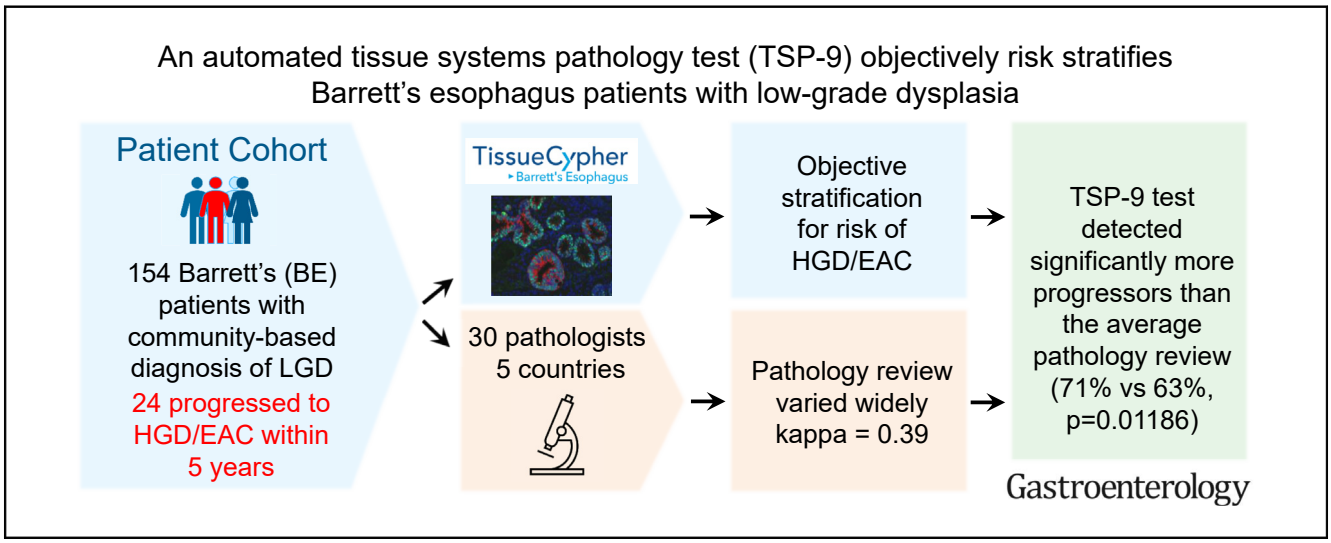
Gastroenterology 2023;■:1–12

The Tissue Systems Pathology Test Outperforms Pathology Review in Risk Stratifying Patients With Low-Grade Dysplasia

Amir M. Khoshiwal,[1] Nicola F. Frei,[1] Roos E. Pouw,[2] TissueCypher SURF LGD Study Pathologists Consortium, Christian Smolko,[3] Meenakshi Arora,[3] Jennifer J. Siegel,[3] Lucas C. Duits,[1] Rebecca J. Critchley-Thorne,[3] and Jacques J. G. H. M. Bergman[1]

[1]Amsterdam UMC location University of Amsterdam, Department of Gastroenterology and Hepatology, Amsterdam, The Netherlands; [2]Amsterdam UMC location Vrije Universiteit Amsterdam, Department of Gastroenterology and Hepatology, Amsterdam, The Netherlands; and [3]Castle Biosciences, Inc, Pittsburgh, Pennsylvania



BACKGROUND & AIMS: Low-grade dysplasia (LGD) is associated with an increased risk of progression in Barrett's esophagus (BE); however, the diagnosis of LGD is limited by substantial interobserver variability. Multiple studies have shown that an objective tissue systems pathology test (TissueCypher Barrett's Esophagus Test, TSP-9), can effectively predict neoplastic progression in patients with BE. This study aimed to compare the risk stratification performance of the TSP-9 test vs benchmarks of generalist and expert pathology. **METHODS:** A blinded cohort study was conducted in the screening cohort of a randomized controlled trial of patients with BE with community-based LGD. Biopsies from the first endoscopy with LGD were assessed by the TSP-9 test and independently reviewed by 30 pathologists from 5 countries per standard practice. The accuracy of the test and the diagnoses in predicting high-grade dysplasia (HGD) and esophageal adenocarcinoma (EAC) were compared. **RESULTS:** A total of 154 patients with BE (122 men), mean age 60.9 ± 9.8 years were studied. Twenty-four patients progressed to HGD/EAC within 5 years (median time of 1.7 years) and 130 did not progress to HGD/EAC within 5 years (median 7.8 years follow-up). The TSP-9 test demonstrated higher sensitivity (71% vs mean 63%, range 33%–88% across 30 pathologists), than the pathology review in detecting patients who progressed (P = .01186). **CONCLUSIONS:** The TSP-9 test outperformed the pathologists in risk stratifying patients with BE with LGD. Care guided by the test can provide an effective solution to variable pathology review of LGD, improving health outcomes by upstaging care to therapeutic intervention for patients at high risk for progression, while reducing unnecessary interventions in low-risk patients.

Keywords: Barrett's Esophagus; Esophageal Adenocarcinoma; High-Grade Dysplasia; TissueCypher; Tissue Systems Pathology test (TSP-9).

Abbreviations used in this paper: **BE, Barrett's esophagus; CI, confidence interval; EAC, esophageal adenocarcinoma; EET, endoscopic eradication therapy; GI, gastrointestinal; H&E, hematoxylin and eosin; HGD, high-grade dysplasia; IHC, immunohistochemistry; IND, indefinite for dysplasia; IQR, interquartile range; LGD, low-grade dysplasia; NDBE, non-dysplastic Barrett's esophagus; NPV, negative predictive value; PPV, positive predictive value; TSP-9, tissue systems pathology test.**

0016-5085
https://doi.org/10.1053/j.gastro.2023.07.029

Barrett's esophagus (BE) is a premalignant condition characterized by the conversion of the esophageal squamous epithelium into metaplastic columnar epithelium. Although the risk of malignant progression of BE is low, if left untreated, it can develop into esophageal adenocarcinoma (EAC) with a 5-year survival rate of less than 20%.[1,2] Current gastrointestinal (GI) society guidelines for the management of BE recommend endoscopic surveillance with biopsies and histologic review to identify dysplasia and EAC at early, treatable stages. However, this approach to endoscopic surveillance is ineffective because of the random biopsy sampling that can miss dysplastic areas and the significant interobserver variability between pathologists in the diagnosis of grades of dysplasia. Current practice guidelines recommend that community-based low-grade dysplasia (LGD) diagnoses be reviewed by a GI subspecialist ("expert") pathologist. If LGD is confirmed, the recommended management strategy is endoscopic eradication therapy (EET) to prevent progression or endoscopic surveillance every 6 to 12 months for close disease monitoring due to the increased risk for neoplastic progression in these patients.[3,4] Although confirmed LGD is a strong predictor of progression to high-grade dysplasia (HGD)/EAC, adherence to the recommended expert review is poor and both generalist pathologists and expert pathologists are prone to significant interobserver variability, making the clinical management of LGD very challenging. LGD is frequently overdiagnosed in the community setting. Studies of community-based cohorts of BE have shown that 73% to 85% of LGD cases are downgraded to a diagnosis of nondysplastic (ND) or indefinite dysplasia (IND) after expert GI pathologist review.[5] Randomized control trials have shown that 22% to 28% of patients with LGD confirmed by an expert GI pathologist appear to regress to NDBE during surveillance without EET.[3,6] It is difficult to determine whether the resolution of these dysplastic changes is due to true histological regression, sampling error, or initial overdiagnosis. Underdiagnosis of LGD also occurs, and a subset of patients who are downstaged to NDBE following expert diagnosis remains at high risk for neoplastic progression and will develop HGD/EAC during their surveillance period.[5,7,8] Because management decisions for LGD involve considering the benefits and risks of intervention with EET vs surveillance, objective risk stratification is needed to identify high-risk patients who will benefit from early EET to prevent progression. Endoscopic eradication therapies are minimally invasive and have been shown to be both safe and highly effective in reducing the risk of progression from NDBE or LGD to EAC by 3.8- to 7.4-fold, leading to improved health outcomes for patients.[9,10] It is also critical to identify low-risk patients who can avoid unnecessary therapy and be managed by a surveillance-only approach, enabling increased quality of life for patients. Health care resources should be targeted at the subset of patients with BE who are at high risk for progression while avoiding overtreatment and overuse of endoscopy in patients who will not benefit from EET.

WHAT YOU NEED TO KNOW

BACKGROUND AND CONTEXT

The optimal clinical management of Barrett's esophagus with low-grade dysplasia is hindered by biopsy sampling errors and pathologists' variations in histologic evaluation.

NEW FINDINGS

The tissue systems pathology test provides an objective risk stratification of community-based low-grade dysplasia, outperforming pathologists in terms of sensitivity, and can be used in conjunction with pathology to provide an effective solution to the subjective and variable pathology review.

LIMITATIONS

Analyses were completed retrospectively, although the included patients were part of a prospectively enrolled study population, and some patients were excluded because of a lack of availability of biopsy tissue and/or clinical data.

CLINICAL RESEARCH RELEVANCE

Care guided by the tissue systems pathology test can provide an effective solution to variable pathology review of low-grade dysplasia, improving health outcomes by upstaging care to therapeutic intervention for patients at high risk for progression, while reducing unnecessary interventions in low-risk patients.

BASIC RESEARCH RELEVANCE

The higher sensitivity of the tissue systems pathology test was due to the test's ability to identify 43% of the progressors who were downstaged to nondysplastic Barrett's esophagus by the pathologists, indicating the test's ability to detect molecular and cellular changes associated with progression that precede morphologic changes that can be observed by traditional histology. Risk stratification by the tissue systems pathology test can provide an objective solution to subjective and variable pathology review in the diagnosis of low-grade dysplasia and can improve health outcomes for these patients.

The TissueCypher Barrett's Esophagus Test is an objective tissue systems pathology test (TSP-9) that was specifically designed to predict the risk of progression of BE to HGD and EAC and has been extensively validated in various multi-institutional studies.[11–15] This test uses a spatialomics-based, multiplexed fluorescent imaging platform to automatically and objectively quantify 9 protein-based biomarkers and nuclear morphology in the context of tissue architecture. The quantitative image analysis is linked to a risk prediction algorithm that integrates 15 quantitative image analysis features to produce a risk score ranging from 0 to 10, which then classifies patients into high, intermediate, and low risk for progression to HGD/EAC within 5 years.[16] The aim of this study was to compare the risk stratification performance of this objective TSP-9 test vs benchmarks of community/generalist and expert pathology in the United States, the Netherlands, Germany,

Belgium, and the United Kingdom in patients with BE with a community-based diagnosis of LGD.

Methods

Population and Setting

The screening cohort for the SURveillance vs Radio-Frequency ablation (SURF) trial consists of patients with BE with a community-based diagnosis of LGD from 9 BE treatment centers and their referring institutions in Europe.[3]

Study Cohort and Design

The study cohort included all patients with BE with a community-based diagnosis of LGD who underwent screening for the SURF trial and in whom the natural history of community-based LGD could be followed (ie, patients randomized to the SURF surveillance arm and patients downgraded to NDBE or IND by an expert pathologist leading to exclusion from the SURF trial). Patients were enrolled between June 2007 and June 2011. Follow-up was completed for each patient in December 2018 by contacting the attending hospital. Endoscopic follow-up of the patients downstaged to NDBE or IND and excluded from the SURF trial was guided by the histological diagnosis per standard of care. Patients with short-segment (<3 cm) NDBE underwent endoscopic follow-up every 5 years, and patients with long-segment (≥3 cm) NDBE underwent endoscopic follow-up every 3 years. In case of IND, patients underwent endoscopic follow-up after 1 year. In case of HGD/EAC during follow-up, this diagnosis was confirmed by at least 1 expert pathologist from the panel. All endoscopies were performed according to the Seattle protocol (4-quadrant biopsies every 1–2 cm BE).[3]

Patients were excluded if informed consent was declined, if the natural outcome was unavailable due to EET, formalin-fixed paraffin-embedded tissue blocks were unavailable, or less than 3 years of disease-free follow-up was available for non-progressors after the first reviewed diagnosis of LGD. Data elements collected were age, sex, segment length, original diagnosis, presence/absence of hiatal hernia, worst histologic diagnosis during follow-up, and HGD/EAC-free surveillance time. The TSP-9 test was run in a blinded manner on all available specimens from the first LGD endoscopy and test results were reported to a statistician who performed the statistical analyses following a prespecified plan. The Institutional Biobank Review Committee of the Academic Medical Center approved the ReBus biobank.

Pathology Review

Hematoxylin and eosin (H&E) staining (2 slides) and p53 immunohistochemistry (IHC) (1 slide) were performed on sections adjacent to those on which TSP-9 testing was performed to ensure direct comparison between pathology review and TSP-9 risk stratification. One researcher (N.F.) sectioned slides from all available tissue. Without reviewing the slides, consecutive sections were randomly selected for TSP-9 testing. Digital slides were made available to pathologists via a web-based review platform (Concentriq for Research) with standard diagnosis categories of ND, IND, LGD, HGD, and EAC per the Vienna Classification, and a category of "other" for use when a diagnosis of BE could not be made. The 14 expert pathologists (3 from the Netherlands, 3 from the United States, 3 from Germany, 3 from the United Kingdom, and 2 from Belgium) and 16 community-based, generalist pathologists (3 from the Netherlands, 3 from the United States, 4 from Germany, 3 from the United Kingdom, and 3 from Belgium) independently reviewed H&E slides from all available biopsy levels per their standard practice. To create a pathology panel that reflects pathology review in the Western world, we included 2 to 3 expert pathologists and 3 community pathologists from 5 different Western countries. The pathologists were aware that each case had an initial community-based diagnosis of LGD but were blinded to clinical data, outcomes, and the TSP-9 test results. US pathologists reviewed only H&E slides per their standard practice, and the pathologists from Europe reviewed H&E slides and p53 IHC slides that were used adjunctively in rendering diagnoses, as per their respective standard practices. For cases with multiple biopsy parts, all parts were evaluated and diagnoses were recorded separately for each part.[17]

Pathologists were considered experts if they had a special interest in the field of BE for more than 10 years with a minimum case load of 5 to 10 mainly dysplastic cases per week, were considered experts by their international peers, co-authored more than 10 peer-reviewed publications in the field, and had been actively involved in pathology training in BE. The community-based generalist pathologists lacked subspecialty expertise and were referring dysplastic BE cases to an expert pathologist for review per society guidelines.

The TissueCypher Barrett's Esophagus Test (the tissue systems pathology test, "TSP-9") was run on sections from each specimen at Castle Biosciences' Clinical Laboratory Improvement Amendments–certified laboratory (Pittsburgh, PA) according to established standard operating procedures. All test parameters, including the 9 biomarkers, image analysis features, risk prediction algorithm, and cut points were prespecified and locked, as previously defined.[11,12] Test results were reported as high, intermediate (int), or low risk for progression to HGD/EAC within 5 years.

Statistical Analyses

Kaplan-Meier curves were used to determine the risk of progression to HGD/EAC as a combined endpoint of the risk classes determined by the test and pathologic diagnoses. Progression was defined as a subsequent diagnosis of HGD/EAC within 5 years. The logrank test was used to evaluate the equality of progression curves of the risk classes and diagnostic groups from the Kaplan-Meier analysis. The rate of progression was similar in the int and high-risk groups of patients, and therefore, the test was evaluated as a binary classifier (int/high risk combined vs low risk, as described previously).[15] Performance metrics for the TSP-9 test were sensitivity (% of progressors scored int/high risk), specificity (% of nonprogressors scored low risk), negative predictive value (NPV, % of patients scored low risk who did not progress), and positive predictive value (PPV, % of patients scored int/high risk who progressed to HGD/EAC) within 5 years with lost-to-follow-up patients analyzed with their last observation carried forward. For evaluation of predictive performance of pathology diagnoses, 2 strategies were used: (1) the pathology reviews were categorized as high risk if IND/LGD/higher and low risk if NDBE, or (2) as high risk if LGD/higher and low risk if NDBE/IND, as specified in the results. When evaluated as IND/LGD/high vs

NDBE, the performance metrics for the pathologists' diagnoses were sensitivity (% of progressors with a diagnosis of IND/LGD/higher), specificity (% of nonprogressors with a diagnosis of NDBE), PPV (% of patients with a diagnosis of IND/LGD/higher who progressed), and NPV (% of patients with a diagnosis of NDBE who did not progress). When evaluated as LGD/higher vs NDBE/IND, the performance metrics for the pathologists' diagnoses were sensitivity (% of progressors with a diagnosis of LGD/higher), specificity (% of nonprogressors with a diagnosis of NDBE/IND), PPV (% of patients with a diagnosis of LGD/higher who progressed), and NPV (% of patients with a diagnosis of NDBE/IND who did not progress). The predictive performance metrics of the various pathologists were summarized as mean and range because a Shapiro-Wilk test indicated that the sensitivity values of the 30 pathologists were normally distributed. Annual progression rates were estimated from 5-year NPV and PPV. The progression rates in this cohort were consistent with reported progression rates in other European BE patient cohorts. However, lower progression rates have been observed in US cohorts, and thus the NPV and PPV of the TSP-9 test and the pathologists were adjusted for prevalence using progression rates to HGD/EAC described in published meta-analysis and population-based studies.[4,18] The prevalence adjustment assumed that the sensitivity and specificity of the test and pathologists' diagnoses are independent of disease prevalence. The percentage of progressors in the cohort after 5 years of follow-up was adjusted to match the expected percentage based on the published progression rates, and then the predictive accuracy of the test and pathologists' diagnoses were calculated. For cases with multiple parts, all parts were evaluated separately and the overall result for the case was based on the highest scoring part by the TSP-9 test and the highest diagnosis per the pathologists. One-sample Wilcoxon rank-sum test was performed to compare the point estimate of the predictive performance of the test vs the pathologists, and subsets of pathologists. A Wilcoxon paired-sample test was used to compare pathologist sensitivities with and without the TSP-9 test adjunct. All analyses were performed using R Statistical Software (v4.1.3; R Core Team 2022).

Interobserver agreement was calculated using the Fleiss' kappa statistic. A kappa of 0 indicates that there is no more agreement than expected by chance alone, whereas a kappa of 1 indicates perfect agreement. Kappa values <0.2 represent poor agreement, 0.21 to 0.40 represent fair agreement, 0.41 to 0.60 represent moderate agreement, and values >0.60 represent good agreement.[19] Fleiss' kappa scores with 95% confidence intervals (CIs) were reported.

Prevalence Adjustment of NPV and PPV

The NPV and PPV for the TSP-9 test and pathology reviewers were adjusted by calculating the number of progressors that would be expected in the US Barrett's patient population based on published estimates of progression rates to HGD/EAC of 0.63% per year for patients with NDBE and 1.7% per year for patients with LGD/IND.[4,18] The adjusted calculations are therefore based on estimating the number of progression events that would have occurred in the sample given the prevalence, and the True/False Positives/Negatives were then calculated using the observed sample Sensitivity/Specificity as shown in the following. Finally, the adjusted NPV and PPV follow logically from the standard formulas shown as follows. Note adj means prevalence-adjusted:

$$Progressors_{adj} = Prevalence * n_{cases}$$

$$TruePositive_{adj} = Progressors_{adj} * Sensitivity$$

$$TrueNegative_{adj} = Specificity * \left(n_{cases} - Progressors_{adj}\right)$$

$$FalsePositive_{adj} = n_{cases} - Progressors_{adj} - TrueNegative_{adj}$$

$$FalseNegatives_{adj} = Progressors_{adj} - TruePositive_{adj}$$

$$PPV_{adj} = \frac{TruePositive_{adj}}{TruePositive_{adj} + FalsePositive_{adj}}$$

$$NPV_{adj} = \frac{TrueNegative_{adj}}{TrueNegative_{adj} + FalseNegative_{adj}}$$

Results

Patient Characteristics

A total of 154 patients met the inclusion criteria out of which 24 progressed, 130 did not progress to HGD/EAC during a 5-year follow-up, 20 had been enrolled into the SURF trial surveillance arm, and 134 were screened but not enrolled in the SURF trial (Table 1, Supplementary Figure 1). Progression was defined as a diagnosis of HGD/EAC within 5 years. Eight patients were diagnosed with HGD/EAC less than 1 year after the baseline endoscopy (prevalent cases), and 16 patients were diagnosed with HGD/EAC between 1 and 5 years after the baseline endoscopy (incident progressors). Progression to HGD/EAC occurred at a median of 1.7 years (interquartile range [IQR], 0.6–2.5) after the baseline endoscopy, whereas nonprogressors had a median HGD/EAC-free follow-up of 7.8 years (IQR, 5.8–10.1), and 110 of the 130 nonprogressors had more than 5 years of follow-up. The median follow-up for the 20 censored patients was 3.78 years (IQR, 3.4–4.3 years). Patients had a median age of 61 ± 10 years and had a median segment length of 5 cm (IQR, 3.0–6.3) and 4 cm (IQR, 3.0–5.5) for progressors and nonprogressors, respectively. Twenty of 24 progressors (83.3%) and 102 of 130 nonprogressors (78.5%) were men, which is consistent with other studies. There were no significant differences in age, sex, and segment length between progressors and nonprogressors.

Risk Stratification Performance of the TSP-9 Test and Pathology Diagnoses

The TSP-9 test scored 45 (29.2%) patients as int/high risk, and 109 (70.8%) patients as low risk for progression to

Table 1. Patient Characteristics

Characteristics		Progressors	Nonprogressors	*P* value
All patients (n = 154)	n	24	130	na
	Male (%)	20 (83.3%)	102 (78.5%)	ns[a]
	Age, *y*, mean ± SD	63.5 ± 9.5	61.0 ± 10.4	ns[b]
	Barrett's segment length, cm, median (IQR)	5.0 (3.0–6.3)	4.0 (3.0–5.5)	ns[c]
	HGD/EAC-free surveillance time,[d] *y*, median (IQR)	1.7 (0.6–2.5)	7.8 (5.8–10.1)	na
Patients screened and enrolled in SURF RCT surveillance arm (n = 20)	n	6	14	na
	Male (%)	5 (83.3%)	13 (93.9%)	ns[a]
	Age, *y*, mean ± SD	58.0 ± 9.5	61.8 ± 9.6	ns[b]
	Barrett's segment length, cm, median (IQR)	5.0 (3.5–5.0)	3.5 (2.3–4.0)	ns[c]
	HGD/EAC-free surveillance time, *y*, median (IQR)	1.6 (1.1–1.8)	7.7(6.7–9.4)	na
Patients screened but not enrolled in SURF RCT surveillance arm (n = 134)	n	18	116	na
	Male (%)	15 (83.3%)	89 (76.7%)	ns[a]
	Age, *y*, mean ± SD	65.3 ± 9.0	60.9 ± 10.6	ns[b]
	Barrett's segment length, cm, median (IQR)	5.0 (3.0–6.8)	4.0 (3.0–6.0)	ns[c]
	HGD/EAC-free surveillance time, *y*, median (IQR)	1.9 (0.5–3.0)	7.9 (5.6–10.1)	na

na, not applicable; ns, nonsignificant; RCT, randomized controlled trial; SD, standard deviation.
[a]Chi-squared test.
[b]Two-sample *t* test performed after verifying normality with Shapiro-Wilk test.
[c]Wilcoxon rank sum test (equivalent to the Mann-Whitney test) performed after violation of normality confirmed by Shapiro-Wilk test.
[d]Time between acquisition of biopsy tested and endoscopy with HGD or EAC (progressors) or last follow-up (nonprogressors).

HGD/EAC within 5 years (Supplementary Figure 2). The pathologists confirmed a mean of 19% of cases (range 8%–41%) to be LGD/higher and downstaged a mean of 13% (range 0%–75%) to IND and a mean of 68% (range 12%–88%) to NDBE (Supplementary Figure 3*A* and *B*). Annual progression rates were 9.2% per year for patients with confirmed LGD/higher, 3% per year for patients downstaged to IND, and 1.7% per year for patients downstaged to NDBE.

Interobserver agreement among the pathologists was assessed using a Fleiss' kappa score. The interobserver agreement was fair to moderate with an overall kappa of 0.39 (95% CI, 0.31–0.45) when evaluating IND/LGD/higher combined vs NDBE (Supplementary Figure 3*C*). The expert pathologists agreed slightly more (kappa of 0.43; 95% CI, 0.34–0.50) than the generalists (kappa of 0.34; 95% CI, 0.26–0.4). Similar results were observed when IND and NDBE cases were grouped (Supplementary Figure 3*C*).

The TSP-9 Test Demonstrated Higher Sensitivity Than Pathology Review in Risk Stratifying Patients With BE With LGD

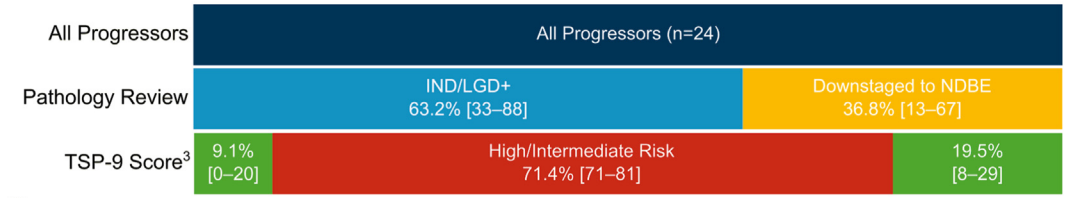
The diagnoses were first evaluated as IND/LGD/higher vs NDBE. The TSP-9 test demonstrated higher sensitivity than the average pathologist ($P = .01186$). The test detected 70.8% of patients who progressed within 5 years, whereas the pathologists detected a mean of 63.2% ($P = .01186$), and a wide range of sensitivity (33%–88%) was observed across the 30 pathologists (Figures 1*A* and 2*A*). The specificity of the TSP-9 test was 78.5% and a mean of 73.5% (range 12%–95%) for the pathologists. There was no statistically significant difference between the sensitivity ($P = .07$) or specificity ($P = .16$) of the expert and generalist pathologists in predicting the risk of progression. Similar results were observed when the 8 prevalent cases were excluded from the analysis (Figure 2*C*). Seven of the 8 prevalent cases were scored TSP-9 high or intermediate risk, indicating that the test has a sensitivity of 87.5% in detecting the presence of prevalent HGD/EAC. The pathologists' diagnoses were also evaluated as LGD/higher vs ND/IND, which resulted in lower sensitivity of the dysplasia diagnosis (mean sensitivity 50.9%, range 25%–71%) and increased specificity (mean 86.7%, range 62%–96%) (Figure 2*B* and *D*, Supplementary Figure 4). The TSP-9 test demonstrated higher sensitivity than all 30 pathologists when only LGD/higher was considered as a true positive diagnosis (Figure 2*B* and *D*, Supplementary Figure 4).

The TSP-9 test demonstrated higher NPV (93.6%) when compared with the pathologists (mean 91.4%, range 84%–95%, $P = .00002$), and the TSP-9 PPV was 37.8% vs mean 35.4% with range 16% to 63% for the pathologists (Figure 1*A*). No statistically significant differences in the predictive value of the expert and generalist pathologist

A

Prediction Metric mean [range]	TSP-9[1]	Pathology Alone
Sensitivity (%)	70.8	63.2 [33–88]
Specificity (%)	78.5	73.5 [12–95]
PPV (%)	37.8	35.4 [16–63]
NPV (%)	93.6	91.4 [84–95]
PPV adj.[2] (%)	23.7	22.6 [9–47]
NPV adj.[2] (%)	96.6	95.4 [91–97]

B



C

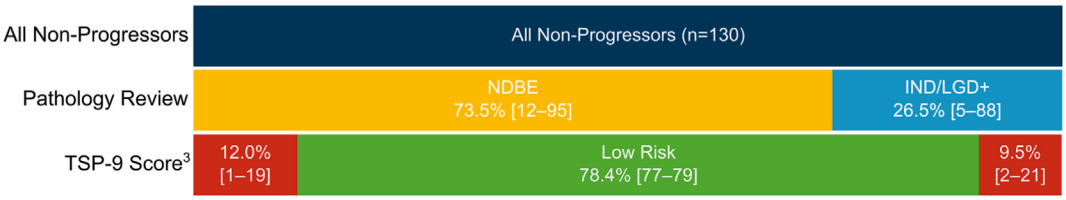


Figure 1. The TSP-9 test finds more progressors than pathologists. (*A*) Prediction metrics for combined use of pathology diagnoses and the TSP-9 test. (*B*) Comparison of diagnoses (IND/LGD/higher vs NDBE) and TSP-9 test results in the subset of patients who progressed to HGD/EAC within 5 years (n = 24). Green segments represent patients who scored low risk by the test. (*C*) Comparison of diagnoses (IND/LGD/higher vs NDBE) and TSP-9 test results in patients who did not progress within 5 years (n = 130). *Red segments* represent patients who scored high/int-risk by the test. Mean and [range] are shown. [1]TSP-9 prediction metrics were calculated in the subset for which each pathologist rendered a BE diagnosis, which ranged from 141 to 154 patients as 5 pathologists scored a subset of cases as non-BE or not reportable. [2]Patients with a TSP-9 low-risk result and a pathology diagnosis of NDBE were considered low risk, and patients with a TSP-9 intermediate/high-risk result or a pathology diagnosis of IND/LGD/higher were considered high risk (see Methods). [3]TSP-9 prediction metrics in the highest scoring case-part from all 154 patients. [4]NPV and PPV were adjusted for prevalence as described in Methods.

diagnoses were observed (PPV, $P = .532$; NPV, $P = .399$). Patients diagnosed as NDBE progressed at a rate of 1.72% per year, whereas patients who scored TSP-9 low risk progressed at a lower rate of 1.28% per year ($P = .000004$), indicating that the test may downstage the risk of progression in patients who are typically managed by surveillance only. Patients diagnosed as IND/LGD/higher and patients who scored TSP-9 int/high risk progressed at a similar rate (7.1% per year and 7.6% per year, respectively, $P = .1978$), indicating that the test's int/high-risk results are clinically actionable.

The rate of progression in patients diagnosed with NDBE was 1.72% per year, which is higher than published estimates of progression rates for patients with NDBE, which may be because of the higher prevalence of HGD/EAC diagnoses in the evaluated cohort.[18] To address this, the NPV and PPV values were also adjusted for prevalence (see Methods) to estimate the predictive performance of the TSP-9 test and the pathologist diagnoses in the US BE population in which lower progression rates have been observed compared with European populations. The prevalence-adjusted PPV of the int/high-risk test result and the pathologists' diagnoses of IND/LGD/higher for predicting progression within 5 years were similar (23.7% for the test compared with a mean of 22.6% [range 9%–47%] for the pathologists) (Figure 1*A*). The prevalence-adjusted NPV of the low-risk test result and the pathologists' diagnosis of NDBE were similar. No differences were observed in the prevalence-adjusted PPV and NPV between the expert and generalist pathologist diagnoses. Similar results were obtained when the pathologists' diagnosis was evaluated as IND/LGD/higher vs NDBE (Supplementary Figure 4).

The Accuracy of Both Expert and Generalist Diagnoses Showed Significant Interobserver Variability

The predictive accuracy of the generalists' diagnoses varied widely, and although the experts' diagnoses demonstrated slightly less variability, the overall predictive accuracy was not significantly higher than the generalists' diagnoses (Figures 2*A* and *C*). A subset of both the generalist and expert pathologists tended to overdiagnose dysplasia with relatively high sensitivity, but clinically unacceptable specificity. Another subset tended to underdiagnose dysplasia, demonstrating high

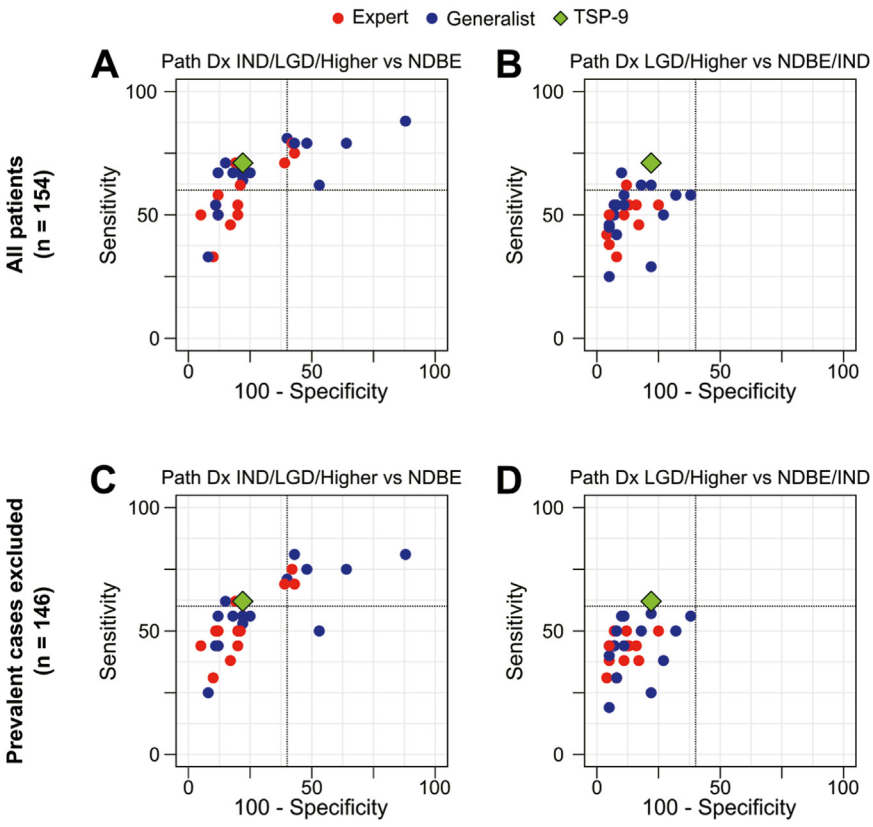


Figure 2. The TSP-9 test provides objective risk stratification that outperforms most pathologists. Predictive accuracy was evaluated using a receiver operating characteristic–like plot showing the point estimates of sensitivity and 1-specificity of the risk classes produced by the TSP-9 test and diagnoses from the pathologists: (*A*) TSP-9 and pathology diagnoses evaluated as IND/LGD/higher vs NDBE in nonprogressors, incident progressors, and prevalent cases (n = 154); (*B*) TSP-9 and pathology diagnoses evaluated as LGD/higher vs NDBE/IND in nonprogressors, incident progressors, and prevalent cases (n = 154); (*C*) TSP-9 and pathology diagnoses evaluated as IND/LGD/higher vs NDBE in nonprogressors and incident progressors (prevalent cases excluded (n = 146); (*D*) TSP-9 and pathology diagnoses evaluated as LGD/higher vs. NDBE/IND in nonprogressors and incident progressors (prevalent cases excluded (n = 146).

specificity but relatively low sensitivity. The predictive accuracy of the diagnoses demonstrated less variability when IND cases were grouped with the NDBE cases (Figure 2*B* and *D*). When evaluated as LGD/higher vs NDBE/IND, most pathologists tended toward underdiagnosis of dysplasia.

Comparison of Pathologic Diagnoses by the Pathologists From the United States and Europe

We compared the predictive performance of diagnoses from the pathologists in Europe who reviewed both p53 IHC and H&E slides vs the pathologists in the United States who only had access to H&E slides per their standard practice (Supplementary Table 2). The US pathologists had slightly higher sensitivity in detecting progressors as compared with the European pathologists (mean 68.8%, range 54%–88% US/H&E only vs mean 61.8%, range 33%–81% for Europe/both H&E and p53, $P = .284$). The pathologists evaluating p53 adjunctively had a marginally higher specificity (mean 75.5%, range 36%–95% vs mean 65.5%, range 12%–89%, $P = .533$, Supplementary Table 2). It should be noted that the European pathologists had access to the p53 IHC slides to use adjunctively per their standard practice, but they did not use a standardized scoring system and did not record p53 IHC scores.

The TSP-9 Test Increases the Detection of Progressors and Reduces Overdiagnosis of Nonprogressors by the Pathologists

The complement and intersect of the TSP-9 test and the pathologists' diagnoses were evaluated. Prediction metrics for the test were calculated in the subset of patients for which each pathologist rendered a BE diagnosis, which ranged from 141 to 154 patients, as 5 pathologists scored a subset of cases as not reportable or not containing BE mucosa. The TSP-9 test scored 71.4% (range 71%–81%, depending on which pathologist scored the slides, as 5 pathologists did not render BE diagnoses for all cases) of the progressors as int/high risk, and the pathologists diagnosed a mean of 63.2% (range 33%–88%) of the progressors as IND/LGD/higher and downstaged a mean of 36.8% (range 13%–67%) of progressors to NDBE (Figure 1*B*), demonstrating that the test detects more progressors. Although there was some overlap in the progressors detected by both methods, the higher sensitivity of the test was due to its ability to detect progressors that the pathologists downstaged from LGD to NDBE. Nine percent (range 0%–20%) of the progressors were scored TSP-9 low risk but IND/LGD/higher by the pathologists, suggesting that the combination of the test with pathology may further increase the detection of progressors. A subset of progressors (mean 19.5%, range 8%–29%) was scored NDBE and TSP-9 low risk. In nonprogressors, the pathologists diagnosed a mean of 73.5% (range 12%–95%) as NDBE and a mean of 26.5% (range 5%–88%) as IND/LGD/higher, demonstrating overdiagnosis of dysplasia by a subset of pathologists. The TSP-9 test scored 78.4% (range 77%–79% depending on which pathologist reviewed the slides) of the nonprogressors as low risk, indicating that the test has a lower overall rate of overscoring than pathology. A subset of patients (9.5%) who were overscored by TSP-9 were also overdiagnosed as IND/LGD/higher by pathology (Figure 1*C*).

When the TSP-9 test was assessed in conjunction with pathology, the sensitivity of the pathologists' diagnoses of IND/LGD/higher increased from a mean of 63.2% (range 33%–88%) to 80.4% (range 71%–92%) ($P = .00000176$), indicating that the test can increase the early detection of progressors when used with pathology (Figure 1*A* and Supplementary Table 3). For the pathologists in the lowest 10th percentile in terms of sensitivity, addition of TSP-9 increased sensitivity from 33%–46% to 71%–79%. The prevalence-adjusted NPV of the NDBE diagnosis increased from a mean of 95.4% (range 91%–97%) to 97% (range 94%–98%) when the TSP-9 test was used as an adjunct. However, the combined predictor had lower specificity and PPV than the test alone and demonstrated significant variability depending on which pathologist reviewed the slides, indicating that the test alone provides overall higher accuracy and reliability in clinical use.

The TSP-9 Test Identifies Progressors in the Subset of Patients Who Were Downstaged From LGD to NDBE or NDBE/IND Upon Pathology Review

The test was evaluated in the subsets of patients downstaged to NDBE or NDBE/IND by the individual pathologists. The number of patients in these downstaged subsets varied widely among the pathologists. The test detected a mean of 43% (range 17%–69%) of the progressors downstaged to NDBE and a mean of 54.4% (range 29%–69%) of the progressors downstaged to NDBE/IND by the pathologists (Table 2 and Supplementary Table 4). The test demonstrated higher sensitivity with a mean of 45.9% (range 17%–69%) in the subset of patients downstaged to NDBE by the expert pathologists as compared with a mean of 40.4% (range 25%–58%) in the subset downstaged by the generalists. This is consistent with the results described previously and prior literature showing that the expert pathologists tend to downstage more patients than the generalist pathologists.[7] The PPV of the test in patients downstaged to NDBE was a mean of 21.4% (range 9%–50%), indicating that the test identifies a subset of missed progressors who progressed at a similar rate to patients with confirmed LGD. Similar results were obtained when the test was evaluated in the subset of patients downstaged to NDBE by at least 15 pathologists, demonstrating that the test detects progressors who are missed by most pathologists (Table 2).

Discussion

This study focused on comparing the performance of an objective tissue systems pathology test (TSP-9) vs benchmarks of generalist and expert pathology in stratifying patients with BE with a community-based diagnosis of LGD for risk of progression to HGD/EAC. The TSP-9 test objectively risk-stratified LGD patients with overall higher sensitivity in detecting progressors than benchmarks of generalist and expert pathologists from 5 different countries. Although this study confirmed the predictive value of confirmation of dysplasia by pathology review, the ranges of confirmation vs downstaging, and the resulting wide ranges in progression rates associated with each diagnosis highlighted the significant interobserver variability in the review of community-based LGD. Although the expert pathologists showed slightly less interobserver variability than the generalist pathologists, which is consistent with previous studies, the overall predictive accuracy of the diagnoses made by the expert pathologists was not significantly higher when compared with diagnoses from the generalist pathologists. The availability of p53 IHC along with the review of H&E slides by pathologists did not show a different predictive accuracy compared with the pathologists who only evaluated H&E slides. However, the practices of the pathologists in Europe and the United States may differ in ways that are independent of use of p53, and those differences were not evaluated in this study.

A subset of both generalist and expert pathologists tended to overdiagnose dysplasia, which can lead to unnecessary therapeutic intervention. Another subset of pathologists tended to downstage a significant proportion of patients to NDBE, which can lead to undertreatment or delayed treatment, as a subset of the downstaged patients will progress to HGD/EAC during a 3- to 5-year surveillance interval. Only a small subset of the 30 pathologists (2 generalists and 1 expert) demonstrated similar overall predictive accuracy to the TSP-9 test. The higher sensitivity of the TSP-9 test was due to the test's ability to identify 43% of the progressors who were downstaged to NDBE by the pathologists. This is a high-risk subset of patients who may be missed by pathology review but can be detected by the TSP-9 test enabling therapeutic intervention or close surveillance. The TSP-9 test also demonstrated higher specificity and a higher NPV than the panel of pathologists, indicating that the low-risk class provided by the test may improve confidence in a surveillance-only management approach for a subset of patients with BE with community-based LGD.

Current guidelines recommend that the diagnosis of BE with LGD should be confirmed by an expert pathologist to better risk-stratify patients and to guide management decisions accordingly.[3,20–22] However, "expert pathology" is not well defined, and confirmation of LGD is challenging, which limits real-world adherence to expert review. One of the main challenges is interobserver variability,[23] which was confirmed in the current study with only moderate agreement (kappa coefficient of 0.43) even among expert pathologists, which limits the effectiveness of the LGD diagnosis to identify patients whose progression risk warrants EET. With all these caveats, a patient's management plan and health outcomes are largely dependent on which pathologist reviews the patient's H&E slides. For example, if reviewed by a pathologist who tends to underdiagnose dysplasia, a subset of patients will be downstaged to NDBE and develop HGD/EAC during their next surveillance interval, which may lead to poor health outcomes. However, if reviewed by a pathologist who tends to overdiagnose, a subset of patients may get EET without being at risk for progressing to HGD/EAC at the expense of adverse events, most commonly esophageal stenosis (15%), post-

Table 2. Predictive Performance of the TSP-9 Test in Patients Downstaged From LGD to NDBE

	TSP-9 Performance in patients who were downstaged to NDBE by:					
	Individual pathologists[a]			At least 15 pathologists		
Prediction metric	All	Experts	Generalists	All	Experts	Generalists
% of progressors who were downstaged to NDBE and scored high/int risk by the test (sensitivity)	43.0 [17–69]	45.9 [17–69]	40.4 [25–58]	37.5	37.5	28.6
% of nonprogressors who were downstaged to NDBE and scored low risk by the test (specificity)	84.5 [79–94]	83.5 [79–91]	85.5 [80–94]	84.3	83.9	83.5

[a]Mean [range].

procedural bleed (4%), and perforation (0.8%).[24,25] The TSP-9 test was specifically designed to stratify patients with BE for risk of progression to HGD/EAC, whereas pathology review provides a diagnosis, which in turn has implications for risk of progression and management of patients with BE. The TSP-9 test provides an effective solution to variable pathology review by objectively detecting more progressors and downstaging more nonprogressors than pathology review as demonstrated in this study. Although the primary aim of the study was to compare the predictive performance of TSP-9 and benchmarks of pathology review, the results demonstrated that the 2 methods can be used in conjunction to improve sensitivity for detection of progressors from 50.9% to 62.3% (depending on whether IND is considered a high-risk diagnosis) up to 80.4% of progressors with a range of 71% to 92%, depending on which pathologist reviewed the slides. In addition to providing risk stratification results to physicians and patients, the TSP-9 test may also provide pathologists with objective information to standardize review of biopsies.

Based on the results of this study, we propose a guide for how the TSP-9 test can be used to improve the management of patients with BE with an initial, community-based diagnosis of LGD (Figure 3). The TSP-9 test may be used after pathology review to further risk-stratify and guide care, enabling clinicians to devise effective strategies aimed at increasing the percentage of progressors whose care is upstaged to EET or short interval surveillance, as well as to increase the percentage of nonprogressors who are managed by surveillance every 3 to 5 years, potentially leading to improved health outcomes. For patients who receive a low-risk score by the TSP-9 test, the optimal clinical management strategy may be surveillance in less than 12 months for patients with a pathology review diagnosis of LGD or IND, and surveillance every 3 to 5 years for patients who have been downstaged to NDBE. For patients who receive a high- or intermediate-risk score by the test, an effective management strategy may be EET to prevent progression to HGD/EAC or surveillance in 6 to 12 months regardless of whether the review diagnosis is LGD, IND, or NDBE (Figure 3). However, given the finding of equivalent overall accuracy between expert and generalist pathologists in this study, this TSP-9 test that objectively risk stratifies with higher overall accuracy may be a logical tool for proper risk stratification of patients with BE with LGD rather than using the test as an adjunct to pathology review. The test can provide a practical solution to observer variability in the diagnosis of LGD and to the challenges in accessing expert reviews in some settings.

The TSP-9 test offers multiple advantages over IHC-based evaluation of biomarkers such as p53 and *Aspergillus oryzae* lectin, as these IHC-based methodologies are limited by qualitative and subjective estimation of single biomarkers per slide.[26,27] The TSP-9 test automatically and objectively quantifies multiple predictive biomarkers without the need for any interpretation by a human, and produces an individualized risk score for progression to HGD/EAC within 5 years. Various other molecular approaches requiring digestion of tissue for risk prediction have been evaluated; however, these result in loss of tissue morphology and spatial relationships that have predictive value, and none have been independently validated so far for risk prediction in BE.[28,29]

The main strengths of this study include the cohort study design derived from the screening cohort of a randomized controlled trial with long-term outcome data.[3,30] In addition, the study includes blinded testing in a Clinical Laboratory Improvement Amendments–certified laboratory and independent blinded review by benchmarks of generalist and expert pathologists from 5 different countries. Furthermore, H&E and p53 IHC staining were performed on sections adjacent to those on which the TSP-9 test was performed such that the morphology reviewed by the pathologists was as close as possible to what was assessed by the test. The limitations are the exclusion of some patients due to lack of availability of biopsy material and/or clinical data, and although the patients

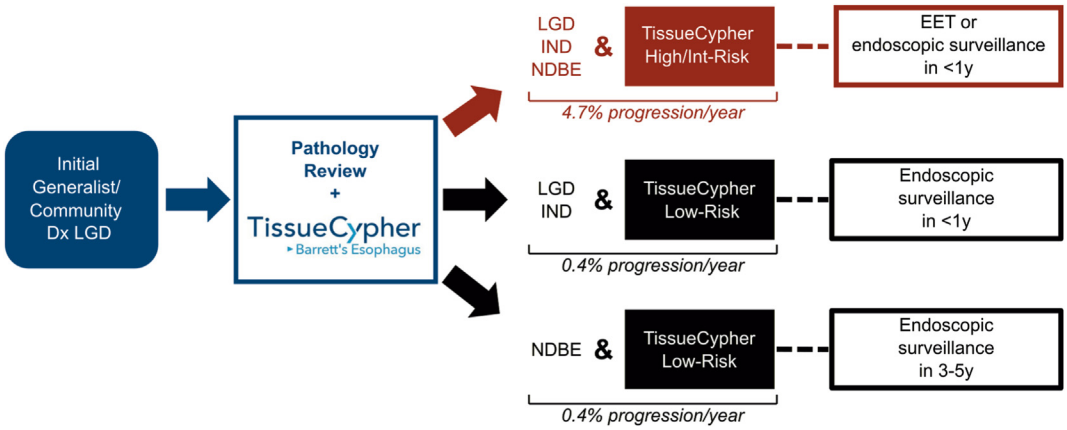


Figure 3. Proposed use of the TSP-9 test to aid the management of patients with BE with a community-based diagnosis of LGD. Adjunctive use of the TSP-9 test results with expert pathology review to guide management decisions.

were followed prospectively as part of the prior study, the analyses were completed retrospectively. However, large prospective studies are very challenging in BE because of the low rate of malignant progression. Although the progression rates reported here are higher than some other reports of clinically observed progression rates, this was corrected for by adjusting the NPV and PPV for both the TSP-9 test and pathology results to the expected prevalence in the general BE population in the United States. Although use of p53 IHC did not appear to increase the predictive accuracy of the pathologic diagnoses, the p53 IHC slides were provided for adjunctive use, and this study did not implement a standardized p53 IHC scoring system or enforce review of those slides by the pathologists.

In conclusion, the TSP-9 test provides objective risk stratification in patients with BE with community-based LGD and outperforms even expert pathologists on average in identifying patients who progress to HGD/EAC. The TSP-9 test is not subject to observer variation and can thus offer an objective solution to the subjective and variable pathology grading. This test can increase the early detection of patients at high risk for progression, while also reducing unnecessary interventions in patients at low risk for progression. The superior predictive accuracy and consistency of the TSP-9 test may allow physicians to make more informed management decisions for their patients that may lead to improved health outcomes.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2023.07.029.

References

1. Thrift AP. Barrett's esophagus and esophageal adenocarcinoma: how common are they really? Dig Dis Sci 2018;63:1988–1996.
2. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 2019; 144:1941–1953.
3. Phoa KN, van Vilsteren FGI, Weusten BLAM, 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.
4. **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. Gastrointest Endosc 2014;79:897–909.e4; quiz 983.e1, 983.e3.
5. 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.
6. Shaheen NJ, Sharma P, Overholt BF, et al. Radiofrequency ablation in Barrett's esophagus with dysplasia. N Engl J Med 2009;360:2277–2288.
7. 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.
8. Hussein M, Sehgal V, Sami S, et al. The natural history of low-grade dysplasia in Barrett's esophagus and risk factors for progression. JGH Open 2021;5:1019–1025.
9. Wani S, Puli SR, Shaheen NJ, et al. Esophageal adenocarcinoma in Barrett's esophagus after endoscopic ablative therapy: a meta-analysis and systematic review. Am J Gastroenterol 2009;104:502–513.
10. Qumseya BJ, Wani S, Gendy S, et al. Disease progression in Barrett's low-grade dysplasia with radiofrequency ablation compared with surveillance: systematic review and meta-analysis. Am J Gastroenterol 2017; 112:849–865.
11. Critchley-Thorne RJ, Duits LC, Prichard JW, et al. A tissue systems pathology assay for high-risk Barrett's esophagus. Cancer Epidemiol Biomarkers Prev 2016; 25:958–968.

12. Critchley-Thorne RJ, Davison JM, Prichard JW, et al. A tissue systems pathology test detects abnormalities associated with prevalent high-grade dysplasia and esophageal cancer in Barrett's esophagus. Cancer Epidemiol Biomarkers Prev 2017;26:240–248.
13. Davison JM, Goldblum J, Grewal US, et al. Independent blinded validation of a tissue systems pathology test to predict progression in patients with Barrett's esophagus. Am J Gastroenterol 2020;115:843–852.
14. Frei NF, Konte K, Bossart EA, et al. Independent validation of a tissue systems pathology assay to predict future progression in nondysplastic Barrett's esophagus: a spatial-temporal analysis. Clin Transl Gastroenterol 2020;11:e00244.
15. Frei NF, Khoshiwal AM, Konte K, et al. Tissue systems pathology test objectively risk stratifies Barrett's esophagus patients with low-grade dysplasia. Am J Gastroenterol 2021;116:675–682.
16. Prichard JW, Davison JM, Campbell BB, et al. TissueCypher(TM): A systems biology approach to anatomic pathology. J Pathol Inform 2015;6:48.
17. Schlemper RJ, Riddell RH, Kato Y, et al. The Vienna classification of gastrointestinal epithelial neoplasia. Gut 2000;47:251–255.
18. Wani S, Falk G, Hall M, et al. Patients with nondysplastic Barrett's esophagus have low risks for developing dysplasia or esophageal adenocarcinoma. Clin Gastroenterol Hepatol 2011;9:220–227; quiz e26.
19. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276–282.
20. 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; quiz 51.
21. Small AJ, Araujo JL, Leggett CL, et al. Radiofrequency ablation is associated with decreased neoplastic progression in patients with Barrett's esophagus and confirmed low-grade dysplasia. Gastroenterology 2015;149:567–576.e3; quiz e13–14.
22. Shaheen NJ, Falk GW, Iyer PG, et al. Diagnosis and management of Barrett's esophagus: an updated ACG guideline. Am J Gastroenterol 2022;117:559–587.
23. Montgomery E, Bronner MP, Goldblum JR, et al. Reproducibility of the diagnosis of dysplasia in Barrett esophagus: a reaffirmation. Hum Pathol 2001;32:368–378.
24. Qumseya BJ, Wani S, Desai M, et al. Adverse events after radiofrequency ablation in patients with Barrett's esophagus: a systematic review and meta-analysis. Clin Gastroenterol Hepatol 2016;14:1086–1095.e6.
25. van Munster S, Nieuwenhuis E, Weusten BLAM, et al. Long-term outcomes after endoscopic treatment for Barrett's neoplasia with radiofrequency ablation ± endoscopic resection: results from the national Dutch database in a 10-year period. Gut 2022;71:265–276.
26. Redston M, Noffsinger A, Kim A, et al. Abnormal TP53 predicts risk of progression in patients with Barrett's esophagus regardless of a diagnosis of dysplasia. Gastroenterology 2022;162:468–481.
27. **Duits LC, Lao-Sirieix P, Wolf WA**, et al. A biomarker panel predicts progression of Barrett's esophagus to esophageal adenocarcinoma. Dis Esophagus 2019;32:doy102.
28. Sepulveda JL, Komissarova EV, Kongkarnka S, et al. High-resolution genomic alterations in Barrett's metaplasia of patients who progress to esophageal dysplasia and adenocarcinoma. Int J Cancer 2019;145:2754–2766.
29. **Stachler MD, Camarda ND,** Deitrick C, et al. Detection of mutations in Barrett's esophagus before progression to high-grade dysplasia or adenocarcinoma. Gastroenterology 2018;155:156–167.
30. **Pouw RE, Klaver E,** Phoa KN, et al. Radiofrequency ablation for low-grade dysplasia in Barrett's esophagus: long-term outcome of a randomized trial. Gastrointest Endosc 2020;92:569–574.

Author names in bold designate shared co-first authorship.

Received October 19, 2022. Accepted July 12, 2023.

Correspondence
Address correspondence to: Rebecca J. Critchley-Thorne, Castle Biosciences, Inc, 100 S Commons, Suite 245, Pittsburgh, Pennsylvania 15212. e-mail: rthorne@castlebiosciences.com; or Jacques J.G.H.M. Bergman, Department of Gastroenterology and Hepatology, Amsterdam University Medical Centers, Academic Medical Center, Meibergdreef 9, Amsterdam1105 AZ, The Netherlands. e-mail: j.j.bergman@amsterdamumc.nl.

Acknowledgments
The TissueCypher SURF LGD Study Pathologists Consortium includes John Goldblum,[1] Elizabeth Montgomery,[2] Jon Davison,[3] Jagjit Singh,[4] Jared Szymanski,[5] Anthony Perry,[5] Kees Seldenrijk,[6] Fiebo ten Kate,[7] G. Johan A. Offerhaus,[7] Paul Drillenberg,[8] Casper Jansen,[9] Natalja Leeuwis-Fedorovich,[10] Runjan Chetty,[11] Roger Feakins,[12] Marnix Jansen,[13] Catherine Chinyama,[14] Edwin Cooper,[15] Reza Vaziri,[16] Gustavo Baretton,[17] Andrea Tannapfel,[18] Michael Vieth,[19] Balint Melcher,[20] Ildiko Mesteri,[21] Heiko Müller,[21] Philipp Wetzel,[21] Anne Hoorens,[22] Stephanie Verschuere,[23] An Tamsin,[24] Kevin Wetzels,[25] and Marie-Astrid van Caillie[26]; from the [1]Cleveland Clinic, Cleveland, Ohio; [2]University of Miami, Miami, Florida; [3]University of Pittsburgh, Pittsburgh, Pennsylvania; [4]University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania; [5]Connect Pathology, Lehi, Utah; [6]St Antonius Hospital, Nieuwegein, The Netherlands; [7]University Medical Center Utrecht, Utrecht, The Netherlands; [8]Onze Lieve Vrouwe Gasthuis, Amsterdam, The Netherlands; [9]Laboratorium Pathologie Oost Nederland, Hengelo, The Netherlands; [10]Deventer Ziekenhuis, Deventer, The Netherlands; [11]Deciphex Ltd, Mornington, Ireland, United Kingdom; [12]Royal Free Hospital, London, United Kingdom; [13]University College London, London, United Kingdom; [14]Princess Elizabeth Hospital, Guernsey, United Kingdom; [15]Yeovil District Hospital, Somerset, United Kingdom; [16]Worcestershire Acute Hospitals, Worcester, United Kingdom; [17]University Hospital Carl Gustav, Dresden, Germany; [18]Ruhr-Universität Bochum, Bochum, Germany; [19]Friedrich-Alexander-Universität, Erlangen, Germany; [20]Institute of Pathology, Koblenz, Germany; [21]Institute of Pathology, Überlingen, Germany; [22]UZ Gent, Gent, Belgium; [23]AZ Groeninge, Kortrijk, Belgium; [24]AZ Delta, Roeselare, Belgium; [25]AZ Sint-Blasius, Dendermonde, Belgium; and [26]AZ Sint-Lucas, Brugge, Belgium.

CRediT Authorship Contributions
Amir M. Khoshiwal, MD (Data curation: Equal; Formal analysis: Equal; Investigation: Supporting; Methodology: Supporting; Writing – review & editing: Equal).

Nicola F. Frei, MD (Data curation: Equal; Formal analysis: Equal; Investigation: Supporting; Methodology: Supporting; Writing – review & editing: Supporting).

Roos E. Pouw, MD, PhD (Investigation: Supporting; Methodology: Supporting; Writing – review & editing: Supporting).

Christian Smolko, PhD (Data curation: Supporting; Formal analysis: Supporting; Writing – review & editing: Supporting).

Meenakshi Arora, PhD (Writing – original draft: Lead; Writing – review & editing: Equal).

Jennifer J. Siegel, MA, PhD (Data curation: Lead; Formal analysis: Lead; Writing – review & editing: Supporting).

Lucas C. Duits, MD, PhD (Data curation: Equal; Formal analysis: Equal; Investigation: Equal; Methodology: Equal; Supervision: Supporting; Writing – review & editing: Equal).

Rebecca J. Critchley-Thorne, PhD (Conceptualization: Equal; Data curation: Equal; Formal analysis: Equal; Funding acquisition: Lead; Investigation: Equal; Methodology: Equal; Project administration: Equal; Resources: Equal; Supervision: Equal; Writing – review & editing: Equal).

Jacques J.G.H.M. Bergman, MD, PhD (Conceptualization: Lead; Investigation: Equal; Methodology: Equal; Supervision: Lead; Writing – review & editing: Equal).

Conflicts of interest

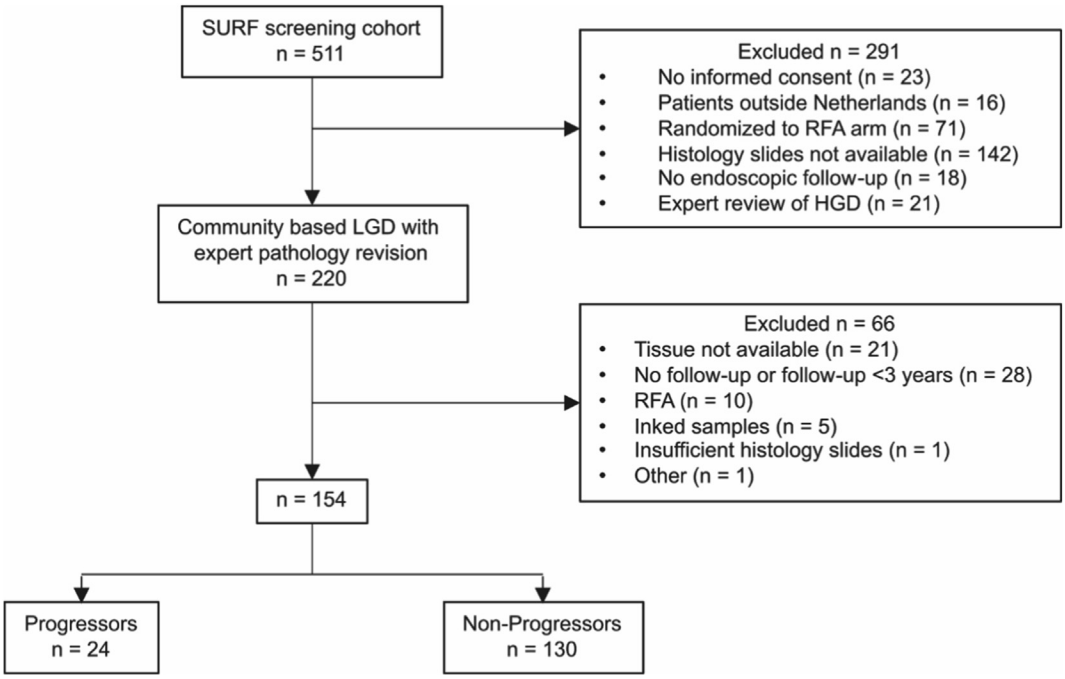
These authors disclose the following: Rebecca J. Critchley-Thorne is a full-time employee of, and holds stock and stock options in Castle Biosciences, Inc, and is an inventor on patents on the TissueCypher Barrett's Esophagus Test. Christian Smolko, Meenakshi Arora, and Jennifer J. Siegel are full-time employees of and hold stock and stock options in Castle Biosciences, Inc. Jacques J.G.H.M. Bergman has received research funding from Cernostics, Inc, Castle Biosciences, Inc, CDx Diagnostics, and Lucid Diagnostics. The remaining authors disclose no conflicts. Corporate Authorship Disclosures: United States: These authors disclose the following: John Goldblum, Jagjit Singh, Jared Szymanski, and Anthony Perry have received consulting income from Cernostics, Inc (a wholly-owned subsidiary of Castle Biosciences, Inc). Jon Davison has received consulting income from Castle Biosciences, Inc and Cernostics, Inc (a wholly-owned subsidiary of Castle Biosciences, Inc). Elizabeth Montgomery discloses no conflicts. The Netherlands: The authors disclose no conflicts. United Kingdom: Roger Feakins received funding from Alimentiv clinical Trials plc. and is consultant for Janssen, Bristol-Myers Squibb, and Decibio. The remaining authors disclose no conflicts. Germany: The authors disclose no conflicts. Belgium: The authors disclose no conflicts.

Funding

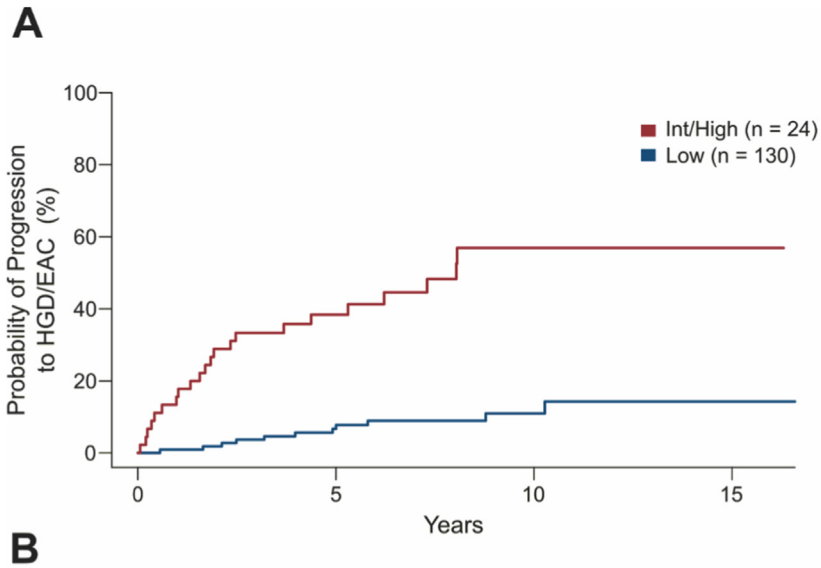
This study was partially funded by Cernostics, Inc, a wholly-owned subsidiary of Castle Biosciences, Inc USA.

Data Availability

The authors confirm that the data supporting the findings of this study are either available within the article or are available from the corresponding author on request.



Supplementary Figure 1. Flowchart of the included progressor and nonprogressor patients from the SURF screening cohort.



B

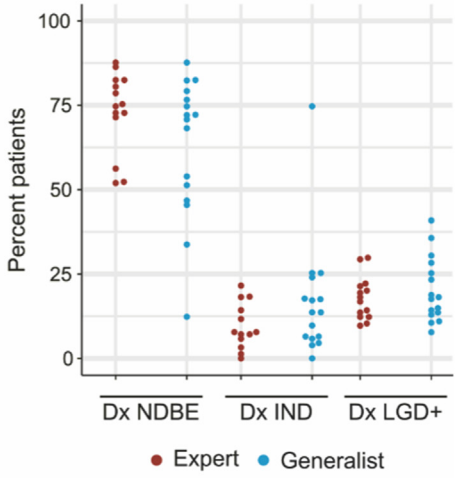
	Progressor	Non-Progressor	All Patients
TSP-9 High/Intermediate Risk	17	28	45
TSP-9 Low Risk	7	102	109
Total	24	130	154

Supplementary Figure 2. Risk stratification by TSP-9. (*A*) Kaplan-Meier (KM) analysis of probability of progression to HGD/EAC in patients with BE stratified low- and intermediate/high-risk classes by the TSP-9 test (n = 154). (*B*) Numbers of progressors and nonprogressors scoring TSP-9 low risk and intermediate/high risk.

A

Diagnosis mean [range]	All Pathologists	Expert Pathologists	Generalist Pathologists
% NDBE	67.8 [12–88]	73.2 [52–88]	63.1 [12–88]
% IND	13.0 [0–75]	8.9 [0–22]	16.6 [0–75]
% LGD+	19.2 [8–41]	17.9 [10–30]	20.3 [8–41]

B



C

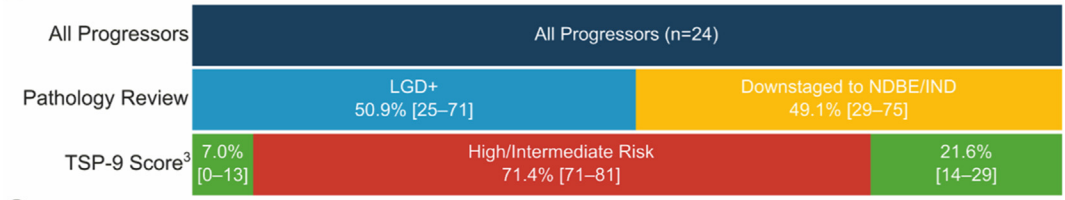
Fleiss' Kappa mean [95% CI]	All Pathologists	Expert Pathologists	Generalist Pathologists
	0.39 [0.31–0.45]	0.43 [0.34–0.50]	0.34 [0.26–0.40]

Supplementary Figure 3. Pathology review demonstrated significant interobserver variability. (*A*) Mean and range of percentage of cases diagnosed as NDBE, IND, or LGD+ by all 30 pathologists, 14 expert pathologists, and 16 generalist pathologists. (*B*) Percentage of cases diagnosed as NDBE, IND, or LGD by expert pathologists (*red*) and generalist pathologists (*blue*). (*C*) Interobserver agreement was calculated using a Fleiss' kappa coefficient.

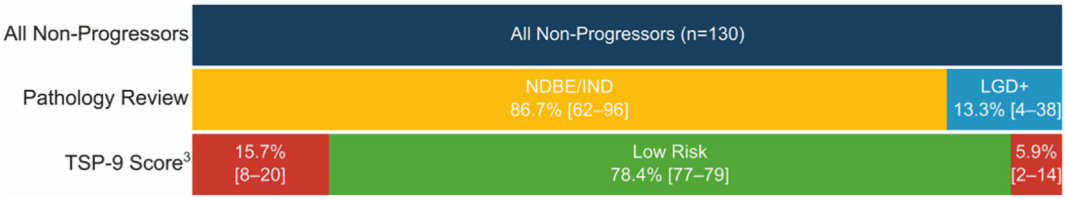
A

Prediction Metric mean [range]	TSP-9[1]	Pathology Alone
Sensitivity (%)	70.8	50.9 [25–71]
Specificity (%)	78.5	86.7 [62–96]
PPV (%)	37.8	46 [19–67]
NPV (%)	93.6	90.7 [86–94]
PPV adj.[2] (%)	23.7	23.3 [3–47]
NPV adj.[2] (%)	96.6	93.1 [88–95]

B



C



Supplementary Figure 4. The TSP-9 test finds more progressors than pathologists when the diagnosis is evaluated as LGD/higher vs NDBE and IND combined. (*A*) Prediction metrics for combined use of pathology diagnoses and the TSP-9 test. (*B*) Comparison of diagnoses (LGD/higher vs NDBE/IND) and TSP-9 test results in the subset of patients who progressed to HGD/EAC within 5 years (n = 24). (*C*) Comparison of diagnoses (LGD/higher vs NDBE/IND) and TSP-9 test results in patients who did not progress within 5 years (n = 130). [1]TSP-9 prediction metrics were calculated in the subset for which each pathologist rendered a BE diagnosis, which ranged from 141–154 patients as 5 pathologists scored a subset of cases as non-BE or not reportable. [2]Patients with TSP-9 low-risk result and pathology diagnosis of NDBE/IND were considered low risk, and patients with TSP-9 intermediate/high-risk result or a pathology diagnosis of LGD/higher were considered high-risk (see Methods). [3]TSP-9 prediction metrics in the highest scoring case-part from all 154 patients. [4]NPV and PPV were adjusted for prevalence as described in Methods.

Supplementary Table 1. Predictive Performance of TSP-9 and Pathologists

Predictor	Country	Sensitivity [95% CI]	Specificity [95% CI]	PPV [95% CI]	NPV [95% CI]
TSP-9	na	0.71 [0.54–0.88]	0.78 [0.72–0.85]	0.38 [0.29–0.49]	0.94 [0.90–0.97]
Expert	US	0.71 [0.50–0.88]	0.81 [0.74–0.87]	0.40 [0.31–0.51]	0.94 [0.90–0.97]
Expert	US	0.54 [0.33–0.75]	0.89 [0.84–0.94]	0.48 [0.34–0.65]	0.91 [0.88–0.95]
Expert	US	0.54 [0.33–0.75]	0.80 [0.73–0.87]	0.33 [0.22–0.46]	0.90 [0.87–0.94]
Generalist	US	0.88 [0.71–1.00]	0.12 [0.07–0.18]	0.16 [0.13–0.18]	0.84 [0.67–1.00]
Generalist	US	0.79 [0.62–0.96]	0.52 [0.43–0.60]	0.23 [0.18–0.28]	0.93 [0.88–0.98]
Generalist	US	0.67 [0.46–0.83]	0.79 [0.72–0.86]	0.37 [0.28–0.49]	0.93 [0.89–0.96]
Expert	Netherlands	0.75 [0.58–0.92]	0.57 [0.48–0.65]	0.24 [0.19–0.30]	0.92 [0.88–0.97]
Expert	Netherlands	0.71 [0.50–0.88]	0.61 [0.53–0.70]	0.25 [0.19–0.32]	0.92 [0.87–0.96]
Expert	Netherlands	0.67 [0.46–0.83]	0.78 [0.72–0.85]	0.36 [0.27–0.48]	0.93 [0.89–0.96]
Generalist	Netherlands	0.81 [0.62–0.95]	0.60 [0.51–0.69]	0.26 [0.20–0.32]	0.95 [0.90–0.99]
Generalist	Netherlands	0.79 [0.62–0.96]	0.36 [0.28–0.45]	0.19 [0.15–0.22]	0.90 [0.83–0.98]
Generalist	Netherlands	0.64 [0.45–0.82]	0.78 [0.71–0.85]	0.33 [0.24–0.44]	0.93 [0.89–0.96]
Expert	Belgium	0.62 [0.42–0.83]	0.79 [0.72–0.86]	0.36 [0.26–0.48]	0.92 [0.88–0.96]
Expert	Belgium	0.58 [0.38–0.79]	0.88 [0.82–0.93]	0.47 [0.33–0.62]	0.92 [0.88–0.96]
Generalist	Belgium	0.79 [0.62–0.96]	0.57 [0.48–0.65]	0.25 [0.20–0.31]	0.94 [0.89–0.99]
Generalist	Belgium	0.67 [0.46–0.83]	0.78 [0.70–0.85]	0.36 [0.26–0.46]	0.93 [0.89–0.96]
Generalist	Belgium	0.54 [0.33–0.75]	0.89 [0.84–0.94]	0.48 [0.33–0.65]	0.91 [0.88–0.95]
Expert	Germany	0.50 [0.29–0.71]	0.95 [0.91–0.98]	0.63 [0.44–0.83]	0.91 [0.88–0.95]
Expert	Germany	0.50 [0.29–0.71]	0.80 [0.73–0.87]	0.32 [0.21–0.44]	0.90 [0.86–0.94]
Expert	Germany	0.40 [0.25–0.67]	0.83 [0.77–0.89]	0.33 [0.21–0.48]	0.89 [0.86–0.93]
Generalist	Germany	0.71 [0.50–0.88]	0.85 [0.79–0.91]	0.47 [0.36–0.61]	0.94 [0.90–0.97]
Generalist	Germany	0.67 [0.46–0.83]	0.88 [0.82–0.93]	0.50 [0.37–0.64]	0.93 [0.90–0.97]
Generalist	Germany	0.67 [0.46–0.83]	0.82 [0.75–0.88]	0.41 [0.30–0.53]	0.93 [0.89–0.97]
Generalist	Germany	0.50 [0.29–0.71]	0.88 [0.82–0.94]	0.44 [0.29–0.62]	0.91 [0.87–0.94]
Expert	UK	0.79 [0.62–0.96]	0.58 [0.50–0.67]	0.26 [0.21–0.32]	0.94 [0.89–0.99]
Expert	UK	0.50 [0.29–0.71]	0.88 [0.82–0.94]	0.44 [0.30–0.62]	0.91 [0.87–0.94]
Expert	UK	0.33 [0.17–0.54]	0.90 [0.85–0.95]	0.38 [0.20–0.58]	0.88 [0.85–0.91]
Generalist	UK	0.67 [0.46–0.83]	0.75 [0.67–0.82]	0.33 [0.24–0.42]	0.92 [0.88–0.96]
Generalist	UK	0.62 [0.46–0.83]	0.47 [0.38–0.55]	0.18 [0.13–0.23]	0.87 [0.81–0.93]
Generalist	UK	0.33 [0.17–0.54]	0.92 [0.86–0.96]	0.42 [0.23–0.65]	0.88 [0.85–0.91]

na, not applicable.

Supplementary Table 2. Predictive Performance of European and US Pathologist Diagnoses (IND/LGD/higher vs NDBE)

European vs US pathologist performance[a]						
	All pathologists		Expert pathologists		Generalist pathologists	
Prediction metric	European	US	European	US	European	US
Sensitivity (%)	61.8 [33–81]	68.8 [54–88]	58.3 [33–79]	59.7 [54–71]	64.7 [33–81]	78.0 [67–88]
Specificity (%)	75.5 [36–95]	65.5 [12–89]	77.9 [57–95]	83.3 [80–89]	73.5 [36–92]	47.7 [12–79]
PPV (%)	36.1 [18–63]	32.8 [16–48]	36.7 [24–63]	40.3 [33–48]	35.5 [18–50]	25.3 [16–37]
NPV (%)	91.6 [87–95]	90.8 [84–94]	91.3 [88–94]	91.7 [90–94]	91.8 [87–95]	90.0 [84–93]
PPV adj.[b] (%)	23.1 [10–49]	20.6 [9–32]	23.6 [14–49]	26.1 [20–32]	22.7 [10–35]	15.1 [9–23]
NPV adj.[b] (%)	95.5 [93–97]	95.1 [91–97]	95.3 [93–97]	95.6 [95–97]	95.6 [93–97]	94.6 [91–96]

[a]US pathologists had access to H&E slides while European pathologists had additional access to p53-stained IHC slides to use adjunctively consistent with their respective standard practices.
[b]NPV and PPV were adjusted for prevalence ad described in Methods.

Supplementary Table 3. Predictive Performance of Pathology Diagnoses and TSP-9 Risk Results Combined

Prediction metrics mean [range]	TSP-9 + pathology diagnosis IND/LGD+ vs NDBE[a]	TSP-9 + pathology diagnosis LGD+ vs NDBE/IND[b]
Sensitivity	80.4 [71–92]	78.3 [71–86]
Specificity	62.7 [13–75]	71.9 [57–78]
PPV	29.8 [16–37]	34.5 [25–39]
NPV	94.5 [89–96]	94.7 [93–96]
PPV adj.[c]	18.0 [9–23]	21.3 [15–26]
NPV adj.[c]	97.0 [94–98]	97.2 [96–98]

[a]Patients with TSP-9 low risk result and pathology Dx of NDBE were considered low risk, and patients with TSP-9 intermediate/high-risk result or a pathology Dx of IND/LGD/higher were considered high risk (see Methods).
[b]Patients with TSP-9 low-risk result and pathology Dx of NDBE/IND were considered low risk, and patients with TSP-9 intermediate/high-risk result or a pathology Dx of LGD/higher were considered high risk (see Methods).
[c]NPV and PPV were adjusted for prevalence as described in Methods.

Supplementary Table 4. Predictive Performance of the TSP-9 Test in Patients Downstaged From LGD to NDBE/IND

	TSP-9 performance in patients who were downstaged to NDBE/IND by:					
	Individual pathologists[a]			At least 15 pathologists		
Prediction Metric	All	Experts	Generalists	All	Experts	Generalists
% of progressors who were downstaged to NDBE/IND and scored high/int-risk by the test (sensitivity)	54.4 [29–69]	54.0 [29–69]	54.8 [33–67]	54.5	61.5	54.5
% of nonprogressors who were downstaged to NDBE/IND and scored low-risk by the test (specificity)	82.1 [79–88]	81.6 [79–85]	82.4 [80–88]	81.8	81.7	82.6

[a]Mean [range].