Pathology

Subtyping intestinal metaplasia in patients with chronic atrophic gastritis: an interobserver variability study

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full length article
Keywords:	Chronic atrophic gastritis; complete intestinal metaplasia; incomplete intestinal metaplasia; interobserver agreement; observer variation; gastric precancerous lesion
Corresponding Author:	Cord Langner, M.D. Medical University of Graz Diagnostics & Research Center for Molecular BioMedicine: Medizinische Universitat Graz Diagnostik und Forschungszentrum fur Molekulare BioMedizin AUSTRIA
First Author:	Julia M. Lerch
Order of Authors:	Julia M. Lerch
	Rish K. Pai
	Ian Brown
	Anthony J. Gill
	Dhanpat Jain
	Bence Kővári
	Ryoji Kushima
	Kieran Sheahan
	Tomas Slavik
	Amitabh Srivastava
	Gregory Y. Lauwers
	Cord Langner
Abstract:	Incomplete gastric intestinal metaplasia (GIM) is associated with an increased risk of gastric cancer. We aimed to examine the interobserver variability of GIM subtyping (incomplete versus complete) in histological diagnosis of patients with chronic atrophic gastritis and to identify factors with potential impact on agreement. Nine international gastrointestinal expert pathologists assessed 46 cases with complete, incomplete or mixed-type GIM on scanned hematoxylin and eosin-stained slides. Results were compared with the consensus diagnosis driven by two experts. Interobserver variability was evaluated by kappa statistics. Focusing on the predominant pattern, the agreement between each observer and the consensus diagnosis ranged from 78% to 98%. The level of agreement was moderate to almost perfect (weighted kappa=0.464-0.984). The participating pathologists reached substantial overall agreement (Fleiss' kappa=0.716, 95% confidence interval 0.677-0.755). Misclassification with potential impact on clinical decision-making occurred in 5.7% of case ratings. The number of biopsy pieces per sample, the portion of mucosal surface involved by GIM and the pattern of GIM (pure GIM versus mixed-type GIM) had potential impact on agreement. Pathologists who apply subtyping in daily routine performed better than those who do not (p=0.040). In conclusion , subtyping GIM on hematoxylin and eosin-stained slides can be achieved satisfactorily with high interobserver agreement. The implementation of GIM subtyping as a risk-stratifying tool in current practice guidelines by the European Society of Gastrointestinal Endoscopy (ESGE) and the American Gastroenterological Association (AGA) carries a low rate of misclassification, at least among gastrointestinal expert pathologists.
Suggested Reviewers:	Michael Vieth, MD vieth.lkpathol@uni-bayreuth.de

	Gastrointestinal expert pathologist
	Neil Shepherd, MD neil.a.shepherd@nhs.net Gastrointestinal expert pathologist
	Pedro Pimentel-Nunes, MD pedronunesml@gmail.com Gastroenterologist, the first author of Updated MAPS Guidelines
	Roberto Fiocca, MD fiocca@unige.it Gastrointestinal expert pathologist
	Nina Zidar, MD Nina.ZIDAR@mf.uni-lj.si Gastrointestinal expert pathologist
Opposed Reviewers:	Massimo Rugge, MD massimo.rugge@unipd.it Prof. Rugge does not really accept gastric intestinal metaplasia as risk stratifying tool (he still promotes gastric atrophy as superior).
	Robert Genta, MD rmgenta@bcm.edu Expert gastrointestinal pathologist from US. He does not like (as most pathologists from Northern America) to perform subtyping, that is, to add subtyping to routine pathology (mainly for legal reasons).
	Elizabeth Montgomery, MD emontgom@jhmi.edu Expert gastrointestinal pathologist from US. She does not like (as most pathologists from Northern America) to perform subtyping, that is, to add subtyping to routine pathology (mainly for legal reasons).
	Rhonda Yantiss, MD rhy2001@med.cornell.edu Expert gastrointestinal pathologist from US. She does not like (as most pathologists from Northern America) to perform subtyping, that is, to add subtyping to routine pathology (mainly for legal reasons).
	Roger Feakins, MD info@rcpath.org Personal issues between Roger Feakins and the corresponding author Cord Langner dating from their joint time at ECCO / H-ECCO.

Subtyping intestinal metaplasia in patients with chronic atrophic gastritis: an interobserver variability study

Running title: Subtyping gastric intestinal metaplasia.

Julia M. Lerch[1], Rish K. Pai,[2] Ian Brown,[3,4] Anthony J. Gill,[5,6,7] Dhanpat Jain,[8] Bence Kővári,[9,10] Ryoji Kushima,[11] Kieran Sheahan,[12,13] Tomas Slavik,[14,15] Amitabh Srivastava,[16] Gregory Y. Lauwers,[10] and Cord Langner[1]

[1] Diagnostic and Research Institute of Pathology, Diagnostic and Research Center for Molecular BioMedicine, Medical University of Graz, Graz, Austria.

[2] Department of Laboratory Medicine and Pathology, Mayo Clinic Arizona, Scottsdale, Arizona, USA.

[3] Envoi Specialist Pathologists, Brisbane, Queensland, Australia.

[4] Faculty of Medicine, University of Queensland, Brisbane, Queensland, Australia.

[5] Cancer Diagnosis and Pathology Group, Kolling Institute of Medical Research, Royal North Shore Hospital, St Leonards, New South Wales, Australia.

[6] New South Wales Health Pathology, Department of Anatomical Pathology, Royal North Shore Hospital, St Leonards, New South Wales, Australia.

[7] Sydney Medical School, University of Sydney, New South Wales, Australia.

[8] Department of Pathology, Yale University School of Medicine, New Haven, Connecticut, USA.

[9] Department of Pathology, University of Szeged, Szeged, Hungary.

[10] H. Lee Moffitt Cancer Center & Research Institute, Tampa, Florida, USA.

[11] Department of Pathology, Shiga University of Medical Science, Shiga, Japan.

[12] Department of Pathology & Centre for Colorectal Disease, St. Vincent's University Hospital, Dublin, Ireland.

[13] University College Dublin, Dublin, Ireland.

[14] Ampath Pathology Laboratories, Pretoria, South Africa.

[15] Department of Anatomical Pathology, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa.

[16] Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA.

Declarations of interest: none

Address for correspondence:

Cord Langner MD

Diagnostic and Research Institute of Pathology

Diagnostic and Research Centre for Molecular BioMedicine

Medical University of Graz

Neue Stiftingtalstraße 6

8010 Graz, Austria

Tel. +43 316 385-71727 / Fax +43 316 385-79000

E-mail: cord.langner@medunigraz.at

Word Count **2578**

Abstract

Incomplete gastric intestinal metaplasia (GIM) is associated with an increased risk of gastric cancer. We aimed to examine the interobserver variability of GIM subtyping (incomplete *versus* complete) in histological diagnosis of patients with chronic atrophic gastritis and to identify factors with potential impact on agreement. Nine international gastrointestinal expert pathologists assessed 46 cases with complete, incomplete or mixed-type GIM on scanned hematoxylin and eosin-stained slides. Results were compared with the consensus diagnosis driven by two experts. Interobserver variability was evaluated by kappa statistics. Focusing on the predominant pattern, the agreement between each observer and the consensus diagnosis ranged from 78% to 98%. The level of agreement was moderate to almost perfect (weighted kappa=0.464-0.984). The participating pathologists reached substantial overall agreement (Fleiss' kappa=0.716, 95% confidence interval 0.677-0.755). Misclassification with potential impact on clinical decision-making occurred in 5.7% of case ratings. The number of biopsy pieces per sample, the portion of mucosal surface involved by GIM and the pattern of GIM (pure GIM *versus* mixed-type GIM) had potential impact on agreement. Pathologists who apply subtyping in daily routine performed better than those who do not (p=0.040). In conclusion**,** subtyping GIM on hematoxylin and eosin-stained slides can be achieved satisfactorily with high interobserver agreement. The implementation of GIM subtyping as a risk-stratifying tool in current practice guidelines by the European Society of Gastrointestinal Endoscopy (ESGE) and the American Gastroenterological Association (AGA) carries a low rate of misclassification, at least among gastrointestinal expert pathologists.

Key Words Chronic atrophic gastritis; complete intestinal metaplasia; incomplete intestinal metaplasia; interobserver agreement; observer variation; gastric precancerous lesion

1. Introduction

Gastric cancer represents the third leading cause of cancer-related mortality and the fifth most common cancer worldwide.[1] Early detection of neoplastic lesions allows for less invasive treatment, such as endoscopic mucosal resection or endoscopic submucosal dissection and decreases gastric cancer mortality. Therefore, international guidelines recommend screening and surveillance of people at risk.[2]

Chronic gastritis, in particular when associated with atrophy and intestinal metaplasia, has been identified as a major risk factor for gastric cancer. A well-defined sequential carcinogenesis, known as Correa's cascade, involves the following steps: non-atrophic chronic gastritis → (multifocal) atrophic gastritis → intestinal metaplasia → dysplasia (low grade and high grade) → invasive adenocarcinoma.[3] Infection with *Helicobacter pylori* is accepted to be the primary driver for this progression, although other triggers such as autoimmune gastritis have been recognized.[4] Atrophy and gastric intestinal metaplasia (GIM), which are collectively known as chronic atrophic gastritis, are pre-neoplastic lesions that can be used for risk stratification.[5-7]

Two types of GIM are recognized (Fig. 1): Complete GIM resembles a small intestinal phenotype with eosinophilic enterocytes displaying a well-defined brush border and well-formed goblet cells. Paneth cells are frequently observed at the base of the crypts. Alternatively, incomplete GIM shows goblet cells of variable size and intervening mucin-secreting columnar cells without brush border. Paneth cells are absent in this subtype. Mixed characteristics of complete and incomplete subtypes are sometimes observed, even in a single biopsy.[8,9] The traditional approach to differentiate GIM into three groups, that is,

type I as complete and types II and III as incomplete, requires sulfomucin or sialomucin staining which is not performed routinely.[8,9]

The extent of atrophy and GIM has been associated with an increased gastric cancer risk in multiple studies, resulting in different staging systems, such as the Operative Link for Gastritis Assessment (OLGA)[10] or the Operative Link on Gastric Intestinal Metaplasia (OLGIM)[11]. In addition, several studies documented a higher risk of progression for incomplete GIM compared to complete GIM.[12-19] According to a recently published technical review of the American Gastrointestinal Association (AGA), incomplete GIM bears a 1.7-fold higher risk of progression to dysplasia and a 3.3-fold higher risk of cancer compared to complete GIM.[4] Consequently, endoscopic surveillance based upon GIM subtyping was implemented in the updated MAPS-Guidelines of the European Society of Gastrointestinal Endoscopy (ESGE)[6] and is also considered as valuable tool in the recently released AGA Clinical Practical Guidelines on the management of GIM.[7]

Despite these recent recommendations, it has not been proven that pathologists can differentiate between the two GIM subtypes with sufficient agreement, as a systematic interobserver variability study is currently lacking. Therefore, we herein aimed to evaluate the interobserver variation in GIM subtyping involving an international group of gastrointestinal expert pathologists. Furthermore, this study addresses the impact of misclassification on patient management and the identification of potential causative factors.

2. Material and Methods

2.1. Cases

The study included antral biopsies from 46 patients with chronic atrophic gastritis, diagnosed at the Diagnostic- and Research- (D&F) Institute of Pathology, Medical University of Graz, Austria, within the period of 11/2019 to 03/2020. All biopsies had been obtained based on Sydney criteria[20], that is, targeting the lesser and greater curvature, excluding the normal gastroduodenal transitional mucosa. It may be of note that corpus and/or fundus biopsies, which had been submitted in separate vials, lacked GIM in all cases and were therefore not part of the evaluation. Since Austria is a country with a low prevalence of *Helicobacter pylori*, resulting in a low incidence of GIM in general and a predominance of complete GIM,[21] we selected the study sample in order to enrich for cases with the incomplete subtype. Table 1 shows the case characteristics including age and gender distribution, *Helicobacter pylori* status and the extent and type of GIM.

Before starting the study, sample size calculation was performed by Dr. Josef Haas, Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, with power of 80% and significance level at 5%. All samples were routinely stained with hematoxylin and eosin (H&E) and scanned thereafter (Pannoramic 1000 Whole-Slide Scanner, 3D Histech Ltd., Budapest, Hungary).

Institutional Review Board approval was obtained from the Ethics Committee of the Medical University of Graz, Austria (EK 33-444 ex20/21).

2.2. Pathologists

Two gastrointestinal expert pathologists (CL and GL), who are used to applying GIM subtyping in routine diagnosis, independently viewed the scans, performed GIM subtyping, and classified the cases into the following categories:

1. Complete GIM alone
2. Mixed, complete exceeding incomplete GIM
3. Mixed, incomplete exceeding complete GIM
4. Incomplete GIM alone

The two pathologists reached initial agreement in 38 cases out of 46 (83%). In case of disagreement, a consensus ("gold standard") was obtained by joint microscopy and case discussion.

Nine international gastrointestinal expert pathologists were invited to participate as observers in the study. Inclusion criteria were completed specialty training for pathology and proven publication activity in gastrointestinal pathology. Two of the observers reported to apply GIM subtyping in their daily routine practice, the other seven did not.

Access to scanned slides was provided by an electronically transferred web link. The assessment was done in a blinded fashion on dynamic images (3D Histech Ltd. Case Viewer, Budapest, Hungary) with viewing between x2 and x40 magnification. Every observer classified each case applying the four categories mentioned above using a standardized evaluation sheet.

2.3. Statistical Analysis

Categorical variables are presented as numbers and percentages. Numerical variables are presented as mean, median and range. Differences in categorical variables were examined using the chi-square test or Fisher's exact test, as appropriate. Differences in numerical

variables were examined using the t-test for two independent samples as parametric test procedure and the Mann-Whitney-U test as non-parametric test procedure.

Interobserver variability was assessed by applying kappa statistics, which are used to quantify the degree of agreement beyond chance.[22] The level of agreement between each observer and the consensus diagnosis was calculated using weighted kappa for all categories (1-4) as well as for combined categories focusing on the predominant GIM pattern (1 and 2 *versus* 3 and 4). Fleiss' kappa was used to calculate the agreement among the pathologists as follows: the scores from the nine observers and the original scores from the two pathologists who provided the consensus diagnosis were included in this calculation.[23]

Kappa values were interpreted according to the scheme of Landis and Koch,[24] modified by Altmann:[25] kappa values <0.00 suggest no agreement, 0.00-0.20 poor agreement, 0.21-0.40 slight agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and 0.81-1.00 almost perfect agreement, respectively.

All statistical operations were performed using IBM SPSS Statistics Version 26, provided by the Medical University of Graz. P-values were two-sided, and values <0.05 were considered statistically significant.

3. Results

3.1. Interobserver variability in GIM subtyping

Table 2 presents the interobserver variability between the nine observers and the consensus diagnosis. For four GIM categories, the interobserver variability ranged from slight to almost perfect agreement (kappa=0.368-0.961). Specifically, two observers reached almost perfect agreement with the consensus, five observers substantial agreement, two observers moderate agreement and one observer slight agreement, respectively. Diagnostic agreement between each observer and the consensus diagnosis ranged from 48% to 96%, with more than 70% agreement for six observers.

For two GIM categories, that is, addressing the predominant pattern (complete, 1 and 2, *versus* incomplete, 3 and 4), the interobserver variability ranged from moderate to almost perfect agreement (kappa=0.464-0.984). Specifically, five observers reached almost perfect agreement, two observers substantial agreement and another two observers moderate agreement, respectively. Diagnostic agreement between each observer and the consensus diagnosis ranged from 78% to 98%.

For four categories, the two observers who apply GIM subtyping in daily routine practice had a higher agreement and higher kappa values (p=0.040) than the other observers, while no significant difference was noted for two GIM categories (p=0.241).

Table 3 shows the interobserver variability among all eleven pathologists (the nine observers and the two pathologists who had agreed on the consensus diagnosis). For four categories, the overall agreement was moderate (kappa=0.447, 95% CI (0.423-0.471)). Applying two categories, the overall agreement was substantial (kappa=0.716, 95% CI (0.677-0.755)).

3.2. Factors with potential impact on agreement

Factors that may be associated with the level of agreement are illustrated in Table 4. Herein, we compared cases with the highest agreement (six cases with 100% agreement) with cases with the lowest agreement (one case with 22%, two cases with 33%, and five cases with 44% agreement). We identified three parameters with potential impact. Agreement was significantly lower (p=0.010) in mixed cases (categories 2 and 3) compared to cases with only one type of GIM (categories 1 and 4). Likewise, the number of biopsy specimens within the sample and the portion of mucosal surface involved by GIM (in particular when more than one biopsy piece was affected) were both higher in the cases with low agreement, but this difference was not statistically significant (p=0.886 and p=0.120, respectively), likely due to small sample size.

3.3. Clinical consequences of misclassification

The updated MAPS Guidelines[6] recommend endoscopic surveillance for all patients with incomplete GIM, while patients with complete GIM do not need follow-up when metaplastic change is found only in the antrum. The recommendation regarding incomplete GIM is based upon the publications by Gonzalez et al., who defined incomplete GIM as either pure incomplete GIM or incomplete GIM as predominant pattern in mixed cases.[13-15]

Thus, patients with pure complete GIM (category 1) may be “overreported” when misclassified as predominantly incomplete or pure incomplete GIM (category 3 or 4), thereby inducing unnecessary follow-up investigations. On the contrary, patients with predominantly incomplete or pure incomplete GIM (category 3 or 4) may be lost to follow-up when misclassified and thereby “underreported” as pure complete GIM (category 1). The

definitive clinical significance of category 2 (incomplete GIM as a minor pattern) still needs to be elucidated.

Table 5 illustrates the clinical consequences of misclassification in our study. Overall, misclassification potentially inducing an erroneous follow-up decision occurred in 19 out of 333 (5.7%) ratings, with “underreporting” occurring more often than “overreporting”.

4. Discussion

The present study demonstrates that differentiating between complete and incomplete GIM subtypes can be achieved satisfactorily on H&E-stained slides, with eleven pathologists reaching substantial overall agreement. When two categories (predominant pattern; categories 1 and 2 *versus* 3 and 4) are applied, subtyping of GIM provided moderate to almost perfect agreement between observers and the consensus diagnosis. Misclassification with potential clinical impact occurred in only about 5% of diagnostic decisions (case ratings).

The presence of incomplete GIM is known to be associated with the extent of total metaplasia[9,26] and may develop from pre-existing complete GIM or develop *de novo*.[8] Thus, complete and incomplete GIM may coexist. While the clinical relevance of cases with a minor incomplete component still needs to be defined,[27] the prognostic relevance of incomplete GIM as a major constituent has been proven on H&E-stained slides.[13-15]

In our study, all patients showed GIM restricted to the antrum (limited GIM). No surveillance would have been scheduled on the basis of GIM extent. However, in 13 out of 46 patients (28%) incomplete GIM is present alone or as predominant component. These patients would have been lost to follow-up, since in this subgroup the GIM subtype represents the only decisive factor.

In contrast to non-metaplastic gastric atrophy, interobserver agreement for GIM is high, with agreement levels among gastrointestinal expert pathologists ranging from 86% to 92% and kappa values ranging from 0.65 to 0.90, respectively. Values for general pathologists are slightly lower.[11, 28-30]

Our study is the first to evaluate the interobserver variability in GIM subtyping. Although all study pathologists were international gastrointestinal expert pathologists, only two of them reported to apply GIM subtyping in their daily routine practice. All nine pathologists reached excellent values; still, the two pathologists who routinely apply GIM subtyping achieved the highest level of agreement and the highest kappa values. This observation supports the study by Kim et al. who demonstrated improved kappa values in GIM diagnosis with increasing experience among the participating pathologists.[30]

While GIM subtyping is routinely performed in several European countries, others, such as the US, have not yet implemented GIM subtyping in daily routine practice on a general scale. This observation has given rise to concerns as to whether the histologic subtype of GIM can be utilized as part of risk stratification without a substantial educational initiative for pathologists.[7] Our study and the study by Kim et al.[30] indicate that histopathological knowledge required for diagnosis on H&E-stained slides can be gained in due time and applied in routine diagnosis without additional costs. This notion complies with the recently published guidelines by ESGE and AGA who both recognized incomplete GIM as an important risk-stratifying feature, but contrasts with the British Society of Gastroenterology (BSG) guidelines that have not yet implemented GIM subtyping in decision-making.[5-7]

According to ESGE and AGA guidelines, patients with complete GIM limited to the antrum do not require surveillance, while patients with incomplete GIM do.[6,7] It appears that the high proportion of mixed GIM cases is not taken into account. The ESGE refers to the publications by Gonzalez et al.[13-15] who established the prognostic impact for patients with pure incomplete or predominant incomplete GIM but did not consider patients with a minor

incomplete component.[6] The AGA states that patients with a higher risk of gastric cancer include those with "at least partial" incomplete GIM.[7]

For our study analysing interobserver variability and potential clinical impact of misclassification, we did not consider cases with a minor incomplete component as this situation warrants further scientific efforts.[27]

In our study, we aimed to identify factors associated with a higher risk of misclassification. Prior studies have established that the likelihood of detecting GIM on gastric biopsies correlates with the number of biopsies obtained.[31] Not surprisingly, the number of biopsy pieces, the portion of mucosal surface involved by GIM and the pattern of GIM (pure GIM *versus* mixed GIM) had potential impact on the pathologists' ratings and thereby on the level of agreement.

Our study has several strengths and limitations. Strengths include the systematic approach involving a large international group of gastrointestinal expert pathologists who analysed a large set of biopsies representing all potential patterns of GIM including mixed cases. The pathologists had varying routine experience in GIM subtyping and originated from different parts of the world, with different education and different approach to diagnosis. Still, the lack of general pathologists in this study may be regarded as a limitation by some. The restriction to H&E-stained slides could likewise be regarded as a limitation. However, H&E histology has proven prognostic impact in prior publications. Furthermore, enzyme-histochemical and/or immunohistochemical staining methods are not generally applied and have no proven additional prognostic impact. Another limitation might be the use of virtual microscopy, which bears specific technical challenges: pathologists may find it harder to move around all biopsy specimens with the same ease they do on a microscope. In addition, the evaluation of scanned slides does not allow the assessment of more than one level,

thereby potentially hampering the identification of brush borders and Paneth cells. However, the findings in our study are still relevant in view of the expected increase in use of virtual diagnostics in the future.

In conclusion, subtyping GIM on H&E-stained slides in patients with chronic atrophic gastritis can be achieved satisfactorily with high interobserver agreement. Pathologists who apply subtyping in daily routine performed better than those who do not. The implementation of GIM subtyping as a risk-stratifying tool in currently updated practice guidelines carries a low rate of misclassification, at least among gastrointestinal expert pathologists.

Table 1: Case characteristics (n=46).

Gender	Female	20 (43%)
	Male	26 (57%)
Age (years)	Mean	65.8
	Median	69
	Range	27-87
Helicobacter pylori status	*Helicobacter pylori* positive	10 (22%)
	Helicobacter pylori negative	36 (78%)
	Reactive gastropathy	9 (25%)
Portion of mucosal surface involved by intestinal metaplasia	Mean	31%
	Median	20%
	Range	10-90%
Type of gastric intestinal metaplasia (consensus diagnosis)	Complete	24 (52%)
	Mixed (complete > incomplete)	9 (20%)
	Mixed (incomplete > complete)	8 (17%)
	Incomplete	5 (11%)
Number of biopsy pieces	Mean	2.7
	Median	2
	Range	1-6

Table 2: Weighted kappa values (95% CI) and agreement (n; %) between the nine observers and the consensus diagnosis. The two observers who perform subtyping of gastric intestinal metaplasia (GIM) in daily routine practice are highlighted in grey.

Observer	Observer *versus* consensus (four GIM categories)	Observer *versus* consensus (two GIM categories, predominant pattern; 1 and 2 *versus* 3 and 4)	Agreement (four GIM categories)	Agreement (two GIM categories, predominant pattern; 1 and 2 *versus* 3 and 4)
#1	0.781 (0.664-0.899)	0.948 (0.846-1.049)	35 (76%)	45 (98%)
#2	0.961 (0.907-1.015)	0.948 (0.846-1.049)	44 (96%)	45 (98%)
#3	0.520 (0.393-0.678)	0.649 (0.415-0.884)	22 (48%)	39 (85%)
#4	0.368 (0.160-0.577)	0.464 (0.180-0.747)	22 (48%)	36 (78%)
#5	0.857 (0.756-0.959)	0.888 (0.736-1.039)	39 (85%)	44 (96%)
#6	0.769 (0.648-0.891)	0.786 (0.586-0.985)	34 (74%)	42 (91%)
#7	0.777 (0.657-0.897)	0.738 (0.523-0.953)	36 (78%)	41 (89%)
#8	0.793 (0.686-0.900)	0.898 (0.759-1.036)	35 (76%)	44 (96%)
#9	0.654 (0.502-0.806)	0.843 (0.671-1.014)	29 (63%)	43 (93%)

Table 3 Fleiss' kappa values (95% CI) including all pathologists (the nine observers and the two pathologists who produced the consensus diagnosis), who participated in subtyping gastric intestinal metaplasia (GIM).

	Overall	Per category
Four GIM categories	0.447 (0.423-0.471)	1 = 0.581 (0.542-0.620) 2 = 0.261 (0.222–0.300) 3 = 0.323 (0.284-0.362) 4 = 0.569 (0.530-0.608)
Two GIM categories (predominant pattern; 1 and 2 *versus* 3 and 4)	0.716 (0.677-0.755)	1 and 2 = 0.716 (0.677-0.755) 3 and 4 = 0.716 (0.677-0.755)

Table 4: Factors with potential impact on the agreement between the observers and the consensus diagnosis in subtyping gastric intestinal metaplasia (GIM), illustrated by cases with highest and lowest agreement (n; %).

		Cases (n=6) with highest agreement	Cases (n=8) with lowest agreement	p-value
GIM categories	1 and 4 (pure GIM types)	6 (100%)	2 (20%)	0.010
	2 and 3 (mixed GIM types)	0 (0%)	6 (80%)	
Mean number of biopsy pieces		2.33	2.75	0.886
Mean portion of mucosal surface involved by intestinal metaplasia		15%	34%	0.120

Table 5: Clinical consequences of misclassification of gastric intestinal metaplasia according to the updated MAPS Guidelines of the European Society of Gastrointestinal Endoscopy (ESGE).[6] Cases in which overreporting (n=5) or underreporting (n=14) may have impact on clinical management are highlighted in grey.

Consensus diagnosis (gold standard)	Observer diagnosis	Observer #1	Observer #2	Observer #3	Observer #4	Observer #5	Observer #6	Observer #7	Observer #8	Observer #9
Category 1 (n=24; endoscopic surveillance not recommended)	1	21	21	14	15	20	22	23	22	15
	2	3	3	10	6	4	2	0	4	8
	3	0	0	0	3	0	0	1	0	1
	4	0	0	0	0	0	0	0	0	0
Categories 3 and 4 (n=13; endoscopic surveillance recommended)	1	0	0	1	3	0	0	0	0	0
	2	0	0	1	2	2	2	2	0	1
	3	4	8	1	7	7	4	8	7	6
	4	9	5	10	1	4	7	3	6	6
Misclassification with potential impact on endoscopic surveillance		0/37 (0%)	0/37 (0%)	2/37 (5%)	8/37 (22%)	2/37 (5%)	2/37 (5%)	3/37 (8%)	0/37 (0%)	2/37 (5%)

5. Figure Legends

Figure 1: Subtypes of gastric intestinal metaplasia (GIM).

(**A**) Complete GIM is characterized by eosinophilic enterocytes with well-defined brush border admixed with well-formed goblet cells (H&E, original x100). (**B**) Incomplete GIM shows goblet cells of variable size and intervening mucin-secreting columnar cells (gastric foveolar cells) without brush border (H&E, original x100). (**C**) Paneth cells are frequently observed at the base of the crypts, note enterocytes with brush border (arrow; H&E, original x150). (**D**) Mixed type with characteristics of complete (arrow) and incomplete (asterisk) subtypes within a single biopsy (H&E, original x100).

6. References

1. Sung H, Ferlay J, Siegel RL, *et al*. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021; 71: 209-49.
2. Matysiak-Budnik T, Camargo MC, Piazuelo MB, *et al*. Recent Guidelines on the Management of Patients with Gastric Atrophy: Common Points and Controversies. *Dig Dis Sci* 2020; 65: 1899-1903.
3. Correa P. A human model of gastric carcinogenesis. *Cancer Res* 1988; 48: 3554-60.
4. Gawron AJ, Shah SC, Altayar O, *et al*. AGA Technical Review on Gastric Intestinal Metaplasia-Natural History and Clinical Outcomes. *Gastroenterology* 2020; 158: 705-31.
5. Banks M, Graham D, Jansen M, *et al*. British Society of Gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma. *Gut* 2019; 68: 1545-75.
6. Pimentel-Nunes P, Libânio D, Marcos-Pinto R, *et al*. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. *Endoscopy* 2019; 51: 365-88.
7. Gupta S, Li D, El Serag HB, *et al*. AGA Clinical Practice Guidelines on Management of Gastric Intestinal Metaplasia. *Gastroenterology* 2020; 158: 693-702.
8. Correa P, Piazuelo MB, Wilson KT. Pathology of gastric intestinal metaplasia: clinical implications. *Am J Gastroenterol* 2010; 105: 493-8.

9. Shah SC, Gawron AJ, Mustafa RA, *et al*. Histologic Subtyping of Gastric Intestinal Metaplasia: Overview and Considerations for Clinical Practice. *Gastroenterology* 2020; 158: 745-50.

10. Rugge M, Meggio A, Pennelli G, *et al*. Gastritis staging in clinical practice: the OLGA staging system. *Gut* 2007; 56: 631-6.

11. Capelle LG, de Vries AC, Haringsma J, *et al*. The staging of gastritis with the OLGA system by using intestinal metaplasia as an accurate alternative for atrophic gastritis. *Gastrointest Endosc* 2010; 71: 1150-8.

12. Dinis-Ribeiro M, Lopes C, da Costa-Pereira A, *et al*. A follow up model for patients with atrophic chronic gastritis and intestinal metaplasia. *J Clin Pathol* 2004; 57: 177-82.

13. González CA, Pardo ML, Liso JM, *et al*. Gastric cancer occurrence in preneoplastic lesions: a long-term follow-up in a high-risk area in Spain. *Int J Cancer* 2010; 127: 2654-60.

14. González CA, Sanz-Anquela JM, Gisbert JP, *et al*. Utility of subtyping intestinal metaplasia as marker of gastric cancer risk. A review of the evidence. *Int J Cancer* 2013; 133: 1023-32.

15. González CA, Sanz-Anquela JM, Companioni O, *et al*. Incomplete type of intestinal metaplasia has the highest risk to progress to gastric cancer: results of the Spanish follow-up multicenter study. *J Gastroenterol Hepatol* 2016; 31: 953-8.

16. Pittayanon R, Rerknimitr R, Klaikaew N, *et al*. The risk of gastric cancer in patients with gastric intestinal metaplasia in 5-year follow-up. *Aliment Pharmacol Ther* 2017; 46: 40-5.

17. Shao L, Li P, Ye J, *et al*. Risk of gastric cancer among patients with gastric intestinal metaplasia. *Int J Cancer* 2018; 143: 1671-7.

18. Chapelle N, Péron M, Quénéhervé L, *et al*. Long-Term Follow-up of Gastric Precancerous Lesions in a Low GC Incidence Area. *Clin Transl Gastroenterol* 2020; 11: e00237. doi: 10.14309/ctg.0000000000000237.
19. Wei N, Zhou M, Lei S, *et al*. A meta-analysis and systematic review on subtypes of gastric intestinal metaplasia and neoplasia risk. *Cancer Cell Int* 2021; 21: 173.
20. Dixon MF, Genta RM, Yardley JH, *et al*. Classification and grading of gastritis. The updated Sydney System. International Workshop on the Histopathology of Gastritis, Houston 1994. *Am J Surg Pathol* 1996; 20: 1161-81.
21. Hooi JKY, Lai WY, Ng WK, *et al*. Global Prevalence of Helicobacter pylori Infection: Systematic Review and Meta-Analysis. *Gastroenterology* 2017; 153: 420-9.
22. Fleiss J. *Statistical methods for rates and proportions*. New York: Wiley, 1981.
23. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Measures of agreement. *Perspect Clin Res* 2017; 8: 187-91.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-74.
25. Altman D. Practical statistics for medical research. London: Chapman and Hall; 1991.
26. Cassaro M, Rugge M, Gutierrez O, *et al*. Topographic patterns of intestinal metaplasia and gastric cancer. *Am J Gastroenterol* 2000; 95: 1431-8.
27. Bhandari P, Longcroft-Wheaton G, Libanio D, *et al*. Revising the European Society of Gastrointestinal Endoscopy (ESGE) research priorities: a research progress update. *Endoscopy* 2021; 53: 535-54.
28. Leja M, Funka K, Janciauskas D, *et al*. Interobserver variation in assessment of gastric premalignant lesions: higher agreement for intestinal metaplasia than for atrophy. *Eur J Gastroenterol Hepatol* 2013; 25: 694-9.

29. Isajevs S, Liepniece-Karele I, Janciauskas D, *et al*. Gastritis staging: interobserver agreement by applying OLGA and OLGIM systems. *Virchows Arch*. 2014; 464: 403-7.

30. Kim SS, Kook MC, Shin OR, *et al*. Factors to improve the interobserver agreement for gastric atrophy and intestinal metaplasia: consensus of definition and criteria. *Histopathology* 2018; 72: 838-45.

31. Altayar O, Davitkov P, Shah SC, *et al*. AGA Technical Review on Gastric Intestinal Metaplasia-Epidemiology and Risk Factors. *Gastroenterology* 2020; 158: 732-44.

7. Acknowledgements

The authors thank Prof. Dr. Josef Haas, Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz for statistical support during the preparation for this study.

Ethics Approval

Institutional Review Board approval was obtained from the Ethics Committee of the Medical University of Graz, Austria (EK 33-444 ex20/21). This study was performed in accordance with the Declaration of Helsinki.

Author Contribution Statement

Cord Langner, Julia M. Lerch and Gregory Y. Lauwers designed the study. Rish K. Pai, Ian Brown, Anthony J. Gill, Dhanpat Jain, Bence Kővári, Ryoji Kushima, Kieran Sheahan, Tomas Slavik, Amitabh Srivastava, Gregory Y. Lauwers and Cord Langner acquired data. Julia M. Lerch, Rish K. Pai, Gregory Y. Lauwers and Cord Langner analyzed and interpreted the data. Julia M. Lerch and Cord Langner drafted the manuscript. Rish K. Pai, Ian Brown, Anthony J. Gill, Dhanpat Jain, Bence Kővári, Ryoji Kushima, Kieran Sheahan, Tomas Slavik, Amitabh Srivastava, Gregory Y. Lauwers critically revised the manuscript for important intellectual content. All authors read and approved the final version of the article.

Funding Statement

There is no extra-institutional funding to declare.

Data Availability Statement

The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

Figure 1