What makes an expert Barrett's pathologist?

Concordance and pathologist expertise within a digital review panel

van der Wel, M.J.

CHAPTER 8

PERFORMANCE OF GASTROINTESTINAL PATHOLOGISTS WITHIN A NATIONAL DIGITAL REVIEW PANEL FOR BARRETT'S ESOPHAGUS IN THE NETHERLANDS: *RESULTS OF 80 PROSPECTIVE BIOPSY REVIEWS*

M. J. van der Wel, E. Klaver, 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, J. G. P. Tijssen, S. L. Meijer, J. J. G. H. M. Bergman

Submitted

ABSTRACT

Background

The histopathological diagnosis of low-grade dysplasia (LGD) in Barrett's esophagus (BE) is associated with a poor inter-observer agreement and guidelines dictate review by an expert pathologist. To facilitate nationwide expert review in the Netherlands a web-based digital review panel has been set up, which currently consists of eight "core" pathologists. The aim of this study was to evaluate if seven other pathologists from the Dutch BE expert centers qualify for the panel by assessing their performance in 80 consecutive LGD reviews submitted to the Dutch BE Pathology panel.

Methods

Pathologists independently assessed digital slides in two phases. Both phases consisted of 40 cases, with a group discussion after phase I. For all cases a previous consensus diagnosis made by five core pathologists was available, which was used as reference. The following criteria were used to assess performance: 1) percentage of 'indefinite for dysplasia' diagnoses, 2) percentage agreement with consensus diagnosis 3) proportion of cases with a consensus diagnosis of LGD or high-grade dysplasia underdiagnosed as non-dysplastic BE. A 95% prediction interval based on the individual scores of the eight core pathologists was used as benchmark range for phase I. Qualified pathologists after phase I were added to the core and incorporated in the benchmark scores for phase II.

Results

After phase I, one of the seven pathologists met the benchmark score for all quality criteria, yet three pathologists only marginally failed the agreement with consensus diagnosis (score 68.3%, benchmark 69%). After a group discussion and phase II, five of the six remaining aspirant panel members qualified for the panel with all scores within the benchmark range.

Conclusion

The Dutch Barrett's Pathology review panel now consists of 14 pathologists, who - after a series of structured assessments and group discussions - can be considered homogeneous in their review of biopsies diagnosed with LGD.

INTRODUCTION

Low-grade dysplasia (LGD) is an important histological risk factor for progression to EAC in patients with Barrett's esophagus (BE). The histological diagnosis of LGD is however challenging, because early morphological changes of dysplasia are difficult to distinguish from reactive atypia of reflux induced inflammation. As a result, the inter-observer agreement for the diagnosis of LGD in the community practice is poor. Two studies from our group have shown that 73-85% of the community LGD cases are down-staged to non-dysplastic BE (NDBE) by BE expert pathologists, and that these down-staged cases have a low progression risk during follow-up. On the other hand, if the diagnosis LGD is confirmed by an expert BE pathologist, the risk of neoplastic progression is approximately 10% per patient-year. [1 2] This risk increases significantly if multiple expert pathologists agree on this diagnosis. [3]

New guidelines dictate review of all dysplastic BE cases by an expert pathologist, [4-7] but the concept of "expert pathologist" is ill-defined and access to expert review is not widely available. To facilitate expert review in the Netherlands, a national, web-based digital histology review platform has been set-up by the eight BE expert centers in the Netherlands. All diagnostic work-up and treatment of early BE neoplasia is centralized in these eight centers. The histology review panel was built around five 'core pathologists'. These 'core' BE expert pathologists had a track record in the field of BE of >10 years (range 10-30 years), had participated in multiple teaching programs (i.e. www.best-academia.eu), and had each co-authored on >10 peer reviewed publications in this field. [8-12]

In a prior study we evaluated if we could expand the number of pathologists in the panel to reduce the individual workload, reduce lead-time, and obtain nationwide coverage while maintaining the assessment quality and homogeneity. [11] For this, 10 other gastrointestinal (GI) pathologists from the eight BE expert centers assessed digitalized slides of 60 endoscopy procedures, enriched for dysplasia. This case set had also been assessed by the five core pathologists. Three of the ten pathologists met the stringent benchmark quality criteria for the case set, as established based on the results of the five core pathologists. Therefore, these three pathologists were considered to be homogeneous in their histological assessment with the five core pathologists and joined the core group of the expert panel. The majority of the other assessors showed only limited deviation from the pre-set benchmark scores.

After this structured assessment, all pathologists participated in face-to-face group meetings where discrepant cases of the assessment rounds were discussed. Although we speculate that these assessments and group discussions will likely have improved their assessments and homogeneity, no formal evaluation has been performed on their performance.

Meanwhile, the national review panel has become operational. This platform is driven by reviews of the eight core pathologists and concentrates on revisions for alleged LGD and IND from centers throughout the Netherlands.

The aim of this study was to evaluate the performance of all pathologists of the eight BE expert centers in the assessment of the first 80 prospective LGD reviews submitted to the Dutch Barrett's Pathology Review panel, and to assess if the "non-core" pathologists were homogeneous in their assessments with the eight core pathologists.

METHODS

Case submission and scanning

Eighty BE surveillance cases with a diagnosis of indefinite for dysplasia (IND) or low-grade dysplasia (LGD) during surveillance endoscopy were submitted to the Dutch Barrett's Pathology Review panel (**Table 1**). Requests for consultation were submitted by gastroenterologists or pathologists via a dedicated website (www.barrett.nl), upon which pathology slides were requested by the back office of the review panel. All slides were digitalized, using a scanner with a x20 microscope objective (Slide, Olympus, Tokyo, Japan). Digitalized slides were checked for focus and acuity by the study coordinator, re-scanned if necessary and stored on a secure server. Subsequently pathologists were invited by e-mail to review the slides.

Table 1: Patient and endoscopy characteristics

	Cases Phase I (n=41)	Cases Phase II (n=39)
Age at endoscopy in years (median, IQR)	67 (61-72)	66 (59-70)
Sex (male) (n, %)	26 (68%)	31 (80%)
Levels biopsied (median, IQR*)	2 (1-4)	2 (1-3)
Total biopsies (median, IQR)	8 (4-12)	7 (3-11)
Initial community diagnosis	16 (39%)	16 (41%)
Indefinite for dysplasia	25 (61%)	23 (59%)
LGD†		

*IQR; interquartile range, †LGD; low-grade dysplasia

Histological assessment

All 15 pathologists independently assessed the digital cases through a virtual slide system (Digital Slidebox 4.5, Leica Microsystems, Dublin, Ireland) allowing them to assess the slides similarly to their conventional microscopic diagnostic practice. Cases were scored according to the modified Vienna criteria for GI neoplasms. [13] Diagnostic categories were: NDBE, LGD, HGD or more, or indefinite for dysplasia (IND). The five initial core expert BE pathologists individually assessed all cases at an earlier stage. Group meetings were organized to discuss all cases in which their individual scores did not show a 4/5 or 5/5 agreement. As such a consensus diagnosis was generated as a reference for all pathologists' individual assessments.

All pathologists assessed the first 41 cases in phase I, after which a group discussion was held to discuss discrepant cases. Cases were considered discrepant if three or more pathologists did not agree with the consensus diagnosis. The group discussion consisted of two sessions with one teleconference and one face-to-face meeting. The cases were shown on screen and discussed plenary. A total of nine cases was discussed. Pathologists had access to their own score and the consensus diagnosis after the group discussion, but were unaware about their relative scores. After the group discussion, pathologists individually assessed another 39 cases in the phase II.

Table 2: Results of all pathologists for phase I (n=41)

Pathologist	Percentage of cases 'indefinite for dysplasia'	Percentage agreement (3 categories)*	Consensus LGD & HGD cases underdiagnosed as NDBE (n (%)) n=28
Core pathologists			
1	24.4	82.9	0
2	9.8	80.5	2 (7.1)
3	22.0	80.5	1 (3.6)
4	7.3	87.8	1 (3.6)
5	9.8	78.0	3 (10.7)
New core pathologists			
B	12.2	78.0	1 (3.6)
E	12.2	75.6	1 (3.6)
J	17.1	73.2	1 (3.6)
Aspirant panel members			
A	12.2	39.0	16 (57.0)
C	14.6	58.5	1 (3.6)
D	31.7	65.9	2 (7.1)
F	12.2	70.7	0
G	12.2	68.3	2 (7.1)
H	7.3	68.3	1 (3.6)
I	9.8	68.3	1 (3.6)
Benchmark value†	≤28%	≥69%	≤ 3 (11%)

*NDBE/IND/LGD+HGD, †Based on 8 core pathologists

Quality criteria

As described in our previous study, [11] we used the following outcome parameters to quantify the quality of the individual histological assessment of each pathologist: 1) the percentage of diagnosis of indefinite for dysplasia per pathologist; 2) the percentage of agreement with the consensus diagnosis per pathologist, using three diagnostic categories (NDBE, IND and LGD+HGD); 3) the percentage of cases with a consensus diagnosis of LGD or HGD underdiagnosed as NDBE per pathologist. For these three quality criteria, we created bench mark scores based on the individual scores of the initial five core pathologists supplemented by the individual scores of

the three pathologists who qualified as a core member based on their scores in the aforementioned study in which digitalized slides of 60 surveillance endoscopies were assessed. [11]

Benchmark ranges for each of the three criteria were based on a 95% Prediction Interval (PI) of the individual scores of these eight core pathologists. The 95% PI was calculated as the mean score from the eight pathologists ±2.365 times the standard deviation (based on a t-distribution with seven degrees of freedom, since n=8 pathologists). The upper or lower limit of these ranges, depending on the criterion, were used as benchmark values. Pathologists who met the benchmark score for all three quality criteria during phase I were added to the core and benchmark scores for phase II were calculated based on the new extended core panel. The statistical analyses were performed using the Statistical Package for the Social Sciences (SPSS 24.0, IBM Corp., Armonk, New York, USA).

RESULTS

Baseline characteristics of samples in case sets

For the case set of phase I median age of patients at endoscopy was 67 years (IQR 61-72) and 68% was male. Cases contained a median of 8 biopsies (IQR 4-12) from a median of 2 levels (IQR 1-4). The initial community diagnosis consisted of LGD in 61% of the cases and of IND in 39%. Baseline characteristics of the case set of phase II were similar to those of phase I (**Table 1**).

Performance of pathologists in phase I

For the percentage of IND cases, six of seven aspirant panel members met the benchmark value of 28% with percentages ranging from 7.3% to 14.6%. Only one pathologist did not meet the required performance score and diagnosed 31.7% of cases as indefinite for dysplasia (**Table 2**). For the percentage of agreement with the consensus diagnosis, one pathologist met the benchmark lower limit of 69%, while three others just fell outside the range with a score of 68.3%. The benchmark value for the percentage of consensus LGD and HGD cases underdiagnosed as NDBE was 11% and six pathologists fell within this value. One pathologist did not qualify for this criterion, underdiagnosing 16 of the 28 (57%) dysplasia cases as NDBE. If all three

quality criteria were taken into account one pathologist met all benchmark scores and thus qualified as a core panel member. This pathologist was added to the core and incorporated in the benchmark ranges calculated for phase II. Four pathologists qualified for two criteria, but showed a small deviation from the benchmark score for the percentage of agreement with the consensus diagnosis. Two pathologists did not meet the required benchmark scores for two of the three quality criteria.

Table 3: Results of all pathologist for phase II (n=39)

Pathologist	Percentage of cases 'indefinite for dysplasia'	Percentage agreement (3 categories)*	Consensus LGD & HGD cases underdiagnosed as NDBE (n (%)) n=23
Core pathologists			
1	17.9	89.7	0
2	10.3	87.2	2 (8.7%)
3	28.2	71.8	2 (8.7%)
4	20.5	87.2	2 (8.7%)
5	17.9	82.1	0
New core pathologists			
B	7.7	74.4	2 (8.7%)
E	25.6	69.2	1 (4.3%)
J	12.8	71.8	1 (4.3%)
New core pathologist after phase 1			
F	33.3	61.5	1 (4.3%)
Aspirant panel members			
A	2.6	56.4	7 (30.4%)
C	7.7	66.7	0
D	12.8	64.1	0
G	12.8	66.7	2 (8.7%)
H	12.8	69.2	1 (4.3%)
I	15.4	76.9	1 (4.3%)
Benchmarkvalue†	≤38%	≥56%	≤ 3 (13%)

*NDBE/IND/LGD+HGD, †Based on 9 core pathologists

Performance of pathologists in phase II

After the group discussion to discuss discrepant cases assessed in phase I and, subsequently, completing assessment of the next 39 cases, all pathologists fell within the benchmark score of ≤38% for the percentage of IND cases in phase II; scores ranged from 2.6% to 15.4% (**Table 3**). All six remaining aspirant members fell within the benchmark range for percentage of agreement with the consensus diagnosis. Five of the six pathologists met the benchmark value for percentage of cases with a consensus diagnosis of LGD or HGD underdiagnosed as NDBE (**Table 3**). They had zero (n=2), one (n=2) or two (n=1) underdiagnosed cases (median percentage 4.3% vs 4.3% for core pathologists).

One pathologist persisted in underdiagnosing a significant number of dysplastic cases as NDBE (7/23: 30.4%). At the end of phase II, five of the six remaining aspirant panel members met all quality criteria.

DISCUSSION

The aim of this study was to evaluate the performance of the GI pathologists of the eight BE expert centers in the assessment of the first 80 consecutive review cases submitted to the Dutch Barrett's Pathology review panel with an initial diagnosis of IND or LGD. In our ambition to expand the current panel of eight core pathologists, while maintaining assessment quality and homogeneity, benchmark quality criteria based on the results of the core pathologists were used. These criteria enabled us to assess the ongoing process of homogenizing future panel members. Assessments were done in two phases, with a group discussion to discuss discrepancies after phase I in order to further homogenize assessments.

When comparing the performance scores of the seven non-core pathologists to the benchmark scores, one out of seven aspirant panel members met all three benchmark quality criteria in phase I and was added to the core. Five out of the remaining six pathologists did not meet the benchmark value for only one of the three performance scores, percentage agreement, with three pathologists scoring only marginally below the required benchmark value (68.3% vs 69%). After the group discussion, five of the six remaining non-core pathologists met the benchmark values for all quality criteria during phase II.

This study is the third in a series of studies with this group of pathologists and includes the third independent set of histology cases. As with the previous studies, individual assessments were complemented with face-to-face group discussions, discussing discrepant cases from the slide set. [9 14] The current study demonstrates that 14 pathologists can be considered homogenous in their assessments, which implies that these 14 pathologists are interchangeable as panel members of the Dutch Barrett's Pathology review panel.

In contrast to our earlier studies, we did not assess intra-observer agreement since the assessors only assessed all cases once. We decided not to use this variable as a benchmark quality criterion. Since kappa resembles the agreement percentage corrected for chance, kappa is influenced by the prevalence of the diagnosis and thus the variance over the different categories. Since the cases submitted for review by Dutch Barrett's Pathology Review panel mainly consisted of dysplastic cases the variance is low, leading to a high agreement by chance. This results in low kappa's and is therefore not a reliable representation of the quality of the assessments.

Strengths of this study are the use of a digitalized case set of consecutive BE cases with IND/LGD submitted for review from all over the Netherlands. Therefore, this case selection reflects the daily workload of the Dutch Barrett's Pathology Review panel. This study builds on three earlier studies in which pathologists were trained through structured assessments and group discussions. All pathologists participating in this study come from the eight Barrett Expert Centers in the Netherlands and thus represent the full potential for histological reviews of the Dutch Barrett's Pathology Review panel.

A limitation is that slides come from different laboratories around the Netherlands. This may have affected slide interpretability, especially for pathologists that are relatively new to reviewing cases from outside their own center. However, this reflects the reality of our review panel, which per definition will receive consultations from different pathology laboratories.

In the future, several important steps will be taken while the Dutch Barrett's Pathology Review panel proceeds. A prediction model will be set up, to establish how many pathologist's assessments are needed in order to obtain a reliable diagnosis. The algorithm will incorporate the performance score of the pathologist in previous

assessments with the outcome of his/her review to decide how many additional pathologists will have to review that specific case in order to retrieve a consensus diagnosis with the same reliability as used in the current study. This will enable an efficient and equal distribution of the workload. Online group discussions will be continued to discuss cases without a majority diagnosis. In addition annual trainings will be held and assessment of homogeneity of all panel members will be renewed periodically. This slide set of the first 80 consecutive review cases will be used in an online training program for other pathologists and pathology residents to improve the histopathologic assessment of BE. This training module will be accredited and freely available. Information from all study sets and group discussions will be incorporated. Pathologists and residents with an interest in BE will be able to improve their skills and compare their performance to our panel.

REFERENCES

1. 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
2. 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* 2015;64(5):700-6. doi: 10.1136/gutjnl-2014-307278
3. 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
4. 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
5. Wani S, Rubenstein JH, Vieth M, et al. Diagnosis and Management of Low-Grade Dysplasia in Barrett's Esophagus: Expert Review From the Clinical Practice Updates Committee of the American Gastroenterological Association. *Gastroenterology* 2016;151(5):822-35. doi: 10.1053/j.gastro.2016.09.040
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. di Pietro M, Fitzgerald RC. Revised British Society of Gastroenterology recommendation on the diagnosis and management of Barrett's oesophagus with low-grade dysplasia. *Gut* 2017 doi: 10.1136/gutjnl-2017-314135 [published Online First: 2017/04/09]
8. 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
9. 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)
10. 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]
11. Van der Wel MJ, Klaver E, Duits LC. Adherence to pre-set benchmark quality criteria to qualify as expert assessor of dysplasia in Barrett's esophagus biopsies. In press
12. van der Wel MJ, Duits LC, Pouw RE, et al. Improved diagnostic stratification of digitised Barrett's oesophagus biopsies by TP53 immunohistochemical staining. *Histopathology* 2018;72(6):1015-23. doi: 10.1111/his.13462 [published Online First: 2018 Feb 28]
13. Schlemper RJ, Kato Y, Stolte M. Diagnostic criteria for gastrointestinal carcinomas in Japan and Western countries: proposal for a new classification system of gastrointestinal epithelial neoplasia. *Journal of gastroenterology and hepatology* 2000;15 Suppl:G49-57.

14. van der Wel MJ, Duits LC, Pouw RE, et al. Improved diagnostic stratification of digitised Barrett's oesophagus biopsies by p53 immunohistochemical staining. *Histopathology* 2018;72(6):1015-23. doi: 10.1111/his.13462 [published Online First: 2018/01/10]