

## Original Article

## Determining the Inter- and Intraobserver Reproducibility of the Diagnosis of Endometrial Hyperplasia Subgroups and Well-Differentiated Endometrioid Carcinoma in Endometrial Curettage Specimens

Narges Izadi-Mood MD\*, Mahmood Khaniki MD\*\*, Guity Irvanloo MD\*\*, Seyed-Ali Ahmadi MD\*\*\*, Hayedeh Hayeri MD\*\*, Alipasha Meysamie MD†

**Background:** Many studies have attempted to identify histologic features that help to distinguish atypical hyperplasia from hyperplasia without atypia and well-differentiated endometrioid carcinoma of the endometrium; however, few have evaluated the reproducibility of these diagnoses.

**Methods:** Five pathologists independently reviewed 100 endometrial curettage specimens chosen to represent the spectrum of proliferative lesions of the endometrium. This included simple hyperplasia, complex hyperplasia, atypical hyperplasia, and well-differentiated endometrioid carcinoma. Slides were reviewed once for interobserver agreement among the five pathologists and twice for intraobserver agreement by one of them.

**Results:** The results were assessed using the weighted kappa statistic. The mean intraobserver kappa value was 0.86. The mean interobserver kappa values by diagnostic category were as follows: simple hyperplasia without atypia: 0.74; complex hyperplasia without atypia: 0.33; atypical hyperplasia: 0.34, and well-differentiated endometrioid carcinoma: 0.64; with a kappa value of 0.53 for all cases combined.

**Conclusion:** A major interobserver discrepancy exists in the diagnosis of complex and atypical hyperplasia which are the most similar mimics of endometrioid carcinoma.

*Archives of Iranian Medicine, Volume 12, Number 4, 2009: 377 – 382.*

**Keywords:** Endometrioid carcinoma • endometrial hyperplasia • pathologic diagnostic agreement • reproducibility

### Introduction

Endometrial carcinoma is the fourth most frequent malignant neoplasm and the most common female gynecologic cancer. This cancer occurs mostly in postmenopausal women and presents as abnormal uterine bleeding (AUB). Endometrial

adenocarcinoma is often well-differentiated and mimics the normal pattern of endometrial glands (endometrioid). This cancer is related to long-term estrogen exposure. Endometrial hyperplasia has a relatively good prognosis. Differential diagnosis between hyperplasia and adenocarcinoma is sometimes difficult, particularly in curettage specimens.<sup>1</sup>

The risk of progression of hyperplasia is highly variable depending on the histopathologic features and is generally about 5 – 10%. Incorrect diagnosis of hyperplasia will lead to mismanagement which includes insufficient treatment for dangerous lesions or over treatment in low-risk lesions, both imposing complications and unnecessary expenditure. Different systems have been proposed

**Authors' affiliations:** Department of Pathology, \*Mirza Koochak Khan Hospital, \*\*Imam Khomeini Hospital, \*\*\*Sina Hospital, Tehran University of Medical Sciences, †Department of Community and Preventive Medicine, Medical School, Tehran University of Medical Sciences, Tehran, Iran.

•**Corresponding author and reprints:** Narges Izadi-Mood MD, Department of Pathology, Mirza Koochak Khan Hospital, Nejatollahi St., Karim Khan Zand Ave., Tehran, Iran.

E-mail: nizadi@sina.tums.ac.ir; nizadimood@yahoo.com

Accepted for publication: 29 April 2009

in recent decades for the classification of endometrial hyperplasia. In 1994, WHO presented its classification of endometrial hyperplasia. This classification was based upon a pilot study<sup>2</sup> indicating the relationship between cytoarchitectural atypia and the increased risk of cancer. The current classification that is based upon Kurman et al.'s definition and is accepted by WHO and ISGP has divided hyperplasia into four groups according to simple or complex glandular architecture and typical and atypical cytology.<sup>2</sup>

Three studies have previously evaluated the reproducibility of the diagnosis of hyperplasia according to the WHO classification. The first study which included the evaluation of the slides of 128 patients by six experienced pathologists indicated that the diagnostic intraobserver reproducibility was moderate and the interobserver agreement was fair.<sup>3</sup> The second study was performed on 100 patients by pathologists with different levels of experience which revealed a notable intraobserver reproducibility and a moderate to substantial interobserver agreement.<sup>4</sup> The third study also indicated a moderate combined interobserver reproducibility and a variable intraobserver agreement.<sup>5</sup> In summary, according to the current classification, the diagnostic agreement is not enough for pathologists and gynecologists to specify the patients' problems easily.<sup>1</sup> The natural history of endometrial hyperplasia is poorly understood due to the difficulty of carrying out prospective follow-up studies that do not interfere with biologic behavior. Also, the initial diagnosis is usually based upon endometrial curettage or biopsy specimens, which may be therapeutic in essence. Conversely, these endometrial specimens may not sample the entire endometrium and the areas of greatest histologic or cytologic severity may thus escape histologic identification.<sup>5</sup>

The present study was a preliminary one which aimed to evaluate the reproducibility of the diagnosis of the endometrial hyperplasia and well-differentiated adenocarcinoma (WDA) in an academic center in Iran and the future aim would be to perform morphometric analysis of the slides to assessing D-score. The general purpose was to determine the extent of interobserver and intraobserver agreement and the extent of reproducibility in different diagnostic groups including simple hyperplasia (SH), complex hyperplasia (CH), atypical hyperplasia (AH), and endometrial carcinoma.

## Materials and Methods

One Hematoxylin and Eosin-stained slide from 100 cases of endometrial curettage from the surgical pathology files of Mirza Koochak Khan Hospital was selected which represented an original diagnosis including SH, CH, AH, and WDA. Twenty-five cases from each category were selected. Slides from each diagnostic category were numbered from 1 to 100. No clinical data was provided for the reviewing pathologists.

Five pathologists from different pathology departments of Tehran University of Medical Sciences including Mirza Koochak Khan, Sina, and Imam Khomeini Hospitals with varying levels of experience participated in this study. In order to evaluate the intraobserver agreement, one of the pathologists re-examined the slides after two months.

Blank questionnaires containing the four diagnostic categories (SH, CH, AH, and WDA) as well as an extra column for additional diagnoses were given to each pathologist.

After the slides were reviewed and coded data were collected, statistical analysis was performed using STATA8 software with weighted kappa statistics. Ten specimens were excluded from the study (three poorly-preserved slides and seven specimens with additional diagnostic suggestions).

Statistical analysis included the evaluation of interobserver and intraobserver agreement using kappa statistics, a measure of agreement between observers that attempts to correct chance agreement. The range of values for kappa is 1.00 or less. 1.00 indicates perfect agreement and 0 indicates the level of agreement expected by chance alone. Negative values indicate less than chance agreement.

The interpretation of the kappa values between 0 and 1 used in this study is classified in Table 1. According to Landis and Koch<sup>6</sup>, the following interpretations of agreement were provided: 0.00–0.20=slight, 0.21–0.40=fair, 0.41–0.60=moderate, 0.61–0.80=substantial, and 0.81–1.00=almost perfect. Ratings weighted by this test are also summarized in Table 2.

For intraobserver agreement, diagnostic

**Table 1.** Results of kappa test for intraobserver agreement.

Diagnostic agreement	Expected agreement	Kappa	prob>z
99.44%	57.60%	0.8690	< 0.0001

**Table 2.** Ratings weighted by kappa test.

Diagnosis	SH	CH	AH	WDA
SH	1.0000	0.6667	0.3333	0.0000
CH	0.6667	1.0000	0.6667	0.3333
AH	0.3333	0.6667	1.0000	0.6667
WDA	0.0000	0.3333	0.6667	1.0000

SH=simple hyperplasia; CH=complex hyperplasia; AH=atypical hyperplasia; WDA=well-differentiated adenocarcinoma.

agreement was analyzed using the diagnosis given by one pathologist in two separate rounds.

## Results

### Intraobserver diagnostic agreement

Kappa statistics were used for analyzing the intraobserver diagnostic reproducibility by one pathologist in two separate rounds, based on four diagnostic categories. Results and interpretation of kappa value are shown in Tables 3 and 1, respectively. Table 1 shows the percentage of agreements to be 99.44% with a corresponding kappa value of 0.86 according to which interpretation of intraobserver agreement was almost perfect.

### Interobserver diagnostic agreement

Kappa statistics were used for analyzing the interobserver diagnostic reproducibility. Results are shown in Table 4. Kappa values revealed a significant difference between the four diagnostic categories according to their diagnostic agreement. The best diagnostic agreement was seen in the diagnosis of SH (kappa 0.74) and WDA (kappa 0.64) in which the interpretation of interobserver agreement was substantial and less agreement was seen in the diagnosis of CH (kappa 0.33) and AH (kappa 0.34) in which interpretation of interobserver agreement was fair. Kappa values are compared between different pathologists in Figure 1. This figure indicates that the more experienced pathologists in the field of gynecologic pathology

(numbers 1, 4, and 5) were more in agreement about the diagnosis.

## Discussion

On November 1994, the Institute of Medicine (IOM) in a report entitled "To Err is Human," discussed medical mistakes and their effects on jeopardizing the patients' lives of the United States of America.<sup>7</sup> One important issue in pathology is the definition of the mistake. Discordance in the diagnosis between pathologists is not necessarily a definite sign of mistake, because it has been proven several times that in the current and accepted taxonomic systems, there are always disagreements between pathologists and even the specialists over specific cases. For this reason, several studies have recently evaluated the reproducibility of the diagnoses in different pathologic fields. These studies have indicated dramatic variability even between specialists in that field.<sup>8-14</sup> Many studies have evaluated the reproducibility of diagnosis of endometrial hyperplasia according to WHO classification. The kappa value for all interobserver agreements in the diagnosis of endometrial hyperplasia reported varying rates ranging from 0.2 – 0.7.<sup>3-5,15-17</sup> Skov et al. who evaluated the slides of 128 patients by six experienced pathologists indicated that the intraobserver reproducibility of the diagnosis was moderate and that the interobserver reproducibility was slight to moderate.<sup>3</sup> Kendall et al. performed the same study on 100 patients using pathologists with different levels of experience and concluded that the intraobserver agreement was from substantial to almost perfect and that the combined interobserver reproducibility was substantial.<sup>4</sup> Bergeron et al. also indicated a moderate combined interobserver agreement.<sup>5</sup> However, it seems that these studies have exaggerated interobserver reproducibility, because using more

**Table 3.** Diagnoses made by one of the pathologists in two different rounds.

Diagnosis	SH	CH	AH	WDA	Sum
SH	23	0	0	0	23
	25.6%	0.0%	0.0%	0.0%	25.6%
CH	0	19	4	0	23
	0.0%	21.1%	4.4%	0.0%	25.6%
AH	0	2	14	3	19
	0.0%	2.2%	15.6%	3.3%	21.1%
WDA	0	1	4	20	25
	0.0%	1.1%	4.4%	22.2%	27.8%
Sum	23	22	22	23	90
	25.6%	24.4%	24.4%	25.6%	100.0%

SH=simple hyperplasia; CH=complex hyperplasia; AH=atypical hyperplasia; WDA=well-differentiated adenocarcinoma.

**Table 4.** Results of kappa test for interobserver diagnostic agreement.

Diagnosis	Kappa	Reproducibility	prob>z
SH	0.7441	Substantial	<0.0001
CH	0.3379	Fair	<0.0001
AH	0.3473	Fair	<0.0001
WDA	0.6428	Substantial	<0.0001
Combined	0.5372	Moderate	<0.0001

SH=simple hyperplasia, CH=complex hyperplasia; AH=atypical hyperplasia; WDA=well-differentiated adenocarcinoma.

experienced pathologists and presenting them with details of classification would increase the reproducibility.

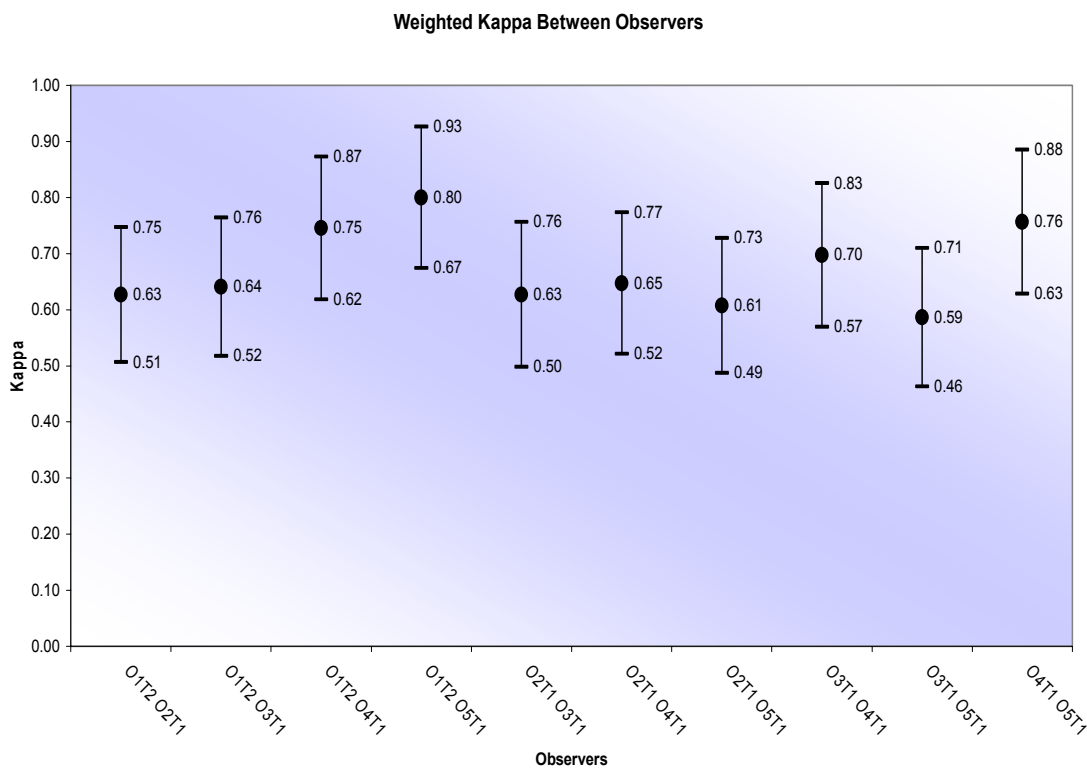
In addition to all these studies, in another study carried out by Winkler and colleagues in 1984, 100 curettage specimens referred to Sloane Gynecologic Hospital were rechecked. In 69% of the reviewed cases, the consultation diagnosis was a downgrade of the original diagnosis.<sup>15</sup>

The Gynecologic Oncology Group has decided to evaluate the hormonal therapy of atypical endometrial hyperplasia (AEH). They managed a panel of three gynecologic pathologists to assign the reproducibility of referring institution's pathologist's diagnosis of AEH. The determined

result by the panel was inappreciable and showed a poor reproducibility. Overestimation of the severity of the lesion was as common as its underestimating.<sup>16</sup>

For determining the causes of diagnostic disagreement, Allison et al. re-examined 2601 endometrial specimens. The cytologic atypia was the most disagreed histologic feature. By the way, architectural crowding and complexity or the presence of endometrial polyp had a great role in making diagnostic disagreements. They revealed that the high disagreement may be a result of both inadequate sample and interpretation of histologic features.<sup>17</sup>

In another study by Sherman et al., 209 slides which were diagnosed with disordered proliferative endometrium, SH, CH, and AH were selected and examined. Also to assess the interobserver agreement, a gynecologic pathology panel composed of two gynecologic pathologists reviewed the slides<sup>18</sup> and the results were compared. The weighted kappa between the original pathologist and the two panel pathologists was 0.267 and 0.633, respectively. In fact, the interobserver agreement between the original and



**Figure 1.** The comparison of kappa values between different observers. [This figure indicates that pathologists (numbers 1, 4 and 5) had more consensus in the diagnosis.] O1=observer 1; O2=observer 2; O3=observer 3; O4=observer 4; O5=observer 5; T1=time 1; T2=time 2

**Table 5.** Comparison of different parameters of the study with other similar studies.

Study	Present study	Skov (1997)	Kendall (1998)	Bergeron (1999)	Zaino (2006)	Allison (2008)	Sherman (2008)
Sample size	90	128	100	56	302	2601	209
Number of pathologists	5	6	5	5	3	2	3
Intraobserver agreement	Almost perfect	Moderate	Substantial /Almost perfect	Moderate / Substantial			Substantial*
Interobserver agreement							
SH	Substantial	—	Substantial	Moderate	Fair	Slight	Slight**
CH	Fair	—	Substantial	Fair	Fair	Fair	Slight**
AH	Fair	—	Moderate	Fair	Fair	Fair	Fair**
WDA	Substantial	—	Substantial	Moderate	Moderate	Moderate	Fair**
Combined	Moderate	Fair	Substantial	Moderate	Fair	Fair	Slight**/ Substantial***

SH=simple hyperplasia; CH=complex hyperplasia; AH=atypical hyperplasia; WDA=well-differentiated adenocarcinoma; \*Panelists' intraobserver agreement; \*\*Agreement between original and panel diagnoses; \*\*\*Agreement between the panelists.

panel diagnoses was slight to fair.<sup>18</sup>

Additionally, there are some other factors that can cause more diagnostic disagreements. The expertise of the pathologist, the presence of borderline disorders, fragmentation of curettage specimens, uncertainty about the importance of focal disorders, inadequate published explanations, terminologic obscurities of architectural and cytologic atypia and incidental problems in the evaluation of interobserver reproducibility according to descriptions of microscopic images can all increase diagnostic disagreement.

It seems that, except for the factors mentioned above or others not stated here and based on the current classification, the diagnostic agreement is not enough for the pathologists and gynecologists to specify patients' problems easily.<sup>1</sup> This study revealed that the reproducibility of the diagnosis of different types of hyperplasia and also the well-differentiated endometrial adenocarcinoma in the biopsy and curettage specimens was more than just an accidental agreement. In general, the intraobserver reproducibility was almost perfect and interobserver reproducibility was moderate. The best interobserver reproducibility was seen in the diagnosis of SH and WDA and the worst was seen in the diagnosis of CH and AH. The general interobserver reproducibility was moderate.

Just similar to most studies in which AH was found to be the least reproducible category with a range of 0.28 to 0.65 in kappa values,<sup>17</sup> the kappa value for AH in our study was also 0.34.

Table 5 compares different parameters of the present study with previous studies and indicates that the results of this study are almost similar to the Bergeron et al's. study which showed that the combined interobserver agreement was moderate

and Kendall et al's. which showed that intraobserver agreement was almost perfect.

It should be emphasized that our aim in this study was similar to a study by Kendall et al.<sup>4</sup> which assessed diagnostic reproducibility rather than the correlation between the diagnosis and the outcome. Therefore, there was no gold standard.

The absence of criteria to predict disease outcome irrespectively may be the significant reason of over and undertreatment.<sup>19</sup>

In pathologic reporting, there is a tendency to combine subgroups of endometrial hyperplasia.<sup>5,20</sup> The European group, without paying attention to the complexity of the glands, refers to both SH and CH as endometrial hyperplasia.<sup>5,20</sup>

The Endometrial Collaborative Group has a similar way for SH and CH, but separates endometrial adenocarcinoma from precancerous lesions and calls it endometrial intraepithelial neoplasia (EIN).<sup>21</sup>

We may express the objective with certainty through combining the diagnostic subgroups, but it would damage the evolution of the scientific knowledge.<sup>22</sup> Although there are vague cases occasionally, there is no evidence concerning the less reproducibility of WHO 94's classification than others, and the obscurities should be proved by young skillful pathologists.<sup>22</sup> Thus, because the cytologic atypia is the most common feature in diagnostic disagreements,<sup>17</sup> when there are large epithelial cells with large vacuolated or dense rounded nuclei in crowded endometrial glands; it should be endorsed to prevent diagnostic mistakes.<sup>20</sup>

Mutter et al. have considered a new classification system which refers to carcinoma precursors as EIN but as determining atypia in the

WHO classification, estimation of Volume Percentage Stroma (VPS) in this new classification could be nonreproducible.<sup>18</sup>

In conclusion, diagnosis of various types of hyperplasia and WDA from endometrial curettage or biopsy specimens are reproducible and agreement is moderate in most studies.

Similar to Sherman et al.,<sup>18</sup> we suggest that there should be more specific efforts to find quantitative and reproducible criteria to correctly diagnose a true precancerous lesion.

## References

- Zaino RJ. Endometrial hyperplasia and carcinoma. In: Haines M, Taylor CW, Fox H, Wells M, eds. *Haines & Taylor Obstetrical and Gynaecological Pathology*. Edinburgh: Churchill Livingstone, 5<sup>th</sup> ed; 2003: 445 – 446.
- Kurman RJ, Kaminski PF, Norris HJ. The behavior of endometrial hyperplasia, a long-term study of "untreated" hyperplasia in 170 patients. *Cancer*. 1985; **56**: 403 – 412.
- Skov BG, Broholm H, Engel U. Comparison of the reproducibility of the 1975 and 1994 WHO classification of endometrial hyperplasia. *Int J Gynecol Pathol*. 1997; **16**: 33 – 37.
- Kendall BS, Ronnett BM, Isacson C. Reproducibility of the diagnosis of endometrial hyperplasia, atypical hyperplasia, and well-differentiated carcinoma. *Am J Surg Pathol*. 1998; **22**: 1012 – 1019.
- Bergeron C, Nogales FF, Masseroli M, Abeler V, Duvillard P, Müller-Holzner E, et al. A multicentric European study testing the reproducibility of the WHO classification of endometrial hyperplasia with a proposal of a simplified working classification for biopsy and curettage specimens. *Am J Surg Pathol*. 1999; **23**: 1102 – 1108.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; **33**: 159 – 174.
- Kohn LT, Corrigan JM, Donaldson MS. *To err is Human: Building a Safer Health System*. Washington DC: National Academy Press, 1999.
- Farmer ER, Gonin R, Hanna MP. Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists. *Hum Pathol*. 1996; **27**: 528 – 531.
- Fitzgibbons PL. Atypical lobular hyperplasia of the breast: a study of pathologists' responses in the College of American Pathologists Performance Improvement Program in Surgical Pathology. *Arch Pathol Lab Med* 2000; **124**: 463 – 464.
- Grenko RT, Abendroth CS, Frauenhoffer EE. Variance in the interpretation of cervical biopsy specimens obtained for atypical squamous cells of undetermined significance. *Am J Clin Pathol*. 2000; **114**: 735 – 740.
- Page DL, Dupont WD, Jensen RA. When and to what end do pathologists agree? *J Natl Cancer Inst*. 1998; **90**: 88 – 89.
- Schnitt SJ, Connolly JL, Tavassoli FA, Fechner RE, Kempson RL, Gelman R, et al. Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria. *Am J Surg Pathol*. 1992; **16**: 1133 – 1143.
- Rosai J. Borderline epithelial lesions of the breast. *Am J Surg Pathol*. 1991; **15**: 209 – 221.
- Wells WA, Carney PA, Eliassen MS. Statewide study of diagnostic agreement in breast pathology. *J Natl Cancer Inst*. 1998; **90**: 142 – 145.
- Winkler B, Alvarez S, Richart RM, Crum CP. Pitfalls in the diagnosis of endometrial neoplasia. *Obstet Gynecol*. 1984; **64**: 185 – 194.
- Zaino RJ, Kauderer J, Trimble CL, Silverberg SG, Curtin JP, Lim PC, et al. Reproducibility of the diagnosis of atypical endometrial hyperplasia: a Gynecologic Oncology Group study. *Cancer*. 2006; **106**: 729 – 731.
- Allison KH, Reed SD, Voigt LF, Jordan CD, Newton KM, Garcia RL. Diagnosing endometrial hyperplasia: why is it so difficult to agree? *Am J Surg Pathol*. 2008; **32**: 691 – 698.
- Sherman ME, Ronnett BM, Ioffe OB, Richesson DA, Rush BB, Glass AG, et al. Reproducibility of biopsy diagnoses of endometrial hyperplasia: evidence supporting a simplified classification. *Int J Gynecol Pathol*. 2008; **27**: 318 – 325.
- Baak JPA, Mutter GL. EIN and WHO94. *J Clin Pathol*. 2005; **58**: 1 – 6.
- Sivridis E, Giatromanolaki A. The endometrial hyperplasias revisited. *Virchows Arch*. 2008; **453**: 223 – 231.
- Mutter GL, Baak JPA, Crum CP, Richart RM, Ferenczy A, Faquin WC. Endometrial precancer diagnosis by histopathology, clonal analysis, and computerized morphometry. *J Pathol*. 2000; **190**: 462 – 469.
- Scully RE, Young RH. Endometrioid neoplasia retrogressive terminology. *Am J Surg Pathol*. 2000; **24**: 753 – 755.