Bone marrow pathology in essential thrombocytemia: inter-observer reliability and utility for identifying disease subtypes

Authors: Bridget S Wilkins (1) *, Wendy N Erber (2) *, David Bareford (3) *, Georgina Buck (4), Keith Wheatley (5), Clare L East (6), Beverley Paul (7), Claire N Harrison (8), Anthony R Green (9) ¥, and Peter J Campbell (9) ¥

* equal first author
¥ equal last author

Institutions:
(1) Department of Cellular Pathology, Newcastle upon Tyne Hospitals NHS Foundation Trust and Newcastle University, Newcastle upon Tyne, UK
(2) Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
(3) Department of Haematology, Russells Hall Hospital, Dudley, Birmingham, UK
(4) Clinical Trial Service Unit, Oxford, UK
(5) Birmingham Clinical Trials Unit, University of Birmingham, Birmingham, UK
(6) Oncology Clinical Trials Unit, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
(7) Department of Haematology, Bassetlaw Hospital, Worksop, UK
(8) Department of Haematology, St Thomas’s Hospital, London, UK
(9) Department of Haematology, University of Cambridge, Cambridge, UK

Addresses for correspondence:
Prof Anthony R Green, Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Hills Rd, Cambridge CB2 2XY, United Kingdom.
arg1000@cam.ac.uk

Dr Bridget S Wilkins, Northern Institute for Cancer Research, Paul O’Gorman Building, Framlington Place, Newcastle University, Newcastle upon Tyne NE2 4HH, United Kingdom.
Bridget.Wilkins@ncl.ac.uk
ABSTRACT

The role of histopathology in the diagnosis of essential thrombocythemia (ET) is controversial and there has been little attempt to quantitate inter-observer variability. Diagnostic bone marrow trephine biopsy specimens from 370 patients with ET by PVSG criteria were assessed by three experienced hematopathologists for 16 different morphological features and overall diagnosis according to the WHO classification. Our results demonstrate substantial inter-observer variability particularly for overall diagnosis and individual cellular characteristics such as megakaryocyte morphology. Reticulin grade was the dominant independent predictor of WHO diagnostic category for all three hematopathologists. Factor analysis identified three independent factors likely to reflect underlying biological processes. One related to overall and lineage-specific cellularity, and was significantly associated with JAK2 V617F status (p=0.0002), a second factor related to megakaryocyte clustering and a third was associated with the fibrotic process. No differences could be discerned between patients labelled as having ‘prefibrotic myelofibrosis’ or ‘true ET’ in clinical and laboratory features at presentation, JAK2 status, survival, thrombosis, major hemorrhage or myelofibrotic transformation. These results demonstrate that histological criteria described in the WHO classification are difficult to apply reproducibly and question the validity of distinguishing ‘true ET’ from ‘prefibrotic myelofibrosis’ on the basis of subjective morphological criteria. This study is registered at http://isrctn.org as #72251782 and at http://eudract.emea.europa.eu/ as #2004-000245-38.
INTRODUCTION

The myeloproliferative disorders (MPDs) are clonal hematological malignancies comprising 3 main disorders: essential thrombocythemia (ET), polycythemia vera (PV) and primary myelofibrosis (MF)\textsuperscript{1-3}. In 2005 the $\text{JAK2}$ V617F mutation\textsuperscript{4-7} was demonstrated in 95% of patients with PV and in just over half of those with ET and MF\textsuperscript{4,8-10}. Prior to this, the diagnosis of these disorders relied on a combination of clinical, laboratory and histological features using one of several different sets of diagnostic criteria. None of these was universally accepted, although the PVSG criteria\textsuperscript{11,12} and modifications\textsuperscript{13} were adopted for major clinical trials such as ECLAP in PV\textsuperscript{14} and PT-1 in ET\textsuperscript{15} both established in the late 1990s. In 2001, the World Health Organisation (WHO) published its criteria for the diagnosis of MPDs\textsuperscript{16-18}. This classification scheme is pathology-based and, compared to the PVSG criteria, introduced a heavy emphasis on bone marrow trephine morphology together with the concept of ‘prefibrotic myelofibrosis’ and ‘true ET’ as distinct disorders. In the last two years, testing for the $\text{JAK2}$ V617F mutation has been incorporated into new diagnostic criteria including a revised version of the WHO criteria\textsuperscript{1,19,20}.

A body of published literature, predominantly originating from the Cologne Group\textsuperscript{21-26}, underpins the histological features described in the WHO classification of patients with thrombocytemia. In recent years, this group has produced a series of publications representing multiple retrospective analyses of an expanding and well characterised archive of trephine biopsy specimens from patients with chronic myeloproliferative diseases. In particular, the authors claim that approximately 40-50% of patients with ET in fact have ‘prefibrotic myelofibrosis’ and that this entity
needs to be distinguished from ‘true ET’ \textsuperscript{21-26}. This claim is based on three main assertions:

(i) The morphological features of bone marrow trephines in patients with thrombocytopenia can be reliably subdivided into two distinct patterns \textsuperscript{22,23}.

(ii) There is minimal development of marrow fibrosis over time in patients diagnosed with ‘true ET’ \textsuperscript{22-24,26}, contrasting with development of at least mild and sometimes severe fibrosis on long-term follow-up in patients with ‘prefibrotic myelofibrosis’ \textsuperscript{24,26}.

(iii) There is an apparent reduction in life expectancy in ‘prefibrotic myelofibrosis’ compared to ‘true ET’ \textsuperscript{22,24}.

However, the interpretation of published data supporting these claims is complicated by the retrospective nature of the patient cohort analyses \textsuperscript{21-26}, the apparently overlapping nature of the cohorts studied in different papers which draw similar conclusions \textsuperscript{22-24,26}, a failure to correct for known prognostic factors in survival analyses \textsuperscript{22,24,26} and lack of details of the causes of death or definition of myelofibrotic transformation \textsuperscript{22,24,26}. In addition, there has been no characterisation of the inter-observer reliability of morphological features used to distinguish the proposed entities of ‘prefibrotic myelofibrosis’ and ‘true ET’, and it has only recently become possible to correlate histological findings with underlying molecular lesions. We have therefore addressed the role of bone marrow histology in a study of patients enrolled in three prospective studies of ET, including the PT-1 trial \textsuperscript{15}.

**METHODS**

**Study population**

Newly diagnosed and previously treated patients, aged 18 years or over, who were judged by local clinicians to meet the Polycythemia Vera Study Group (PVSG)
criteria for essential thrombocythemia, were recruited into one of three multicenter studies: the Medical Research Council PT-1 trial, in which high-risk patients were randomly assigned to either hydroxyurea plus aspirin or to anagrelide plus aspirin; the National Cancer Research Institute study for intermediate risk patients (no high risk features and age 40-60), a randomization between aspirin alone or hydroxyurea plus aspirin; or the National Cancer Research Institute study, for low risk patients (no high risk features and age <40) a prospective observational study of patients receiving aspirin alone. Patients entered a higher risk study if they developed appropriate features. Follow-up procedures and definitions of end-points have been detailed previously, but importantly all data were collected prospectively with >99% patients having complete follow-up. All end-point events were validated prospectively by a central clinical committee without knowledge of treatment allocation, and all relevant histological material from patients who suffered myelofibrotic or leukemic transformation was reviewed by a histology committee. Events occurring before 31st January, 2006 that were notified before 30th June, 2006 were included in the analysis, meaning that the median follow-up for the cohort from trial entry was 68 months. The study protocol was approved by institutional ethics committees in all centers, and written informed consent was obtained from all patients.

**Bone Marrow Trephine Specimens**

Bone marrow trephine biopsy specimens were requested from all patients enrolled in the three trials upon patient registration. Although these were not a requirement of trial entry, 636 trephine specimens were received at St Thomas’ Hospital from the 1022 patients enrolled before July 2005. For assessment, trephine biopsy sections were stained with hematoxylin and eosin (H&E) and Gordon and Sweets’ silver stain for reticulin. Staining was performed in a single laboratory for consistency.
Only bone marrow trephines taken at diagnosis were considered. Trephine biopsies that were embedded in resin without decalcification had morphology that could not be compared directly with the majority of decalcified, wax embedded specimens and these were not included in the statistical analysis. Not all paraffin-embedded sections were of sufficient length or quality to enable all parameters to be assessed. The statistical analysis therefore included a core set of 370 trephine specimens greater than 5mm in length, for which all criteria could be assessed by all three hematopathologists.

**Assessment of bone marrow trephine slides**

Trephine sections were assessed by three hematopathologists, each with >10 years of consultant-level experience and a subspecialist interest in the myeloproliferative disorders. Consensus discussions were held to agree the criteria for assessment (table 1 and figure 1) and how to assess them. Each of the 3 hematopathologists assessed the sections independently and without knowledge of patient outcomes, with only the age and sex of the patient provided for each trephine specimen (to allow determination of cellularity). An overall diagnosis was made according to the WHO criteria and recorded on a five-point scale: ‘true ET’ (0), ‘prefibrotic myelofibrosis’ (1), and manifest myelofibrosis of increasing severity (2-4). Overall cellularity and erythroid and granulocytic cellularity were scored as reduced (-1), normal (0) or increased (+1) relative to expectation for the patient’s age. Megakaryocyte cellularity was scored as normal (0), mildly increased (+1), moderately increased (+2) or severely increased (+3). Individual features of megakaryocyte morphology were scored as absent (0), present (+1) or predominant (+2). These features were staghorn megakaryocytes,
cloud-like megakaryocytes, dysplastic megakaryocytes, pyknotic megakaryocytes, and bare megakaryocytic nuclei. Megakaryocyte size was classed as predominantly small (0), mixed small and large (+1) and predominantly large (+2). Clustering of megakaryocytes was recorded as absent (0), loose (1) or tight (2) depending on assessment of the predominant pattern found. The size of clusters was recorded as no clusters (0), predominantly small clusters of <6 cells (1) or predominantly large clusters of ≥6 cells (2). In addition, the number of clusters was scored on a semi-quantitative scale as absent (0), occasional (1) or predominant (2). New bone formation and presence of paratrabecular megakaryocytes were scored as absent (0) or present (+1). Finally, reticulin staining was scored using a scale from 0 to 4:

0  Almost complete absence of fibres
1  A few scattered fibres, predominantly around stromal vessels
2  An incomplete meshwork of randomly orientated fibres; relatively few intersections
3  A more dense and complete meshwork, still with randomly orientated fibres but with many intersections
4  A denser meshwork still, with organisation of fibres into parallel bands and areas within which organisation of these parallel fibres into thicker bands is found.

Statistical analysis

Inter-observer agreement for each of the individual morphological criteria was assessed using log-linear modelling of the second-order marginal tables from the 3 pairwise comparisons among the 3 hematopathologists, as described. The model controlled for the marginal distributions of each of the pathologists, and fitted a linear-by-linear association term in order to measure the strength of inter-observer
agreement. Since samples in different pairwise marginal tables are not truly independent, the jack-knife procedure was used to correct the point estimates and 95% confidence intervals of the parameters for this dependence in the data structure, as described. Under fairly general assumptions, the linear-by-linear association term is implied by a latent structure model, suggesting that estimates of the strength of association across criteria are comparable even if criteria are scored on different scales.

The independent predictors of WHO classification score for each of the 3 hematopathologists were analysed by Bayesian proportional odds logistic regression. Firstly, the hematopathologists provided subjective assessments of the importance of individual morphological features by apportioning 100 points amongst the 16 criteria (figure 3A). Then, to identify which combination of variables optimally predicted WHO classification score, the stochastic search variable selection method was used, with prior probabilities for each variable being included derived from the subjective assessment of weights (0.05 for variables given no weight by the pathologist, and 0.3+0.03x(weight) for other variables, although the results were not sensitive to the priors used). The WHO classification score was assumed to follow a latent normally-distributed variable, and data augmentation used to estimate this underlying metric from the observed discrete response. A proportional odds probit model was estimated by constrained Gibbs sampling of the cut-points, augmented data and model parameters, as described. Finally, the model with the highest posterior probability was fitted to data scaled so that all predictors had the same range, in order to allow estimation of the relative contributions of each variable to the WHO score.
For clinical outcome, we used a composite clinical end-point of time to first arterial or venous thrombosis; major hemorrhage; myelofibrotic, leukaemic or myelodysplastic transformation; or death. Comparison of end-point rates between ‘prefibrotic myelofibrosis’ and ‘true ET’ was performed using Kaplan-Meier analysis for univariate analysis. Cox proportional hazards modelling was used for multivariate analysis, with age, sex, treatment allocation (hydroxyurea, anagrelide, or intermediate/low risk trial), prior cytoreductive therapy and history of end-point events prior to trial entry added as covariates.

To apply exploratory factor analysis, we took the consensus score from the 3 hematopathologists for each of the 16 criteria on all 370 diagnostic bone marrow trephines. Where there was disagreement, the median of the scores from the 3 hematopathologists was taken. Kaiser’s criterion (number of eigenvalues >1) was used to determine the number of factors to fit, and the model was estimated using the varimax rotation.

**RESULTS**

To establish the role of bone marrow histopathology in the diagnostic evaluation of a patient presenting with thrombocytosis, bone marrow biopsy specimens were obtained from patients enrolled in three prospective trials of ET. 370 diagnostic trephine specimens were studied independently by three experienced hematopathologists for 16 morphological criteria (table 1, figure 1). Each hematopathologist made an overall diagnosis from the trephine histology according to the WHO criteria.

**Inter-observer agreement when assessing morphological features**
Many measures of inter-observer agreement, such as Cohen’s kappa score, fail to correct for differences in the pattern of scores for each observer. We therefore used log-linear modelling of pairwise inter-observer agreements, a well-established method that explicitly models the strength of association among the observers after correcting for the distribution of scores in the cohort studied. The method estimates the odds that if two observers score two trephine specimens in adjacent categories, then they agree on which biopsy is in which category. An estimate of 1 implies no agreement beyond chance, and the greater the estimate is above 1, the more agreement there is among the observers.

A number of interesting patterns emerged from these inter-observer comparisons (table 1). Firstly, agreement on a single marker of marrow fibrosis, the reticulin grade (strength of association, 5.1; 95% CI, 4.0-6.4), was much greater than agreement on the WHO diagnosis (strength of association, 2.1; 95% CI, 1.8-2.4). The three hematopathologists agreed to within one grade of one another in 69% of cases when scoring reticulin, compared to 53% of cases when assigning WHO diagnosis (p<0.0001). Across the three hematopathologists the frequency of patients with ‘true ET’ ranged from 10% to 48%, ‘prefibrotic myelofibrosis’ from 9% to 28%, and higher levels of fibrosis from 37% to 76%. Secondly, the strength of association for cellularity criteria was generally greater than that for megakaryocyte morphological criteria, with the exception of bare megakaryocyte nuclei. In particular, agreement on the frequency of dysplastic megakaryocytes was very poor, barely above chance, with agreement on staghorn and cloud-like megakaryocytes little better. Thirdly, agreement for both the number of megakaryocyte clusters and the size of clusters was greater than for whether clusters were tight or loose. Fourthly, inter-observer
agreement on the presence or absence of new bone formation was excellent. We considered the possibility that poor inter-observer agreement reflected one discrepant observer. However pairwise comparisons were similar for all pairs (data not shown) suggesting that differences in agreement shown in table 1 were not due to an ‘outlier’ observer.

In summary, agreement was better for measures of general morphological patterns such as cellularity, number of clusters, and reticulin grade, and weaker for measures of individual cellular features such as megakaryocyte morphology and whether clustering is tight or loose. In addition, the hematopathologists showed poor agreement in synthesising the various parameters when assigning cases to individual diagnostic categories using WHO criteria.

**Relative importance of different morphological criteria: significance of reticulin grade**

The WHO monograph lists several histological features that are said to be characteristic of ‘true ET’, such as the presence of staghorn megakaryocytes, normal overall cellularity and loose megakaryocyte clustering. In contrast, ‘prefibrotic myelofibrosis’ is said to be characterised by the presence of tight megakaryocyte clusters, cloud-like, dysplastic or pyknotic megakaryocytes and abnormal cellularity. The poor inter-observer agreement for the WHO classification that we have found could result from two potential causes, which are not mutually exclusive. The hematopathologists may not agree on the interpretation of the individual criteria themselves, as demonstrated in the previous section. In addition, they may put
differing emphasis or weight on the relative importance of the various morphological
criteria.

The WHO monograph provides minimal guidance as to the relative importance of the
various morphological features as individual contributions to reaching a diagnosis.
We found this to be problematic since we not infrequently found examples of bone
marrow histology with some of the morphological features said to reflect ‘true ET’
and coexistent with changes thought to imply ‘prefibrotic myelofibrosis’ or even overt
myelofibrosis. For example, there were sections with pyknotic megakaryocytes, a
feature of myelofibrosis, in a loose megakaryocyte cluster and a normocellular
background, both features said to be suggestive of ‘true ET’ (figure 2A). Similarly,
we found biopsies with large numbers of staghorn megakaryocytes (a feature of ‘true
ET’) together with hypercellularity (figure 2B), cloud-like megakaryocytes (figure
2C) and tight megakaryocyte clusters (figure 2D); all features more suggestive of
‘prefibrotic myelofibrosis’. Of course, these examples do not mean that the overall
patterns and associations of morphological features described in the WHO monograph
are invalid. In fact, many of the features did show significant correlations with one
another, as shown in table 2. However, while many of the correlations among
individual morphological features are statistically significant, they are far from fully
concordant, suggesting that bone marrow biopsy specimens with conflicting features,
such as those in figure 2, are reasonably common. This underscores the difficulty of
combining multiple morphological features into a single diagnosis without explicit
guidance as to which factors are most significant or characteristic.
We therefore explored whether the three hematopathologists employed different weighting schemes in using the 16 morphological criteria to arrive at a final diagnosis according to the WHO monograph. Each hematopathologist independently generated subjective assessments of which morphological features he or she found important in determining the overall WHO diagnosis by distributing an arbitrary 100 points among the 16 variables (figure 3A). From this subjective assessment, two important points emerge. Firstly, the three hematopathologists show different patterns of emphasis, with, for example, pathologists 1 and 3 putting more weight on cellularity criteria than pathologist 2, and pathologist 3 rating staghorn and dysplastic megakaryocytes as more important than do pathologists 1 and 2. Secondly, each of the three hematopathologists believed that reticulin grade was individually the most important criterion for determining the WHO classification score.

To provide a more objective assessment, a Bayesian proportional odds logistic regression was undertaken (figure 3B). The reason for applying this methodology was to identify, for each hematopathologist, which factors were independently predictive of his or her WHO diagnosis. This demonstrated that reticulin grade was the dominant independent predictor of WHO diagnosis for all 3 hematopathologists. Of the other criteria, there was little concordance as to which factors were independently informative. Many of the criteria that each pathologist identified as important for determining WHO diagnosis in the subjective weightings (figure 3A) were not independently associated, after controlling for the correlation of WHO diagnosis with reticulin grade.
These results demonstrate that reticulin grade was the major factor determining WHO classification assignment by all three hematopathologists. Since inter-observer agreement was quite high for reticulin grade, this suggests that the poor inter-observer agreement for WHO diagnosis was largely driven either by differences in the interpretation of the other morphological criteria, and/or by the relative importance applied to them.

**No difference in presenting blood counts or clinical outcome between ‘true ET’ and ‘prefibrotic myelofibrosis’**

One of the central claims of the WHO classification is that ‘prefibrotic myelofibrosis’ and ‘true ET’ are biologically distinct disorders with different prognoses. Due to the poor inter-observer reliability with which these putative entities could be identified, the only means to assess the reproducibility of these claims in our cohort was to compare presenting blood counts and clinical outcomes between ‘prefibrotic myelofibrosis’ and ‘true ET’ for each hematopathologist separately.

There were no differences in hemoglobin, platelet count or white cell count at diagnosis between cases labelled as ‘true ET’ and ‘prefibrotic myelofibrosis’ for any of the hematopathologists (p>0.1 for all 3 hematopathologists and each blood count variable). Similarly, there were no differences between ‘true ET’ and ‘prefibrotic myelofibrosis’ in age, sex or rates of splenomegaly, leukoerythroblastic blood film and cytogenetic abnormalities for any of the hematopathologists (p>0.1 all variables).

There was a weak association between JAK2 positivity and ‘prefibrotic myelofibrosis’ for one of the pathologists (‘true ET’ 38% V617F-positive; ‘prefibrotic myelofibrosis’ 61% V617F-positive; p=0.04), but this was not found for the other two (p=0.9 and
p=0.1). Given the large number of hypothesis tests performed in this section, this single significant test is likely to be due to the play of chance. Finally, we were unable to identify any distinguishing diagnostic clinical or laboratory features even when subsets of patients identified as ‘prefibrotic myelofibrosis’ by any two or all three hematopathologists were considered (data not shown).

In total, 143 of the 370 patients were identified by at least one of the hematopathologists as having ‘prefibrotic myelofibrosis’ (32 hematopathologist 1, 101 hematopathologist 2, 39 hematopathologist 3). Of these, with a median follow-up of 68 months from trial entry, not a single patient underwent myelofibrotic transformation. Moreover, only one of the 194 patients labelled as ‘true ET’ by any of the hematopathologists (173 hematopathologist 1, 36 hematopathologist 2, 40 hematopathologist 3) transformed to myelofibrosis.

On univariate analysis, there was no difference for any of the 3 hematopathologists between ‘prefibrotic myelofibrosis’ and ‘true ET’ in the rate of the composite end-point of time to first arterial or venous thrombosis; major hemorrhage; disease transformation; or death (Hazard ratio (HR) for hematopathologist 1, 1.16; 95% CI, 0.4-3.2; p=0.7. HR for hematopathologist 2, 0.61; 95% CI, 0.2-1.8; p=0.4. HR for hematopathologist 3, 1.24; 95% CI, 0.39-3.9; p=0.7). After controlling for age, sex, treatment allocation, prior cytoreductive therapy and a history of previous end-point events, multivariate survival analysis similarly revealed no differences in this composite end-point between ‘true ET’ and ‘prefibrotic myelofibrosis’ (HR for hematopathologist 1, 1.03; 95% CI, 0.4-2.8; p=0.9. HR for hematopathologist 2, 0.70; 95% CI, 0.2-2.0; p=0.5. HR for hematopathologist 3, 0.48; 95% CI, 0.1-1.8; p=0.3).
We next compared ‘prefibrotic myelofibrosis’ and ‘true ET’ for each of the individual end-point categories that comprised the composite end-point, noting that numbers of events were generally low in these individual categories. There were no differences in overall survival between patients labelled as ‘true ET’ and ‘prefibrotic myelofibrosis’ for any of the three hematopathologists and furthermore, there were no differences in rates of arterial thrombosis, rates of venous thrombosis or rates of major hemorrhage between patients labelled as ‘true ET’ and those labelled as ‘prefibrotic myelofibrosis’ (p>0.1 for all three hematopathologists on univariate and multivariate analysis on each individual end-point).

**Exploratory factor analysis identifies megakaryocyte clustering, degree of fibrosis and cellularity as independent underlying processes in ET**

The practice of histopathology is, to a great extent, concerned with the recognition and description of patterns of morphological features. Individual morphological features of the marrow are of little intrinsic value in isolation, but represent manifestations of underlying pathophysiological factors. In our analysis, we found many significant correlations among the individual morphological features (table 2), suggesting that many features coexist and there may be coherent patterns of histological abnormalities. Factor analysis is a multivariate statistical method that seeks to identify these unobserved pathophysiological processes by picking out patterns of correlations among the morphological features that suggest common underlying, independent processes (or factors).
Initial screening analysis suggested that a model with 3 independent factors was the most appropriate (by Kaiser’s criterion), and this was therefore fitted. The five most important morphological features contributing to each factor are presented in figure 4, and suggest that the factors have relatively straightforward biological interpretations. We start with the third factor, since it has the most straightforward interpretation. It is particularly weighted towards cellularity criteria, with emphasis on overall cellularity as well as the degree of hypercellularity for each of the three lineages separately. An association between \textit{JAK2} status and marrow cellularity has been demonstrated previously in patients with ET, with V617F-positive ET showing greater overall, erythroid and granulocytic cellularity than V617F-negative ET\textsuperscript{10}. Consistent with this observation, we found a significant association between the \textit{JAK2 V617F} mutation and scores for this cellularity factor (p=0.0002). In contrast, there were no significant associations between \textit{JAK2} status and the other 2 factors, discussed below (p=0.1 for the clustering factor and p=0.4 for the fibrosis factor).

Considering now the other two factors, the first is particularly weighted towards number, size and type of megakaryocyte clusters together with megakaryocyte size and cellularity, and captures an underlying process related to megakaryocyte clustering. The second independent factor appears to relate to the extent of fibrosis. Trephine specimens with extensive reticulin fibrosis, new bone formation, frequent pyknotic or dysplastic megakaryocytes, and bare megakaryocyte nuclei would score particularly highly on this factor. These are all features the WHO monograph identifies as suggestive of a diagnosis of overt myelofibrosis.
This factor analysis suggests that three underlying processes describe many of the morphological patterns evident in the bone marrow trephine histology of patients diagnosed with ET (by PVSG criteria), namely cellularity, megakaryocyte clustering and extent of fibrosis. The cellularity factor shows significant correlation with JAK2 status, but the other two reflect unknown biological mechanisms.

**DISCUSSION**

ET has long been thought to represent a heterogeneous disorder, likely to contain pathogenetically distinct subgroups united by the lack of positive diagnostic markers. In keeping with this concept, histological features of bone marrow in ET, as defined by PVSG criteria, show substantial variability. This has led to attempts to subclassify the disorder on the basis of trephine morphology, most recently in the WHO classification. Many claims have been made about the clinical utility of such classifications, but generally there has been little detailed assessment of inter-observer reliability or inter-correlations among variables, and the clinical validation has tended to be retrospective and uncorrected for other risk factors. In this study, we have evaluated the reproducibility with which individual histological criteria can be assessed and contribute to the definition of subtypes of ET in a large, prospective, multi-center cohort of patients.

Our results demonstrate that several histological features of the marrow can be ascertained with reasonable reproducibility, specifically those associated with marrow topography, cellularity and degree of fibrosis (eg reticulin grade, megakaryocyte clustering, new bone formation). However, the assessment of other cytological features was much less reliable (particularly megakaryocyte morphology), as was
classification according to WHO disease category including the distinction between ‘prefibrotic myelofibrosis’ and ‘true ET’. These data are consistent with at least two interpretations. It is possible that even experienced hematopathologists need special training to distinguish subtypes of ET. Our study was designed to assess the utility of the WHO criteria in a “real world” pathology setting. Biopsies were assessed by experienced hematopathologists (but not directly involved in the development of the WHO criteria) working without a training set of slides, evaluating trephines from all patients in all ET risk categories. It remains possible that the pathologists involved in the WHO classification may have better reproducibility or that a training set of slides may have enhanced inter-observer agreement for our hematopathologists. However, neither is available in routine diagnostic practice and it is hard to see the general utility of criteria the application of which is so difficult even for experienced hematopathologists. Moreover our results were obtained as part of a focused assessment in a finite period during which a large number of MPD trephine biopsy specimens were reviewed, a situation which is likely to enhance intra-observer reproducibility. Application of the criteria is likely to be significantly more difficult for most histopathologists who see such specimens relatively infrequently.

An alternative explanation for our results, and the one we favor, is that the current WHO histological criteria are not sufficiently robust to define subtypes of ET. There was poor inter-observer agreement on what is represented by the terms ‘prefibrotic myelofibrosis’ and ‘true ET’, and there were striking differences in the emphasis each of the hematopathologists placed on different morphological criteria when arriving at a diagnosis (figure 3). These results demonstrate that the published histological criteria for these proposed entities are difficult to apply in a reproducible manner. It
has been suggested that patients labelled as having ‘prefibrotic myelofibrosis’ have a worse outcome compared to those said to have ‘true ET’\textsuperscript{21-26}, a key argument supporting the existence of these putative entities. However, we have been unable to reproduce these findings. There were no differences in the rates of thrombosis, major hemorrhage, myelofibrotic transformation or survival between ‘prefibrotic myelofibrosis’ and ‘true ET’ as labelled by any of the hematopathologists. We cannot exclude the possibility that prolonged follow up or greater numbers of patients might reveal differences in outcome. However, if such large sample sizes are required to show statistical significance, such differences are unlikely to be clinically relevant.

Our favored interpretation is also consistent with recent molecular genetic insights. The subgroup of ET patients who carry the \textit{JAK2} V617F mutation are biologically distinct from those lacking the mutation, both in presenting features and in clinical outcome\textsuperscript{10,33-35}. The \textit{JAK2} V617F negative subgroup is also heterogeneous. An activating mutation in \textit{MPL} occurs in approximately 10\% of this subgroup\textsuperscript{36} but the molecular mechanisms responsible for the rest remain unclear. However, there is no evidence for any correlation between the molecular subtypes of ET and the proposed histological subtypes, ‘true ET’ and ‘prefibrotic myelofibrosis’. In patients with ET, the \textit{JAK2} V617F mutation was associated with increased overall cellularity, increased erythropoiesis and increased granulopoiesis but there was no association between \textit{JAK2} status and reticulin grade, megakaryocyte clustering or the presence of staghorn, cloud-like, dysplastic or pyknotic megakaryocytes\textsuperscript{10}.

The results presented here suggest that current histological criteria are not sufficient to permit routine separation of ET into biologically distinct subsets. However,
exploratory multivariate analysis did identify at least three independent processes underpinning the extensive variability of bone marrow histology in patients with thrombocythemia (figure 4). One of these processes, the cellularity factor, correlates with whether or not the patient has the JAK2 V617F mutation, but the pathophysiology underlying the other two processes is less clear. The second factor, scoring highly on reticulin, new bone formation, and pyknotic, dysplastic or bare megakaryocyte nuclei, is similar to the descriptions of overt myelofibrosis in the WHO monograph. The molecular mechanisms underlying the development of fibrosis and the other process identified by our factor analysis, megakaryocyte clustering, are unclear. TGF-beta, NF-kB, low levels of GATA-1 and excessive MPL signalling have all been implicated in the development of fibrosis and genetic or environmental factors influencing these pathways may influence the degree of fibrosis and associated morphological features. Little is known about the molecular regulation of megakaryocyte location and clustering, but it may be relevant that megakaryocyte clusters are observed in mice treated with SDF-1, the ligand for the CXCR4 receptor.

It is generally accepted that there is histological heterogeneity within the group of patients labelled as ET using PVSG criteria. The molecular basis for this heterogeneity is unclear and our data cast doubt on the concept of using current histological criteria to divide ET into ‘true ET’ and ‘prefibrotic myelofibrosis’. However, our results are consistent with a recent molecular classification of the MPDs in which it is suggested that reticulin accumulates to a variable extent in patients with ET (figure 5). In this model, patients with both ET and PV gradually accumulate reticulin fibrosis as an inherent part of their disease, with the degree of
fibrosis reflecting an interplay between the duration of the disease and physiological or genetic modifiers. This concept is supported by histological studies in patients with these disorders, and the frequent development of post-polycythemic myelofibrosis in mouse models. The development of fibrosis is likely to be influenced by inherited genetic modifiers (as evidenced by differences in the rates of myelofibrosis among different strains of mice expressing JAK2 V617F), environmental factors and acquired genetic or epigenetic changes. In a proportion of patients the accumulation of these genetic/epigenetic results in an acceleration of their disease which may present as myelofibrotic transformation. Several observations are consistent with the model shown in figure 5. V617F-positive patients with PV and ET share several features, and represent a phenotypic continuum, with homozygosity for the V617F mutation strongly favoring a polycythemic phenotype. Patients labelled as primary myelofibrosis are clinically indistinguishable from those with myelofibrotic transformation of a preceding MPD. The model also predicts that the genetic lesions responsible for V617F-negative primary myelofibrosis will be found in V617F-negative ET, as has been found for MPL W515 mutations.
Acknowledgements

We would like to thank Jon van der Walt for his assistance with curation of the biopsy samples and organising the cutting and staining of specimens. We would also like to acknowledge Pat Collins for her assistance with managing the trephine slides. The research has been supported by grants from the UK Medical Research Council, the Leukaemia Research Fund and the Kay Kendall Leukaemia Fund.

Author contributions:
BSW, WNE and DB contributed equally to the design of the study, the interpretation and scoring of the bone marrow trephine biopsies and the subsequent analysis. GB and CLE collected clinical outcome data and samples from patients in the three trials, under the oversight of KW, CNH and ARG. BP contributed to collection of patient samples. Statistical analyses were performed by PJC. ARG and PJC equally co-ordinated and directed the research. All authors have had the opportunity to contribute to the drafting of the paper.

The authors declare no competing financial interests.
Table 1. Assessment of inter-observer reliability for bone marrow histology criteria.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Strength of association</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Myelofibrosis criteria (5 point scale)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WHO diagnosis</td>
<td>2.1</td>
<td>1.8 – 2.4</td>
</tr>
<tr>
<td>Reticulin grade</td>
<td>5.1</td>
<td>4.0 – 6.4</td>
</tr>
<tr>
<td><strong>Cellularity criteria (3 point scale)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>8.3</td>
<td>5.5 – 12.6</td>
</tr>
<tr>
<td>Erythroid</td>
<td>3.9</td>
<td>2.8 – 5.3</td>
</tr>
<tr>
<td>Granulocytic</td>
<td>5.0</td>
<td>3.4 – 7.2</td>
</tr>
<tr>
<td>Megakaryocytic</td>
<td>5.1</td>
<td>4.0 – 6.4</td>
</tr>
<tr>
<td><strong>Megakaryocyte morphology (3 point scale)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staghorn megakaryocytes</td>
<td>2.5</td>
<td>1.9 – 3.3</td>
</tr>
<tr>
<td>Cloud-like megakaryocytes</td>
<td>2.2</td>
<td>1.7 – 2.8</td>
</tr>
<tr>
<td>Dysplastic megakaryocytes</td>
<td>1.4</td>
<td>1.1 – 1.7</td>
</tr>
<tr>
<td>Pyknotic megakaryocytes</td>
<td>3.2</td>
<td>2.4 – 4.2</td>
</tr>
<tr>
<td>Bare megakaryocyte nuclei</td>
<td>7.7</td>
<td>4.3 – 13.6</td>
</tr>
<tr>
<td>Megakaryocyte size</td>
<td>3.2</td>
<td>2.3 – 4.5</td>
</tr>
<tr>
<td><strong>Megakaryocyte clustering (3 point scale)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of clusters</td>
<td>9.1</td>
<td>6.0 – 13.9</td>
</tr>
<tr>
<td>Type of clusters (none, loose, tight)</td>
<td>2.7</td>
<td>2.2 – 3.3</td>
</tr>
<tr>
<td>Size of clusters (none, small, large)</td>
<td>6.8</td>
<td>4.7 – 9.7</td>
</tr>
<tr>
<td><strong>Miscellaneous (2 point scale)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paratrabeucular megakaryocytes</td>
<td>4.4</td>
<td>3.1 – 6.1</td>
</tr>
<tr>
<td>New bone formation</td>
<td>10.1</td>
<td>4.8 – 21.8</td>
</tr>
</tbody>
</table>

1 Higher scores for strength of association represent stronger inter-observer reliability, with a score of 1 indicating no agreement beyond chance.

2 The categories of normal (score=0) and mildly increased (score=1) were combined because of very small numbers in the normal category and to allow comparison with the other criteria for cellularity (all on a 3 point scale).
<table>
<thead>
<tr>
<th>WHO diagnosis</th>
<th>Reticulin grade</th>
<th>Overall cellularity</th>
<th>Erythroid cellularity</th>
<th>Granulocyte cellularity</th>
<th>Mega cellularity</th>
<th>Staghorn megas</th>
<th>Cloud-like megas</th>
<th>Dysplastic megas</th>
<th>Pyknotic megas</th>
<th>Bare mega nuclei</th>
<th>Mega size</th>
<th>Cluster numbers</th>
<th>Cluster type</th>
<th>Cluster size</th>
<th>Paratrabec megas</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHO diagnosis</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reticulin grade</td>
<td>0.7 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall cellularity</td>
<td>0.3 0.3 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erythroid cellularity</td>
<td>NS NS 0.3 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Granulocyte cellularity</td>
<td>0.2 0.3 0.6 0.2 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mega cellularity</td>
<td>0.3 0.3 0.5 0.2 0.4 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staghorn megas</td>
<td>-0.3 NS NS 0.1 NS 0.1 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloud-like megas</td>
<td>NS 0.1 0.2 0.2 0.2 0.3 NS 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dysplastic megas</td>
<td>0.4 0.3 0.2 0.1 0.3 0.3 -0.1 0.1 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pyknotic megas</td>
<td>0.2 0.3 0.3 0.1 0.2 0.3 NS 0.2 0.3 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bare mega nuclei</td>
<td>0.1 0.2 0.3 NS 0.1 0.3 0.1 0.2 0.3 0.6 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mega size</td>
<td>NS 0.3 0.3 NS 0.2 0.4 0.3 0.2 NS 0.2 0.2 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster numbers</td>
<td>NS 0.3 0.2 NS 0.2 0.4 0.2 0.2 0.1 0.3 0.2 0.4 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster type</td>
<td>0.2 0.4 0.1 NS NS 0.3 NS 0.1 0.2 0.2 0.2 0.4 0.7 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster size</td>
<td>0.1 0.3 0.3 NS 0.2 0.6 0.2 0.2 0.2 0.3 0.2 0.5 0.7 0.6 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paratrabec megas</td>
<td>0.2 0.4 0.2 NS 0.1 0.4 NS 0.1 0.2 0.2 0.1 0.3 0.3 0.4 0.4 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New bone formation</td>
<td>0.2 0.3 NS -0.1 NS 0.1 NS NS NS 0.2 NS 0.2 0.2 0.2 0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

Figure 1. Examples of morphological features scored on 370 diagnostic bone marrow trephine specimens from patients enrolled in prospective clinical trials of ET. (A) H&E stained section (x400 magnification) showing a large, staghorn megakaryocyte in the center of the field. (B) H&E stained section (x1000) showing a cloud-like megakaryocyte in a background of with increased cellularity. (C) A loose cluster of megakaryocytes showing other marrow cells between individual megakaryocytes (H&E stain x400). (D) A tight cluster of megakaryocytes (H&E stain x1000), showing moulding of the juxtaposed cell surfaces between adjacent megakaryocytes, and no intervening marrow cells. (E) – (H) Reticulin stains (all x400) showing increasing reticulin, graded 1-4 respectively.

Figure 2. Trephine histology showing morphological features suggesting different WHO diagnoses in the same medium- (x400, A & B) or high-power (x1000, C & D) field. (A) Pyknotic megakaryocytes (indicative of myelofibrosis) in a field otherwise suggestive of ‘true ET’ (normal cellularity, loose megakaryocyte cluster). (B) Staghorn megakaryocytes in a loose cluster (characteristic of ‘true ET’) in a hypercellular marrow (suggesting ‘prefibrotic myelofibrosis’). (C) Staghorn megakaryocyte (‘true ET’) immediately adjacent to a cloud-like megakaryocyte (‘prefibrotic myelofibrosis’). (D) Staghorn megakaryocyte (‘true ET’) in a tight cluster (‘prefibrotic myelofibrosis’).

Figure 3. (A) Subjective estimates of the relative importance of 16 morphological features in determining WHO diagnosis for each of the three hematopathologists. (B) Factors independently associated with WHO diagnosis for the three hematopathologists. The relative proportion of the pie-chart for each variable was calculated as the ratio of its regression coefficient to the sum of all regression coefficients in the model, after all variables were scaled to the same 0-1 range.

Figure 4. Exploratory factor analysis identified 3 underlying processes contributing to the morphological patterns of ET. Shown for each factor are the 5 variables that contribute most heavily to each factor, with the y axis representing their relative weight (contribution to the factor score). Factor 1 emphasises morphological features describing megakaryocyte clusters. Factor 2 captures the process of marrow fibrosis, weighting reticulin, new bone formation and pyknotic megakaryocytes. Factor 3, in picking out the cellularity criteria, correlates significantly with \( JAK2 \) V617F status (p=0.0002).

Figure 5. A model for the myeloproliferative diseases. \( JAK2 \)-positive polycythemia and thrombocythemia overlap, and can progress to an accelerated phase. This can be clinically variable, with myelofibrotic transformation, cytopenias, increased blasts and increased white cells potentially present. \( JAK2 \)-negative disease follows similar patterns of progression, and is biologically distinct from V617F-positive disease. Under this model, patients currently labelled as having primary myelofibrosis may in fact represent individuals presenting in accelerated phase of a pre-existing MPD.
REFERENCES


Figure 3
Factor 1: Clustering

Relative weight

Cluster size  Cluster numbers  Cluster type  Mega size  Mega cellularity

Factor 2: Degree of myelofibrosis

Relative weight

Pyknotic megas  Reticulin  Bare nuclei  Dysplastic megas  New bone

Factor 3: Cellularity or JAK2 status

Relative weight

Overall  Gran  Mega  Erythroid  Cloud-like megas

Figure 4
Bone marrow pathology in essential thrombocythemia: inter-observer reliability and utility for identifying disease subtypes

Bridget S Wilkins, Wendy N Erber, David Bareford, Georgina Buck, Keith Wheatley, Clare L East, Beverley Paul, Claire N Harrison, Anthony R Green and Peter J Campbell