EHA2025 Congress
June 12-15 | Milan, Italy

International, Multicentre Evaluation of AI Augmented Fibrosis Grading in a Real-World, Clinical Cohort of 1000 Patients

TIMOTHY J EBSWORTH[1], SHARON RUANE[2], NEHA BHARDWAJ[3], ROSALIN COOPER[3,4], ANNA GREEN[5], DEBORAH HAY[3], JANA IHLOW[6], FATIMA JELLOUL[11], RASHMI KANAGAL[11], PHILIP MACKLIN[3], VIDHYA MANOHAR[3], EDOARDO OLMEDA[8,9], MARK ONG[5], ANDRES QUESADA[11], CARLO PESCIA[4], SAAD BASHIR[2], SAM MAXWELL[2], KORSUK SIRNUKUNWATTANA[2], KA HO TAM[2], ALAN ABERDEEN[2], JENS RITTSCHER[10], DANIEL ROYSTON[3,4]

1Nuffield Division of Clinical Laboratory Sciences, John Radcliffe Hospital, Oxford, United Kingdom, 2 Ground Truth Labs, Oxford, United Kingdom, 3 Department of Cellular Pathology, John Radcliffe Hospital, Oxford, United Kingdom, 4 Nuffield Division of Clinical Laboratory Science, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom, 5 Guy's and St Thomas' NHS Foundation Trust, Department of Cellular Pathology, London, UNited Kingdom, 6 Berlin Institute of Health at Charite-Universitatsmedizin Berline, BIH Biomedical Innovation Academy, BIH Charite Clinician SCientist Program, Berline, Germany, 7 Institute of Pathology, Charite-Universitatsmedizin Berlin, Corporate Member of Freie Universitat Berlin, HUmboldt-Universitat zu Berlin, and Berline Institute of Health, NBerlin, Germany, 8 European Institute of Oncology, IRCCS, Division of Pathology, Milan, Italy, 9 University of Milan, Milan, Italy, 10 Institute of Biomedical Engineering and the Nuffield Department of Medicine, University of Oxford, Department of Engineering Science, Oxford, United Kingdom, 11 M.D. Anderson Cancer Center, Houston, Texas, United States of America

Radcliffe Department of Medicine
Medical Sciences Division
Ground Truth Labs
University of Oxford

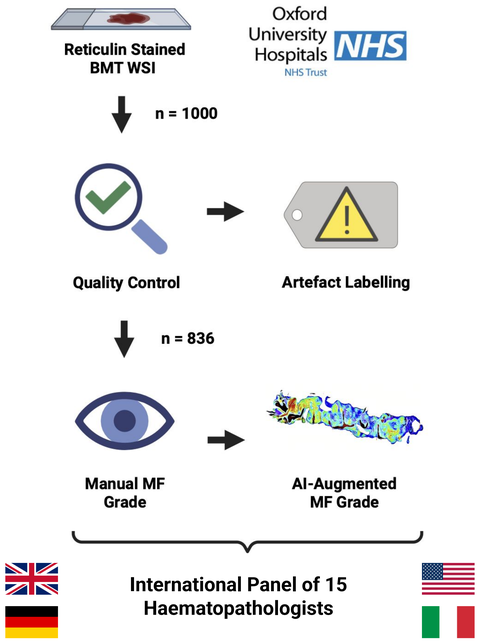
INTRODUCTION

Fibrosis grading is critical to the classification, prognostication and monitoring of myeloproliferative neoplasms (MPN). Manual estimation of fibrosis is subjective and only semi-quantitative. AI-augmented evaluation of fibrosis using *continuous indexing of fibrosis* (CIF) improves accuracy of fibrosis grading in clinical trials.[1] Here we evaluate its performance to support deployment in clinical practice.

AIMS

- Determine the quality of digital whole slide images (WSI) of reticulin-stained bone marrow trephines (BMT) from a large, tertiary referral laboratory.
- Measure the concordance of reticulin quality control (QC) assessments between expert Haematopathologists.
- Evaluate the performance of an AI fibrosis algorithm (CIF v1.5.3) on a large cohort of routine clinical BMT samples.

METHOD



MODEL TRAINING

CIF v1.5.3 is a ranking convolutional neural network, trained using a pairwise ranking strategy (*RankNet*), incorporating human-in-the-loop manual image ranking.

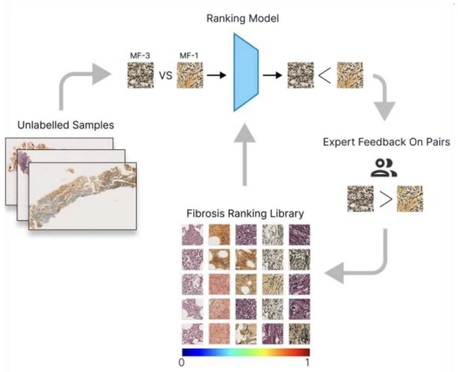
WSIs are divided into uniform tiles then segmented to exclude non-haematopoeitic areas. The model output is the CIF score for each tile: 0 (no fibrosis) to 1 (maximal fibrosis).

Training: 476 BMTs (23,504 tiles)

Independent Testing: 105 BMTs (6,090 tiles)

Accuracy 0.891 | Precision 0.904 | Recall 0.885 | F1 0.895

Performance based on pairwise ranking agreement vs. human annotations.



Overview of the computational steps for detection and quantification of reticulin fibrosis from WSIs of BMTs. From Ryou, H et al Leukaemia (2023.)

VALIDATION DATASET

Digital WSIs of 1,000 sequential BMTs were identified from Oxford University Hospitals pathology archive (April 2023-July 2024.)

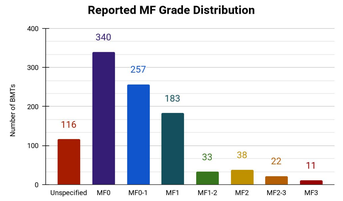
Inclusion: H&E + reticulin WSIs available

Exclusion: patient registered for NHS data opt out

Age (mean): 60.6 years

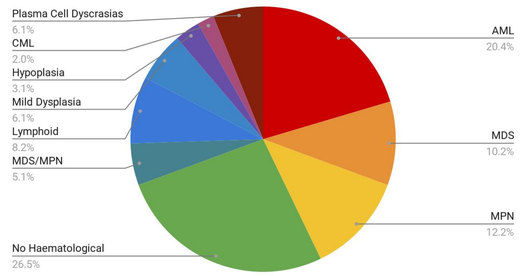
Age (range): 1-95 years

Sex: ♂ 56%, ♀ 43%



Histogram of reported MF-grades for sample BMTs

Ethnicity: White 50%, Not Stated 41%, Asian 6%, Black 1%, Other 1%



Pie chart of histological diagnoses for sample BMTs as reported by Oxford University Hospitals NHSFT Pathology Department

PRELIMINARY ANALYSIS I

On viewing the same sample on separate occasions, pathologists assigned a different MF-grade in 1/3 cases (intraobserver agreement 66.26%)

Access to heatmaps significantly reduced interobserver variability:

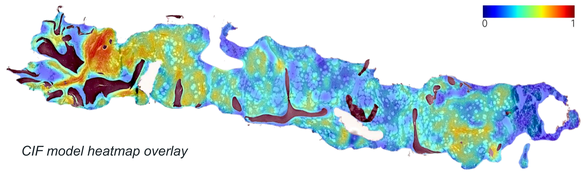
Manual Read: 0.51-0.84 (weighted kappa range)
CIF Assisted Read 0.60-0.84
} p=0.0017*

Access to heatmaps significantly improved MF-grade consensus:

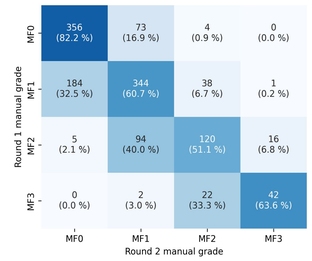
CIF Sequential Read OR 1.43 (1.19-1.71), p=0.0001*

CIF Concurrent Read OR 2.41 (1.93-3.02), p<0.0001*

(Odds ratio of consensus agreement relative to manual read, based on generalised linear modelling)



CIF model heatmap overlay



LEFT: confusion matrix for overall intraobserver variation: manual MF-grade round 1 vs. manual MF-grade round 2 (identical slide sets.)

Weighted kappa = 0.746 (95% CI: 0.720-0.772)

PRELIMINARY ANALYSIS II

Interpretation of CIF heatmaps was in agreement with ground truth*:

Round	Agree (n)	Disagree (n)	Agreement with ground truth*
Manual Read	771	384	66.8%
CIF Assisted Sequential Read	798	384	69.1%
CIF Assisted Concurrent Read	719	342	67.8%

*ground truth was the manual consensus MF-grade due to conventional acceptability.

CONCLUSIONS

Access to a CIF generated heatmap significantly improved interobserver agreement and MF-grading consensus, whilst remaining in agreement with our ground truth (manual consensus).

The variability we have described in human assigned MF-grades supports the need to pursue a more objective approach to quantifying reticulin fibrosis.

REFERENCES

1. Ryou, H., Sirinukunwattana, K., Aberdeen, A. *et al.* Continuous Indexing of Fibrosis (CIF): improving the assessment and classification of MPN patients. *Leukemia* 37, 348–358 (2023). https://doi.org/10.1038/s41375-022-01773-0

ACKNOWLEDGEMENTS

This project received funding from:

CANCER RESEARCH UK
Blood cancer UK
British Society for Haematology
Listening • Learning • Leading

CONTACT INFORMATION

Dr. Timothy J Ebsworth: timothy.ebsworth@ouh.nhs.uk
Dr Sharon Ruane: sharon@groundtruthlabs.com
Prof. Daniel Royston: daniel.royston@ndcls.ox.ac.uk