BONE MARROW FIBROSIS UNDER THE LENS OF AI

Timothy Ebsworth[1], Sharon Ruane[2], Ros Cooper[1], Korsuk Sirinukunwattana[2], Alan Aberdeen[2], Jens Rittscher[3], Daniel Royston[1]

1) Nuffield Division of Clinical Laboratory Sciences, John Radcliffe Hospital, Oxford, United Kingdom, 2) Ground Truth Labs, Oxford, United Kingdom, 3) Institute of Biomedical Engineering and the Nuffield Department of Medicine, University of Oxford, Department of Engineering Science, Oxford, United Kingdom
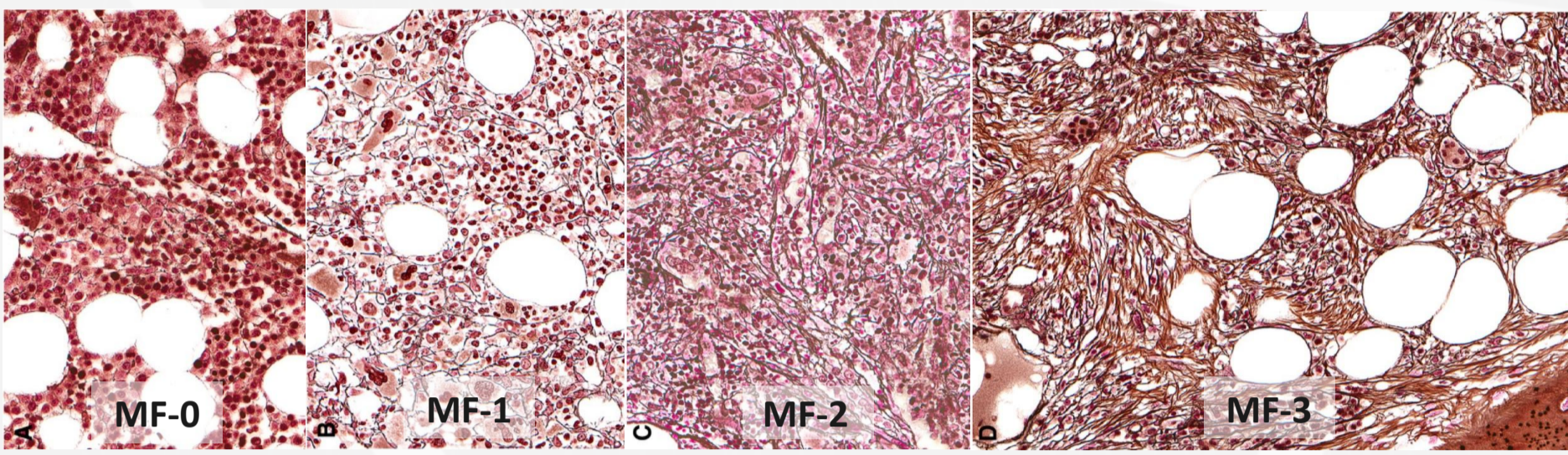
1. CONTEXT

Myeloproliferative neoplasms (MPNs) are blood cancers that cause overproduction of blood cells from bone marrow.

Bone marrow fibrosis (scarring) is a key feature for diagnosis, prognosis and guiding treatment.

Doctors examine bone marrow (BM) under a microscope and grade the severity of fibrosis on a scale of 0 to 3.

2. PROBLEM
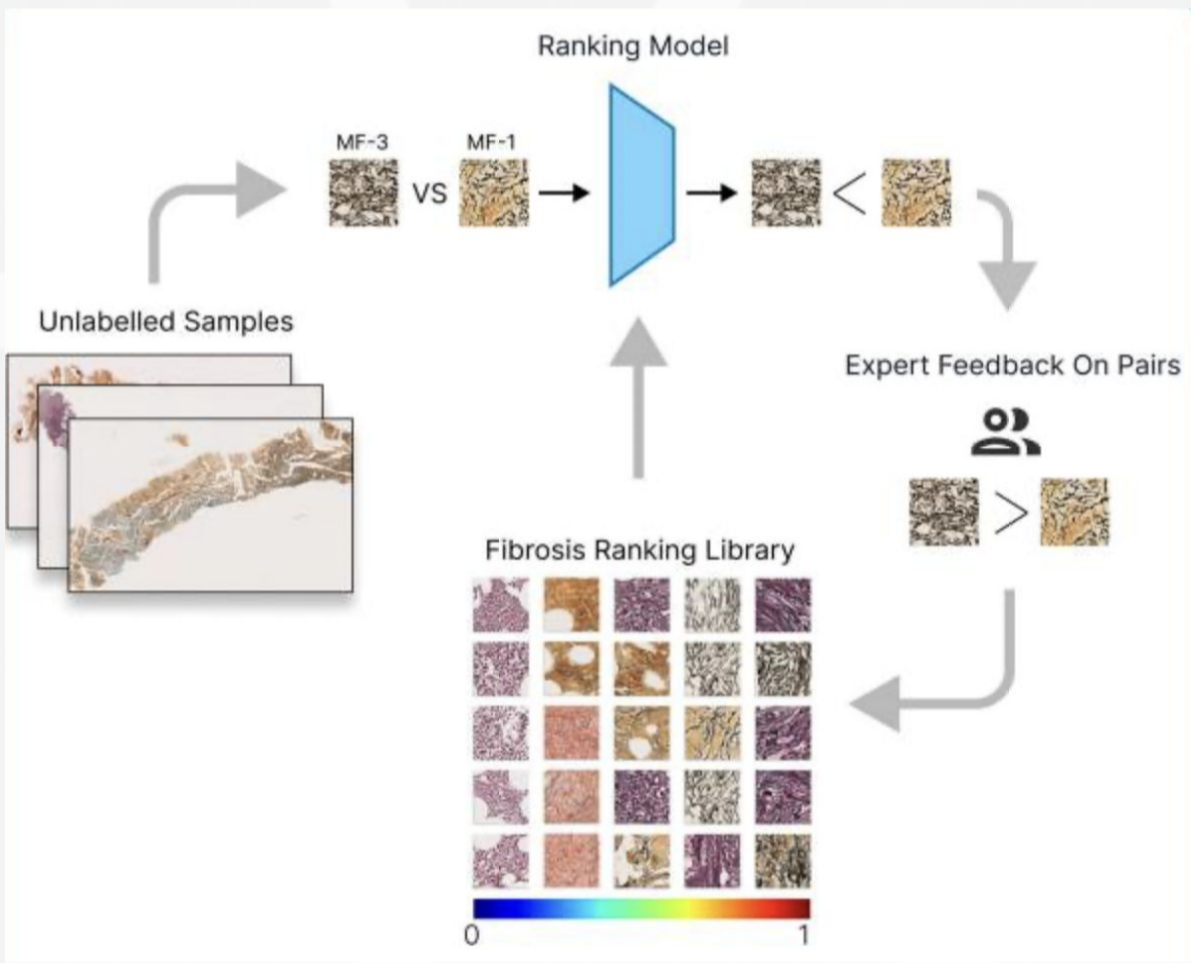

MF-0 · MF-1 · MF-2 · MF-3
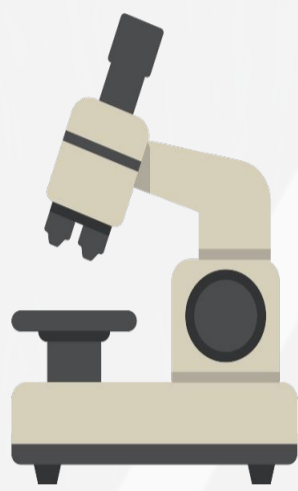
Severity of Fibrosis

Subjective → poor intra-/interobserver variability when grading

3. MODEL DEVELOPMENT

CIF is a ranking convolutional neural network, trained using a pairwise ranking strategy.

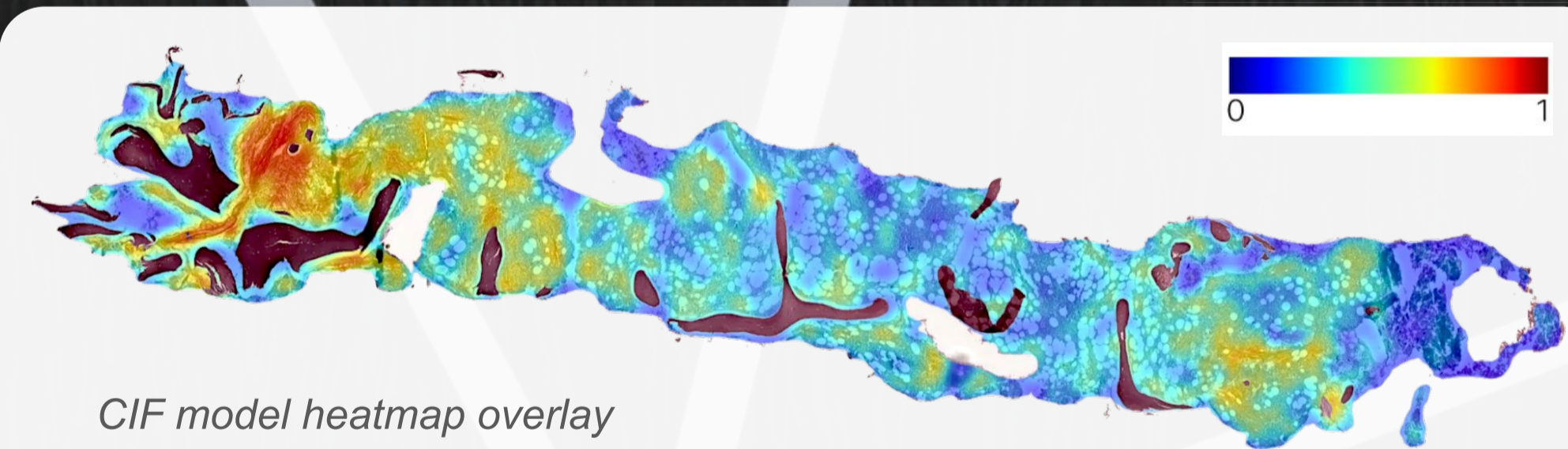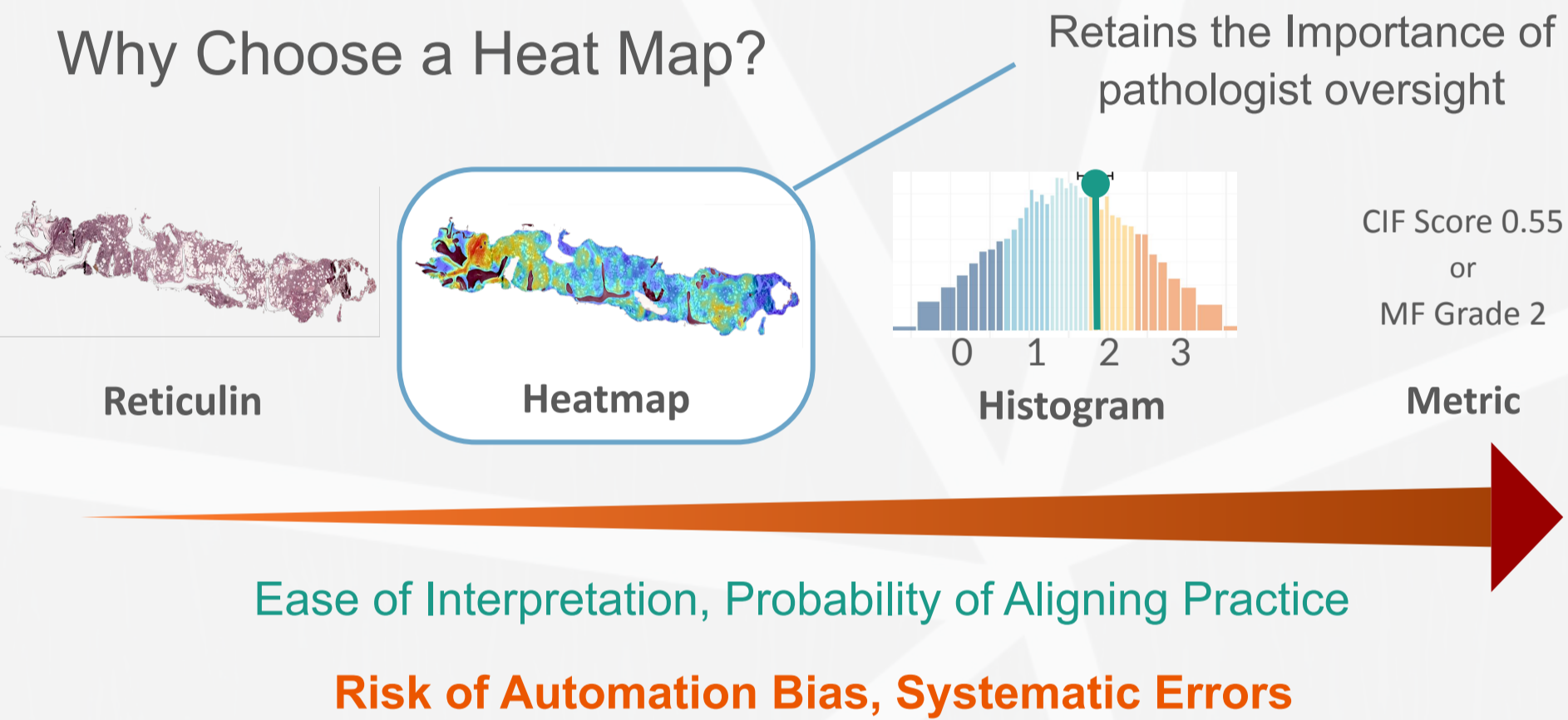**Training:** 476 BM
Testing: 105 BM

Accuracy: 0.891
Precision: 0.904
Recall: 0.885
F1: 0.895

Overview of the computational steps for detection and quantification of reticulin fibrosis from whole slide images of BM. From Ryou, H et al Leukaemia (2023.)

CIF model heatmap overlay

4. CLINICAL TRANSLATION

Why Choose a Heat Map?

Retains the Importance of pathologist oversight

Reticulin · Heatmap · Histogram · Metric

CIF Score 0.55 or MF Grade 2

Ease of Interpretation, Probability of Aligning Practice

Risk of Automation Bias, Systematic Errors

5. METHOD

International Panel of 15 Expert Haematopathologists

1000x BMT WSI — Oxford University Hospitals' NHSFT Digital Pathology Archive

Quality Control · n = 836 · Manual MF Grade · AI-Augmented MF Grade

Artefact Labelling

- MF-Grade Concordance
- MF-Grade Accuracy
- Review Time
- User Experience

Visual abstract from Ebsworth et al. (June 2025). International Multicentre Evaluation of AI Augmented Fibrosis Grading in a Real World Clinical Cohort of 1000 Patients. EHA Congress, Milan

6. RESULTS

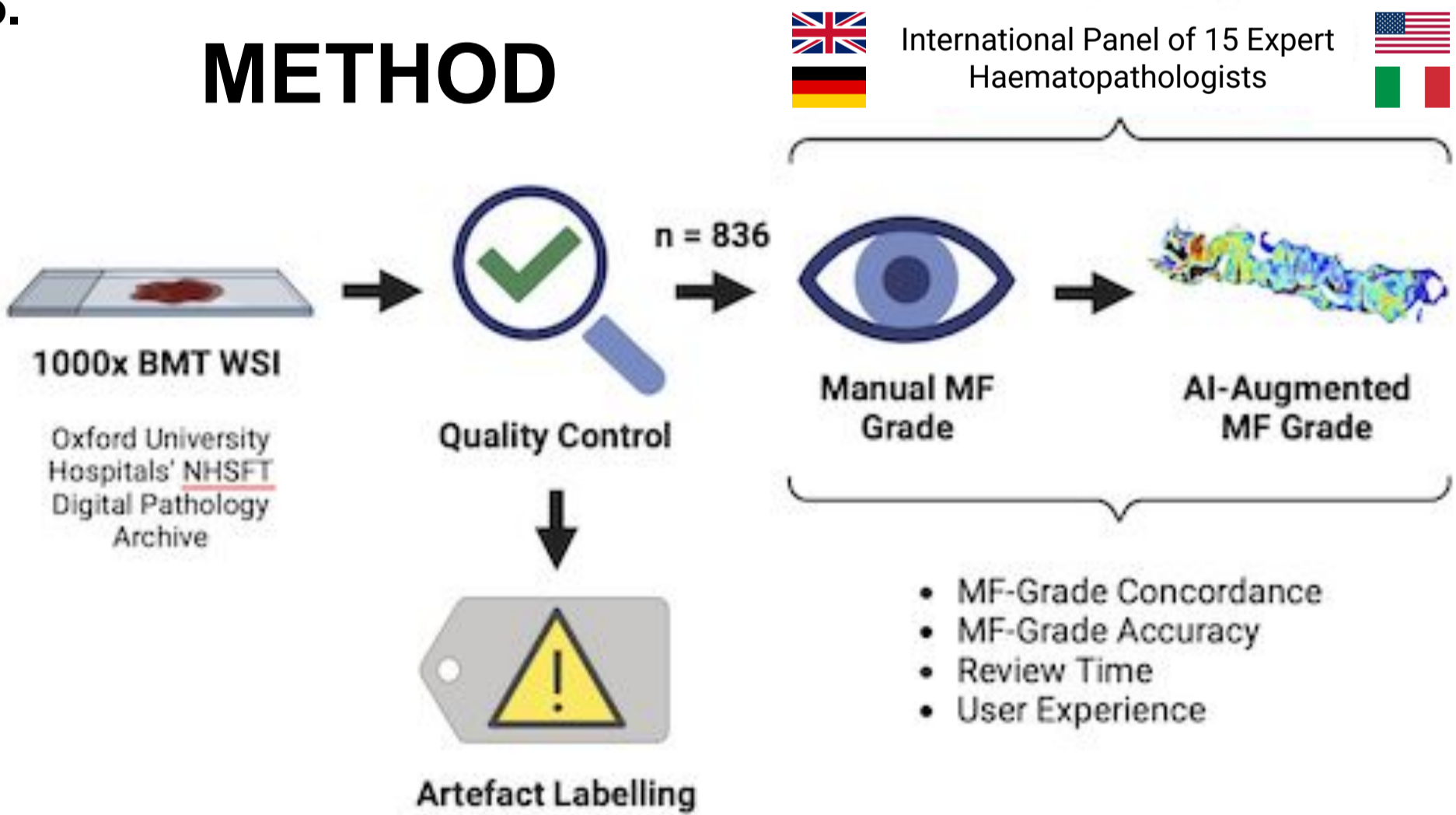Access to heatmaps significantly reduced interobserver variability (weighted kappa ranges):

Manual: 0.51-0.84
CIF: 0.60-0.84 → p=0.0017*

Results remained in agreement with ground truth (manual consensus)

Round 1 manual grade	MF0	MF1	MF2	MF3
MF0	356 (82.2 %)	73 (16.9 %)	4 (0.9 %)	0 (0.0 %)
MF1	184 (32.5 %)	344 (60.7 %)	38 (6.7 %)	1 (0.2 %)
MF2	5 (2.1 %)	94 (40.0 %)	120 (51.1 %)	16 (6.8 %)
MF3	0 (0.0 %)	2 (3.0 %)	22 (33.3 %)	42 (63.6 %)

Round 2 manual grade

Confusion matrix: manual intraobserver variation

7. NEXT STEPS

In Progress
- MPN Enriched Cohort
- Effect of a Summary Metric
- Impact on Pathology Trainees
- Patient and Public Involvement

TBC
- Prospective Validation
- Investigate Risk of Ethnic Bias