Interobserver agreement in assessment of Rutgeerts' score of endoscopic recurrence of ileal Crohn's disease

Kennedy NA[1,2], Ennis H[3], Gaya D[4], Mowat C[5], Arnott I[2], TOPPIC trial investigators, Satsangi J[1,2]

[1]GI Unit, Western General Hospital, Edinburgh; [2]GI Unit, CGEM, IGMM, University of Edinburgh; [3]Edinburgh Clinical Trials Unit
[4]GI Unit, Glasgow Royal Infirmary; [5]GI Unit, Ninewells Hospital, Dundee
Email: nick.kennedy@ed.ac.uk

Introduction

- Rutgeerts' score is widely used for the assessment of endoscopic recurrence following ileocaecal (IC) resection for Crohn's disease (CD).
- Higher scores have been shown to be associated with an increased risk of clinical recurrence.[1]
- TOPPIC is a double-blind randomised, placebo-controlled trial of mercaptopurine for the prevention of post-operative recurrence after IC resection for CD and includes a secondary endpoint of endoscopic recurrence.[2]
- Few published data are available on interobserver agreement of Rutgeerts' score.

Aim

- This study aimed to assess the interobserver agreement of Rutgeerts' score on images from endoscopies carried out as part of the TOPPIC trial.

Methods

- Five TOPPIC trial investigators (NK, DG, CM, IA, JS) were shown endoscopic images taken from 43 colonoscopies performed in Edinburgh as part of the TOPPIC trial.
- The investigators were blinded to the original report and were shown only the images with a description of the anatomical location from which each image was taken.
- Each investigator was independently asked to score each colonoscopy using the Rutgeerts' score and a custom-designed application.
- Statistical analysis was performed using R (R Foundation for Statistical Computing, Vienna) and the psych package.[3]
- The five scores for each colonoscopy were compared with each other and the score made by the original endoscopist. Shrout and Fleiss' intraclass correlation[4] and pairwise weighted Cohen's kappa[5] were calculated.

Results

- The original scores for the colonoscopies were spread across the possible scores, with 11 i0, 10 i1, 7 i2, 10 i3 and 5 i4.
- Intraclass correlation for single ratings (ICC3) was 0.82 (95% confidence interval 0.74-0.88).
- Pairwise weighted Cohen's kappa ranged from 0.72 to 0.86:

	CM	DG	IA	NK	JS
Original	0.75	0.74	0.84	0.78	0.75
CM		0.72	0.79	0.85	0.77
DG			0.9	0.85	0.86
IA				0.88	0.86
NK					0.84

- When scores were stratified into endoscopic recurrence or not, as defined by a score of i2 or greater, all five scorers agreed with the original score in 34/43 (79%).
- There was no significant difference in this agreement between those procedures with an original score ≥i2 or <i2.

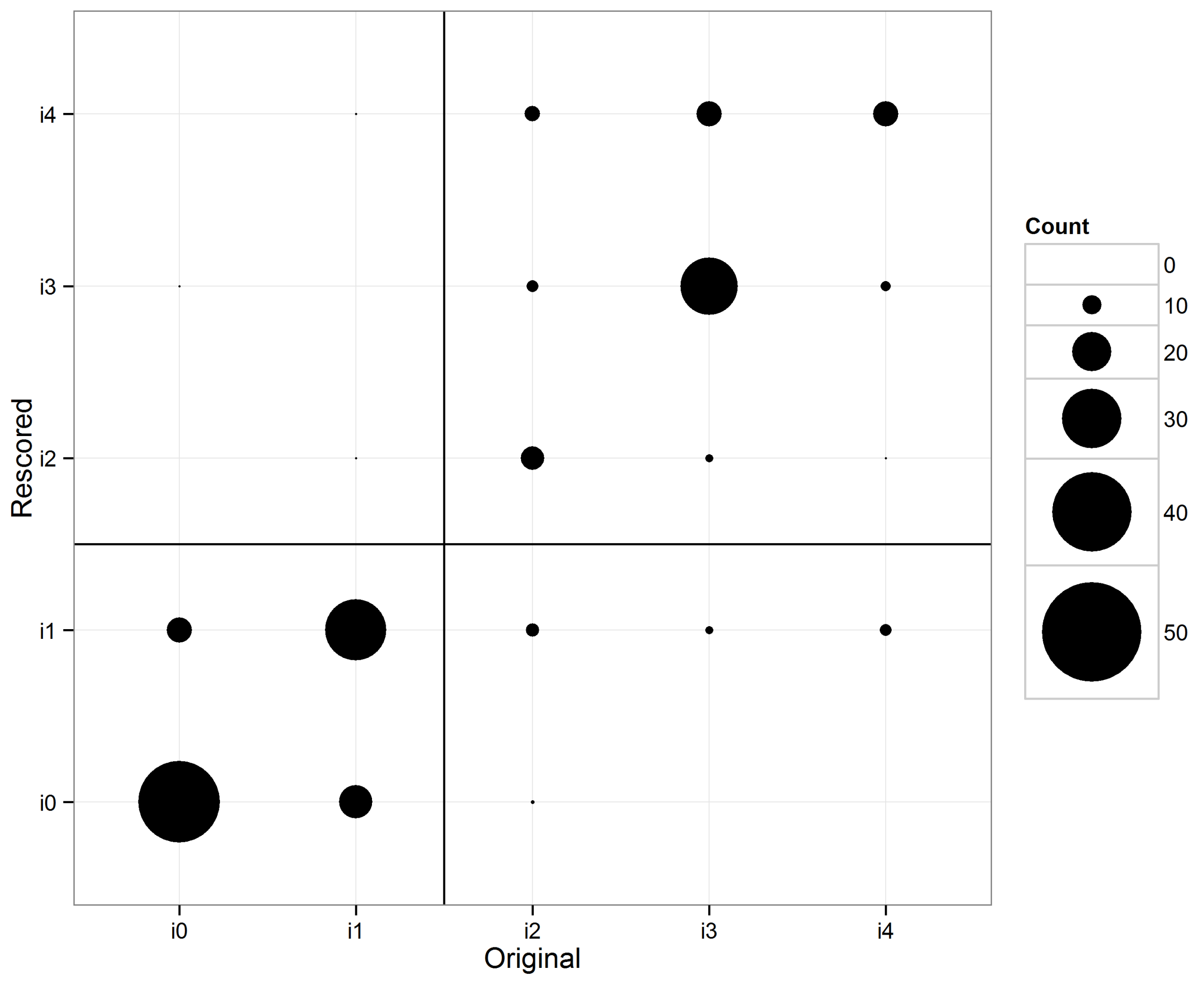
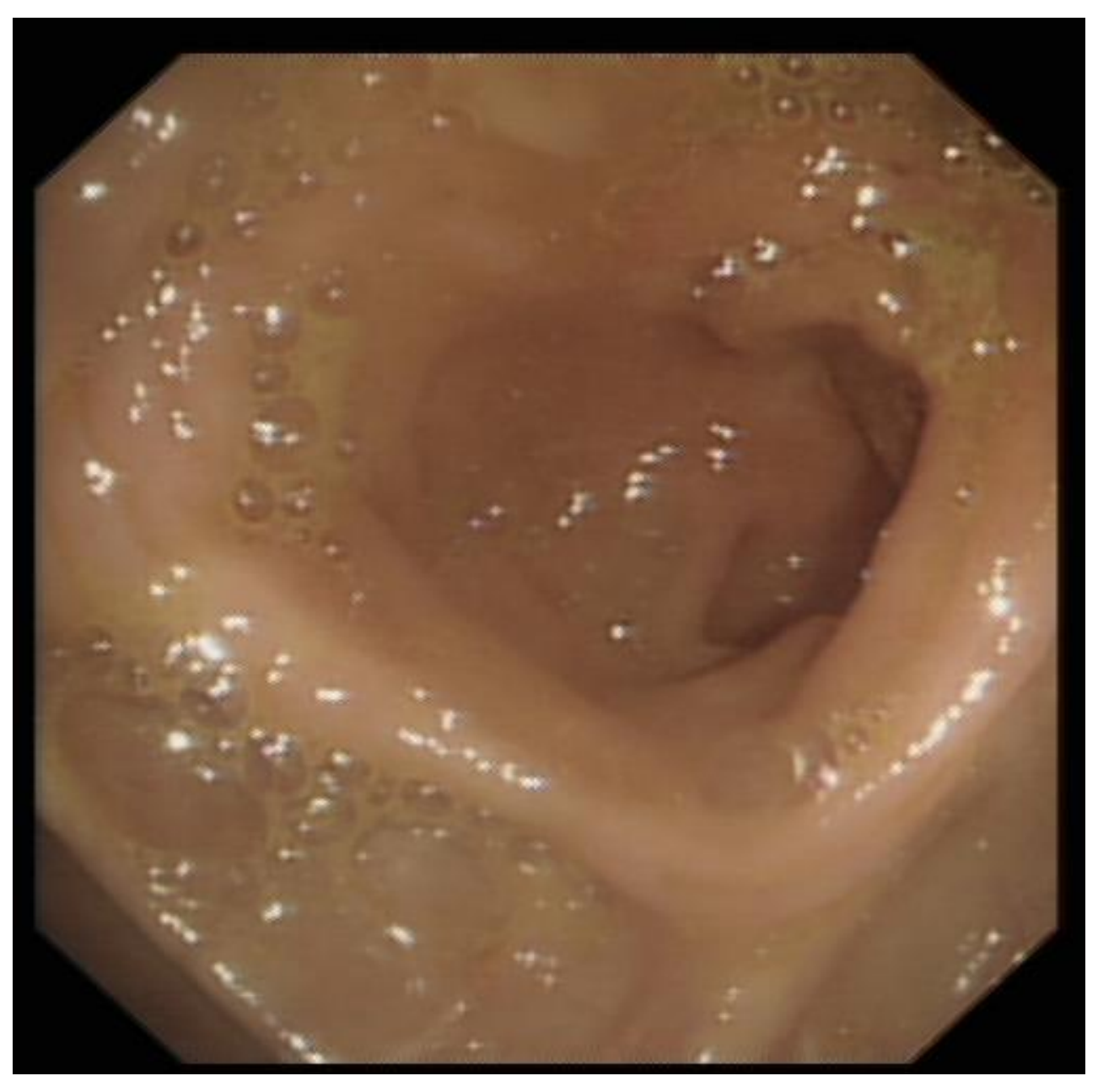
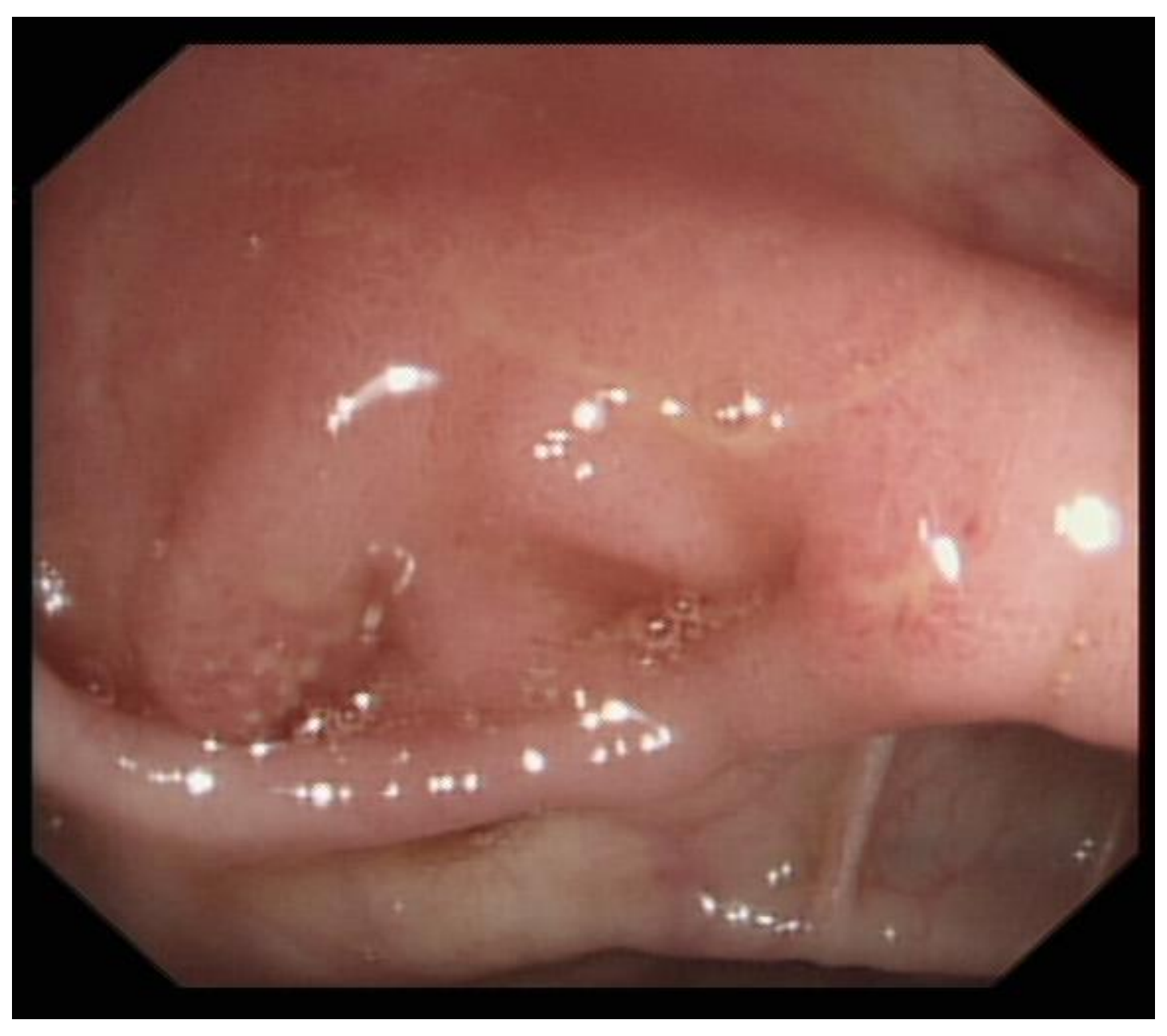


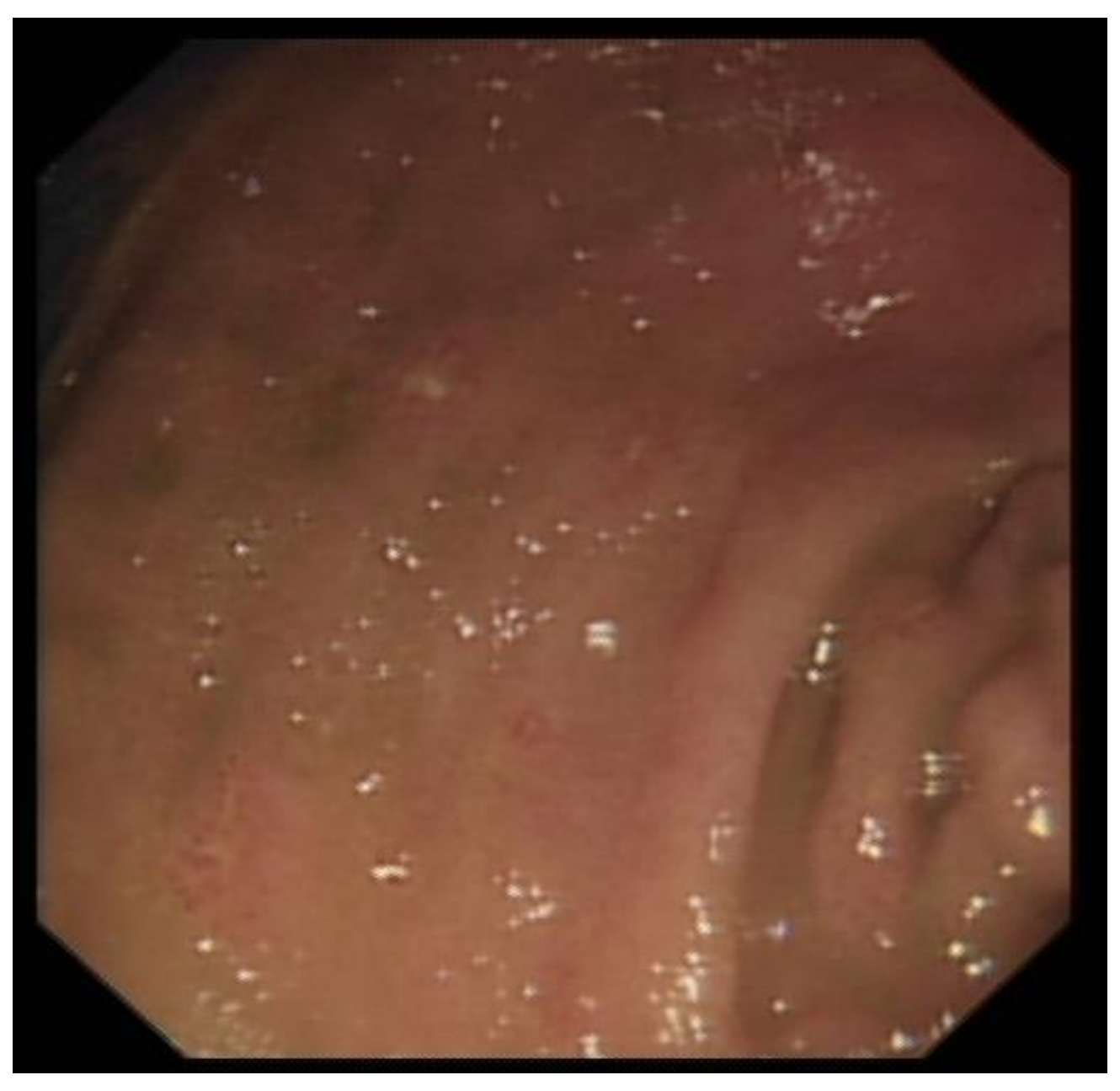
Figure 1: Original Rutgeerts' scores compared to those when endoscopic images were rescored; lines indicate the cut-off between recurrence (≥i2) and not.



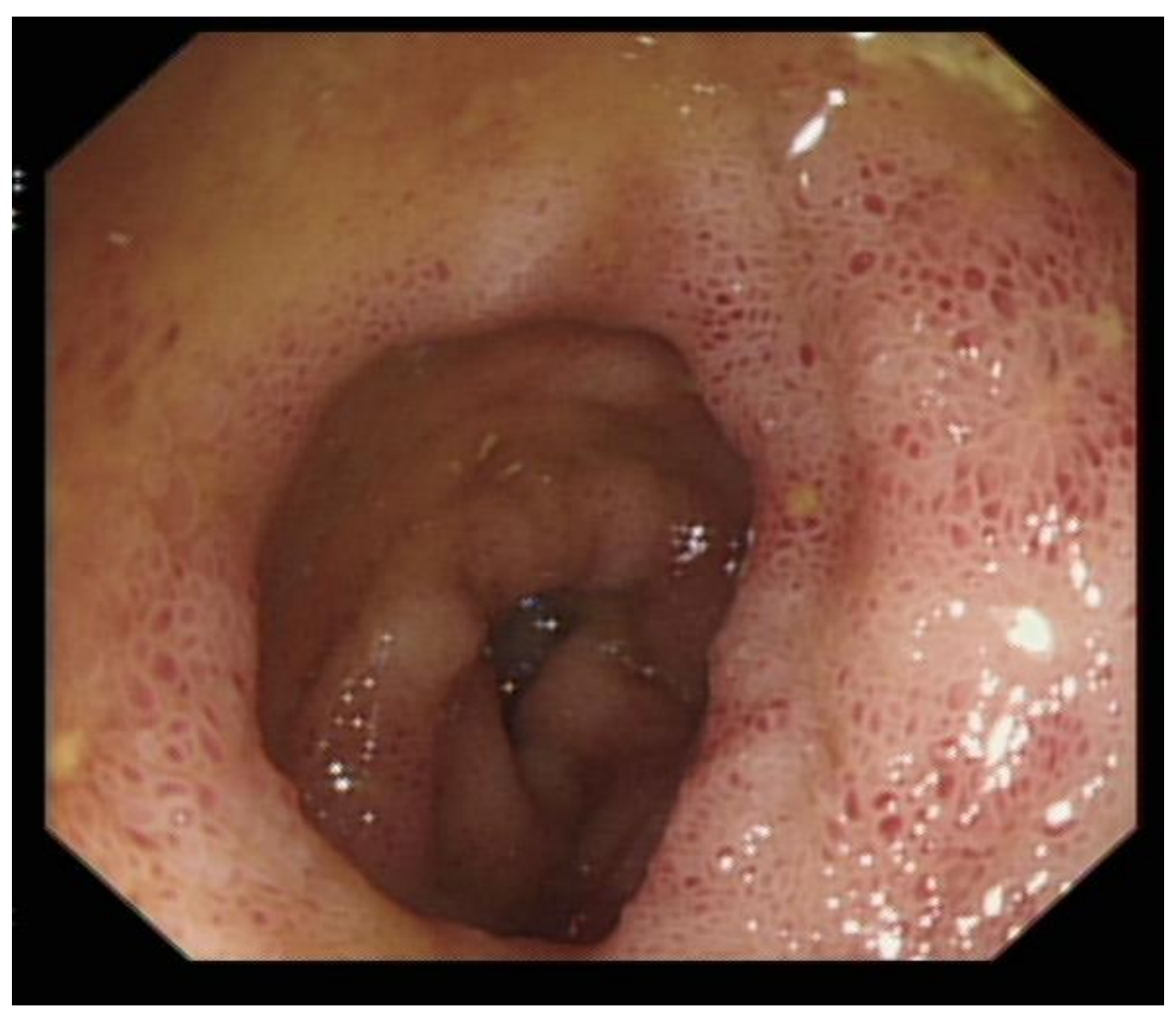
i0 No lesions in distal ileum



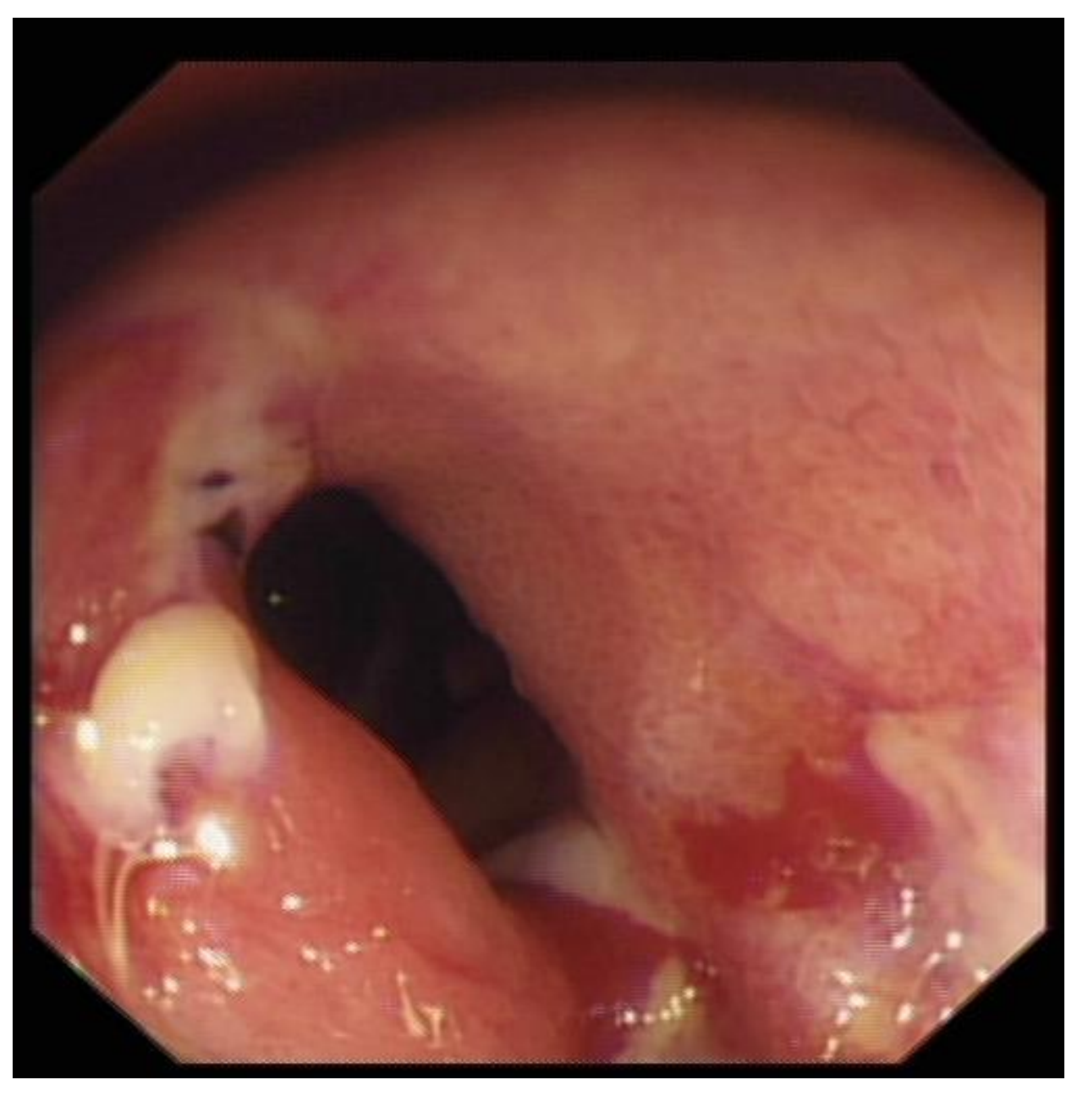
i1 ≤5 apthous lesions



i2 >5 apthous lesions with normal mucosa between the lesions, or skip areas of larger lesions or lesions confined to ileocolonic anastomosis



i3 Diffuse apthous ileitis with diffusely inflamed mucosa



i4 Diffuse inflammation with already larger ulcers, nodules, and/or narrowing

Conclusions

- Interobserver agreement for Rutgeerts' score of endoscopic recurrence was generally good in this cohort of patients, but there was some variation in assessment even when assessing the presence/absence of endoscopic recurrence.
- These findings are important when considering the reliability of outcome data in multicentre clinical trials.

References

1. Rutgeerts P, Geboes K, Vantrappen G, Beyls J, Kerremans R, Hiele M. Predictability of the postoperative course of Crohn's disease. Gastroenterology. 1990 Oct;99(4):956–63.
2. Edinburgh Clinical Trials Unit. TOPPIC Trial. Available from: http://www.clinicaltrials.ed.ac.uk/TrialsPortfolio.aspx
3. Revelle, W. (2014) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, http://CRAN.R-project.org/package=psych Version = 1.4.8.
4. Shrout, Patrick E. and Fleiss, Joseph L. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 1979, 86, 420-3428.
5. Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220.

TOPPIC is funded by a grant from the Medical Research Council.

Dr Nick Kennedy is funded by a research training fellowship from the Wellcome Trust (grant number WT097943MA).