of the examined cohorts and the wide variety in methods used for determining cellularity. Ogawa *et al.*,[3] Hagiya *et al.*,[4] and the present study all measured cellularity in patients without haematopoietic malignancies, while Hartsock *et al.*[2] measured cellularity in cases of sudden death. Ogawa *et al.*[3] studied bone marrow biopsies predominantly taken from the sternum, while other studies used biopsies taken from the iliac crest, two sites with different marrow distributions with age.[33] Ethnic origin seems to matter as well.[5] Perhaps most importantly, some papers investigated cellularity by visual estimation,[34] while some used point counting[2,5] and others applied analog or digital image analysis.[3,4] There is a noticeable difference in accuracy between these methods.[35]

Taken together, the present study and previous literature on bone marrow cellularity show that the cellularity range in biopsies considered to be normal (negative staging biopsies) is extremely broad, roughly ranging from 30 to 70% in the age range 50–80 years, the general population of patients requiring a bone marrow biopsy. Therefore comparing individual patients or characterising biopsies as 'hypo' or 'hypercellular' requires caution and consideration of the clinical context.

This study was subject to a number of limitations. Our cohort consisted predominantly of lymphoma patients, which may inadvertently have affected haematopoiesis, for example by cytokine production or autoimmune effects. We did not control for other factors that might affect haematopoiesis, such as smoking, alcohol consumption, pharmaceutical agents or systemic illness. An inevitable technical limitation is that the neural network had to be evaluated on exhaustively annotated areas, while it is impossible to determine the lineage of each cell in a biopsy with certainty (as opposed to an aspirate, because of the sectioning of the tissue). Therefore, some predictions of the neural network that were incorrect according to the annotation could indeed be correct.

Due to the current availability of digital pathology in an increasing number of pathology departments,[36] the quantification of cells in bone marrow using digital image analysis techniques will become feasible to support routine diagnostics in the near future. This analysis can be rapid and high-throughput, exhaustive and quantify all different cell types present within the marrow. We show how this quantification can be used to calculate relevant diagnostic metrics such as bone marrow cellularity. Future work includes adapting this technique for more routinely used stains such as H&E (possibly with the use of a stain-transforming CycleGAN),[37] using it for tasks such as the detection of deviant morphology in primary bone marrow disease or subtyping cells according to their morphology (similar to how Sirinukunwattana *et al.*[38] use megakaryocyte morphology for phenotyping), detecting the presence of infiltrating cells and performing longitudinal studies.

To conclude, a neural network was developed for the segmentation of cells and tissues, and the quantification of cellularity in digitised bone marrow biopsies. Cellularity measurements agreed well with visual estimates by experienced pathologists. The relationship between age and cellularity was examined in a cohort of 130 patients, showing a significant downward trend from an average of 50% in the third decade of life to approximately 40% in the seventh and eighth decade with a steady decrease rate of ±2% per decade. The wide normal range of bone marrow cellularity in the adult population (30–70%) has to be taken into account during clinical consideration of hypo- and hypercellularity.

Conflicts of interest and sources of funding: This study was funded by a grant from the Dutch MPN Foundation. The authors state that there are no conflicts of interest to disclose.

APPENDIX A. SUPPLEMENTARY DATA

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pathol.2021.07.011.

Address for correspondence: Dr Konnie M. Hebeda, Radboud University Medical Center, Department of Pathology (812), PO Box 9101, 6500 HB, Nijmegen, the Netherlands. E-mail: konnie.hebeda@radboudumc.nl

References

1. Riley RS, Williams D, Ross M, *et al*. Bone marrow aspirate and biopsy: a pathologist's perspective. II. Interpretation of the bone marrow aspirate and biopsy. *J Clin Lab Anal* 2009; 23: 259–307.
2. Hartsock RJ, Smith EB, Petty CS. Normal variations with aging of the amount of hematopoietic tissue in bone marrow from the anterior iliac crest: a study made from 177 cases of sudden death examined by necropsy. *Am J Clin Pathol* 1965; 43: 326–31.
3. Ogawa T, Kitagawa M, Hirokawa K. Age-related changes of human bone marrow: a histometric estimation of proliferative cells, apoptotic cells, T cells, B cells and macrophages. *Mech Ageing Dev* 2000; 117: 57–68.
4. Hagiya AS, Etman A, Siddiqi IN, *et al*. Digital image analysis agrees with visual estimates of adult bone marrow trephine biopsy cellularity. *Int J Lab Hematol* 2018; 40: 209–14.
5. Schnitzler CM, Mesquita J. Bone marrow composition and bone microarchitecture and turnover in blacks and whites. *J Bone Miner Res* 1998; 13: 1300–7.
6. Kerndrup G, Pallesen G, Melsen F, *et al*. Histomorphometrical determination of bone marrow cellularity in iliac crest biopsies. *Scand J Haematol* 1980; 24: 110–4.
7. Ho-Yen DO, Slidders W. Bone marrow cellularity assessed by point-counting. *J Clin Pathol* 1978; 31: 753–6.
8. Al Adhadh AN, Cavill I. Assessment of cellularity in bone marrow fragments. *J Clin Pathol* 1983; 36: 176–9.
9. Tuzuner N, Bennett JM. Reference standards for bone marrow cellularity. *Leuk Res* 1994; 18: 645–7.
10. Nielsen FS, Pedersen MJ, Olsen MV, *et al*. Automatic bone marrow cellularity estimation in H&E stained whole slide images. *Cytometry A* 2019; 95: 1066–74.
11. Tratwal J, Bekri D, Boussema C, *et al*. MarrowQuant across aging and aplasia: a digital pathology workflow for quantification of bone marrow compartments in histological sections. *Front Endocrinol* 2020; 11: 480.
12. Kim Y, Kim M, Kim Y, *et al*. Estimation of bone marrow cellularity using digital image nucleated cell counts in patients receiving chemotherapy. *Int J Lab Hematol* 2014; 36: 548–54.
13. Litjens G, Kooi T, Bejnordi BE, *et al*. A survey on deep learning in medical image analysis. *Med Image Anal* 2017; 42: 60–88.
14. Song T-H, Sanchez V, Ei Daly H, *et al*. Simultaneous cell detection and classification in bone marrow histology images. *IEEE J Biomed Health Inform* 2019; 23: 1469–76.

Please cite this article as: van Eekelen L et al., Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation, Pathology, https://doi.org/10.1016/j.pathol.2021.07.011

15. de Laak–de Vries I, Siebers AG, Burgers L, *et al*. How we do: optimizing bone marrow biopsy logistics for sign-out within 2 days. *J Hematop* 2016; 9: 67–71.
16. Bándi P, Balkenhol M, Van Ginneken B, *et al*. Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks. *PeerJ* 2019; 2019: https://doi.org/10.7717/peerj.8242.
17. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 8 Mar 2015; cited Aug 2020: http://arxiv.org/abs/1411.4038
18. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 10 Apr 2015; cited Aug 2020. http://arxiv.org/abs/1409.1556
19. Tellez D, Litjens G, Bándi P, *et al*. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med Image Anal* 2019; 58: 101544.
20. Keras. Cited Aug 2020. https://keras.io
21. Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945; 26: 297–302.
22. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; 15: 155–63.
23. Harris CR, Jarrod Millman K, van der Walt SJ, *et al*. Array programming with NumPy. *Nature* 2020; 585: 357–62.
24. McKinney W. Data Structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. SciPy, 2010. http://conference.scipy.org/proceedings/scipy2010/mckinney.html
25. Virtanen P, Gommers R, Oliphant TE, *et al*. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; 17: 261–72.
26. Vallat R. Pingouin: statistics in Python. *J Open Source Softw* 2018; 3: 1026.
27. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng* 2007; 9: 90–5.
28. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference. SciPy, 2010. https://conference.scipy.org/proceedings/scipy2010/seabold.html
29. Alemi Koohbanani N, Jahanifar M, Zamani Tajadin N, *et al*. NuClick: a deep learning framework for interactive segmentation of microscopic images. *Med Image Anal* 2020; 65: 101771.
30. Burkhardt R, Kettner G, Böhm W, *et al*. Changes in trabecular bone, hematopoiesis and bone marrow vessels in aplastic anemia, primary osteoporosis, and old age: a comparative histomorphometric study. *Bone* 1987; 8: 157–64.
31. Marcus R, Kosek J, Pfefferbaum A, *et al*. Age-related loss of trabecular bone in premenopausal women: a biopsy study. *Calcif Tissue Int* 1983; 35: 406–9.
32. Mosekilde L, Mosekilde L. Iliac crest trabecular bone volume as predictor for vertebral compressive strength, ash density and trabecular bone volume in normal individuals. *Bone* 1988; 9: 195–9.
33. Cristy M. Active bone marrow distribution as a function of age in humans. *Phys Med Biol* 1981; 26: 389–400.
34. Friebert SE, Shepardson LB, Shurin SB, *et al*. Pediatric bone marrow cellularity: are we expecting too much? *J Pediatr Hematol Oncol* 1998; 20: 439–43.
35. Fong TP, Okafor LA, Schmitz TH, *et al*. An evaluation of cellularity in various types of bone marrow specimens. *Am J Clin Pathol* 1979; 72: 812–6.
36. Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. *Histopathology* 2012; 61: 1–9.
37. de Bel T, Bokhorst J-M, van der Laak J, *et al*. Residual cyclegan for robust domain transformation of histopathological tissue slides. *Med Image Anal* 2021; 70: 102004.
38. Sirinukunwattana K, Aberdeen A, Theissen H, *et al*. Artificial intelligence-based morphological fingerprinting of megakaryocytes: a new tool for assessing disease in MPN patients. *Blood Adv* 2020; 4: 3284–94.

Please cite this article as: van Eekelen L et al., Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation, Pathology, https://doi.org/10.1016/j.pathol.2021.07.011