What makes an expert Barrett's pathologist?

Concordance and pathologist expertise within a digital review panel

van der Wel, M.J.

CHAPTER 5

IMPROVED DIAGNOSTIC STRATIFICATION OF DIGITISED BARRETT'S OESOPHAGUS BIOPSIES BY *TP53* IMMUNOHISTOCHEMICAL STAINING: *IMPROVED HOMOGENEITY SUPPORTS P53 USE IN GUIDELINES*

M. J. van der Wel, L. C. Duits, R. E. Pouw, C. A. Seldenrijk, G. J. A. Offerhaus, M. Visser, F. J. W. ten Kate, K. Biermann, L. A. A. Brosens, M. Doukas, C. Huysentruyt, A. Karrenbeld, G. Kats-Ugurlu, J. S. van der Laan, G. van Lijnschoten, F. C. P. Moll, A. H. A. G. Ooms, H. van der Valk, J. G. P. Tijssen, J. J. Bergman, S. L. Meijer

Histopathology. 2018 May;72(6):1015-1023

ABSTRACT

Aims

Interobserver agreement for dysplasia in Barrett's oesophagus (BO) is low and guidelines advise expert review of dysplastic cases. We assessed the added value of p53 immunohistochemistry (IHC) on the homogeneity within a group of dedicated gastro-intestinal (GI) pathologists.

Methods and results

Sixty single hematoxylin & eosin (HE) slide referral BO cases (20 low-grade dysplasia (LGD); 20 high-grade dysplasia (HGD) and 20 non-dysplastic BO (NDBO) reference cases) were digitalised and independently assessed twice in a random order by 10 dedicated GI pathologists. After a 'wash-out' period, cases were re-assessed with the addition of a corresponding p53 IHC slide. Outcomes were 1) proportion of 'indefinite for dysplasia' (IND) diagnoses, 2) interobserver agreement and 3) diagnostic accuracy compared to a consensus 'gold standard' diagnosis defined at an earlier stage by 5 core expert BO pathologists after their assessment of this case set. Addition of p53-IHC decreased the mean proportion of IND diagnoses from 10/60 to 8/60 (p=0.071). Mean interobserver agreement increased significantly from 0.45 to 0.57 (p=0.0021). The mean diagnostic accuracy increased significantly from 72% to 82% (p=0.0072) after addition of p53 IHC.

Conclusion

Addition of p53-IHC significantly improves the histological assessment of BE biopsies, even within a group of dedicated GI pathologists. It decreases the proportion of IND diagnoses and increases interobserver agreement and diagnostic accuracy. This justifies the use of accessory p53 IHC within our upcoming national digital review panel for BO biopsy cases.

INTRODUCTION

Barrett's oesophagus (BO) is a known precursor lesion for the development of oesophageal adenocarcinoma (OAC) and is characterised by the replacement of stratified squamous epithelium with metaplastic intestinal epithelium at the distal oesophagus. OAC can eventually develop through a sequence of events, the metaplasia-dysplasia-carcinoma sequence. [1] Current guidelines recommend endoscopic surveillance of patients with BO with biopsies for proper risk stratification. Histopathological diagnosis of low-grade dysplasia (LGD) is the only accepted predictor for progression. [2] However, this is complicated by interobserver variability between pathologists, [3-6] and conflicting results in reporting rates of progression to OAC for LGD have been reported.[7-11] Studies suggest that when LGD is confirmed by an expert pathologist, risk of progression increases sharply. [9 10 12] Various international BO guidelines also advocate the use of expert review of dysplastic cases, by a pathologist with special interest and extensive experience in interpretation of BO associated neoplasia. [3 5 6 13] To facilitate these revisions in the Netherlands, our goal is to set up a national review panel consisting of expert BO pathologists. For practical purposes we want to solely use digitalised slides. The use of digital microscopy has been validated earlier. [14] To optimize the panel (consensus) diagnosis we wanted to investigate the use of an adjunct diagnostic marker. Earlier studies have shown that aberrant nuclear immunohistochemical staining of the protein encoded for by the tumour suppressor gene *TP53* (p53) may improve the histological assessment of BO.[15-28] *TP53* is a tumour suppressor gene acting as the guardian of genetic stability. The half-life of a normal p53 protein is short and results in a weak immunohistochemical staining (wild-type expression). However, the half-life of mutated *TP53* is prolonged, which together with cellular feedback mechanisms attempting DNA damage repair, lead to nuclear accumulation of (mutated) p53 protein, visible on immunohistochemistry (IHC). In the case of a nonsense mutation or homozygous deletion of the *TP53* locus, a failed translation of the protein, or translation into a truncated protein, lacking the p53 antibody epitope and therefore a complete absence of staining (null mutation) occurs. [29] Currently, guidelines are contradictory concerning the use of p53 IHC in BO diagnostic work up and their advice to use p53 IHC as a diagnostic aid is tentative at most.

Presently, ten dedicated gastro-intestinal (GI) pathologists from the Netherlands are participating in a standardised self-assessment training program, using study sets

of BO biopsy cases enriched for dysplasia. We want to know if p53 IHC can improve the homogeneity of this group of pathologists, defined by the following outcome parameters: 1) the proportion of diagnoses 'indefinite for dysplasia' per pathologist, 2) the interobserver agreement per pathologist, 3) the diagnostic accuracy per pathologist compared to the consensus gold standard diagnosis. The main aim of this study was therefore to test the added value of p53 IHC in the diagnostic work up of a case set consisting of mainly dysplastic BO cases.

MATERIALS AND METHODS

The Medical Ethical Committee of the AMC waived the need for approval for this study.

Case selection, assessors and study design

For this diagnostic study, we selected 60 single hematoxylin & eosin (HE) stained slides from 60 individual BO biopsy cases referred to the BO surveillance program at the Amsterdam Academic Medical Center for pathology review by the local Barrett expert panel [30 31] between 2007 and 2013. The referring diagnosis was LGD in 20 cases and high-grade dysplasia (HGD) or esophageal adenocarcinoma (EAC) in 20 cases. These cases were supplemented with 20 non-dysplastic BO (NDBO) reference cases. From each individual case, a single representative HE and concomitant p53 IHC slide was selected and digitalised by the study coordinator, based on detailed reviewing of the pathology report. For each case, a consensus gold standard diagnosis had been generated previously by five 'core' expert BO pathologists through multiple group discussions (NDBO; n=20, LGD; n=29; HGD; n=11). [14] These core experts currently constitute the national digital review panel for dysplastic BO. They are all working at one of the eight BO expert centers in the Netherlands (range of experience: 10-30 years). In the Netherlands, the care for patients with dysplastic BO is centralised in these eight centers. The expert BO pathologists handle a case load of 5-10 BO cases per week of which 25% is dysplastic. They are considered experts among their peers and each has co-authored more than ten peer-reviewed publications in this field. [14 32] The assessors were ten other dedicated gastro-intestinal (GI) pathologists, also working at the eight BO expert centers in the Netherlands. All ten pathologists independently assessed the case set twice; with a wash-out time of at least one month between the assessment rounds. During the assessments they were blinded to the consensus gold standard diagnoses. The first assessment round consisted of only the HE slide and in

the second assessment round the HE slides were examined in tandem with the p53 immunohistochemically stained slides. All individual diagnoses from each round were recorded on a case record form (CRF).

Histology, immunohistochemistry and digitalising slides

The processs of staining and digitalising of slides is described in the Supplementary Methods. The p53 IHC slides were scored according to international reported criteria as one of 3 staining patterns: normal background staining (wild-type expression), overexpression (nuclear accumulation) or complete absent staining (null-mutation). Normal background staining served as internal control in cases with completely negative staining. Examples of different staining patterns can be appreciated in **Figure 1**.

Outcome measurements

The outcomes of this study were: the proportion of diagnoses 'indefinite for dysplasia' (IND) without and with p53 IHC; the interobserver agreement of the pathologists without and with p53 IHC; and the diagnostic accuracy compared to the consensus gold standard diagnosis, without and with p53 IHC.

Group discussion

After finishing the two assessment rounds, the pathologists met in a group discussion to discuss discrepant cases in relation to the gold standard diagnosis. They especially discussed the interpretation of different p53 staining patterns and formulated a consensus decision rule, which can be appreciated in **Figure 2**.

Figure 1: Examples of p53 immunohistochemistry wild-type expression (A), overexpression (B) and null-mutation (C) from the case set. Note the normal background staining in C (arrow head) serving as internal control.

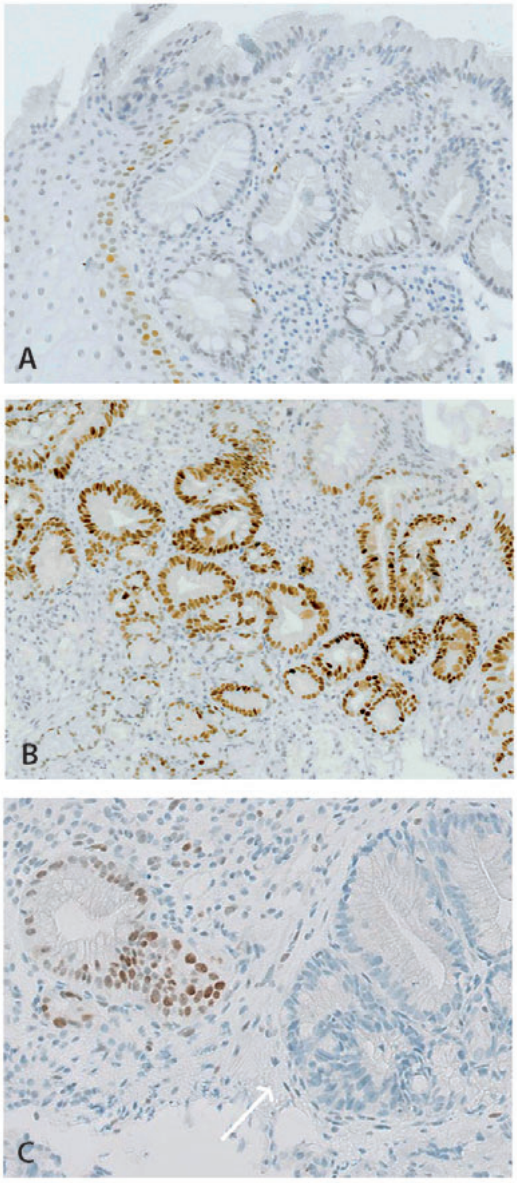
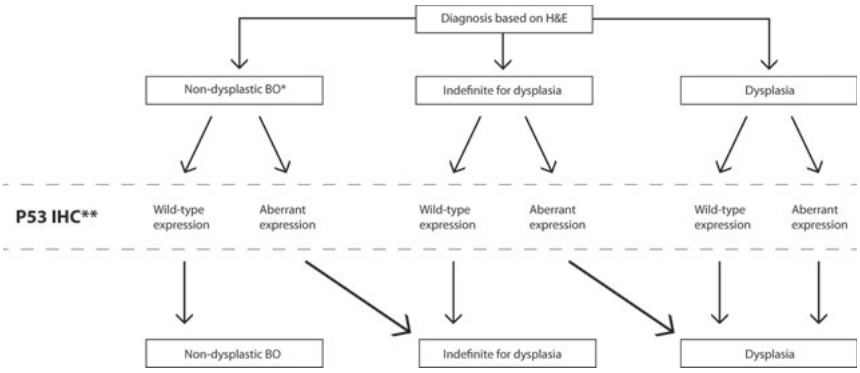


Figure 2: Consensus decision rule for p53 immunohistochemistry interpretation. Aberrant expression includes both overexpression (nuclear accumulation) and complete absent staining (null-mutation).



*Barrett's oesophagus, **immunohistochemistry.

Statistical analysis

The proportion of IND diagnoses per pathologist was counted per assessment round and the difference between the two rounds was calculated. Statistical significance of this difference was calculated using the paired t-test, dichotomizing the outcome to compare the proportion of IND diagnoses per pathologist per round. The statistical significance of the median difference was also calculated using the paired t-test. The interobserver agreement is measured in weighted Cohen's kappa (K) using two diagnostic categories (NDBO + IND and LGD + HGD). [14 33 34] The ten pathologists assessed the case set twice, yielding a total of twenty assessment rounds with nine pairwise assessments per pathologist (pathologist 1 against 2, 1-3, 1-4, 1-5, 1-6, 1-7, 1-8, 1-9 and 1-10). The interobserver agreement per pathologist was defined as the mean kappa of these nine pairwise assessments, per round. The mean interobserver agreement was defined as the mean of all 45 pairwise assessments. Due to the possibility of skewed marginal totals, the maximum possible kappa per cross table does not always equal one. Therefore, the agreement calculated as fraction of maximum possible kappa is also depicted. Strength of agreement was traditionally categorised as: a value of zero or less indicates agreement no better than chance alone ('poor'); 0.00-0.20, 'slight'; 0.21-0.40, 'fair'; 0.41-0.60, 'moderate'; 0.61-0.80, 'substantial'; 0.81-1.00, 'almost perfect'. [35] The difference before and after p53 addition was calculated.

Statistical significance of this difference was calculated using the paired t-test for paired samples, either comparing two times nine pairs per pathologist, or two times 45 pairs for the mean interobserver agreement. The diagnostic accuracy was calculated for the discrimination between non-dysplastic and dysplastic BO of the individual pathologists with respect to the consensus gold standard diagnosis and compared between the two assessment rounds. The outcomes were first dichotomised into 'dysplasia' (LGD + HGD cases) and 'no dysplasia' (NDBO + IND). The 'true positive cases' and 'true negative cases' combined were compared to the outcomes of the pathologists.

A p-value of ≤ 0.05 was considered statistically significant for all tests, and a reason to reject H0. Statistical analyses were performed using the Statistical Package for the Social Sciences (SPSS 24.0, IBM Corp., Armonk, New York, USA). The weighted kappa was developed using the self-automated program Agreestat (version 24, Advanced Analytics, LCC, Gaithersburg, USA).

RESULTS

Proportion of IND diagnoses

Table 1 shows the proportion of IND diagnoses per pathologist. It can be appreciated that the proportion of IND diagnoses decreased for all pathologists except one (pathologist six) after addition of p53 IHC. The proportional differences between assessment round one and two of pathologist six and seven were significant (p=0.034). The median proportion of IND diagnoses was 10/60 cases (17%) in the first assessment round, before p53 IHC addition. After addition of p53 IHC, this proportion decreased to a median of 7/60 (12%) IND diagnoses in the second assessment round. There was a trend towards a lower proportion of IND diagnoses after addition of p53 IHC (p=0.071).

Table 1: Proportion of cases diagnosed as 'indefinite for dysplasia' before and after addition of p53 immunohistochemistry

Pathologist	IND* before p53 IHC†	%	IND after p53 IHC	%	Difference	Difference (%)	p-value‡
1	8/60	13%	5/60	8%	-3	-5%	0.26
2	8/60	13%	3/60	5%	-5	-8%	0.13
3	10/60	17%	5/60	8%	-5	-8%	0.13
4	11/60	18%	9/60	15%	-2	-3%	0.62
5	2/60	3%	1/60	2%	-1	-2%	0.57
6	8/60	13%	15/60	25%	7	12%	0.034
7	9/60	15%	2/60	3%	-7	-12%	0.034
8	22/60	37%	17/60	28%	-5	-8%	0.096
9	13/60	22%	10/60	17%	-3	-5%	0.41
10	10/60	17%	9/60	15%	-1	-2%	0.74
Mean	10/60	17%	8/60	13%	-2	-3%	0.071

*indefinite for dysplasia, †immunohistochemistry, ‡paired t-test (2-tailed), significant when p≤ 0.05

Interobserver agreement

Table 2 shows the interobserver agreement for dysplasia vs no dysplasia before and after the addition of p53 IHC. It can be noted that all individual kappa's improved after the addition of p53 IHC. Six out of ten kappa's improved significantly (see **Table 2** for values). Before p53 IHC addition, the mean interobserver agreement was 0.45, increasing to 0.57 after p53 IHC addition (both 'moderate', p=0.0021), a significant improvement of 0.12. The maximum possible kappa's for these weighted kappa's were all below one. After correction, the mean fraction of maximum possible kappa was 0.59 ('moderate') for round one and 0.73 ('substantial') for round two.

Diagnostic accuracy compared to gold standard diagnosis

Table 3 depicts the diagnostic accuracy per pathologist compared to the gold standard diagnosis, for the distinction of dysplasia from non-dysplastic BO (NDBO + IND vs LGD + HGD). The mean accuracy was 72% for round one (without p53 IHC) and 82% for round two (with p53 IHC), a significant difference of 10% (p-value=0.0072). The diagnostic accuracy improved for all pathologists after the addition of p53 IHC, except for pathologist six (see **Table 3**). Because there was only one cut-off point, the results were depicted in a table instead of in an ROC curve.

Table 2: Interobserver agreement in 2 diagnostic categories*, before and after the addition of p53 immunohistochemistry

Pathologist	Before p53 IHC†			After p53 IHC				
	Mean weighted kappa	Mean max kappa	Mean weighted / mean max kappa	Weighted kappa	Mean max kappa	Mean weighted / mean max kappa	Difference (mean weighted kappa)	p-value‡
1	0.52	0.81	0.62	0.64	0.86	0.76	+0.12	0.0069
2	0.58	0.77	0.75	0.64	0.77	0.84	+0.06	0.069
3	0.52	0.83	0.63	0.58	0.85	0.67	+0.06	0.13
4	0.35	0.83	0.43	0.54	0.77	0.73	+0.19	0.0007
5	0.54	0.81	0.66	0.63	0.71	0.89	+0.09	0.025
6	0.42	0.64	0.65	0.49	0.71	0.71	+0.07	0.18
7	0.43	0.83	0.54	0.62	0.86	0.73	+0.19	0.0001
8	0.26	0.71	0.39	0.44	0.83	0.55	+0.18	0.0002
9	0.46	0.77	0.62	0.52	0.85	0.62	+0.06	0.11
10	0.42	0.81	0.53	0.59	0.85	0.70	+0.17	0.004
Mean§	0.45	0.78	0.59	0.57	0.81	0.73	+0.12	0.0021§

*non-dysplastic BO + indefinite for dysplasia; low-grade dysplasia + high-grade dysplasia, †immunohistochemistry, ‡paired t-test (2-tailed), significant when p≤ 0.05, §total of 45 pairwise kappas

Table 3: Diagnostic accuracy of non-dysplastic BE versus dysplasia, compared to the gold standard diagnosis before and after the addition of p53 immunohistochemistry

Pathologist	Before p53 IHC* (%)	After p53 IHC (%)	Difference (%)	p-value†
1	82	83	1	0.57
2	85	93	8	0.096
3	77	85	8	0.096
4	63	73	10	0.14
5	82	95	13	0.01
6	80	70	-10	0.16
7	65	85	20	0.002
8	63	78	15	0.002
9	63	77	14	0.004
10	63	83	20	0.001
Mean	72	82	10	0.0072

*IHC = immunohistochemistry, †paired t-test (2-tailed), significant when p≤ 0.05

DISCUSSION

In this diagnostic study, ten dedicated GI pathologists assessed a histological single-slide case set of 60, mainly dysplastic, BO biopsy cases. The addition of p53 IHC in the second assessment round decreases the median proportion of IND diagnoses (p=0.071), significantly increases the mean interobserver agreement (p=0.0021) and significantly increases the mean diagnostic accuracy (p=0.0072) of this large group of GI pathologists. This signifies that p53 IHC appears to aid pathologists in accurate stratification of BO patients compared to a consensus gold standard diagnosis generated by five core expert pathologists that are current members of the national digital review panel for BO. [14] The positive effect on interobserver agreement observed in our study is comparable to existing literature. Kaye et al. performed two histopathological studies to investigate the added value of p53 IHC on observer agreement amongst two groups of pathologists. The first study consisted of a group of five pathologists assessing 186 single-slide cases of BO, [17] and the second study of a group of ten pathologists assessing 72 BO cases. [18] The first study showed improvement of interobserver weighted kappa scores after addition of p53 IHC, from 0.42 to 0.48 (both 'moderate'). [17] The second study investigated generalizability of p53 IHC, showing that p53 IHC interpretation was more reliable than HE interpretation

for diagnosing BO + dysplasia cases and improved the mean interobserver weighted kappa from 0.47 to 0.55 (both 'moderate'). [18] Both studies also showed aberrant p53 expression as an independent predictor of disease progression. However, compared to our study, the pathologists did not meet in group discussions to discuss discrepant cases and did not have a consensus diagnosis for comparison. Despite the positive effect of p53 IHC on BO diagnostics in these studies, guidelines still only half-heartedly advise the use of p53 IHC.

This study has a number of unique features. First, a face-to-face group discussion with all dedicated GI pathologists was held after the two assessment rounds were finished. This gave the group the opportunity to discuss all discrepant cases in relation to p53 IHC expression and establish a consensus decision rule for the interpretation of p53 IHC, if it is performed, which can be appreciated in **Figure 2**. Second, the set-up of this study was thorough. The study set was enriched for difficult dysplastic cases, it was carried out with the help of digitalised slides for maximum efficiency, and all pathologists assessed the study set twice and independently in a randomised fashion, with a 'wash-out' phase. Third, to calculate the accuracy of the pathologists, we used a consensus gold standard diagnosis, generated earlier by the five 'core' expert pathologists of the national digital review panel. They are considered true experts by their peers and have proven this in many earlier studies. [14]

The first limitation of this study concerns interpretability of p53 IHC. International guidelines are not uniform in the use of p53 IHC and no objective parameters for the morphological assessment, nor for the application frequency of p53 IHC in BO biopsies, exist. It is both a quantitative and qualitative interpretation that is combined with the assessment of morphological features on the HE slide. There is no compelling evidence to perform p53 IHC on every single BO biopsy. However, since it is useful in diagnostic work-up and improves diagnostic agreement, we do advocate a low threshold for p53 IHC use, while still recognising its limits. The morphological diagnosis on HE should be leading, and aberrant staining can be used to support a diagnosis of dysplasia. In this setting, aberrant p53 IHC staining can also be used as a diagnostic biomarker for comparison of future biopsies, or as a marker for progression in clinical studies. [16-18] A non-aberrant p53 IHC staining pattern in a morphologically dysplastic specimen, however, does not exclude the presence of dysplasia. When p53 IHC exhibits a focus of unexpectedly aberrant staining, this may cause diagnostic confusion. The

exact clinical meaning of such a focus is unknown, but when unequivocal aberrant, we consider it best to regard it as IND, which has been found to be associated with progression. [36] Using the results from our group discussion, we summarised the above findings in a p53 IHC decision rule (**Figure 2**), in order to increase uniformity of p53 IHC interpretation in the future. For optimal p53 IHC staining, adhering to external quality programs and using standardized staining protocols with adequate controls is a prerequisite. Furthermore, training in interpretation of the different staining patterns compared to background staining is necessary. Second, we are aware that our study shows a large range of outcome measures in both the proportion of IND diagnoses and the interobserver agreement. We attribute this to the fact, that the group of dedicated GI pathologists has just started a standardised training program for dysplastic BO, but has never before received training together as a group. Nonetheless, in all pathologists except one we see a trend towards less IND diagnoses when p53 IHC is added. However, when the pathologist with more IND diagnoses after p53 IHC is excluded from the analysis, the rest of the group exhibits significantly less IND diagnoses after p53 IHC addition (p=0.0009, **Supplementary Table 1**). These results confirm the need for uniform interpretation of p53 IHC. The mean interobserver agreement and mean diagnostic accuracy of pathologist six does show a statistical significant improvement when p53 IHC is added. The last limitation concerns the artificial set-up of the study, because each case only consisted of a single slide. Most endoscopic procedures generate biopsies on more than one level, leading to more than one slide per case. Therefore, the ten GI pathologists will now assess a second study set, consisting of all slides of all tissue blocks generated during one endoscopic procedure.

In conclusion, our results show that addition of p53 IHC significantly improves the proportion of IND diagnoses, the interobserver agreement, and the diagnostic accuracy within a large group of ten dedicated GI pathologists. This justifies the use of p53 IHC within our upcoming national digital review panel for BO biopsy cases. It can hereby improve the stratification of patients in order to optimize diagnostic work-up and (endoscopic) treatment.

REFERENCES

1. Buttar NS, Wang KK. Mechanisms of disease: Carcinogenesis in Barrett's esophagus. *Nat Clin Pract Gastroenterol Hepatol* 2004;1(2):106-12. doi: 10.1038/ncpgasthep0057

2. Wang KK, Sampliner RE, Practice Parameters Committee of the American College of G. Updated guidelines 2008 for the diagnosis, surveillance and therapy of Barrett's esophagus. *The American journal of gastroenterology* 2008;103(3):788-97. doi: 10.1111/j.1572-0241.2008.01835.x

3. Fitzgerald RC, di Pietro M, Ragunath K, et al. British Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus. *Gut* 2014;63(1):7-42. doi: 10.1136/gutjnl-2013-305372

4. Fock KM, Talley N, Goh KL, et al. Asia-Pacific consensus on the management of gastro-oesophageal reflux disease: an update focusing on refractory reflux disease and Barrett's oesophagus. *Gut* 2016;65(9):1402-15. doi: 10.1136/gutjnl-2016-311715

5. Whiteman DC, Appleyard M, Bahin FF, et al. Australian clinical practice guidelines for the diagnosis and management of Barrett's Esophagus and Early Esophageal Adenocarcinoma. *Journal of gastroenterology and hepatology* 2015 doi: 10.1111/jgh.12913

6. Shaheen NJ, Falk GW, Iyer PG, et al. ACG Clinical Guideline: Diagnosis and Management of Barrett's Esophagus. *The American journal of gastroenterology* 2016;111(1):30-50. doi: 10.1038/ajg.2015.322

7. Wani S, Falk GW, Post J, et al. Risk factors for progression of low-grade dysplasia in patients with Barrett's esophagus. *Gastroenterology* 2011;141(4):1179-86, 86 e1. doi: 10.1053/j.gastro.2011.06.055

8. Hvid-Jensen F, Pedersen L, Drewes AM, et al. Incidence of adenocarcinoma among patients with Barrett's esophagus. *The New England journal of medicine* 2011;365(15):1375-83. doi: 10.1056/NEJMoa1103042

9. Duits LC, van der Wel MJ, Cotton CC, et al. Patients With Barrett's Esophagus and Confirmed Persistent Low-Grade Dysplasia Are at Increased Risk for Progression to Neoplasia. *Gastroenterology* 2017;152(5):993-1001 e1. doi: 10.1053/j.gastro.2016.12.008

10. Duits LC, Phoa KN, Curvers WL, et al. Barrett's oesophagus patients with low-grade dysplasia can be accurately risk-stratified after histological review by an expert pathology panel. *Gut* 2014 doi: 10.1136/gutjnl-2014-307278

11. Phoa KN, van Vilsteren FG, Weusten BL, et al. Radiofrequency ablation vs endoscopic surveillance for patients with Barrett esophagus and low-grade dysplasia: a randomized clinical trial. *JAMA : the journal of the American Medical Association* 2014;311(12):1209-17. doi: 10.1001/jama.2014.2511

12. Curvers WL, ten Kate FJ, Krishnadath KK, et al. Low-grade dysplasia in Barrett's esophagus: overdiagnosed and underestimated. *The American journal of gastroenterology* 2010;105(7):1523-30. doi: 10.1038/ajg.2010.171

13. Weusten B, Bisschops R, Coron E, et al. Endoscopic management of Barrett's esophagus: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. *Endoscopy* 2017;49(2):191-98. doi: 10.1055/s-0042-122140

14. Van der Wel MJ, Duits LC, Seldenrijk CA, et al. Digital microscopy as valid alternative to conventional microscopy for histological evaluation of Barrett's esophagus biopsies. *Diseases of the Esophagus* 2017;30(7)

15. di Pietro M, Boerwinkel DF, Shariff MK, et al. The combination of autofluorescence endoscopy and molecular biomarkers is a novel diagnostic tool for dysplasia in Barrett's oesophagus. *Gut* 2015;64(1):49-56. doi: 10.1136/gutjnl-2013-305975

16. Kastelein F, Biermann K, Steyerberg EW, et al. Aberrant p53 protein expression is associated with an increased risk of neoplastic progression in patients with Barrett's oesophagus. *Gut* 2013;62(12):1676-83. doi: 10.1136/gutjnl-2012-303594

17. Kaye PV, Haider SA, Ilyas M, et al. Barrett's dysplasia and the Vienna classification: reproducibility, prediction of progression and impact of consensus reporting and p53 immunohistochemistry. *Histopathology* 2009;54(6):699-712. doi: 10.1111/j.1365-2559.2009.03288.x

18. Kaye PV, Ilyas M, Soomro I, et al. Dysplasia in Barrett's oesophagus: p53 immunostaining is more reproducible than haematoxylin and eosin diagnosis and improves overall reliability, while grading is poorly reproducible. *Histopathology* 2016;69(3):431-40. doi: 10.1111/his.12956

19. Younes M, Brown K, Lauwers GY, et al. p53 protein accumulation predicts malignant progression in Barrett's metaplasia: a prospective study of 275 patients. *Histopathology* 2017 doi: 10.1111/his.13193

20. Weston AP, Banerjee SK, Sharma P, et al. p53 protein overexpression in low grade dysplasia (LGD) in Barrett's esophagus: immunohistochemical marker predictive of progression. *The American journal of gastroenterology* 2001;96(5):1355-62. doi: 10.1111/j.1572-0241.2001.03851.x

21. Murray L, Sedo A, Scott M, et al. TP53 and progression from Barrett's metaplasia to oesophageal adenocarcinoma in a UK population cohort. *Gut* 2006;55(10):1390-7. doi: 10.1136/gut.2005.083295

22. Timmer MR, Martinez P, Lau CT, et al. Derivation of genetic biomarkers for cancer risk stratification in Barrett's oesophagus: a prospective cohort study. *Gut* 2016;65(10):1602-10. doi: 10.1136/gutjnl-2015-309642

23. Bird-Lieberman EL, Dunn JM, Coleman HG, et al. Population-based study reveals new risk-stratification biomarker panel for Barrett's esophagus. *Gastroenterology* 2012;143(4):927-35 e3. doi: 10.1053/j.gastro.2012.06.041

24. Bani-Hani K, Martin IG, Hardie LJ, et al. Prospective study of cyclin D1 overexpression in Barrett's esophagus: association with increased risk of adenocarcinoma. *Journal of the National Cancer Institute* 2000;92(16):1316-21.

25. Sikkema M, Kerkhof M, Steyerberg EW, et al. Aneuploidy and overexpression of Ki67 and p53 as markers for neoplastic progression in Barrett's esophagus: a case-control study. *The American journal of gastroenterology* 2009;104(11):2673-80. doi: 10.1038/ajg.2009.437

26. Skacel M, Petras RE, Rybicki LA, et al. p53 expression in low grade dysplasia in Barrett's esophagus: correlation with interobserver agreement and disease progression. *The American journal of gastroenterology* 2002;97(10):2508-13. doi: 10.1111/j.1572-0241.2002.06032.x

27. Gimenez A, de Haro LM, Parrilla P, et al. Immunohistochemical detection of p53 protein could improve the management of some patients with Barrett esophagus and mild histologic alterations. *Arch Pathol Lab Med* 1999;123(12):1260-3. doi: 10.1043/0003-9985(1999)123<1260:IDOPPC>2.0.CO;2

28. Younes M, Ertan A, Lechago LV, et al. p53 Protein accumulation is a specific marker of malignant potential in Barrett's metaplasia. *Digestive diseases and sciences* 1997;42(4):697-701.

29. van der Wel MJ, Jansen M, Vieth M, et al. What Makes an Expert Barrett's Histopathologist? *Adv Exp Med Biol* 2016;908:137-59. doi: 10.1007/978-3-319-41388-4_8

30. Hulscher JB, Haringsma J, Benraadt J, et al. Comprehensive Cancer Centre Amsterdam Barrett Advisory Committee: first results. *Neth J Med* 2001;58(1):3-8.

31. Offerhaus GJ, Correa P, van Eeden S, et al. Report of an Amsterdam working group on Barrett esophagus. *Virchows Archiv : an international journal of pathology* 2003;443(5):602-8. doi: 10.1007/s00428-003-0906-z

32. van der Wel MJ, Duits LC, Klaver E, et al. Development of benchmark quality criteria for assessing whole-endoscopy Barrett's esophagus biopsy cases. *United European gastroenterology journal* 2018;6(6):830-37. doi: 10.1177/2050640618764710 [published Online First: 2018/07/20]

33. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 1968;70(4):213-19.

34. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;20(1):37-45.

35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159-74.

36. Kestens C, Leenders M, Offerhaus GJ, et al. Risk of neoplastic progression in Barrett's esophagus diagnosed as indefinite for dysplasia: a nationwide cohort study. *Endoscopy* 2015;47(5):409-14. doi: 10.1055/s-0034-1391091

SUPPLEMENTARY MATERIAL

Supplementary Table 1: Proportion of cases diagnosed as 'indefinite for dysplasia' before and after addition of p53 immunohistochemistry after excluding pathologist six

Pathologist	IND* before p53 IHC†	%	IND after p53 IHC	%	Difference	Difference (%)	p-value‡
1	8/60	13%	5/60	8%	-3	-5%	0.26
2	8/60	13%	3/60	5%	-5	-8%	0.13
3	10/60	17%	5/60	8%	-5	-8%	0.13
4	11/60	18%	9/60	15%	-2	-3%	0.62
5	2/60	3%	1/60	2%	-1	-2%	0.57
7	9/60	15%	2/60	3%	-7	-12%	0.034
8	22/60	37%	17/60	28%	-5	-8%	0.096
9	13/60	22%	10/60	17%	-3	-5%	0.41
10	10/60	17%	9/60	15%	-1	-2%	0.74
Mean	10/60	17%	8/60	13%	-3	-5%	0.0009

*indefinite for dysplasia, †immunohistochemistry, ‡paired t-test (2-tailed), significant when $p \leq 0.05$