Histopathologist Features Predictive of Diagnostic Concordance at Expert Level Amongst a Large International Sample of Pathologists Diagnosing Barrett's Dysplasia Using Digital Pathology

Histopathologist Features Predictive of Diagnostic Concordance at Expert Level Amongst a Large International Sample of Pathologists Diagnosing Barrett's Dysplasia Using Digital Pathology

RUNNING HEAD: Quantitative model of Barrett's histopathology expert review

Authors: Myrtle J. van der Wel[1,2*], Helen G. Coleman[3*], Jacques JGHM Bergman[2], Marnix Jansen[4,5,§], Sybren L. Meijer[1,§], on behalf of the BOLERO working group[#]

Affiliations: [1]Amsterdam UMC, University of Amsterdam, Dept. of Pathology, Amsterdam, The Netherlands; [2]Amsterdam UMC, University of Amsterdam, Dept. of Gastroenterology and Hepatology, Amsterdam, The Netherlands; [3]Cancer Epidemiology Research Group, Centre for Public Health, Queen's University Belfast, Belfast, Northern Ireland, UK; [4]University College London Hospitals NHS Trust, Dept. of Pathology, London, UK; [5]UCL Cancer Institute, London, UK

[*] These first authors contributed equally
[§] Shared corresponding and senior authors
[#] List of working group members appears at the end of this manuscript

Keywords: Barrett's oesophagus; Oesophageal neoplasms; Digital pathology; Health Services Research.

Disclosures: No conflicts of interest to declare

Author contributions: Study concept and design (MW, MJ, SM); acquisition of data (MW, MJ, SM); analysis and interpretation of data (MW, HC); drafting of the manuscript (MW, HC, MJ, SM); critical revision of the manuscript (MW, HC, JB, MJ, SM); study supervision (SM, MJ).

Sources of funding: Cancer Research UK and Dutch Cancer Society

[§]Corresponding authors:
Marnix Jansen
UCL Cancer Institute, room 234D
72 Huntley Str, London WC1E 6AG, United Kingdom
Email: m.jansen@ucl.ac.uk

Sybren L. Meijer
Academic Medical Center
Meibergdreef 9, 1105 AZ Amsterdam, the Netherlands
Tel: +31 20 5665648; Fax: +31 20 6917033;
Email: s.l.meijer@amc.uva.nl

Word Count: 4795

LIST OF ABBREVIATIONS

BO; Barrett’s oesophagus
BMI; body mass index
CI; confidence interval
CRF; case record form
OAC; oesophageal adenocarcinoma
HGD; high-grade dysplasia
IHC; immunohistochemistry
IMC; intramucosal carcinoma
IND; indefinite for dysplasia
IQR; interquartile range
K; kappa value
LGD; low-grade dysplasia
NDBO; non-dysplastic Barrett’s oesophagus
OR; odd’s ratio
WSI; whole slide imaging

ABSTRACT

Objective: Guidelines mandate expert pathology review of Barrett's oesophagus (BO) biopsies that reveal dysplasia, but there are no evidence-based standards to corroborate expert reviewer status. We investigated BO concordance rates and pathologist features predictive of diagnostic discordance.

Design: Pathologists (n=51) from over 20 countries assessed 55 digitised BO biopsies from across the diagnostic spectrum, before and after viewing matched p53 labelling. Extensive demographic and clinical experience data were obtained via online questionnaire. Reference diagnoses were obtained from a review panel (n=4) of experienced Barrett's pathologists.

Results: We recorded over 6,000 case diagnoses with matched demographic data. Of 2,805 H&E diagnoses, we found excellent concordance (>70%) for non-dysplastic BO and high-grade dysplasia, and intermediate concordance for low-grade dysplasia (42%) and indefinite for dysplasia (23%). Major diagnostic errors were found in 248 diagnoses (8.8%), which reduced to 232 (8.3%) after viewing p53 labelled slides. Demographic variables correlating with diagnostic proficiency were analysed in multivariate analysis, which revealed that at least 5 years of professional experience was protective against major diagnostic error for H&E slide review (OR 0.48, 95%CI 0.31-0.74). Working in a non-teaching hospital was associated with increased odds of major diagnostic error (OR 1.76, 95%CI 1.15-2.69), however this was neutralised when pathologists viewed p53 labelled slides. Notably, neither case volume nor self-identifying as an expert predicted diagnostic proficiency. Extrapolating our data to real-world case prevalence suggests that 92.3% of major diagnostic error is due to overinterpreting non-dysplastic Barrett's oesophagus.

Conclusion: Our data provide evidence-based criteria for diagnostic proficiency in Barrett's histopathology.

What is already known about this subject?

Pathology evaluation of Barrett's patients' surveillance biopsies is poorly reproducible. Guidelines mandate that biopsies with dysplasia be reviewed by an expert, but there are no evidence-based criteria to corroborate expert reviewer status.

What are the new findings?

Through a large online consensus study amongst more than 50 pathologists in over 20 countries we reveal histopathologist-dependent predictors of major diagnostic error in the assessment of Barrett's biopsies. The size of our dataset allows us to quantify the impact of these variables, such as experience commensurate with age and professional setting, in multivariate analysis.

How might it impact on clinical practice in the foreseeable future?

Our data provide evidence-based criteria for diagnostic proficiency in Barrett's histopathology and will help facilitate training and support to reduce diagnostic variability.

INTRODUCTION

Barrett's oesophagus (BO) is a premalignant condition, which predisposes to oesophageal adenocarcinoma (OAC), with a reported annual conversion rate of 0.1 - 0.2%. [1-3] BO is defined histopathologically as the replacement of normal stratified squamous epithelial lining of the distal oesophagus with columnar epithelium that can contain intestinal metaplasia. The implementation of formal surveillance strategies and widespread adoption of endoscopic treatment techniques, such as endoscopic resection and ablation for dysplastic BO, have led to a surge in diagnostic pathology workload. The goal of endoscopic surveillance and biopsy verification is objective risk stratification for patients according to their perceived progression risk to OAC.

Previous studies have revealed, however, that diagnostic reproducibility (inter-observer agreement) amongst pathologists grading dysplastic BO biopsy material is moderate to poor, even amongst expert reviewers (**Supplementary Table 1**). [4-17] Previous work from our group has shown that central pathology review by a dedicated panel within the context of prospective intervention trials failed to confirm an initial diagnosis of low-grade dysplasia (LGD) in over three-quarters of cases submitted for panel review. On follow up, cases that had been downgraded to non-dysplastic BO (NDBO) revealed a nominal progression risk of about 0.5% per patient/year, whilst cases that had been confirmed LGD on central review showed a progression risk of about 10% per patient/year. These data clearly attest to the clinical return of dedicated pathology review. [18 19] International BO management guidelines now mandate histopathology review of all BO biopsy cases found to reveal dysplasia by an independent expert pathologist. [20 21] However, whilst major society guidelines have qualitatively defined an expert BO pathologist as 'a pathologist with a special interest in BO-related neoplasia who is recognised as an expert in this field by their peers', we lack firm evidence-based standards to corroborate expert reviewer status. [21-26] This now represents an acute unmet need as these considerations also carry important medico-legal implications.

Recently, the US Food and Drug Administration has approved the use of whole slide imaging (WSI) for primary diagnostic use. [27] The advantages of WSI are numerous and include simultaneous assessment by multiple pathologists, streamlined expert consultation, and digital image analysis. It is expected that digital pathology will rapidly gain widespread acceptance in the coming years, in particular in the context of distant case review. A number of large-scale diagnostic consensus studies have been performed, which have broadly suggested that the diagnostic discordance rate between pathologists using digital slide review is non-inferior to conventional glass slide diagnosis. [28-30] However, these studies generally examined a large number of diagnostic categories without focusing on a particular category of known diagnostic discordance such as Barrett's dysplasia. Establishing the validity of this new technology to BO histopathologic workup is therefore a clear priority.

Here we set out to develop quantitative standards of expert reviewer status for guideline development purposes using massive online digital pathology reporting. We define expert reviewer status as evidence of diagnostic concordance on a par with consensus within an expert review panel, acknowledging that, in lieu of an objective biomarker of progression risk, there will be diagnostic variation amongst expert pathologists. We collected extensive demographic information of participating pathologists to understand operator-dependent predictors of diagnostic variation.

METHODS

Ethical considerations

This study utilised anonymised archived formalin-fixed, paraffin embedded material and did not require approval from the relevant Institutional Ethics Committee under applicable local regulatory law ('Code of conduct', FEDERA).

Assessors

Sixty-five gastrointestinal pathologists worldwide were approached to join this study through either professional gastrointestinal pathology working groups or direct professional contacts. Fifty-nine pathologists responded positively to our enquiries and were recruited to this study of which 51 pathologists completed the entire case set of 55 H&E-stained and 55 matching p53 immunohistochemistry (IHC) labelled slides (110 slides total). These 51 pathologists are henceforth referred to as participating pathologists. Participating pathologists received emails detailing the study objectives and were provided with personal log-in credentials to the purpose-built online scoring environment described below. Lead study author (MvdW) provided assistance with participating pathologists' log-in queries, evaluated study progress, and chaired the panel consensus meeting.

Four BO pathologists (including two study authors, MJ and SM) with extensive experience in BO dysplasia assessment reviewed all slides as a reference pathologist panel. This group has successfully collaborated on previous BO intervention studies where patient outcome has been evaluated prospectively [18 19 31-37] as well as on the Amsterdam Barrett's Advisory Committee. [31] These four pathologists are henceforth referred to as reference pathologists.

Slide selection and scanning

The lead study author selected a representative case-mix of 55 BO biopsy cases from across the diagnostic spectrum (**Supplementary Table 2**). Inclusion criteria were: diagnosis confirmed by a second gastrointestinal pathologist; documented clinical follow-up of at least one year available; and tissue block available. All cases were treatment-naïve. Per case, immunohistochemical staining for p53 was performed using a Ventana Benchmark XT autostainer (Ventana Medical Systems, Tucson, AZ). Antigen retrieval was performed with CC1 mild. P53 was detected with p53 Antibody (Mouse DO-7 + BP 53-12, Thermo Scientific) and the sections were incubated in a 1:500 dilution for 32 min at room temperature. Bound antibody was detected using the Biotin free Ultraview Universal DAB Detection Kit (Roche Diagnostics) and slides were counterstained with Hematoxylin (Roche Diagnostics). [38] One H&E slide and one consecutive section p53 labelled slide were digitised from each case using a scanner with a 20x microscope objective (Slide, Olympus, Tokyo, Japan). Scans were checked for focus and acuity by the study coordinator and re-scanned if necessary. Subsequently, slides were anonymised, randomised, renamed, and stored on a secure server. The 'Digital Slidebox 4.5' (https://dsb.amc.nl/dsb/login.php, Slidepath, Leica Microsystems, Dublin, Ireland) virtual slide viewing software was used to evaluate the digital slides during the study. EMR specimens were not included in our study cohort.

Electronic scoring environment

Template electronic Case Record Forms (CRFs) were custom built within a web-based software tool designed to capture clinical study data (OpenClinica v3.6, an open source CTMM

TraiT project, LLC, Waltham, USA). One CRF consists of an extensive questionnaire documenting pathologist characteristics such as age, sex, host institution, and experience in reporting BO biopsies and digital pathology (full questionnaire details in **Supplementary Table 3**). The second CRF was built to record individual case diagnoses. Importantly, this second CRF consists of separate parts to record H&E and H&E plus p53 labelled slide diagnoses independently. The first part of the case diagnosis CRF contains a dynamic URL link to the scanned H&E slide and includes questions about the slide quality and diagnosis, and whether the assessor would require a p53 labelled slide. Importantly, the second part of the templated CRF that contains a dynamic link to the p53 labelled slide alongside the matching H&E slide, only opens after the study pathologist has completed assessment of the H&E-stained slide and saved their case diagnosis for this slide. This second part of the templated CRF, in addition to a dynamic link to the matching p53 labelled slide, again included corresponding slide assessment questions.

Digital case assessments

Reference and participating pathologists were asked to assess each case, according to the modified Vienna classification for gastrointestinal neoplasia. [39 40] Reference pathologists first assessed all cases individually and completed the questionnaire. An online consensus meeting was then convened after a two-month wash out period to discuss discrepancies and produce reference diagnoses for each of the 110 assessments (55 H&E-stained slides and 55 matching p53 labelled slides). The panel assessment was taken forward as the reference diagnosis without further discussion if reference panel members achieved a majority diagnosis (i.e. concordance between either 3 out of 4 or 4 out of 4 pathologists) on a case directly from their independent scoring. Group discussions were held between these four pathologists to review and discuss cases for which there was no majority diagnosis to mimic real-world practice. The discrepancies where a majority diagnosis had not been reached after individual slide review encompassed 21 cases based on H&E slide viewing, and 13 cases based on the p53 labelled slide. These cases were reviewed during the panel discussion (21 H&E slides reviewed without matching p53 labelled slide, and 13 cases with H&E-stained slide and matching p53 labelled slides) to arrive at a consensus diagnosis for all 110 assessments.

From the case assessments by the participating pathologists, two post-p53 labelled case assessments were inadvertently left blank by individual participating pathologists (one each) after evaluating the case H&E slide. Results from the matching H&E slides were imputed as post-p53 case diagnosis in these cases, based on the H&E slide score, corresponding to 2 HGD diagnoses.

Population estimates

To extrapolate our findings to the proportional prevalence of Barrett's dysplasia in real-world practice, we used incident and surveillance reports from the population-based Northern Ireland Barrett's oesophagus register, methods of which have been described elsewhere. [41 42] The prevalence for the most recently available data in 2014 were applied, in which n=2,872 patients received a pathology diagnosis of NDBE (n=2,627, 91.5%), IND (n=36, 1.2%), LGD (n=85, 3%) or HGD (n=124, 4.3%). These values were then used to estimate the population impact of interpretation discordance for each diagnostic category.

Statistical analysis

Characteristics of the four reference pathologists and the 51 participating pathologists were compared informally. We examined the overall concordance of the study pathologists compared to the consensus reference diagnosis per case. This process was conducted for each of the four individual members of the reference panel against the final consensus diagnosis of this panel, as well as for the overall sample of 51 pathologists against the consensus diagnosis. Per pathologist scores were not calculated, since we aimed to study the cohort behavior rather than the individual pathologist. Concordance was initially compared based on four relevant diagnostic categories (NDBO, IND, LGD, HGD), and then compared based on three relevant diagnostic categories (NDBO, IND, LGD or HGD) to reflect the fact that HGD and LGD are now treated endoscopically in some settings. [32] We calculated 95% CIs for overall concordance and per diagnostic category. Since this cohort was strongly enriched for dysplasia, we did not use kappa statistics, since these are less reliable when cross tables are skewed.

To evaluate the potential clinical impact of discordant interpretations across the cohort of participating pathologists, we then reclassified all discordant assessments as either major or minor discordances. Major overinterpretation is defined as NDBO reference diagnosis overinterpreted as either LGD or HGD, whereas, vice versa, major underinterpretation is LGD or HGD reference diagnosis underinterpreted as NDBO by the participating pathologist. These discordant interpretations would bear major consequences in clinical practice. All other discordant interpretations were classified as minor discordant interpretations. A tabular overview of interpretation classifications as major or minor is shown in **Supplementary Table 4**. Since both major overinterpretation and major underinterpretation can have negative implications for patient management, these were further combined for the purposes of some analyses, as indicated.

Unadjusted logistic regression analyses were then conducted to identify any pathologist characteristics that were associated with overall and major over or underinterpretation of BO cases compared to the consensus diagnosis. Considering that age and professional experience are inextricably linked, we evaluated individual combinations of age and experience for odds of major over and underinterpretations, and combined these into three categories in whom similar odds ratios were observed (**Supplementary Table 5**). Forward selection of significant factors was used to create multivariable-adjusted logistic regression models of characteristics associated with misinterpretation. Although routine use of p53 immunohistochemistry was not associated with diagnostic errors, this was retained in multivariate models for p53 stained slides. All statistical analyses were performed using Stata version 14.2 (StataCorp., College Station, TX, USA).

RESULTS

Study design

This study is based on assessments of digitised slides to investigate diagnostic concordance of BO biopsies amongst a large and heterogeneous sample of gastrointestinal pathologists. We investigated rates and features predictive of diagnostic concordance amongst these pathologists, with a particular focus on the demographic characteristics of the pathologists, the impact of viewing p53 labelled slides alongside H&E-stained slides, and on features associated with major diagnostic discordance that would negatively impact upon patient stratification and treatment pathways. The purpose of this study was to build a quantitative model of expert BO pathologist review characteristics, and to provide practical recommendations that could minimize errors in the interpretation of BO biopsies in the routine setting.

The study flowchart is shown in **Figure 1A**. All pathologists first filled out a baseline questionnaire for detailed demographic and clinical experience data. Pathologists then assessed the 110 digitised slides (55 H&E slides and matching p53 labelled slides) and recorded their answers on dedicated electronic CRFs. As detailed in the methods section, diagnostic entries were recorded after viewing the H&E-stained slide and again after the matched p53 labelled slide was revealed alongside the case H&E slide.

The entire study set was completed by fifty-five pathologists working in over 20 countries and 5 continents (**Figure 1B**). Of these fifty-five pathologists, 4 pathologists with extensive and published experience in BO dysplasia assessment were designated beforehand as reference pathologists. [18 19 32 43 44] In sum, with 55 pathologists reviewing 55 biopsy cases, each of which includes one H&E-stained slide and a matched p53 labelled slide, this generated a massive dataset of over 6,000 case diagnoses with matched demographic data as input data for our Barrett's digital pathology (BOLERO) consensus study, one of the largest digital pathology consensus studies reported thus far. Case diagnoses were compared to reference diagnoses and we searched for pathologist demographic features that predict diagnostic consensus at expert level.

Patient characteristics of BO biopsy samples

Patient characteristics of the sample biopsies are shown in **Supplementary Table 2**. Of these patients, 94.5% was male (52/55). The median age at diagnosis was 65, the median BMI was 27, the median BO segment length was Circumferential (C) 4 cm, Maximum (M) 5 cm. Patients had a history of smoking in 63.6% of cases (35/55), a history of heartburn symptoms in 89% of cases (49/55), and used anti-reflux medication in 96.4% of cases (53/55).

Pathologist characteristics

Baseline characteristics of the pathologists taking part in the study are displayed in **Table 1** and **Supplementary table 6**. Participating pathologists represented a heterogeneous sample comprising a wide range of ages, workplace settings (academic teaching, private and/or district general hospital settings) and years of professional experience. Just over 50% of participating pathologists reported dedicated fellowship experience, whilst the majority (72%) worked in a large laboratory with ≥10 pathologist colleagues. The most commonly reported guidelines to which pathologists adhered were North American, British, or Japanese, however a quarter of pathologists reported using other guidelines in their clinical practice. Two thirds of participating pathologists self-identified as expert gastrointestinal pathologists. Note that although pathologists were approached through professional societies, no effort was made to purposely recruit experts onto the study. Pathologists also reported on other parameters and working practices in their laboratories, such as typical numbers of BO cases reported per week, confidence and enjoyment in reporting BO, reporting of endoscopic resection specimens, frequency of adjunct p53 labelled slide use in BO reporting, participation in double-reporting, multi-disciplinary team meetings, and use of WSI, as well as typical interactions and perceptions of practices of their endoscopy colleagues (**Table 1 and Supplementary table 6**). Participating and reference pathologists were generally well matched for age ranges and professional experience although all four reference pathologists were male, whereas 22 of 51 (43.1%) participating pathologists in the larger cohort were female.

Case assessment overview

A total of 3,025 diagnoses were generated based on H&E-stained slide case review and another 3,025 diagnoses were recorded after viewing the matching p53 labelled slides for study cases (**Figure 2A and B**). The corresponding waterfall plots showing the ranked distribution of assessments reveal a gradual transition from NDBO examples with high interobserver concordance to HGD cases with similarly high interobserver concordance and diagnostic categories where concordance gradually transitions between these extremes. These plots also confirm that our case set includes representative biopsies from across the diagnostic spectrum of BO pathology. Relevant examples of study cases are shown in **Figure 2C**.

Concordance of reference pathologists vs. consensus diagnosis on H&E and p53 labelled slides

Consensus diagnoses were generated following panel review. The reference panel consensus diagnoses for the H&E-stained slide case review included 16 NDBO, 6 IND, 18 LGD, and 15 HGD case diagnoses. After the addition of matched p53 labelled slides and reference panel review a small number of cases were reclassified, including 1 NDBO diagnosis as LGD, 1 LGD diagnosis as NDBO, and 4 IND diagnoses as LGD, thus totaling 16 NDBO, 2 IND, 22 LGD and 15 HGD after p53 labelled slide review.

Individual consensus panel member diagnoses were then compared to the final consensus panel diagnosis to obtain concordance rates between the 4 reference pathologists. This revealed excellent diagnostic agreement when reporting NDBO, LGD and HGD on H&E-stained slides alone (84.4%, 65.3% and 78.3%, respectively), rising to 89.4% when LGD and HGD diagnoses were combined. After revealing the matching p53 labelled slide for the 55 cases, agreement further improved to 85.9% for ND, 72.7% for LGD, and 76.7% for HGD, rising to 91.9% when LGD and HGD were combined (**Supplementary Tables 7A and B**).

Concordance of participating pathologists vs. consensus diagnosis on H&E and p53 stained slides

The complete set of 5,610 case assessments recorded by the 51 participating pathologists was then compared to the reference panel diagnoses to obtain concordance rates and compare diagnostic agreement within and between categories. The diagnostic agreement between 51 participating pathologists for H&E-stained slide diagnoses is depicted in **Figure 3A-C** and **Supplementary Figure 1A**, while concordance percentages are shown in **Table 2A**. We found excellent concordance between the participating pathologists for NDBO reference diagnosis cases (643 of 816 diagnoses; 78.8%) and HGD reference diagnosis cases (544 of 765 diagnoses; 71.1%). As expected, there was moderate concordance for LGD reference diagnosis cases (382 of 918; 41.6%) and poor concordance for IND reference diagnosis cases (70 of 306; 22.9%). However, if dysplastic assessments were grouped (i.e. combining LGD and HGD reference diagnosis cases) then 77.5% (1,305 of 1,683) of cases were concordant. Major over or underinterpretation was found in 8.8% of assessments (248 of 2,805 diagnoses).

Addition of matched p53 labelled slides improved diagnostic concordance (**Figure 3D-F** and **Supplementary Figure 1B)** with small but clinically meaningful improvements seen in the diagnostic concordance between participating pathologists for NDBO reference diagnosis cases (83.8% v. 78.8% on H&E slide) and LGD/HGD combined reference diagnosis cases (79.3% v. 77.5% on H&E slide), **Table 2B**. In addition to this, p53 labelled slides also had a small but beneficial impact on reducing the number of major over and underinterpretations (8.3%, 232 of 2,805 diagnoses), representing 0.5% fewer overall major misinterpretations compared to H&E-stained slide diagnosis alone.

Characteristics associated with concordance on H&E slides

This massive dataset was then interrogated to reveal histopathologist predictors of over or underreporting and major diagnostic errors in univariate analysis. To this end all diagnostic discordances within our dataset (i.e. case diagnoses not matching reference diagnosis) were first reclassified as major or minor over or underinterpretation (see Methods and **Supplementary Table 4**). Factors associated with reduced odds of major diagnostic errors included: ≥5 years of experience commensurate with age (OR 0.65, 95%CI 0.45-0.93); working in an academic teaching hospital (OR 0.59, 95%CI 0.43-0.81); routinely double reporting indefinite for dysplasia cases (OR 0.70, 95%CI 0.52-0.94); working in a larger lab (≥10 versus <10 pathologists OR 0.72, 95%CI 0.54-0.96) and using digital pathology (OR 0.63; 95%CI 0.47-0.89). In contrast, working within a district general hospital (OR 1.72, 95%CI 1.30-2.26) or private hospital (OR 1.41, 95%CI 1.04-1.91), or not using major society guidelines (OR 1.43, 95%CI 1.06-1.94) were all associated with increased odds of major diagnostic errors **(Supplementary Tables 8A-C**).

Several factors were not associated with major diagnostic error, including pathologist sex. Participating in upper gastrointestinal multidisciplinary team meetings was not associated with reduced odds of major diagnostic error, although it was associated with reduced odds of overreporting. Notably, self-identifying as a Barrett's pathology expert, holding a dedicated fellowship, or reporting greater enjoyment or confidence in Barrett's reporting were not associated with decreased odds of major over or underinterpretation (**Supplementary Table 8A**). Finally, reporting ≥20 cases per week was associated with reduced odds of over or under-interpretation of Barrett's dysplasia (OR 0.69, 95%CI 0.53-0.89), although this association was attenuated when investigating major diagnostic errors (**Supplementary Table 8B**).

Multivariate analyses before and after revealing matched p53 labelled slides

Multivariable models were then applied, including all factors associated with collective over and underinterpretation on H&E digital slide review in univariate analysis, as shown in **Figure 4**. At least 5 years of experience commensurate with age was the strongest protective factor against major diagnostic error on H&E slide review (OR 0.48, 95%CI 0.31-0.74). In contrast, working in a district general hospital was associated with increased odds of major diagnostic error (OR 1.76, 95%CI 1.15-2.69). Importantly, this effect was neutralised if pathologists in these settings viewed cases with additional p53 labelled slides (OR 1.44, 95%CI 0.92-2.28). As expected, routine use of p53 labelled slides was associated with reduced odds of major diagnostic error. Viewing 5-19 BO cases with p53 stained slides per week was associated with increased odds of major diagnostic errors, which was neutralised when viewing ≥20 cases per week. Most other results showed similar trends to those seen in univariate analysis, but these were no longer statistically significant (**Figure 4**).

Population estimates

To determine the impact of our results in a real-world clinical setting, we extrapolated the results from this case set (in which dysplastic biopsies were purposely over-represented) to the Barrett's dysplasia prevalence reported from the population-based Northern Ireland

Barrett's oesophagus register. As shown in **Figure 5**, 18.6% of all Barrett's cases would be classified as having a major over- or under-interpretation, based on the findings of this study as applied to the real word clinical setting of H&E slide plus adjunct p53 labelled slide viewing. The majority of these would be attributed to potential overinterpretation of NDBO (426 out of 461 cases, or 92.3%, **Figure 5**).

DISCUSSION

We have carried out the largest investigation of diagnostic concordance of BO biopsy reporting amongst gastrointestinal pathologists to date. Previous studies had been limited to a small number of expert pathologists, which meant findings were not necessarily generalizable to real-world settings. This work has revealed several novel findings.

First, overall concordance for H&E digital slide review of NDBO and LGD/HGD as a combined outcome was excellent (exceeding 77%), although concordance for IND and LGD as a stand-alone diagnosis was lower (23-42%). These test characteristics replicate known glass slide test characteristics (**Supplementary Table 1**), suggesting that distant BO biopsy slide review is reproducible and safe.

Second, our multivariate analyses revealed several pathologist characteristics and working practices independently associated with the risk of misinterpretations. Reassuringly, pathologist experience commensurate with age was most protective against major over- or underinterpretation, confirming the validity of our experimental strategy. Our multivariate regression analyses also confirm that working within a teaching hospital environment protects against major diagnostic error. This provides supportive evidence for guideline statements that BO complicated by dysplasia is best managed within an expert center. [21-23 26] Importantly, self-identifying as an expert was not associated with decreased odds of major over or underinterpretation.

Lastly, our study design sheds light on the context-dependent impact of p53 labelled slides. We find that the overall prevalence of major misinterpretations (NDBO classified as LGD/HGD, or vice versa) across this biopsy series enriched for IND/LGD/HGD cases was 8.8%, which was reduced, marginally, by the addition of p53 labelled slides (8.3%). Although this would suggest a limited impact of the adjunct use of p53 labelled slides, our multivariate analysis allows us to unpack this figure and reveals that major discordance was reduced by viewing matched p53 labelled slides specifically for those pathologists working away from teaching hospital settings. This demonstrates that the beneficial impact of adjunct p53 labelled slides is dependent on context and is greatest outside expert centre settings where, indeed, most primary dysplasia diagnoses in surveillance are made. Extrapolating our concordance data to real-world dysplasia prevalence shows that the majority of major misdiagnoses in real world practice overinterpret NDBO (426 out of 461 cases, or 92.3%, **Figure 5**). In these cases, routine addition of adjunct p53 labelled slides may have substantial impact towards limiting overdiagnosis, although our study was not designed to examine the latter point. Routine use of p53 labelled slides is supported by several national guidelines, [21 23 26] and our study confirms that this is appropriate.

Taken together, our study for the first time provides an evidence-based quantitative model of BO histopathology diagnosis at expert consensus level. Our data reassuringly suggest that BO reporting on a par with expert consensus is not limited to a small league of experienced histopathologists but can be predicted from a small number of intuitive demographic predictors (experience, professional setting, use of p53 labelled slides). This suggests practical interventions to reduce diagnostic variability are feasible, through improved training and support. To implement routine external review of dysplastic BO biopsies, as mandated by several major society guidelines, requires regional or national teams of dedicated gastrointestinal pathologists with Barrett's expertise. Combined with our observation that concordance rates for digital slide viewing were not inferior to conventional glass slide pathology review, [18 19] together these data suggest that distant digital review of challenging BO biopsy cases is safe to formally implement within current care delivery systems, provided quality benchmarks are met. In the Netherlands, such a set-up has been successfully implemented over the past five years, to accommodate nationwide digital expert review of all dysplastic BO biopsies. [44 45]

Our study has considerable strengths compared to previous interobserver variation studies of BO reporting. We have evaluated diagnostic concordance for dysplastic BO amongst the largest group of gastrointestinal pathologists worldwide. The heterogeneous mix of pathologists involved in this study also enabled novel investigations into pathologist-dependent predictors associated with diagnostic discordance. The online reporting strategy mimicked routine workflow and facilitated data collection and curation in a flexible manner. The case set was purposely enriched for dysplastic cases in order to attain sufficient statistical power in our downstream regression analyses. Diagnostic concordance within a large group of pathologists with different levels of gastrointestinal pathology expertise was excellent for LGD and HGD combined.

This study also has limitations that are important to note. One caveat to our study design is the original dataset which is skewed towards the inclusion of dysplastic biopsies. Our case-mix therefore does not represent a cross-section of diagnostic biopsy cases encountered in daily practice, which would be heavily weighted towards the NDBO end of the spectrum. Because a complete revision study whereby all consecutive surveillance biopsies are prospectively reviewed by a consensus panel of experienced pathologists is not practically feasible, we set out to extrapolate the population impact of histopathologist diagnostic variation from our dataset. To this end, we exploited the dysplasia population prevalence from the Northern Ireland Barrett's register (see Methods) and modelled the impact of diagnostic variation using our concordance data (**Figure 5**). We found that, across all diagnostic categories, 81.4% of all diagnoses would be confirmed by consensus of four experienced Barrett's pathologists. Given the fact that the overbearing majority of Barrett's surveillance biopsies were reported to contain non-dysplastic Barrett's mucosa, proportionally the largest share of diagnostic discordance is seen in this category (92.3%). Vice versa a small number of biopsies in routine practice (estimated at 1.3% of total) will initially be reported as non-dysplastic Barrett's mucosa, whereas consensus panel review would reveal high-grade dysplasia. These data suggest that the population impact of diagnostic variation is real and is most prominent for non-dysplastic Barrett's biopsies that are overinterpreted, which may lead to overtreatment. A small number of patients would be undertreated despite the presence of abnormalities that mandate invasive management.

A second limitation is that while our heterogeneous global group of pathologists allowed us to interrogate associations of a host of operator-dependent characteristics with diagnostic consensus (case volume, practice setting, diagnostic experience, etc.), this study feature may limit the generalizability of our findings within the national setting. Replication of our findings in samples of pathologists within particular geographic regions adhering to one diagnostic guideline will be required to determine whether the quantitative predictive features described here are similarly applicable in that setting. Given that the majority of pathologists participating in this study were based either in Europe or North America, greater representation from low to middle income settings would be particularly welcome. This could further enhance the value of this recursive exercise for teaching and registration purposes.

In conclusion, using this rich dataset of case assessments by a large, heterogeneous sample of gastrointestinal pathologists, we have evaluated diagnostic concordance for BO diagnosis using digital case review. Our results reveal quantitative predictors of diagnostic performance that will aid formulation of quality assurance criteria for guideline development and standard implementation of digital pathology in BO biopsy review.

REFERENCES

1. 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
2. Desai TK, Krishnan K, Samala N, et al. The incidence of oesophageal adenocarcinoma in non-dysplastic Barrett's oesophagus: a meta-analysis. *Gut* 2012;61(7):970-6. doi: 10.1136/gutjnl-2011-300730 [published Online First: 2011/10/15]
3. Singh S, Manickam P, Amin AV, et al. Incidence of esophageal adenocarcinoma in Barrett's esophagus with low-grade dysplasia: a systematic review and meta-analysis. *Gastrointestinal endoscopy* 2014;79(6):897-909 e4; quiz 83 e1, 83 e3. doi: 10.1016/j.gie.2014.01.009
4. Coco DP, Goldblum JR, Hornick JL, et al. Interobserver variability in the diagnosis of crypt dysplasia in Barrett esophagus. *The American journal of surgical pathology* 2011;35(1):45-54. doi: 10.1097/PAS.0b013e3181ffdd14
5. Horvath B, Singh P, Xie H, et al. Risk for esophageal neoplasia in Barrett's esophagus patients with mucosal changes indefinite for dysplasia. *Journal of gastroenterology and hepatology* 2015;30(2):262-7. doi: 10.1111/jgh.12696
6. 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
7. 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
8. Kerkhof M, van Dekken H, Steyerberg EW, et al. Grading of dysplasia in Barrett's oesophagus: substantial interobserver variation between general and gastrointestinal pathologists. *Histopathology* 2007;50(7):920-7. doi: 10.1111/j.1365-2559.2007.02706.x
9. Lim CH, Treanor D, Dixon MF, et al. Low-grade dysplasia in Barrett's esophagus has a high risk of progression. *Endoscopy* 2007;39(7):581-7. doi: 10.1055/s-2007-966592

10. Montgomery E, Bronner MP, Goldblum JR, et al. Reproducibility of the diagnosis of dysplasia in Barrett esophagus: a reaffirmation. *Human pathology* 2001;32(4):368-78. doi: 10.1053/hupa.2001.23510
11. Pech O, Vieth M, Schmitz D, et al. Conclusions from the histological diagnosis of low-grade intraepithelial neoplasia in Barrett's oesophagus. *Scandinavian journal of gastroenterology* 2007;42(6):682-8. doi: 10.1080/00365520601075803
12. Sanders DS, Grabsch H, Harrison R, et al. Comparing virtual with conventional microscopy for the consensus diagnosis of Barrett's neoplasia in the AspECT Barrett's chemoprevention trial pathology audit. *Histopathology* 2012;61(5):795-800. doi: 10.1111/j.1365-2559.2012.04288.x
13. Sangle NA, Taylor SL, Emond MJ, et al. Overdiagnosis of high-grade dysplasia in Barrett's esophagus: a multicenter, international study. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 2015;28(6):758-65. doi: 10.1038/modpathol.2015.2
14. Skacel M, Petras RE, Gramlich TL, et al. The diagnosis of low-grade dysplasia in Barrett's esophagus and its implications for disease progression. *The American journal of gastroenterology* 2000;95(12):3383-7. doi: 10.1111/j.1572-0241.2000.03348.x
15. 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
16. Sonwalkar SA, Rotimi O, Scott N, et al. A study of indefinite for dysplasia in Barrett's oesophagus: reproducibility of diagnosis, clinical outcomes and predicting progression with AMACR (alpha-methylacyl-CoA-racemase). *Histopathology* 2010;56(7):900-7. doi: 10.1111/j.1365-2559.2010.03571.x
17. 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
18. 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
19. 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
20. Sharma P, Katzka DA, Gupta N, et al. Quality indicators for the management of Barrett's esophagus, dysplasia, and esophageal adenocarcinoma: international consensus recommendations from the American Gastroenterological Association Symposium. *Gastroenterology* 2015;149(6):1599-606. doi: 10.1053/j.gastro.2015.08.007 [published Online First: 2015/08/25]
21. 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
22. 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
23. 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
24. 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

25. 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
26. 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
27. Abels E, Pantanowitz L. Current State of the Regulatory Trajectory for Whole Slide Imaging Devices in the USA. *J Pathol Inform* 2017;8:23. doi: 10.4103/jpi.jpi_11_17
28. Snead DR, Tsang YW, Meskiri A, et al. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology* 2016;68(7):1063-72. doi: 10.1111/his.12879 [published Online First: 2015/09/27]
29. Goacher E, Randell R, Williams B, et al. The Diagnostic Concordance of Whole Slide Imaging and Light Microscopy: A Systematic Review. *Arch Pathol Lab Med* 2017;141(1):151-61. doi: 10.5858/arpa.2016-0025-RA [published Online First: 2016/07/12]
30. Mukhopadhyay S, Feldman MD, Abels E, et al. Whole Slide Imaging Versus Microscopy for Primary Diagnosis in Surgical Pathology: A Multicenter Blinded Randomized Noninferiority Study of 1992 Cases (Pivotal Study). *The American journal of surgical pathology* 2018;42(1):39-52. doi: 10.1097/PAS.0000000000000948 [published Online First: 2017/09/30]
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. 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
33. 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(2):216-22.
34. Phoa KN, Pouw RE, Bisschops R, et al. Multimodality endoscopic eradication for neoplastic Barrett oesophagus: results of an European multicentre study (EURO-II). *Gut* 2015 doi: 10.1136/gutjnl-2015-309298
35. 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(6):765-73. doi: 10.1136/gut.2010.229310
36. Pouw RE, Gondrie JJ, Sondermeijer CM, et al. Eradication of Barrett esophagus with early neoplasia by radiofrequency ablation, with or without endoscopic resection. *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract* 2008;12(10):1627-36; discussion 36-7. doi: 10.1007/s11605-008-0629-1
37. Duits LC, van der Wel MJ, Cotton CC, et al. Barrett's esophagus patients with confirmed and persistent low-grade dysplasia are at increased risk of neoplastic progression. *Gastroenterology* 2017 doi: 10.1053/j.gastro.2016.12.008 [published Online First: 21 Dec 2016]
38. 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]

39. 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.
40. Reid BJ, Haggitt RC, Rubin CE, et al. Observer variation in the diagnosis of dysplasia in Barrett's esophagus. *Human pathology* 1988;19(2):166-78.
41. Bhat S, Coleman HG, Yousef F, et al. Risk of malignant progression in Barrett's esophagus patients: results from a large population-based study. *Journal of the National Cancer Institute* 2011;103(13):1049-57. doi: 10.1093/jnci/djr203
42. Coleman HG, Bhat S, Murray LJ, et al. Increasing incidence of Barrett's oesophagus: a population-based study. *Eur J Epidemiol* 2011;26(9):739-45. doi: 10.1007/s10654-011-9596-z [published Online First: 2011/06/15]
43. 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
44. 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]
45. 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 : official journal of the International Society for Diseases of the Esophagus / ISDE* 2017;30(11):1-7. doi: 10.1093/dote/dox078

ACKNOWLEDGEMENTS

We sincerely thank Joann Elmore and Gary Longdon for their helpful additions to our study protocol. We sincerely thank Alden van Putten, Rudy Scholten, Rene Breet and David de Koning (Open Clinica) for their help in building and maintaining the online study environment. We sincerely thank Onno de Boer, Eelco Roos, and Wim van Est for their help with scanning of the slides.

# BOLERO STUDY PARTICIPANTS (in alphabetical order)

Dr. Junko Aida, Tokyo Metropolitan Institute of Gerontology, Tokyo, Japan
Dr. Rossana Baiocco, General Hospital of Desenzano del Garda, Desenzano, Italy
Dr. Camille Boulagnon-Rombi, Université de Reims Champagne-Ardenne, Reims, France
Dr. Iva Brcic, Medical University of Graz, Graz, Austria
Dr. Lodewijk Brosens, University Medical Center Utrecht, Utrecht, the Netherlands
Dr. Fátima Carneiro, IPATIMUP, Porto, Portugal
Dr. Gieri Cathomas, Kantosspital Baselland, Liestal, Switzerland
Dr. Denis Chatelain, CHU Amiens-Picardie, Amiens, France
Dr. Allison Cluroe, Addenbrookes Hospital, Cambridge, United Kingdom
Dr. Parag Dabir, Regional Hospital, Randers, Denmark
Dr. Giovanni De Petris, Penrose Hospital, Colorado Springs, United States of America
Dr. Michael Doukas, Erasmus Medical Center, Rotterdam, the Netherlands
Dr. Hala El-Zimaity, Toronto General Hospital, Toronto, Canada
Dr. Matteo Fassan, University of Padua, Padua, Italy

Dr. Roberto Fiocca, University of Genova, Genova, Italy
Dr. Jean-François Fléjou, Saint Antoine Hospital, Paris, France
Dr. Alejandro García Varona, Hospital El Bierzo, Leon, Spain
Dr. Elvira Gonzalez Obeso, Hospital Clinico Universitario, Valladolid, Spain
Dr. Heike Grabsch, 1. Division of Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds,UK, 2. Department of Pathology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, NL
Dr. Federica Grillo, University of Genova, Genova, Italy
Dr. Barbara Gruber, Patologia Bariloche, San Carlos de Bariloche, Argentina
Dr. Laura Guerra Pastrian, University Hospital La Paz, Madrid, Spain
Dr. Anne Hoorens, University Hospital Gent, Gent, Belgium
Dr. Marnix Jansen, University College Hospital, London, United Kingdom
Dr. Katerina Kamaradova, Charles University Hospital, Hradec Kralove, Czech Republic
Dr. Ryoji Kushima, Shiga University of Medical Science, Shiga, Japan
Dr. Cord Langner, Medical University of Graz, Graz, Austria
Dr. Rupert Langer, University of Bern, Bern, Switzerland
Dr. Felix Lasitschka, Universitätsklinikum Heidelberg, Heidelberg, Germany
Dr. Ester Lörinc, University Hospital Lund and Malmö, Lund, Sweden
Dr. Luca Mastracci, University of Genova, Genova, Italy
Dr. Damian McManus, Belfast HSC Trust, Belfast, Northern Ireland
Dr. Sybren Meijer, Academic Medical Center Amsterdam, the Netherlands
Dr. Carmen Mendez, University Hospital La Paz, Madrid, Spain
Dr. Anya Milne, Diakonessenhuis, Utrecht, the Netherlands
Dr. Miriam Mitchison, University College Hospital London, United Kingdom
Dr. Masoud Mireskandari, Jena University Hospital, Jena, Germany
Dr. Elizabeth Montgomery, Johns Hopkins Medical Institute, Baltimore, United States of America
Dr. Cian Muldoon, St. James's Hospital, Dublin, Ireland
Dr. Maria O'Donovan, Cambridge Cancer Centre, Cambridge, United Kingdom
Dr. Rob Odze, Brigham and Women's Hospital, Boston, United States of America
Dr. Johan Offerhaus, University Medical Center Utrecht, the Netherlands
Dr. Gabriel Olmedilla, University Hospital La Paz, Madrid, Spain
Dr. John Pauli, The Prince Charles Hospital, Brisbane, Australia
Dr. Rachel S. van der Post, Radboud university medical centre, Nijmegen, the Netherlands
Dr. Bob Riddell, Mount Sinai Hospital, Toronto, Canada
Dr. Ari Ristimaki, Haartman Institute, Helsinki, Finland
Dr. Ana Rodriguez, University Hospital La Paz, Madrid, Spain
Dr. Manual Rodriguez-Justo, University College Hospital, London, United Kingdom
Dr. Shigeki Sekine, National Cancer Center Hospital, Tokyo, Japan
Dr. Kees Seldenrijk, St. Antonius Hospital, Nieuwegein, the Netherlands
Dr. Tulio Souza, Hospital Aliança, Salvador, Brazil
Dr. Matt Stachler, Brigham and Women's Hospital, Boston, United States of America
Dr. Michael Vieth, Klinikum Bayreuth, Bayreuth, Germany
Dr. Vincenzo Villanacci, Spedali Civili di Brescia, Brescia, Italy
Dr. Rhonda Yantiss, Weill Cornell Medical College, New York, United States of America

FIGURE LEGENDS

Figure 1: Study design and study participants | A) Fifty-five representative BO biopsies with H&E slide and consecutive p53 labelled slides were collected and scanned for digital diagnostic review. Each pathologist on the study first completed a detailed demographic questionnaire (Supplementary Table 3). Pathologists then assessed 55 biopsy cases whereby diagnostic entries on H&E slide alone and after revealing matched p53 labelled slides were recorded separately allowing detailed insight into the added benefit of p53 labelled slides on diagnostic agreement. Reference diagnoses were established after consensus panel meeting. Within-group interobserver agreement was established for reference panel (n=4) and participating pathologists (N=51) and multivariate regression analyses were carried out to interrogate demographic predictors of diagnostic concordance, as detailed in the text. B) Map showing geographical dispersion of pathologists participating in the BOLERO study.

Figure 2: Diagnostic variation across the study cohort | A) Waterfall plot showing the ranked distribution of case assessments (n=3,025) based on H&E slides alone for the entire cohort of pathologists. X-axis shows diagnostic concordance in percentages and y-axis shows ranked cases 1-55. Color coding as in B. B) Same visualisation for case assessments (n=3,025) after revealing matched p53 labelled slides. C) Four representative examples of the study set. Consensus diagnosis and cohort diagnoses are shown.

Figure 3: Diagnostic variation per reference diagnoses | A-F) Waterfall plots showing the ranked distribution of case assessments by participating pathologists per diagnostic category, as indicated. Left column (A-C) shows diagnostic variation per reference diagnosis based on H&E slide review alone and right column (D-F) shows diagnostic variation per reference diagnosis after revealing matched p53 labelled slides. X-axis shows diagnostic concordance in percentages and y-axis shows ranked cases. Color coding as in Figure 2B. Diagnostic variation for indefinite for dysplasia cases is shown in Supplementary Figure 1.

Figure 4: Characteristics associated with odds of major over- or under-interpretation of Barrett's oesophagus with dysplasia in multivariable adjusted analysis | *All characteristics factors mutually adjusted for each other, **Additional adjustment for p53 labelled slides in routine pathology practice

Figure 5: Population level impact of diagnostic variation for Barrett's oesophagus surveillance biopsies. | X-axis shows population prevalence of diagnostic classes where the width of each class is consistent with its proportional prevalence (total 100%) and Y-axis shows diagnostic concordance with the total surface area adding up to all diagnoses made in one year. Diagnostic concordance is shown as either concordant (in white), overinterpreted (in blue), and underinterpreted (in magenta), where % shown reveal concordant diagnoses that would be confirmed for each diagnostic class upon review by an expert pathologist panel (Table 2).

Supplementary Figure 1: Diagnostic variation for indefinite for dysplasia diagnoses before (A) and after (B) revealing matched p53 labelled slide. X-axis shows diagnostic concordance in percentages and y-axis shows ranked cases. See text for details.

Table 1: Demographics of pathologists reporting in the BOLERO study

Characteristics	**Participating pathologists n=51 (%)**	**Reference panel pathologists n=4 (%)**
Pathologist specific characteristics		
Age, years		
30-39	13 (25.5)	1 (25.0)
40-49	17 (33.3)	1 (25.0)
50-59	14 (27.5)	1 (25.0)
60+	7 (13.7)	1 (25.0)
Gender		
Male	29 (56.9)	4 (100.0)
Female	22 (43.1)	0 (0.0)
Experience, years		
0-4	8 (15.7)	1 (25.0)
5-9	9 (17.7)	1 (25.0)
10-19	18 (35.3)	0 (0.0)
20+	16 (31.4)	2 (50.0)
Considered BE* expert?		
Yes	34 (66.7)	4 (100.0)
No	8 (15.7)	0 (0.0)
Don't know	9 (17.7)	0 (0.0)
Confidence of assessment of BE biopsies		
1 (very confident)	10 (19.6)	1 (25.0)
2	25 (49.0)	3 (75.0)
3	13 (25.5)	0 (0.0)
4	3 (5.9)	0 (0.0)
5 (not confident)	0 (0.0)	0 (0.0)
Fellowship undertaken in GI-pathology	28 (54.9)	2 (50.0)
Pathology/endoscopy practice characteristics		
Work Setting (can be multiple settings)		
Academic teaching hospital	42 (82.4)	3 (75.0)
District general hospital	16 (31.4)	1 (25.0)
Private hospital	11 (21.6)	1 (25.0)
Mean number of BE cases assessed per week	11 (21.6)	0 (0.0)
0-4	16 (31.4)	3 (75.0)
5-9	14 (27.5)	1 (25.0)
10-19	8 (15.7)	0 (0.0)
20+	2 (3.9)	0 (0.0)
Don't know		
Lab size, number of reporting pathologists		
<10	14 (27.4)	0 (0.0)
10+	37 (72.6)	4 (100.0)

Table 1 continued: Demographics of pathologists reporting in the BOLERO study

Characteristics	**Participating pathologists n=51 (%)**	**Reference panel pathologists n=4 (%)**
Pathology/endoscopy practice characteristics		
Guidelines adhered to:		
North American	23 (45.1)	2 (50.0)
British	10 (19.6)	2 (50.0)
Japanese	3 (5.9)	0 (0.0)
Australian	1 (2.0)	0 (0.0)
Other	14 (27.4)	0 (0.0)
p53 IHC staining routinely used?		
Always	1 (2.0)	1 (25.0)
Most times	11 (21.6)	1 (25.0)
Sometimes	32 (62.8)	2 (50.0)
Never	7 (13.7)	0 (0.0)
Digital pathology characteristics		
Use of whole slide imaging		
Yes	22 (43.1)	4 (100.0)
No	29 (56.9)	0

Table 2: Cross table comparing the 51 participating pathologists' diagnoses to the consensus derived reference diagnoses for 55 esophageal biopsy cases (a) on H&E stained slides and (b) on H&E and p53 labelled slides for 5,610 total case interpretations*

	Consensus reference panel**	**Participating pathologists' individual diagnoses (preconsensus)**				**% Concordance (95% CI)**		
						Under-interpretation	**Over-interpretation**	**Concordance**
a. Before addition of p53 labelled slides								
Diagnosis		ND	IND	LGD	HGD			
NDBO	816	643	93	71	9	/	21.2 (18.4-24.0)	78.8 (0.70-81.6)
IND	306	59	70	110	67	19.2 (14.8-23.6)	57.8 (52.3-63.3)	22.9 (18.2-27.6)
LGD	918	151	165	382	220	34.4 (31.3-37.5)	24.0 (21.2-26.8)	41.6 (38.4-44.8)
HGD	765	17	45	159	544	28.9 (25.7-32.1)	/	71.1 (25.6-32.2)
LGD or HGD	1683	168	210	1305		22.5 (20.4-24.5)	/	77.5 (75.5-79.5)
Total	2805							
	Consensus reference panel***	**Participating pathologists' individual diagnoses (preconsensus)**				**% Concordance (95% CI)**		
						Under-interpretation	**Over-interpretation**	**Concordance**
b. After addition of p53 labelled slides								
Diagnosis		ND	IND	LGD	HGD			
NDBO	816	684	74	53	5	/	16.2 (13.7-18.7)	83.8 (81.3-86.3)
IND	102	36	24	27	15	35.3 (26.0-44.6)	41.2 (31.6-50.8)	23.5 (15.3-31.7)
LGD	1122	153	178	516	275	29.5 (26.8-32.2)	24.5 (22.0-27.0)	46.0 (43.7-49.5)
HGD	765	21	38	165	541	29.3 (26.1-32.5)	/	70.7 (67.8-73.9)
LGD or HGD	1887	174	216	1497		20.7 (18.9-22.5)	/	79.3 (77.5-81.1)
Total	2805							

Table 2 Legend: *Overall concordance for 1639/2805 diagnoses (58.4%, 95%CI 56.6-60.2%); increasing to 2018/2805 (71.9%, 95%CI 70.2-73.6%) when LGD and HGD were combined, **Note consensus reference panel results are scaled x51 to allow for comparison versus the 51 participating pathologists. Results represent 5,610 diagnoses in 55 oesophageal biopsy cases. ***Overall concordance for 1765/2805 diagnoses (62.9%, 95% CI61.1-64.7%); increasing to 2205/2805 (78.6%, 95%CI 77.1-80.1%) when LGD and HGD were combined.