USCAP

ARTICLE

Reproducibility of scoring criteria for HER2 immunohistochemistry in endometrial serous carcinoma: a multi-institutional interobserver agreement study

Natalia Buza [1] · Elizabeth D. Euscher [2] · Xavier Matias-Guiu[3] · Austin McHenry[1] · Esther Oliva[4] ·
Zehra Ordulu [4] · Carlos Parra-Herran [5] · Douglas Rottmann [1] · Bradley M. Turner[6] · Serena Wong[1] · Pei Hui [1]

Abstract

Targeted anti-human epidermal growth factor receptor 2 (HER2) therapy has recently been proven to improve progression-free and overall survival of patients with advanced stage or recurrent endometrial serous carcinoma. To date, no specific pathology HER2 testing or scoring guidelines exist for endometrial cancer. However, based on evidence from the recent successful clinical trial and comprehensive pre-trial pathologic studies, a new set of HER2 scoring criteria have been proposed for endometrial serous carcinoma—distinct from the existing breast and gastric cancer-specific criteria. We present the first study assessing interobserver agreement of HER2 scores using the proposed serous endometrial cancer-specific scoring system. A digitally scanned set of 40 HER2-immunostained slides of endometrial serous carcinoma were sent to seven gynecologic pathologists, who independently assigned HER2 scores for each slide following a brief tutorial. Follow-up fluorescent in situ hybridization (FISH) for *HER2* gene amplification was performed on cases with interobserver disagreement when a 2+ HER2 score was assigned by at least one observer. Complete agreement of HER2 scores among all 7 observers was achieved on 15 cases, and all but one case had an agreement by at least 4 observers. The overall agreement was 72.3% (kappa 0.60), 77.5% (kappa 0.65), and 83.3% (kappa 0.65), using four (0 to 3+), three (0/1+, 2+, 3+), or two (0/1+, 2/3+) HER2 scoring categories, respectively. Based on the combination of HER2 immunostaining scores and FISH, the interobserver disagreement may have potentially resulted in a clinically significant difference in HER2 status only in three tumors. We conclude, that the proposed serous endometrial cancer-specific HER2 scoring criteria are reproducible among gynecologic pathologists with moderate to substantial interobserver agreement rates comparable to those of previously reported in breast and gastric carcinomas. Our findings significantly strengthen the foundation for establishing endometrial cancer-specific HER2 scoring guidelines in the future.

Introduction

Targeted therapy against human epidermal growth factor receptor 2 (HER2) has been a successful treatment option for patients with HER2-positive breast cancer for more than two decades and for patients with gastric carcinoma since 2010 [1]. Recently, it was shown that addition of

✉ Natalia Buza
    natalia.buza@yale.edu

1   Department of Pathology, Yale School of Medicine, New Haven, CT, USA

2   Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

3   Departments of Pathology, Hospital U Arnau de Vilanova and Hospital U de Bellvitge, IRBLleida, IDIBELL, Universities of Lleida and Barcelona, AECC grupos estables, CIBERONC, Lleida, Spain

4   Department of Pathology, Massachusetts General Hospital, Boston, MA, USA

5   Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA

6   Department of Pathology and Laboratory Medicine, University of Rochester Medical Center, Rochester, NY, USA

Table 1 Proposed HER2 scoring system for endometrial serous carcinoma based on the recent clinical trial patient enrollment criteria.

HER2 score	Staining pattern
0	No staining in tumor cells
1+	Faint/barely perceptible, incomplete membrane staining in any proportion, or weak complete staining in <10% of tumor cells
2+	Strong complete or basolateral/lateral membrane staining in ≤30%, or weak to moderate staining in ≥10% of tumor cells
3+	Strong complete or basolateral/lateral membrane staining in >30% of tumor cells

trastuzumab (a humanized monoclonal anti-HER2 anti-body) to standard chemotherapy resulted in significant improvement in progression-free and overall survival in HER2-positive advanced stage and recurrent endometrial serous carcinoma [2, 3], a highly aggressive subtype of endometrial cancer. This therapeutic breakthrough was endorsed by the National Comprehensive Cancer Network Guidelines for Uterine Neoplasms in 2019 [4]. Pathologic evaluation of tumor HER2 protein expression and gene amplification is a critical part of appropriate patient selection for this targeted therapeutic approach.

Tumors in different organ systems have distinct characteristics of HER2 protein expression and gene amplification, which should be taken into account for successful targeted therapy. Thus, tumor-specific HER2 testing algorithm and scoring guidelines have been developed for breast and gastric carcinomas based on clinical trial experience incorporating the unique features of HER2 expression/amplification in these tumor types [5–8]. Similarly, endometrial serous carcinoma has been found to have characteristic HER2 protein expression and gene amplification patterns, which were incorporated in the recent clinical trial patient enrollment criteria, using a 30% tumor cell staining cut-off for a 3+ score [2, 9]. To date, the clinical trial reported by Fader et al. is the only study with successful HER2-targeted therapy for endometrial serous carcinoma, and their patient enrollment HER2 testing and scoring protocol is currently the only one proven to correlate with therapeutic response in this tumor type [2, 3]. Based on the trial criteria, specific HER2 testing algorithm has been recently proposed for routine pathology evaluation of endometrial serous carcinoma [10, 11].

As an important step toward the development of standard HER2 scoring guidelines for endometrial serous carcinoma, we set out to evaluate the reproducibility of the proposed criteria among academic gynecologic pathologists using a virtual slide set of digitally scanned HER2-immunostained slides.

Our study demonstrated moderate to substantial inter-observer agreement, comparable to prior literature in breast and gastric cancer [12–17] and highlighted specific challenges in interpretation of HER2 immunohistochemistry in endometrial serous carcinoma.

Materials and methods

Case selection

A set of 40 HER2-immunostained slides of pure endometrial serous carcinoma was selected from a prior study [18] by one of the authors (NB) to include a wide range of staining patterns and HER2 scores. The HER2 score distribution mirrored that of the previously observed large series ($n = 108$) of endometrial serous carcinoma using the modified ASCO/CAP 2007 guidelines: $3 + 23\%$, $2 + 21\%$, $1+$, 40%, and 0 16% [9]. Patient clinicopathologic characteristics were retrieved from the archived pathology reports. Pathology slide review (H&E and diagnostic immunostains) was performed as part of the prior study [18]. The EP3 antibody clone (Abcam; Cambridge, MA, USA) was used for HER2 immunohistochemistry for all cases, the technical details of staining have been previously published [18]. On-slide positive controls were included on two slides, the remaining 38 cases had batch controls only. HER2-immunostained whole slides for each case were scanned with Aperio slide scanner (Leica Biosystems, Inc., Buffalo Grove, IL, USA) and the images were collated into a virtual slide set (https://pathpresenter.net/#/public/display?token=269097f2). A brief tutorial was prepared by two of the authors (NB, DR) using the recently proposed specific HER2 scoring criteria for endometrial serous carcinoma [10, 11] (Table 1; Supplementary material 1). A link to the slide set was sent to co-authors along with the HER2 scoring tutorial.

Evaluation of HER2 immunohistochemistry

Seven gynecologic pathologists (ZO, SW, BMT, CP-H, EDE, XM-G, and EO) with different levels of academic clinical practice experience were asked to assign a HER2 score (0, 1+, 2+, or 3+) for each slide blinded to the original HER2 score. Heterogeneity of HER2 protein expression was defined as a two-point or greater difference in HER2 staining intensity (e.g., no staining to moderate or strong staining, or weak to strong staining) over more than 5% of the tumor area [9].

Fluorescent in situ hybridization

Cases with a disagreement in the HER2 score and a 2+ score by any of the observers were subjected to fluorescent in situ hybridization (FISH), using PathVysion (Abbott Molecular, Inc., Abbott Park, IL) according to the manufacturer's instructions, to assess the potential clinical significance of disagreement in HER2 scores among pathologists. FISH was performed in direct correlation with HER2 immunohistochemistry: a tumor area of at least 1 cm^2 was selected on the immunostained slide with the most protein expression. FISH scoring was performed by one of two authors (PH or NB). A HER2/CEP17 signal ratio of ≥2.0, or <2.0 with ≥6 HER2 signals on average/ nucleus was considered positive for gene amplification [10].

Statistical analysis

Interobserver agreement was analyzed using SAS software for multioperator variability correlation (Fleiss's fixed marginal kappa). The first-order agreement coefficient (AC1) value was calculated by Gwet's method [19]. Kappa or AC1 values for interobserver agreement were interpreted as follows: kappa/AC1 < 0 less than chance agreement; 0.01–0.20 slight agreement; 0.21–0.40 fair agreement; 0.41–0.60 moderate agreement; 0.61–0.80 substantial agreement; 0.81–0.99 almost perfect agreement [20]. The overall percent agreement was calculated by the number of pairwise agreements between all observer-pairs for all cases divided by the total number of observer-pairs for all cases. P values of <0.05 were considered statistically significant.

Results

Details of patient clinicopathologic characteristics are presented in Table 2. Forty HER2 immunostained tumor slides from 31 patients were included in the study. The patient age at diagnosis ranged between 50 and 91 years (mean: 69.3). The specimen type was endometrial biopsy/curettage in 18 tumors and the remaining 22 were from hysterectomy specimens, with both specimen types from the same patient tested in 9 cases. Nineteen patients (61.3%) had early stage (FIGO stage I or II) tumors and twelve patients (38.7%) had advanced stage (stage III or IV) disease. P53 status by immunohistochemistry and/or mutation analysis showed aberrant immunostaining pattern and/or TP53 gene mutation on all cases. Tumor mismatch repair (MMR) status by immunohistochemistry and/or microsatellite instability (MSI) PCR was available in 23 of 31 patients: 21 tumors were MMR proficient, and 2 tumors showed loss of MSH6 expression and were MSI-high by PCR. POLE mutation status by next generation sequencing was available in 11

Table 2 Clinicopathologic characteristics of cases (total patients $n =$ 31, total tumor specimens $n = 40$).

Clinicopathologic characteristics	
Patient age at diagnosis, years: mean (range)	69.3 (50–91)
Specimen type (n, %)	
Endometrial biopsy/curettage	18 (45)
Hysterectomy	22 (55)
FIGO tumor stage (n, %)	
Early stage (I, II)	19 (61.3)
Advanced stage (III, IV)	12 (38.7)
HER2 IHC score assigned in previous study (n, %)	
0	8 (20)
1+	14 (35)
2+	9 (22.5)
3+	9 (22.5)
Heterogeneity of HER2 expression (n, %)	
Present	11 (27.5)
Absent	29 (72.5)

IHC immunohistochemistry, FIGO International Federation of Gynecology and Obstetrics.

tumors and all but one case showed absence of POLE mutation. HER2 scores assigned in the prior study for the same set of slides were as follows: score 0 $n = 8$ (20%), score $1 + n = 14$ (35%), score $2 + n = 9$ (22.5%), score 3 $+ n = 9$ (22.5%) [18]. Intratumoral heterogeneity of HER2 protein expression was present in 11 cases (27.5%).

Analysis of interobserver agreement for 4 HER2 score categories (0, 1+, 2+, 3+) showed 72.3% overall agreement and a kappa value of 0.60 (moderate agreement) (Table 3). The overall agreement increased to 77.5% (kappa 0.65; substantial agreement) and 83.3% (kappa 0.65) when combining scores 0/1+ and $2 + /3 +$, respectively. Using Gwet's method [19], the first-order agreement coefficient (AC1) values are 0.64, 0.67, and 0.68 for four, three and two HER2 score categories, respectively. Thirty-nine of 40 slides showed an agreement between at least 4 observers. Complete agreement among all 7 observers was achieved in 15 cases with the following HER2 IHC scores: score 0 $n = 1$, score $1 + n = 3$, score $2 + n = 5$, score $3 + n = 6$ (Table 4) [Fig. 1]. Among the remaining 25 cases, all but one (case #26) case had a score agreement by at least 4 observers, and a 2-score difference was seen in only 2 cases (#26 and #35): in case #26 HER scores 3+ and 2+ were assigned by three observers each, and one observer scored it as 1+; case #35 was scored as 1+ by four observers, 2+ by two observers and 0 by one observer (Table 5). HER2 protein expression was heterogeneous in 7 of the 25 cases (28%) with any degree of disagreement, versus in 4 of 15 cases (26.7%) in cases with perfect agreement.

Among the 25 cases with any degree of disagreement, FISH was performed on 21 tumors with a 2+ HER2 score

Table 3 Statistical analysis of interobserver agreement.

	Overall agreement (%)	Kappa (95% CI)	AC1 (95% CI)	P value
40 slides, 4 scoring categories (0, 1+, 2+, 3+)	72.3	0.60 (0.49–0.72)	0.64 (0.54–0.74)	<0.0001
40 slides, 3 scoring categories (0/1+, 2+, 3+)	77.5	0.65 (0.54–0.77)	0.67 (0.56–0.77)	<0.0001
40 slides, 2 scoring categories (0/1+, 2/3+)	83.3	0.65 (0.51–0.79)	0.68 (0.54–0.82)	<0.0001

AC1 Gwet's first agreement coefficient, *CI* confidence interval.

Table 4 Detailed HER2 scores of slides with agreement between at least 4 of 7 observers.

HER2 score (number of cases)	Number of observers with agreement (n)			
	7	6	5	4
0	1	1	0	1
1+	3	3	3	3
2+	5	4	5	1
3+	6	1	1	1
All scores (total, n)	15	9	9	6

assigned by at least one observer (Table 5). In one tumor (case #32) FISH was technically unsuccessful as no hybridization signal was detected. Among the remaining cases, based on the combined evaluation of HER2 scores and FISH results the interobserver disagreement may have been clinically significant in three tumors (#25, 26, and 37) [Figs. 2 and 3].

When comparing interobserver agreement based on specimen type, endometrial biopsies/curettings showed a higher agreement rate (kappa 0.69, AC1 0.70) than hysterectomy specimens (kappa 0.52, AC1 0.59).

Discussion

Endometrial serous carcinoma accounts for only ~10% of all endometrial carcinomas, yet it is responsible for as much as 40% of all endometrial cancer mortality [21]. A large proportion of patients present at advanced stage and often have a poor response to traditional platinum-based chemotherapy [22–24]. A recent phase II clinical trial demonstrated significant improvement in progression-free and overall survival of patients with HER2-positive advanced stage or recurrent endometrial serous carcinoma when trastuzumab was added to standard chemotherapy [2, 3]. The HER2 status also has prognostic implications, as HER2 positivity in early stage endometrial serous carcinoma has been found to be associated with a three-fold greater risk of disease recurrence and a significantly worse progression-free and overall survival [25]. Targeted anti-HER2 therapy now plays an important role in treatment planning of these

aggressive tumors, and accurate determination of HER2 status by pathologic evaluation is paramount for successful clinical management of patients.

Tumors in different organ systems have distinct characteristics of HER2 protein expression and gene amplification, leading to development of different HER2 testing guidelines for breast and gastric carcinomas [5–8]. More recently, yet another specific set of HER2 scoring criteria has been established in the HERACLES clinical trial for colorectal cancer [26]. For endometrial serous carcinoma, two recent studies evaluated the characteristics of HER2 protein expression and gene amplification, and observed higher concordance between immunohistochemistry and FISH when a 30% tumor cell staining cut-off was used, compared with a 10% cut-off [9, 27]. The same studies also reported significant intratumoral heterogeneity of HER2 protein expression in over 50% of HER2 positive tumors by immunohistochemistry, directly correlating with heterogeneous *HER2* gene amplification by FISH [27]. In addition, similar to gastric and colorectal adenocarcinomas, lack of apical membrane staining was frequently observed in endometrial serous carcinoma, resulting in a basolateral/lateral staining pattern. Patient enrollment for the clinical trial reported by Fader et al. [2] began in 2011, when the 2007 ASCO/CAP HER2 scoring guidelines were in use for breast cancer, employing a 30% tumor cell staining cut-off for a 3+ immunohistochemical score [5]. Thus, the trial enrollment criteria were based on the 2007 ASCO/CAP breast guidelines with specific modifications to incorporate the observations by the above two studies: first H&E stained tumor sections were reviewed to confirm serous histology, followed by HER2 immunohistochemistry. Tumors with intense complete or lateral/ basolateral membranous HER2 staining in more than 30% of tumor cells were scored 3+, while those with intense complete or lateral/basolateral membrane staining in ≤30%, or weak to moderate staining in ≥10% of tumor cells were scored 2+ and subjected to reflex *HER2* FISH in direct correlation with the HER2 immunostained slide [2, 3]. A *HER2/CEP17* ratio of ≥2.0 was considered amplified. Since the publication of the clinical trial results, the same HER2 testing algorithm and scoring criteria have been applied in three other studies [18, 25, 28] and have been proposed for use in clinical

Fig. 1 Representative examples of cases with complete agreement among the seven observers. A HER2 score 3+, strong complete or lateral/basolateral membranous staining in >30% of tumor cells; **B** HER2 score 2+, weak to moderate complete or lateral/basolateral membranous staining in ≥10% of tumor cells; **C** HER2 score 1+, faint, barely perceptible incomplete membranous staining in tumor cells; **D** HER2 score 0, no staining in tumor cells.
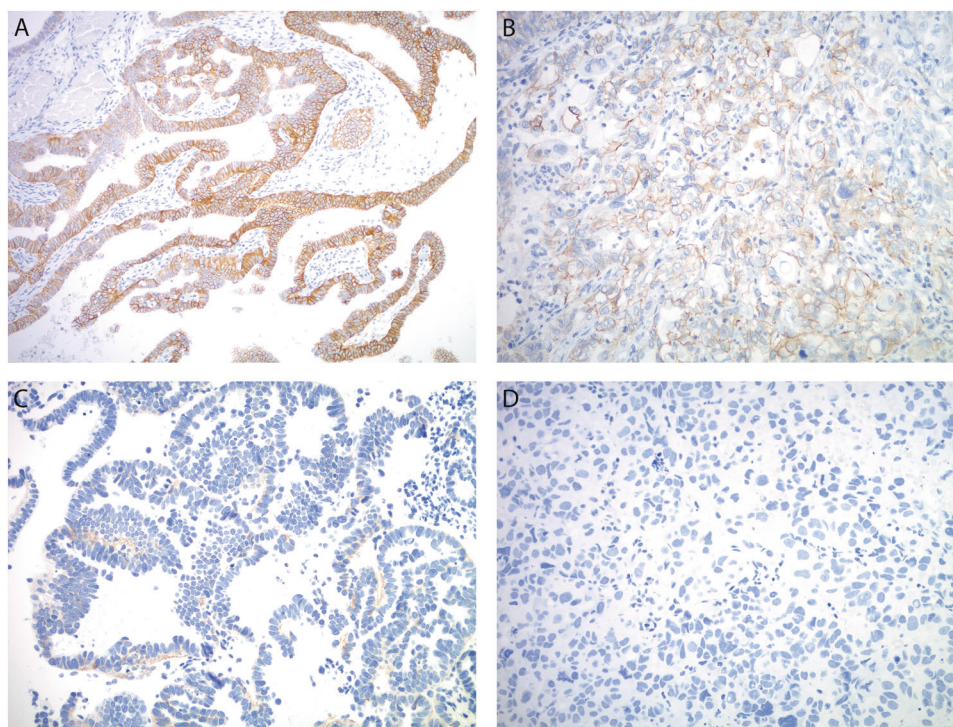
practice [10, 11]. To date, this is the only set of HER2 scoring criteria proven to predict clinical response to targeted therapy in endometrial cancer patients.

Our study represents the first multi-institutional effort to analyze the reproducibility of the proposed criteria in endometrial serous carcinoma. We observed a high level of overall score agreement among seven gynecologic pathologists, ranging from 72.3 to 83.3% for all cases, depending on the number of scoring categories. At least 4 of the 7 observers agreed on the HER2 immunohistochemical score in 39 of 40 cases (97.5%) and the interrater agreement ranged from moderate to substantial (kappa values of 0.60 to 0.65; AC1 of 0.64 to 0.68). Importantly, we observed substantial interobserver agreement (kappa 0.65, AC1 0.67) when grouping HER2 scores 0 and 1+ together. Distinguishing between scores 0 and 1+ does not have a clinical significance according to the current treatment recommendations.

In previous breast cancer series, the interobserver agreement of HER2 scores ranged between slight to substantial, with kappa values showing a wide range from the lowest 0.19 between observer-pairs up to 0.80 among multiple observers [12–14, 17, 29, 30]. Fewer studies evaluated the reliability of HER2 scoring in gastric cancer, reporting kappa values from 0.61 to 0.73 for all HER2 score categories [15, 16, 31]. Several factors have been found to play a role in interobserver variability in these cancer types, including the individual pathologist's experience and prior training in evaluation of HER2 immunohistochemistry, the specific antibody clone used, tumor histologic subtype, and

the scoring criteria applied. Layfield et al. observed 85% absolute interobserver agreement rate in breast cancer with Herceptest™ (kappa = 0.74), while the agreement rate was only 69% for the 4B5 HER2 antibody clone (kappa = 0.57) [13]. The authors noted that the staining characteristics were different between the two antibodies; staining with 4B5 did not appear as crisp as those with the Herceptest™, resulting in greater variability among observers [13]. Similarly, another study compared four different HER2 antibodies (Herceptest™, CB11, TAB250, and A0485) in breast cancer and concluded that staining heterogeneity, and cytoplasmic, pseudomembranous, and non-specific stromal staining likely all contributed to interobserver variability of interpretation [32]. The slides in the current study were stained with a clinically validated EP3 antibody clone, which has not been evaluated in prior interobserver analyses. The type of positive controls, on-slide versus batch, may also have an impact on interobserver variability. On-slide controls are preferred for biomarker testing in clinical practice, and may improve interobserver agreement of HER2 scoring. However, most prior interobserver studies on HER2 scoring in breast and gastric cancer did not specify which type of positive control was used [12–17, 29–31]. One study reported the use of positive tissue controls for "each run" [32]. In our series on-slide controls were used on two slides, the remaining 38 cases had batch controls only, a potential weakness in our study design.

The proposed HER2 scoring algorithm for endometrial serous carcinoma is primarily based on the ASCO/CAP

Table 5 Characteristics of cases with interobserver disagreement (n, number of observers).

Case #	HER2 score 0	HER2 score 1+	HER2 score 2+	HER2 score 3+	Specimen type	Heterogeneity of HER2 expression (Yes/No)	HER2 FISH
2	0	0	6	1	TH	Y	positive (ratio 2.0, HER2 CN 5.4)
4	0	6	1	0	TH	N	negative (ratio 1.0, HER2 CN 2.5)
7	0	2	5	0	EMB	N	negative (ratio 1.0, HER2 CN 1.8)
8	0	3	4	0	TH	N	negative (ratio 1.0, HER2 CN 2.1)
9	0	0	6	1	TH	Y	positive (ratio 3.5, HER2 CN 5.0)
10	0	5	2	0	EMB	N	negative (ratio 0.9, HER2 CN 1.6)
11	0	2	5	0	EMB	N	negative (ratio 1.6, HER2 CN 1.9)
12	0	0	1	6	TH	Y	positive (ratio 2.2, HER2 CN 9.5)
14	0	2	5	0	TH	N	negative (ratio 1.4, HER2 CN 2.6)
17	6	1	0	0	TH	N	ND
18	0	1	6	0	EMB	Y	negative (ratio 0.9, HER2 CN 1.5)
19	2	5	0	0	EMB	N	ND
22	4	3	0	0	TH	N	ND
23	0	6	1	0	TH	N	negative (ratio 1.1, HER2 CN 1.7)
24	0	4	3	0	TH	N	negative (ratio 0.9, HER2 CN 1.6)
25[a]	0	1	6	0	TH	N	positive (ratio 2.3, HER2 CN 4.1)
26[a]	0	1	3	3	EMC	Y	positive (ratio 3.8, HER2 CN 10.5)
27	0	2	5	0	TH	N	negative (ratio 1.2, HER2 CN 1.9)
32	0	2	5	0	TH	Y	no signal
34	0	5	2	0	TH	N	negative (ratio 0.6, HER2 CN 1.2)
35	1	4	2	0	TH	N	negative (ratio 0.9, HER2 CN 1.5)
36	3	4	0	0	EMB	N	ND
37[a]	0	0	2	5	TH	N	negative (ratio 1.3, HER2 CN 3.4)
39	0	6	1	0	TH	N	negative (ratio 1.3, HER2 CN 4.8)
40	0	0	3	4	EMC	Y	positive (ratio 4.2, HER2 CN 12.4)

CN copy number, ND not done, TH total hysterectomy, EMB endometrial biopsy, EMC endometrial curettage.

[a]Potentially clinically significant disagreement.

2007 breast HER2 scoring guidelines, which was used in the recent clinical trial with specific modifications. Weak to moderate incomplete membranous staining in <10% of tumor cells was not specifically addressed in the ASCO/CAP 2007 breast guidelines and is a limitation of the proposed criteria for endometrial serous carcinoma. Currently we do not have sufficient data to incorporate this specific staining pattern in the algorithm, and it may have been a contributing factor to interobserver disagreement in our study. However, it will be an important parameter for future studies in a prospective clinical trial setting to assess the correlation of this pattern with treatment response. Ideally as more data become available the scoring algorithm will continue to evolve to address all HER2 staining scenarios.

Another limitation of our study is that HER2 FISH was performed only on a subset of cases to assess the potential clinical significance of disagreement in HER2 scores among pathologists. Thus, we do not have comprehensive data on the HER2 immunohistochemistry -FISH concordance in the

current paper, but it will be an important area to explore in future studies. Of note, the clinical trial by Fader et al. followed the same algorithm that we used in the current study (HER2 immunohistochemistry followed by HER2 FISH only in tumors with a 2+ IHC score), and did not include patients with HER2 gene amplification without HER2 protein expression (scores 0 and 1+). Only patients whose tumors showed a 3+ HER2 IHC score, or a 2+ IHC score with gene amplification by FISH were eligible for enrollment [2]. Future clinical trials may be able to address the clinical response in patients with HER2 gene amplification in the absence of protein expression (scores 0 and 1+) and potentially expand the group of patients benefitting from HER2 targeted therapy.

The importance of the pathologists' training in the reproducibility of HER2 scoring has been long recognized in breast and gastric cancer, and focused training, quality assurance and proficiency testing programs are available for these tumor types. Interestingly, Koopman et al. reported

Fig. 2 Case #25 (in Table 5) with interobserver disagreement. HER2 immunostain (**A**) was scored as 2+ by 6 observers, and as 1+ by 1 observer. FISH (**B**) demonstrated *HER2* amplification (*HER2/CEP 17* ratio = 2.3, average *HER2* copy number/cell = 4.1).
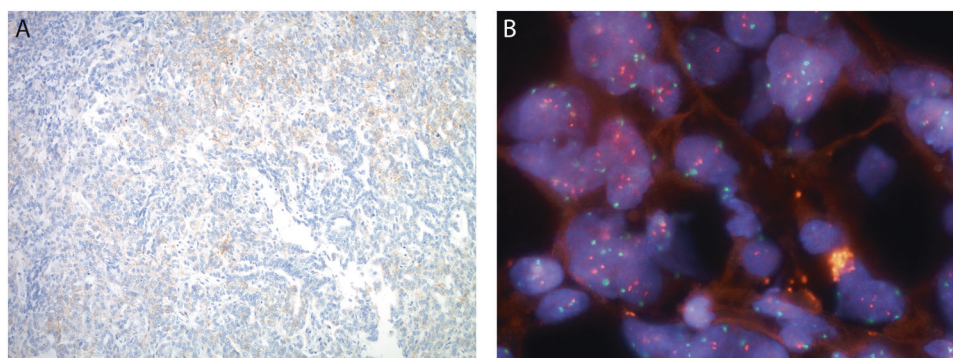
Fig. 3 Case #26 (in Table 5) with interobserver disagreement. HER2 immunostain (**A**, **B**, **C**) was scored as 3+ by 3 observers, as 2+ by 3 observers, and as 1+ by 1 observer. HER2 expression showed significant intratumoral heterogeneity (**A**). Several tumor fragments demonstrated strong complete (**B**) or lateral/basolateral (**C**) HER2 immunoreactivity. *HER2* amplification was present by FISH (**D**) (*HER2/CEP 17* ratio = 3.8, average *HER2* copy number/cell = 10.5).
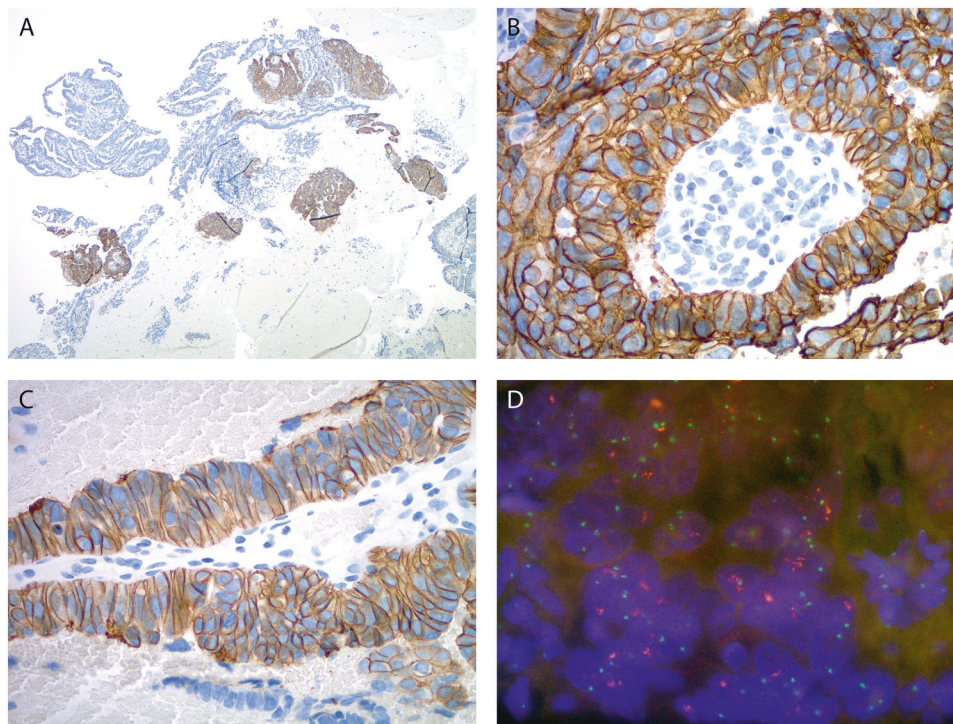
that the interobserver agreement was also significantly affected by the tumor histology in gastric cancer, with intestinal tumor types showing almost perfect agreement (kappa = 0.815) while only moderate agreement (kappa = 0.566) was achieved in diffuse and mixed carcinomas [15]. Additional parameters with potential impact on interobserver variability include specimen type and size of tumor tissue (i.e., tissue microarray or biopsy versus whole slide tumor sections), the number of scoring categories (2, versus 3, or 4 HER2 scoring categories), and the methods used for statistical analysis. Furthermore, representation of different HER2 scoring categories in the study may also play a significant role: agreement on scores 1+ and 2+ have been reported to be poor in both breast and gastric cancer [15, 32].

Our study set mirrored the previously reported HER2 score distribution of endometrial serous carcinoma: the original HER2 score assigned was 3+, 2+, 1+, and 0 in 22.5%, 22.5%, 35% and 20% of cases, respectively. Thus, the proportion of combined 1+ and 2+ scores in our series was 57.5%, likely contributing to the overall interobserver variability among our cases. In fact, in more than half of the cases (14/25; 56%) with interobserver disagreement, the disagreements were between scores 1+ and 2+ only. We performed FISH on all cases with disagreement and a 2+ HER2 score assigned by any observer. Based on the combined immunohistochemistry and FISH results, the scoring disagreement resulted in a potentially clinically significant change in HER2 status in three tumors: in two tumors the HER2 scores ranged between 1+ to 2+ and 1+ to 3+, with *HER2* gene amplification by FISH, and in one case the HER2 score was 2+ to 3+ with no *HER2* gene amplification. A two-degree difference in HER2 scores was seen in two cases: scores 0 to 2+ in one case, and 1+ to 3+ in another; the latter showing *HER2* amplification by FISH

(see above). All but one of the 14 cases with 1+ to 2+ HER2 score disagreement were FISH negative.

We used scanned whole slide images, on which small foci of faint membranous staining may be easier to miss, potentially resulting in a disagreement between scores 0 and 1+, although this would not be clinically significant. Similarly, scanning a larger tumor area on whole tissue sections in the current study, compared to a smaller amount of tumor on a tissue microarray or core needle biopsy, coupled with the frequent intratumoral heterogeneity of HER2 protein expression in endometrial serous carcinoma may impact evaluation of percent of tumor staining and decrease agreement between scores 2+ and 3+. On the other hand, the 30% tumor cell staining cut-off for a 3+ HER2 score in our study may improve interobserver agreement compared with a 10% staining cut-off, as was previously reported in breast cancer [30]. We also observed a difference in interobserver agreement among different specimen types: endometrial biopsies/curettings showed a higher agreement rate (kappa 0.69, AC1 0.70) compared with hysterectomy specimens (kappa 0.52, AC1 0.59), which may be explained by a generally higher proportion of tumor tissue on slides from biopsies/curettings and/or better tissue fixation and staining quality.

In conclusion, our study demonstrates that the recently proposed, clinical trial-based serous endometrial cancer-specific HER2 scoring criteria are reproducible among gynecologic pathologists with moderate to substantial interobserver agreement rates, comparable to those of previously reported in breast and gastric carcinomas. Corroborating existing literature, our findings provide strong support toward establishing serous endometrial cancer-specific HER2 scoring guidelines for maximal clinical benefit for patients.

Compliance with ethical standards

References

1. Twomey JD, Brahme NN, Zhang B. Drug-biomarker co-development in oncology - 20 years and counting. Drug Resist Updat. 2017;30:48–62.
2. Fader AN, Roque DM, Siegel E, Buza N, Hui P, Abdelghany O, et al. Randomized phase II trial of carboplatin-paclitaxel versus carboplatin-paclitaxel-trastuzumab in uterine serous carcinomas that overexpress human epidermal growth factor receptor 2/neu. J Clin Oncol. 2018;36:2044–51.
3. Fader, AN, Roque, DM, Siegel, E, Buza, N, Hui, P, Abdelghany, O et al. Randomized phase II trial of carboplatin-paclitaxel compared with carboplatin-paclitaxel-trastuzumab in advanced (stage III-IV) or recurrent uterine serous carcinomas that overexpress Her2/Neu (NCT01367002): updated overall survival analysis. Clin Cancer Res. https://doi.org/10.1158/1078-0432.CCR-20-0953 (2020).
4. National Comprehensive Cancer Network. Uterine Neoplasms (Version 4.2019), https://www.nccn.org/professionals/physician_gls/pdf/uterine.pdf Accessed October, 2019.
5. Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. Arch Pathol Lab Med. 2007;131:18–43.
6. Wolff AC, Hammond ME, Hicks DG, Dowsett M, McShane LM, Allison KH, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. Arch Pathol Lab Med. 2014;138:241–56.
7. Wolff AC, Hammond MEH, Allison KH, Harvey BE, Mangu PB, Bartlett JMS, et al. Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. Arch Pathol Lab Med. 2018;142:1364–82.
8. Bartley AN, Washington MK, Ventura CB, Ismaila N, Colasacco C, Benson AB 3rd, et al. HER2 testing and clinical decision making in gastroesophageal adenocarcinoma: guideline from the College of American Pathologists, American Society for Clinical Pathology, and American Society of Clinical Oncology. Arch Pathol Lab Med. 2016;140:1345–63.
9. Buza N, English DP, Santin AD, Hui P. Toward standard HER2 testing of endometrial serous carcinoma: 4-year experience at a large academic center and recommendations for clinical practice. Mod Pathol. 2013;26:1605–12.
10. Buza, N HER2 Testing in endometrial serous carcinoma. Arch Pathol Lab Med. https://doi.org/10.5858/arpa.2020-0207-RA (2020).
11. Buza N. HER2 testing and reporting in endometrial serous carcinoma: practical recommendations for HER2 immunohistochemistry and fluorescent in situ hybridization: proceedings of the ISGyP companion society session at the 2020 USCAP annual meeting. Int J Gynecol Pathol. 2021;40:17–23.
12. Lacroix-Triki M, Mathoulin-Pelissier S, Ghnassia JP, Macgrogan G, Vincent-Salomon A, Brouste V, et al. High inter-observer agreement in immunohistochemical evaluation of HER-2/neu expression in breast cancer: a multicentre GEFPICS study. Eur J Cancer. 2006;42:2946–53.
13. Layfield LJ, Frazier S, Esebua M, Schmidt RL. Interobserver reproducibility for HER2/neu immunohistochemistry: a comparison of reproducibility for the HercepTest and the 4B5 antibody clone. Pathol Res Pr. 2016;212:190–5.
14. Castera C, Bernet L. HER2 immunohistochemistry inter-observer reproducibility in 205 cases of invasive breast carcinoma additionally tested by ISH. Ann Diagn Pathol. 2020;45:151451.
15. Koopman T, Louwen M, Hage M, Smits MM, Imholz AL. Pathologic diagnostics of HER2 positivity in gastroesophageal adenocarcinoma. Am J Clin Pathol. 2015;143:257–64.
16. Kushima R, Kuwata T, Yao T, Kuriki H, Hashizume K, Masuda S, et al. Interpretation of HER2 tests in gastric cancer: confirmation of interobserver differences and validation of a QA/QC educational program. Virchows Arch. 2014;464:539–45.
17. Tsuda H, Kurosumi M, Umemura S, Yamamoto S, Kobayashi T, Osamura RY. HER2 testing on core needle biopsy specimens from primary breast cancers: interobserver reproducibility and

concordance with surgically resected specimens. BMC Cancer. 2010;10:534.

18. Rottmann, D, Assem, H, Matsumoto, N, Wong, S, Hui, P & Buza, N Does specimen type have an impact on HER2 status in endometrial serous carcinoma? Discordant HER2 status of paired endometrial biopsy and hysterectomy specimens in the presence of frequent intratumoral heterogeneity. Int J Gynecol Pathol. https://doi.org/10.1097/PGP.0000000000000690 (2020).

19. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61:29–48.

20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.

21. Hamilton CA, Cheung MK, Osann K, Chen L, Teng NN, Longacre TA, et al. Uterine papillary serous and clear cell carcinomas predict for poorer survival compared to grade 3 endometrioid corpus cancers. Br J Cancer. 2006;94:642–6.

22. Homesley HD, Filiaci V, Gibbons SK, Long HJ, Cella D, Spirtos NM, et al. A randomized phase III trial in advanced endometrial carcinoma of surgery and volume directed radiation followed by cisplatin and doxorubicin with or without paclitaxel: a Gynecologic Oncology Group study. Gynecol Oncol. 2009;112:543–52.

23. de Boer SM, Powell ME, Mileshkin L, Katsaros D, Bessette P, Haie-Meder C, et al. Adjuvant chemoradiotherapy versus radiotherapy alone in women with high-risk endometrial cancer (PORTEC-3): patterns of recurrence and post-hoc survival analysis of a randomised phase 3 trial. Lancet Oncol. 2019;20:1273–85.

24. Tortorella L, Langstraat CL, Weaver AL, McGree ME, Bakkum-Gamez JN, Dowdy SC, et al. Uterine serous carcinoma: reassessing effectiveness of platinum-based adjuvant therapy. Gynecol Oncol. 2018;149:291–6.

25. Erickson BK, Najjar O, Damast S, Blakaj A, Tymon-Rosario J, Shahi M, et al. Human epidermal growth factor 2 (HER2) in early stage uterine serous carcinoma: a multi-institutional cohort study. Gynecol Oncol. 2020;159:17–22.

26. Valtorta E, Martino C, Sartore-Bianchi A, Penaullt-Llorca F, Viale G, Risio M, et al. Assessment of a HER2 scoring system for colorectal cancer: results from a validation study. Mod Pathol. 2015;28:1481–91.

27. Buza N, Hui P. Marked heterogeneity of HER2/NEU gene amplification in endometrial serous carcinoma. Genes Chromosomes Cancer. 2013;52:1178–86.

28. Rottmann D, Snir OL, Wu X, Wong S, Hui P, Santin AD, et al. HER2 testing of gynecologic carcinosarcomas: tumor stratification for potential targeted therapy. Mod Pathol. 2020;33:118–27.

29. Hsu CY, Ho DM, Yang CF, Lai CR, Yu IT, Chiang H. Interobserver reproducibility of Her-2/neu protein overexpression in invasive breast carcinoma using the DAKO HercepTest. Am J Clin Pathol. 2002;118:693–8.

30. Hameed O, Adams AL, Baker AC, Balmer NE, Bell WC, Burford HN, et al. Using a higher cutoff for the percentage of HER2+ cells decreases interobserver variability in the interpretation of HER2 immunohistochemical analysis. Am J Clin Pathol. 2008;130:425–7.

31. Ruschoff J, Dietel M, Baretton G, Arbogast S, Walch A, Monges G, et al. HER2 diagnostics in gastric cancer-guideline validation and development of standardized immunohistochemical testing. Virchows Arch. 2010;457:299–307.

32. Thomson TA, Hayes MM, Spinelli JJ, Hilland E, Sawrenko C, Phillips D, et al. HER-2/neu in breast cancer: interobserver variability and performance of immunohistochemistry with 4 antibodies compared with fluorescent in situ hybridization. Mod Pathol. 2001;14:1079–86.