UNITED EUROPEAN GASTROENTEROLOGY
ueg journal

Original Article

United European Gastroenterology Journal
2018, Vol. 6(6) 830–837
© Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2050640618764710
journals.sagepub.com/home/ueg
SAGE

Development of benchmark quality criteria for assessing whole-endoscopy Barrett's esophagus biopsy cases

MJ van der Wel[1,2], LC Duits[2], E Klaver[2], RE Pouw[2], CA Seldenrijk[3], GJA Offerhaus[4], M Visser[5], FJW ten Kate[4], JG Tijssen[6], JJGHM Bergman[2] and SL Meijer[1]

Abstract

Background: Dysplasia in Barrett's esophagus (BE) biopsies is associated with low observer agreement among general pathologists. Therefore, expert review is advised. We are developing a web-based, national expert review panel for histological review of BE biopsies.

Objective: The aim of this study was to create benchmark quality criteria for future members.

Methods: Five expert BE pathologists, with 10–30 years of BE experience, weekly handling 5–10 cases (25% dysplastic), assessed a case set of 60 digitalized cases, enriched for dysplasia. Each case contained all slides from one endoscopy (non-dysplastic BE (NDBE), $n = 21$; low-grade dysplasia (LGD), $n = 20$; high-grade dysplasia (HGD), $n = 19$). All cases were randomized and assessed twice followed by group discussions to create a consensus diagnosis. Outcome measures: percentage of 'indefinite for dysplasia' (IND) diagnoses, intra-observer agreement, and agreement with the consensus 'gold standard' diagnosis.

Results: Mean percentage of IND diagnoses was 8% (3–14%) and mean intra-observer agreement was 0.84 (0.66–1.02). Mean agreement with the consensus diagnosis was 90% (95% prediction interval (PI) 82–98%).

Conclusion: Expert pathology review of BE requires the scoring of a limited number of IND cases, consistency of assessment and a high agreement with a consensus gold standard diagnosis. These benchmark quality criteria will be used to assess the performance of other pathologists joining our panel.

Keywords

Barrett's esophagus, review panel, low-grade dysplasia, digital microscopy, whole slide imaging, intra-observer agreement, benchmark quality criteria

Received: 11 December 2017; accepted: 12 February 2018

Key summary

- Barrett's esophagus (BE) with low-grade dysplasia (LGD) is an independent risk factor for the development of oesophageal cancer.
- Interobserver agreement for the diagnosis of LGD by general pathologists is low.
- Review of LGD cases by expert pathologists can accurately stratify patients according to progression risk.
- However, what constitutes an expert pathologist has not currently been defined.

[1]Department of Pathology, Academic Medical Center, Amsterdam, The Netherlands
[2]Department of Gastroenterology and Hepatology, Academic Medical Center, Amsterdam, The Netherlands
[3]Pathology-DNA, Department of Pathology, St. Antonius Hospital, Nieuwegein, The Netherlands
[4]Department of Pathology, University Medical Center, Utrecht, The Netherlands
[5]Symbiant BV, Department of Pathology, Alkmaar, The Netherlands
[6]Department of Cardiology, Academic Medical Center, Amsterdam, The Netherlands

Corresponding author:
Myrtle J van der Wel, AMC, Meibergdreef 9, Amsterdam 1105, AZ, The Netherlands.
Email: vanderwelmj@gmail.com

- We propose to quantify expertise of pathologists assessing dysplastic BE, through the establishment of benchmark values for four quality criteria.
- Adhering to these benchmark quality criteria can improve the uniformity of interpretation of dysplastic BE and serve as useful criteria in a teaching environment.

Introduction

In BE, the normal stratified squamous epithelium of the distal esophagus has been replaced by columnar epithelium containing intestinal metaplasia. BE is a known risk factor for esophagus adenocarcinoma (EAC), especially when dysplasia is present. BE biopsies are graded according to the modified Vienna criteria for gastrointestinal neoplasms.[1] The grading of dysplasia in BE biopsies is difficult and associated with low observer agreement, because the morphological changes are gradual in the metaplasia-dysplasia-carcinoma sequence. Since endoscopic management of BE patients depends on the dysplasia grade,[2–6] BE guidelines advise that all diagnoses of dysplasia should be reviewed by an expert gastrointestinal (GI) pathologist.[2–7] We have shown that such an expert pathology review of BE biopsies has a significant impact on the management and outcome of patients.[8–10] Based on this, and to implement recent BE guidelines, we set up a national digital review panel for dysplastic BE biopsy cases. This panel makes use of digital microscopy slides and is supported by all 15 expert BE pathologists from the eight BE expert centres in the Netherlands. The core of the panel consists of five pathologists who have been working together as a group for many years and all have extensive experience in the field of BE neoplasia.[11–13] One of the problems in creating such an expert panel is that expert pathology is not easily quantified. In earlier publications, we have used the following qualifications for an expert BE pathologist: an actively practising histopathologist who is dedicated to the field of Barrett's for a minimum of 5 years, has a minimum BE biopsy caseload of five cases per week of which ≥25% are dysplastic, has participated in multiple training programmes, is considered an expert by his or her peers and has co-authored or peer-reviewed publications in this field.[9,10,14–20]

There are plans to expand the panel to include 10 other dedicated GI pathologists, working at the eight BE expert centres in the Netherlands. These pathologists have not been collaborating as intensively as the core group; therefore, they are currently participating in a structured self-assessment programme with multiple group discussions before joining the review panel. The goal of the current study was to establish quality parameters for our national digital BE review panel. For this, the five core expert BE pathologists reviewed all slides from all biopsies taken from 60 BE whole-endoscopy cases, followed by group discussions to create a consensus 'gold standard' diagnosis for all cases. The aim was to define benchmark quality criteria for future pathologists who wish to join this panel.

Materials and methods

Slide selection and scanning

We selected all formalin-fixed, paraffin-embedded tissue blocks and/or slides of 60 BE endoscopy procedures. The case set was enriched for dysplastic cases. Thirty-nine cases with an original diagnosis of LGD ($n = 20$) or HGD ($n = 19$) had been sent to our centre for consultation between 2012 and 2014. These 39 dysplastic cases were supplemented with 21 consecutive NDBE cases from a community hospital in the Amsterdam region. All cases were anonymized. Every case contained at least an Hematoxylin & Eosin (HE) and corresponding p53 immunohistochemically stained slide (clone DO-7+BP53-12, #MS-738-P, Thermo Fisher Scientific, Waltham, MA, USA). For each case, all slides were fully digitalized, using a scanner with a ×20 microscope objective (Slide, Olympus, Tokyo, Japan). They were checked for focus and acuity by the study coordinator and re-scanned if necessary. Subsequently, the slides were anonymized, randomized, renamed and stored on a secure server. The viewing software used to view the digital slides during the study was the virtual slide system 'Digital Slidebox 4.5' (http://dsb.amc.nl/dsb/login.php, Slidepath, Leica Microsystems, Dublin, Ireland).

Assessors

The core expert pathology panel consisted of five pathologists (FJWtK, CAS, SLM, MV, GJAO). They have been dedicated to the field of BE for a minimum of 10 years (range 10–30 years) and have a minimum caseload of 5–10 cases per week of which 25% are dysplastic. All pathologists have participated in the Dutch Barrett advisory committee for many years[9,11,12] and are actively practising pathologists. All pathologists participated in multiple training programmes for endoscopists and pathologists (www.best-academia.eu) and each has co-authored more than 10 peer-reviewed publications in this field.[8,9,12,15,17,18,20–25]

Histologic assessment and earlier joint assessments and group discussions

The expert BE pathologists scored cases according to the modified Vienna criteria for gastrointestinal neoplasms.[1,26] In a previous comparative study, they demonstrated that their histological assessment of glass slides and digitalized slides yielded comparable results.[13] For the current study, the pathologists independently assessed all cases twice in random order, with a wash-out time of at least 1 month between the two rounds. They individually logged onto the virtual slide system to assess the cases. The study coordinator supervised all assessments and recorded the pathologists' answers on a case record form. Diagnostic possibilities were: NDBE; LGD; HGD; or 'indefinite for dysplasia' (IND). After the two assessment rounds, a group discussion was held in which cases that did not have an agreement of 4/5 or 5/5 pathologists were discussed. After discussion, all cases had a diagnostic agreement of 4/5 or 5/5 pathologists, and these diagnoses were considered as the consensus gold standard diagnosis of each case.

Outcome measurements

The outcome measurements were: (1) the percentage of diagnoses 'indefinite for dysplasia' per pathologist; (2) the intra-observer agreement per pathologist; and (3) the percentage agreement with the consensus gold standard diagnosis per pathologist. The percentage of IND diagnoses was depicted as the mean percentage over two assessment rounds, per pathologist. The intra-observer agreement was measured in kappa (see below) and was calculated by comparing each pathologist's first and second assessment per case. The percentage agreement with the consensus gold standard diagnosis was defined as the proportion of correct diagnoses per pathologist when comparing these to the consensus gold standard diagnosis (Supplementary Figure 1, diagonal). The cases that were not in agreement with the consensus gold standard diagnosis were either overdiagnosed (i.e. given a higher diagnosis by the pathologist than the consensus gold standard diagnosis) or underdiagnosed (i.e. given a lower diagnosis by the pathologist than the consensus gold standard diagnosis; see Supplementary Figure 1, lower left and upper right triangles). An additional focus was put on the cases diagnosed in consensus as HGD that were misdiagnosed as NDBE by the individual pathologist (see the darker square at the top right corner of Supplementary Figure 1). All calculations were carried out by using the mean of the two assessment rounds.

Statistical analysis

We studied the variation in the outcome parameters among our five core expert pathologists in order to use these as benchmark quality criteria for other pathologists joining the expert panel. For this, we considered them as a random sample taken from a hypothetical population of expert BE pathologists and assumed a normal distribution for the values. Therefore, we calculated the mean and standard deviation (SD), from which a 2.776*SD range around the mean was calculated ($n = 5$ pathologists yields 4 degrees of freedom) for the 95% prediction interval (PI). We assumed no statistical difference between pathologists if all values fell within this prediction interval. For the calculation of the intra-observer agreement, we used Cohen's kappa. This is a statistical measure for agreement adjusted for chance agreement.[27,28] We used three diagnostic categories (NDBE; LGD + HGD; IND) and assigned custom weights to (dis)agreements, since the spectral changes do not necessarily follow the diagnostic categories 1 to 4. For example, IND is ranked '2' but is not always situated between NDBE (1) and LGD (3).[13,29] Agreement was assigned a score of 1, disagreements between NDBE and HGD were assigned a score of 0, all other disagreements a score of 0.5. Due to the possibility of skewed marginal totals, the maximum possible kappa per cross table does not always equal 1. Therefore, the agreement calculated as a fraction of maximum possible kappa is also depicted. The agreement was traditionally categorized as follows: 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'.[30] The percentage agreement with the consensus gold standard diagnosis was calculated by correlating the pathologist's diagnoses and consensus gold standard diagnoses in a 4 × 4 table (NDBE; IND; LGD; HGD, see Supplementary Figure 1 and Supplementary Tables 1 and 2). Since the management of both LGD and HGD as cancer precursors is the same in the Netherlands, these two categories were grouped. The statistical analyses were performed using the Statistical Package for the Social Sciences (SPSS 24.0, IBM Corp., Armonk, New York, USA). The custom weighted kappa was developed using the self-automated program Agreestat (version 2013.2, Advanced Analytics, LCC, Gaithersburg, USA).

Results

Baseline characteristics of samples in case set

Median age of patients at diagnosis was 66 years (IQR 58–71) and 73% were male. Cases contained a median of five slides (IQR 3–9), from a median of two levels (IQR 1–4) with four biopsies per level (IQR 3-4.5).

Percentage of diagnoses 'indefinite for dysplasia'

Table 1 shows the percentage of IND diagnoses per expert BE core pathologist for the complete case set ($n = 60$ cases). The mean percentage of IND diagnoses over both rounds was 8% (95% PI: 3–14%).

Table 1. Percentage of cases diagnosed as 'indefinite for dysplasia' for the five core pathologists (mean over two assessment rounds) for the complete case set ($n = 60$).

Pathologist	Percentage of cases 'indefinite for dysplasia' (95% PI)[a]
1	6
2	12
3	7
4	8
5	9
Mean	8 (3–14)

[a] 95% prediction interval.

Intra-observer agreement over all cases ($n = 60$)

Table 2 shows the intra-observer agreement of the five core pathologists. The assessments are categorized into three categories according to the Vienna criteria (NDBE; IND; LGD + HGD). The panel displayed 'almost perfect' agreement for the distinction of dysplasia versus no dysplasia, with a mean intra-observer weighted kappa score of 0.84 (95% PI: 0.66–1.02). Due to skewed marginal totals, the maximum kappas per pathologist were lower than 1, which makes mean weighted kappas less representative for the true agreement between the pathologists. Therefore, the fraction of maximum kappa ('weighted/max kappa') was also calculated. With a value of 0.89 (95% PI: 0.62–1.15), the mean fraction of maximum kappa was also 'almost perfect'.

Table 2. Intra-observer agreement of five core pathologists for the complete case set ($n = 60$) in three categories.[a]

Pathologist	Weighted kappa[b] (95% PI)[c]	Max kappa[d]	Weighted/max kappa (95% PI)
1	0.91	0.92	1.00
2	0.77	0.86	0.79
3	0.89	0.92	0.96
4	0.87	0.91	0.92
5	0.75	0.87	0.76
Mean	0.84 (0.66–1.02)	0.90	0.89 (0.62–1.15)

[a] Non-dysplastic BE; indefinite for dysplasia; low-grade dysplasia/high-grade dysplasia.
[b] Custom-weighted Cohen's kappa.
[c] 95% prediction interval.
[d] Maximum possible kappa per cross table.

Percentage agreement of the five core pathologists with the gold standard diagnosis

Table 3 shows the mean agreement of the five core pathologists with the consensus gold standard diagnosis over two assessment rounds. The mean percentage of cases where the diagnosis was in agreement with the consensus gold standard diagnosis was 90% (95% PI: 82–98%). The mean percentage of overdiagnosed cases was 3%, and the mean percentage of underdiagnosed cases was 8%. The mean percentage of consensus gold standard HGD diagnosed cases that were misdiagnosed as NDBE by the pathologists was 0.17%, with a maximum number of 0.8% for pathologist 2, or 1/120 assessments. The cross tables of the consensus gold standard diagnoses versus every pathologist separately are depicted in Supplementary Table 1.

Post-hoc analysis on cases with a baseline diagnosis of dysplasia ($n = 39$)

We observed that for almost all cases with a baseline diagnosis of NDBE, the agreement of the panel was 4/5 or 5/5 on that diagnosis (results not shown). Since these NDBE cases increase the overall agreement and the case load presented to the future panel will presumably include mostly dysplastic cases, we performed a post hoc analysis on only those cases with a baseline diagnosis of LGD or HGD ($n = 39$). Table 4 displays the percentage of diagnoses 'indefinite for dysplasia' for cases with a baseline diagnosis of LGD or HGD. The mean percentage of diagnoses 'indefinite for dysplasia' decreased to 7% (95% PI: –2 to 16%), compared with when the whole case set was taken into account. The mean intra-observer agreement (Table 5) for cases with a baseline diagnosis of LGD or HGD ($n = 39$) was 'fair' with a value of 0.56 (95% PI: 0.39–0.73). When corrected for the maximum possible kappa (given skewed marginal totals in some cross tables), the mean fractions of maximum kappa are again 'almost perfect' with a value of 0.81 (95% PI: 0.35–1.27). When looking

Table 3. Percentage agreement of five core pathologists with consensus gold standard diagnosis (mean over two assessment rounds) for the complete case set ($n = 60$).

Pathologist	Agreement (%; 95% PI[a])	Overdiagnosis (%)	Underdiagnosis (%)	HGD[b] cases misdiagnosed as NDBE[c] (%; fraction)
1	92	2	7	0 (0/120)
2	84	2	14	0.8 (1/120)
3	93	2	6	0 (0/120)
4	91	2	8	0 (0/120)
5	91	6	3	0 (0/120)
Mean	90 (82–98)	3	8	0.17

[a] 95% prediction interval.
[b] High-grade dysplasia.
[c] Non-dysplastic BE.

Table 4. Percentage of cases diagnosed as 'indefinite for dysplasia' for the five core pathologists (mean over two assessment rounds) for cases with a baseline diagnosis of low-grade dysplasia or high-grade dysplasia ($n = 39$).

Pathologist	Percentage of cases 'indefinite for dysplasia' (%; 95% PI)[a]
1	4
2	13
3	5
4	8
5	5
Mean	7 (−2 to 16)

[a] 95% prediction interval.

Table 5. Intra-observer agreement of five core pathologists for cases with a baseline diagnosis of low-grade dysplasia or high-grade dysplasia ($n = 39$) in three categories.[a]

Pathologist	Weighted kappa[b] (95% PI)[c]	Max kappa[d]	Weighted/max kappa (95% PI)
1	0.59	0.59	1.00
2	0.52	0.87	0.60
3	0.64	0.64	1.00
4	0.58	0.86	0.67
5	0.46	0.82	0.56
Mean	0.56 (0.39–0.73)	0.75	0.81 (0.35–1.27)

[a] Non-dysplastic BE; indefinite for dysplasia; low-grade dysplasia/high-grade dysplasia.
[b] Custom-weighted Cohen's kappa.
[c] 95% prediction interval.
[d] Maximum possible kappa per cross table.

at the agreement with the consensus gold standard diagnosis, the mean proportion of cases in agreement with the consensus gold standard diagnosis is 89% (95% PI: 73–104, Table 6). The mean percentage of overdiagnosed cases was 2%, and the mean percentage of underdiagnosed cases was 10%. The mean percentage of consensus gold standard HGD cases that were misdiagnosed as NDBE by the pathologists was 0.25%, with a maximum percentage of 1.3% for pathologist 2, or 1/78 assessments. These results are also visualized in cross tables per pathologist in Supplementary Table 2.

Discussion

The aim of this study was to define benchmark quality criteria for the assessment of BE biopsies for our national digital review panel. For this purpose, our five core expert BE pathologists reviewed all slides of 60 whole-endoscopy BE cases enriched for dysplasia. After their individual assessments, they discussed discrepant cases and agreed on a consensus gold standard diagnosis for all cases. Our five core expert BE pathologists were found to have a mean percentage of IND diagnoses of 8% (95% PI: 3–14), a mean intra-observer agreement of 0.84 (95% PI: 0.66–1.02) and a mean agreement with consensus gold standard diagnosis of 90% (95% PI: 82–98). The scenario with the largest clinical consequences, i.e. a consensus diagnosis of HGD but misdiagnosed as NDBE, was a rare event. When we focused on those cases relevant to our future panel, namely the cases with a baseline diagnosis of LGD or HGD ($n = 39$), results were similar. For clinical decision making, the distinction between LGD and HGD has limited consequences. After all, confirmed LGD has the same management as HGD. The distinction between NDBE-LGD-IND is the one that

Table 6. Percentage agreement of five core pathologists with consensus gold standard diagnosis (mean over two assessment rounds) for cases with a baseline diagnosis of low-grade dysplasia or high-grade dysplasia ($n = 39$).

Pathologist	Agreement (%; 95% PI[a])	Overdiagnosis (%)	Underdiagnosis (%)	HGD[b] cases misdiagnosed as NDBE[c] (%; fraction)
1	90	3	8	0 (0/78)
2	78	1	21	1.3 (1/78)
3	94	1	5	0 (0/78)
4	87	3	10	0 (0/78)
5	94	1	5	0 (0/78)
Mean	89 (73–104)	2	10	0.25

[a] 95% prediction interval.
[b] High-grade dysplasia.
[c] non-dysplastic BE.

Table 7. Values for benchmark quality criteria based on 95% prediction interval of five core pathologists.

Quality criterium	95% PI[b] core pathologists all cases ($n = 60$)	Benchmark value	95% PI core pathologists' dysplastic cases ($n = 39$)	Benchmark value
Percentage of IND[a] cases (%)	3–14%	≤14%	-2 to 16%	≤16%
Intra-observer agreement in three categories (K)	0.66–1.02	≥0.66	0.39–0.73	≥0.39
Agreement with consensus gold standard diagnosis (%)	82–98%	≥82%	73–104%	≥73%
Consensus HGD[c] cases misdiagnosed as NDBE[d] (%; fraction)	0.8% (1/120)	≤0.8% (1/120)	1.3% (1/78)	≤1.3% (1/78)

[a] Indefinite for dysplasia.
[b] 95% prediction interval.
[c] High-grade dysplasia.
[d] Non-dysplastic BE.

pathologists find most difficult and also the one that has the biggest impact on further patient management.

While developing the digital review panel for BE, we ran into the problem that validated benchmark quality criteria of an 'expert BE pathologist' do not exist. We worked around this problem by evaluating the performance of expert BE pathologists with an international reputation in this field. Based on their performance and the current case set of 60 cases enriched for dysplasia, we propose the following four benchmark quality criteria: first, a low proportion of cases diagnosed as 'indefinite for dysplasia', signifying a sufficient contribution of the expert pathologist to panel decision making; second, a high intra-observer agreement signifying consistency of the pathologist in his/her diagnoses over different assessment rounds; third and fourth, a high agreement with the consensus gold standard diagnosis, with an additional focus on a low rate of misdiagnosed consensus gold standard HGD cases as NDBE, both signifying a high reliability of the pathologist concerning panel output. In this current study, not all expert BE pathologists performed equally in all categories. These five were selected based on their experience and track record within the field of BE neoplasia and clearly met all subjective criteria for qualifying as an expert in the field. Since one of the purposes of this study was to create quantitative benchmark values for the assessment of BE biopsies, we do have to accept a certain range of values within our highly selected group of experts, in order to allow other pathologists to work towards an achievable goal. For the first three criteria, the future panel pathologists are required to fall within the 95% PI of our five core pathologists, when assessing the complete case set but also only the subset of dysplastic cases (post hoc analysis). Additionally, the maximum number of HGD cases misdiagnosed as NDBE per pathologist is maximized at one case in two assessment rounds (i.e. 1/120 assessments, Table 7). We are planning to use the criteria and the current case set to evaluate future GI pathologists aiming to join the panel. Our future plans also include expansion of the panel to an international level.

This study has a number of unique features. First, the pathologists participating in this study are the top BE pathologists of the Netherlands, all with an international reputation in this field. This is the second study in the line of the national digital BE review panel that they are performing as a group. Second, the case set consists of whole-endoscopy cases (all slides from all biopsy levels of one endoscopy), was fully digitalized and only contains review cases from clinical practice. There were two assessment rounds with an adequate wash-out time and the pathologists held group discussions afterwards to discuss all discrepant cases and create a consensus gold standard diagnosis for every case. This digital case set of dysplastic BE cases will be made available in a teaching and testing environment to allow pathologists in- or outside the Netherlands to evaluate whether or not they meet the aforementioned benchmark quality criteria.

A limitation of our study is that the benchmark values for the chosen quality criteria generated in this study are only applicable to this particular case set, since they depend on this particular distribution of diagnoses. In addition, although we feel that our choice of criteria (how often indefinite, how confident, i.e. intra-observer agreement, and how accurate compared with a consensus diagnosis) is logical, some may argue that this choice is subjective. We feel that these benchmark quality criteria are currently the best to quantify expertise in diagnosing BE dysplasia in biopsy samples. In conclusion, our study shows that expert BE pathologists reach high levels of agreement when assessing a dysplastic, whole-endoscopy case set of BE cases. Their agreement scores have generated benchmark values for four quality criteria, namely: (1) the percentage of IND diagnoses; (2) the intra-observer agreement; (3) the percentage agreement compared with a consensus gold standard diagnosis; and (4) the percentage of cases of HGD misdiagnosed as NDBE. The values for these benchmark quality criteria set by our five core pathologists and digital dysplastic BE case set will be used to assess if other pathologists can join our national digital review panel.

Declaration of conflicting interests

None declared.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Ethics approval

Since the materials used in this study were anonymized, the medical ethical committee of the AMC waived the need for approval.

Informed consent

Since the materials used in this study were anonymized, no informed consent was obtained.

Supplementary materials

The research materials supporting this publication can be accessed through the supplementary materials and/or by contacting Myrtle J van der Wel at m.j.vanderwel@amc.uva.nl.

References

1. Schlemper RJ, Kato Y and Stolte M. Diagnostic criteria for gastrointestinal carcinomas in Japan and Western countries: proposal for a new classification system of gastrointestinal epithelial neoplasia. *J Gastroenterol Hepatol* 2000; 15(Suppl): G49–57.
2. 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: 7–42.
3. Shaheen NJ, Falk GW, Iyer PG, et al. ACG Clinical guideline: diagnosis and management of Barrett's esophagus. *Am J Gastroenterol* 2016; 111: 30–50.
4. American Gastroenterological Association; Spechler SJ, Sharma P, Souza RF, et al. American Gastroenterological Association medical position statement on the management of Barrett's esophagus. *Gastroenterology* 2011; 140: 1084–1091.
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. *J Gastroenterol Hepatol* 2015; 30: 804–820.
6. 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: 191–198.
7. 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: 1402–1415.
8. Curvers WL, ten Kate FJ, Krishnadath KK, et al. Low-grade dysplasia in Barrett's esophagus: overdiagnosed and underestimated. *Am J Gastroenterol* 2010; 105: 1523–1530.
9. 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: 700–706.
10. 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: 993–1001.e1.
11. Hulscher JB, Haringsma J, Benraadt J, et al. Comprehensive Cancer Centre Amsterdam Barrett

Advisory Committee: first results. *Neth J Med* 2001; 58: 3–8.

12. Offerhaus GJ, Correa P, van Eeden S, et al. Report of an Amsterdam working group on Barrett esophagus. *Virchows Arch* 2003; 443: 602–608.
13. 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. *Dis Esophagus* 2017; 30: 1–7.
14. 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: 822–835.
15. Polkowski W, Baak JP, van Lanschot JJ, et al. Clinical decision making in Barrett's oesophagus can be supported by computerized immunoquantitation and morphometry of features associated with proliferation and differentiation. *J Pathol* 1998; 184: 161–168.
16. Phoa KN, Pouw RE, Bisschops R, et al. Multimodality endoscopic eradication for neoplastic Barrett oesophagus: results of an European multicentre study (EURO-II). *Gut* 2016; 65: 555–562.
17. Phoa KN, Pouw RE, van Vilsteren FG, et al. Remission of Barrett's esophagus with early neoplasia 5 years after radiofrequency ablation with endoscopic resection: a Netherlands cohort study. *Gastroenterology* 2013; 145: 96–104.
18. 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* 2014; 311: 1209–1217.
19. van Vilsteren FG, Phoa KN, Alvarez Herrero L, et al. A simplified regimen for focal radiofrequency ablation of Barrett's mucosa: a randomized multicenter trial comparing two ablation regimens. *Gastrointest endoscopy* 2013; 78: 30–8.
20. van Sandick JW, Baak JP, van Lanschot JJ, et al. Computerized quantitative pathology for the grading of dysplasia in surveillance biopsies of Barrett's oesophagus. *J Pathol* 2000; 190: 177–83.
21. Curvers WL, van Vilsteren FG, Baak LC, et al. Endoscopic trimodal imaging versus standard video endoscopy for detection of early Barrett's neoplasia: a multicenter, randomized, crossover study in general practice. *Gastrointest Endosc* 2011; 73: 195–203.
22. van Sandick JW, van Lanschot JJ, Kuiken BW, et al. Impact of endoscopic biopsy surveillance of Barrett's oesophagus on pathological stage and clinical outcome of Barrett's carcinoma. *Gut* 1998; 43: 216–222.
23. Alvarez Herrero L, van Vilsteren FG, Pouw RE, et al. Endoscopic radiofrequency ablation combined with endoscopic resection for early neoplasia in Barrett's esophagus longer than 10 cm. *Gastrointest Endosc* 2011; 73: 682–690.
24. van Vilsteren FG, Pouw RE, Seewald S, et al. Stepwise radical endoscopic resection versus radiofrequency ablation for Barrett's oesophagus with high-grade dysplasia or early cancer: a multicentre randomised trial. *Gut* 2011; 60: 765–773.
25. Peters FP, Brakenhoff KP, Curvers WL, et al. Histologic evaluation of resection specimens obtained at 293 endoscopic resections in Barrett's esophagus. *Gastrointest Endosc* 2008; 67: 604–609.
26. Reid BJ, Haggitt RC, Rubin CE, et al. Observer variation in the diagnosis of dysplasia in Barrett's esophagus. *Hum Pathol* 1988; 19: 166–178.
27. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968; 70: 213–219.
28. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20: 37–45.
29. 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: 699–712.
30. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.