



Ali, A. et al. (2016) Investigating various thresholds as immunohistochemistry cutoffs for observer agreement. *Applied Immunohistochemistry & Molecular Morphology*.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/119057/>

Deposited on: 21 June 2016

Enlighten – Research publications by members of the University of Glasgow
<http://eprints.gla.ac.uk>

Investigating various thresholds as immunohistochemistry cut-offs for observer agreement

Authors:

Asif Ali^{1,2} PhD, Sarah Bell³MRCPath, Alan Bilisland¹PhD, Jill Slavin³MRCPath, Victoria Lynch³ MRCPath, Maha Elgoweini³ MRCPath, Mohammad H Derakhshan⁴ PhD, Nigel B Jamieson^{5,6} PhD, David Chang¹ PhD, Victoria Brown⁷ MSc, Simon Denley⁵ MD, Clare Orange¹ PhD, Colin McKay⁵ FRCS, Ross Carter⁵ FRCS, Karin A Oien^{1,3} PhD and Fraser R Duthie³ FRCPath

Affiliations:

¹Institute of Cancer Sciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, G61 1QH, UK

²Institute of Basic Medical Sciences, Khyber Medical University, Phase 5, Hayatabad, KPK, Peshawar, Pakistan

³Department of Pathology, Laboratory Medicine Building, Queen Elizabeth University Hospital, Greater Glasgow & Clyde NHS, Glasgow, G51 4TF, UK

⁴Institute of Cardiovascular and Medical Sciences, University of Glasgow, Western Infirmary, Glasgow, G11 6NT, UK

⁵West of Scotland Pancreatic Unit and Glasgow Royal Infirmary, Alexandra Parade, Glasgow G31 2ER, UK

⁶Academic Unit of Surgery, School of Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow Royal Infirmary, Glasgow, G4 OSF, UK

⁷Pathology Laboratory, Forth Valley Royal Hospital, Stirling Road, Larbert FK5 4WR, UK

Email Addresses

Asif Ali: draliasif7@gmail.com (corresponding author)

Sarah Bell: sarah.bell5@nhs.net

Alan Bilisland: Alan.Bilisland@glasgow.ac.uk

Jill Slavin: jillslavin@nhs.net

Victoria Lynch: victoria.jeffrey@gmail.com

Maha Elgoweini: mahaelgoweini@doctors.org.uk

Mohammad H Derakhshan: Mohammad.Derakhshan@glasgow.ac.uk

Nigel B Jamieson: Nigel.Jamieson@glasgow.ac.uk

David Chang: David.chang@glasgow.ac.uk

Victoria Brown: v.brown3@nhs.net

Simon Denley: smdenley@hotmail.com

Clare Orange: Clare.Orange@glasgow.ac.uk

Colin McKay: Colin.McKay@ggc.scot.nhs.uk

Ross Carter: rosscarterno1@gmail.com

Karin A Oien: Karin.Oien@glasgow.ac.uk

Fraser R Duthie: fraser.duthie@ggc.scot.nhs.uk

Acknowledgements

We thank Professor Alan Foulis for participating in the study and providing invaluable feedback on study design. AA is supported by a PhD studentship from the Khyber Medical University, Peshawar, Pakistan under the Higher Education Commission of Pakistan grants.

Abstract

Background

Clinical translation of immunohistochemistry (IHC) biomarkers requires reliable and reproducible cut-offs or thresholds for interpretation of immunostaining. Most IHC biomarker research focuses on the clinical relevance (diagnostic, prognostic or predictive utility) of cut-offs, with less emphasis on observer agreement using these cut-offs. From the literature, we identified three commonly used cut-offs of 10% positive epithelial cells, 20% positive epithelial cells and moderate to strong staining intensity (+2/+3 hereafter) to use for investigating observer agreement.

Materials and Methods

A series of 36 images of microarray cores stained for four different IHC biomarkers, with variable staining intensity and percentage of positive cells, was used for investigating inter- and intra-observer agreement. Seven pathologists scored the immunostaining in each image using the three cut-offs for positive and negative staining. Kappa statistic was used to assess the strength of agreement for each cut-off.

Results

The inter-observer agreement between all seven pathologists using the three cut-offs was reasonably good, with mean κ scores 0.64, 0.59 and 0.62

respectively for 10%, 20% and +2/+3 cut-offs. A good agreement was observed for experienced pathologists using the 10% cut-off and their agreement was statistically higher than for junior pathologists ($p=0.02$). In addition, the mean intra-observer agreement for all seven pathologists using the three cut-offs was reasonably good, with mean κ scores 0.71, 0.60 and 0.73 respectively for 10%, 20% and +2/+3 cut-offs. For all three cut-offs, a positive correlation was observed with perceived ease of interpretation ($p<0.003$). Finally, cytoplasmic-only staining achieved higher agreement using all three cut-offs than mixed staining patterns.

Conclusions

All three cut-offs investigated achieve reasonable strength of agreement modestly decreasing inter and intra-observer variability in IHC interpretation. These cut-offs have previously been used in cancer pathology and this study provides evidence that these cut-offs can be reproducible between practising pathologists.

Keywords

Observer agreement, Kappa, immunohistochemistry, biomarker, cut-off

Introduction

The use of immunohistochemistry (IHC) biomarkers for clinical decision making is an important research field with significant translational potential. A multitude of biomarkers for a variety of cancers is available and a large literature exists on novel biomarker discovery, but only a minority impact upon patient care.

Amongst other reasons, one barrier to clinical translation of biomarkers is the lack of a standardised cut-off or threshold for interpretation of IHC staining^{1,2}.

Evaluation of immunostaining is important in translational studies assessing biomarker expression for diagnostic, prognostic or predictive purposes.

Biomarker expression assessment usually employs a continuous or ordinal scale; but for meaningful clinical use it is usually dichotomised and a cut-off established for assigning a patient into either positive/negative expression category or high/low expression category³. In addition, for some biomarkers, more than two categories may be required for example the 'Allred score' for estrogen receptor positivity⁴. For clinical translation, there are two main issues in the development and application of a standardised cut-off for IHC biomarkers. One is the identification of an appropriate cut-off that provides suitable sensitivity/specificity for diagnostic biomarkers or that stratifies patients based on survival and response to treatment for prognostic and predictive biomarkers respectively. The other issue is to assess the inter- and intra-observer agreement in the interpretation of a cut-off threshold. One potential strategy to address the former is the use of a receiver operating characteristics (ROC) curve that can

help to identify an appropriate cut-off^{1,5}. The latter issue can be answered by assessing the level of agreement between pathologists⁶⁻⁸.

There is currently no standardized cut-off for diagnostic IHC biomarkers. Most of the reported cut-offs are purposive that best fit cancer or normal groups. These cut-offs are based on the intensity of staining or percentage of positive cells or on a combination of both intensity and percentage in terms of immunoreactive scores, H scores and "quick" scores^{7,9-13}.

Two widely used cut-offs reported in the literature for IHC diagnostic biomarkers are positive/negative staining (e.g. p16/Ki-67 staining for the diagnosis of cervical intra-epithelial neoplasia 3) and 10% positive epithelial cells (e.g. a panel of napsin A, TTF 1, CK 5, and p63 in differentiating adenocarcinoma from squamous cell carcinoma of the lung)¹⁴⁻¹⁹. Other reported cut-offs are: more than 30% cells with uniform, intense membranous staining of invasive tumour cells for human epidermal growth factor receptor 2 (HER2) (positive HER2 staining in breast cancer)²⁰; and more than 5% positive tumour cells for CK 7 and CK 20 (differential diagnosis in metastatic carcinoma of unknown origin)²¹.

These scoring systems and cut-offs have been adopted for research purposes. Some of them are used in clinical practice; but studies looking at their reproducibility between pathologists are few. A cut-off should be both clinically relevant and easily interpretable by pathologists. There is a tendency to focus more on the clinical relevance of the cut-off for a biomarker with less focus on the level of agreement between pathologists when they use it for scoring purposes^{6,22}. Inter- and intra-observer variation of a cut-off is infrequently

analysed despite the fact that it is recognised as a potential barrier to clinical translation.

We selected three cut-offs for investigation based on our diagnostic IHC work²³ and the wider IHC literature. These cut-offs are 10% positive epithelial cells (10% hereafter), 20% positive epithelial cells (20% hereafter) and moderate to strong staining intensity with any proportion of positive cells (+2/+3 hereafter)^{8,14,17,23-27}. These cut-offs are clinically relevant and we postulated that they are easily interpretable and reproducible amongst pathologists. The aims of the current study were to investigate the cut-offs (10%, 20% and +2/+3) for inter- and intra-observer agreement; and to explore factors influencing agreement between pathologists for IHC cut-offs.

Materials and Methods

Immunohistochemistry images and participants

A series of 36 images of pancreatic ductal adenocarcinoma (PDAC) tissue microarray cores for four diagnostic IHC biomarkers (nine images each from KOC, maspin, mesothelin and S100P) were used for this study. These cores have previously been studied for diagnostic utility²³. These cores were carefully selected for each biomarker based on a variable range of staining intensity and proportion of positive cells. Some cores with no immunostaining were also included. The purpose of using images from one type of tumour i.e. PDAC was to allow the observers to concentrate on the immunostaining cut-offs rather than interpreting the morphology of different tumours. KOC expression was cytoplasmic; maspin has both cytoplasmic and nuclear expression but the pathologists were asked to score only cytoplasmic staining for maspin and disregard nuclear staining; mesothelin expression was cytoplasmic and/or membranous; and S100P expression was cytoplasmic and/or nuclear. Seven pathologists (three experienced pathologists and four junior pathologists) participated in the study. Experienced pathologists have clinical pathology experience of more than 15 years, while junior pathologists have 3-7 years of experience. All pathologists were sufficiently trained to evaluate pancreatic tumours. Pathologists were coded as A, B, C, D, E, F and G. Ethical approval has been granted by the North Glasgow University Hospitals NHS Trust Ethics Committee and by the National Health Service Greater Glasgow and Clyde Ethics Committee. This ethics approval includes the use of archival pathology

specimens, where the patients were not given the opportunity to donate their tissue.

Scoring the IHC cut-offs

The 36 images were shown via projection on Powerpoint, and were arranged based on biomarkers with reference staining intensities (weak, moderate, severe) provided at the start for each biomarker. A scoring sheet with instructions on scoring was prepared with the help of pathology colleagues (Supplementary Table 1). All the participating pathologists participated in one session for the inter-observer part of the study. After a short presentation (5-10 min) on the purpose of this study, the scoring sheets were distributed between all seven pathologists. Each image was shown for one minute. The pathologists were asked to interpret the immunostaining of each image for the three cut-offs as a binary categorical variable, "present" or "absent". The three cut-offs were: 10%, 20% and +2/+3 cut-offs. For example, a 10% cut-off is "present" when more than or equal to 10% epithelial cells are positive in the desired subcellular compartment and is "absent" when fewer than 10% epithelial cells are positive. Each core was also recorded as being easy (1) or challenging (2) to score.

All seven pathologists participated in the intra-observer part of the study three weeks after the inter-observer session. The tissue core images shown were the same, but arranged in a different order to minimise recall bias.

Statistics and data analysis

We used kappa (κ) scores as a measure of the strength of agreement between pathologists for all three cut-offs. Kappa scores reflect the strength of agreement between observations, adjusted for chance agreement, and can range from 0 to 1. We used the standards suggested by Landis and Koch²⁸ for the interpretation of strength of agreement. Kappa scores are shown in six categories from 0-1 and each category is colour coded (Supplementary Table 2).

Inter-observer agreement: to determine inter-observer agreement for each of the three cut-offs, each pathologist's interpretation of immunostaining was compared with that of the other pathologists in a pair-wise manner. 21 inter-observer (AB, AC, AD and so on...) κ scores were generated for each of the three cut-offs (10%, 20% and +2/+3). Finally a mean inter-observer κ score for each cut-off was used as a measure of strength of agreement between pathologists.

Impact of pathologists' experience and antibody staining pattern on inter-observer agreement: for each cut off, mean inter-observer κ scores were calculated for experienced pathologists and for junior pathologists and then compared. The staining pattern was noted for the antibody used in each slide and mean κ scores calculated for each staining pattern. The aim was to determine if the pathologists tend to have more agreement for a particular staining pattern (cytoplasmic, nuclear and/or membranous).

To determine whether these three cut-offs are statistically different from each other, the paired sample t test (for large sample size) and Wilcoxon signed

ranked test (for small sample size) were used to compare the pairwise κ scores. To determine which cut-off is most easily scored, these three cut-offs as predictor variables were put in a linear regression model against perceived ease of scoring as a dependent variable.

Intra-observer agreement: to determine reproducibility of these three cut-offs, kappa scores were generated comparing scoring and re-scoring of the same image arranged in different orders three weeks apart for each pathologists. Kappa scores were generated for all seven pathologists (A-A, B-B, C-C, D-D, E-E, F-F and G-G) using the three cut-offs. Seven intra-observer agreements were generated for each cut-off. A mean intra-observer κ score for each cut-off was then used as a measure of strength of agreement.

A p value <0.05 was considered statistically significant. SPSS version 21 was used for statistical analyses.

Results

Taken together, 1512 evaluations were made in the inter- and intra-observer sessions by the pathologists. The average time for interpretation of an image was roughly 30-45 seconds. Results are divided into three parts: inter-observer agreement; perceived ease of scoring; and intra-observer agreement.

Inter-observer agreement

All seven pathologists

The mean inter-observer κ scores were 0.64, 0.59 and 0.62 for 10%, 20% and +2/+3 cut-offs respectively (Table 1). The mean κ score agreement for 10% and +2/+3 cut-offs is in the 'substantial' agreement category and for the 20% cut-off is in the 'moderate' agreement category. However, the κ score agreements between the three cut-offs were not statistically different from each other (Supplementary Figure 1).

Figure 1 shows examples of IHC images used in this study. Images with low observer agreement have either weak staining intensity, or the proportion of positively stained cells is lower compared to images with high level agreement. In fact, tissues with both strong staining intensity and a higher percentage of positive cells have higher agreement regardless of the biomarker and staining pattern.

In summary, the inter-observer agreements between all seven pathologists for the three cut-offs were reasonably good. In addition, the agreements for the cut-offs were not statistically different from each other.

Impact of pathologists' experience on inter-observer agreement

The mean inter-observer κ scores for experienced pathologists were 0.81, 0.70 and 0.55 for 10%, 20% and +2/+3 cut-offs respectively (Table 2A). The mean inter-observer κ scores for junior pathologists were 0.61, 0.60 and 0.73 for 10%, 20% and +2/+3 cut-offs respectively (Table 2B). The agreement on 10% cut-off is statistically higher for the experienced pathologists than the junior pathologists ($P=0.02$, Mann-Whitney U test). However, no statistically significant difference between experienced and junior pathologists was observed for 20% and +2/+3 cut-offs.

In summary, a higher level of agreement was observed for experienced pathologists using 10% cut-off and this was statistically higher than junior pathologists.

Impact of antibody staining pattern on inter-observer agreement

In the images studied, there were three staining patterns: cytoplasmic only staining; cytoplasmic and/or nuclear staining (CN); and cytoplasmic and/or membranous staining (CM).

The mean κ scores for cytoplasmic only staining were higher than the other staining patterns. More specifically, a statistically higher agreement for

cytoplasmic only staining was observed in the following scenarios: cytoplasmic compared to CN category using +2/+3 cut-off; and cytoplasmic compared to CM category using 20% and +2/+3 cut-offs.

Moreover, a statistically higher agreement for CN staining was observed in the following scenarios: CN compared to CM category using 20% and +2/+3 cut-offs. No statistically significant difference between different staining patterns was observed for the 10% cut-off (Table 3).

In summary, there is more inter-observer agreement for cytoplasmic only staining followed by CN and CM. Finally the 10% cut-off appears to yield good inter-observer agreement irrespective of the staining compartment of cell.

Relationship between cut-offs and perceived ease of scoring

A positive correlation was observed between all three cut-offs and perceived ease of scoring ($p < .0001$). However, in a multivariate analysis the 10% cut-off ($\beta = 0.41$, $p < 0.001$) was more easily scored as compared to the +2/+3 cut-off ($\beta = 0.38$, $p = 0.001$) or the 20% cut-off ($\beta = 0.34$, $p = 0.004$) (Table 4).

Interestingly, the pattern emerging from this correlation, that 10% is relatively more easily scored, followed by +2/+3 and 20%, supports the mean inter- and intra-observer κ scores for these cut-offs (Table 1 and Table 5).

Intra-observer agreements

The mean intra-observer κ scores were 0.71, 0.60 and 0.73 for 10%, 20% and +2/+3 cut-offs respectively (Table 5). The κ score agreement for 10% and +2/+3 cut-offs is in the 'substantial' agreement category and for the 20% cut-off is in the 'moderate' agreement category. However, the κ score intra-observer agreements between all seven pathologists for the three cut-offs were not statistically different (Supplementary Figure 2).

In summary, the intra-observer agreements for the three cut-offs were reasonably good. In addition, the agreements for the three cut-offs were not statistically different from each other. Thus a good intra-observer agreement confirms the reproducibility of these cut-offs by pathologists and again this supports their use for IHC biomarkers.

The inter- and intra-observer agreements follow the same pattern i.e. 'substantial' agreement for 10% and +2/+3 and 'moderate' agreement for 20% cut-offs.

Discussion

Three IHC cut-offs, namely 10%, 20% and +2/+3 were assessed for observer agreement between pathologists. All cut-offs demonstrated good inter- and intra-observer agreement between pathologists. Similarly, all three cut-offs showed high correlation with perceived ease of scoring. Finally, the observer agreement for cytoplasmic only staining was higher than cytoplasmic/nuclear staining and cytoplasmic/membranous staining.

Establishing a cut-off for biomarker assessment is an essential pre-requisite for clinical translation. A wide range of cut-offs have been used for diagnostic, prognostic and predictive IHC biomarkers in research and clinical settings. The purpose of a cut-off for a diagnostic biomarker is to assign patients into positive or negative categories with reasonable sensitivity without compromising specificity²⁹. Based on the expression level for a candidate biomarker in cancer and normal tissue, a cut-off is established. A good diagnostic cut-off has a low probability of false positivity and false negativity²⁹. The purpose of a cut-off for a prognostic biomarker is to divide the population into categories of longer and shorter survival for the outcome. In research settings a cut-off based on percentage of positive tumour cells is mostly used^{30,31}. Similarly, the aim of a cut-off for predictive biomarkers is to stratify patients into likely responders and non-responders to treatment and intervention³².

IHC cut-offs used for prognostic and predictive biomarkers have been investigated for observer agreement but such studies are limited for diagnostic

biomarkers. The cut-offs of 10% and 30% positive cells with strong membranous staining for HER2 have been investigated for reproducibility amongst pathologists⁸. In addition, for oestrogen receptor (ER) and progesterone receptor (PR), the continuous H-score (range 0-300) and categorical scores (negative: H-score<1, positive: H-score ≥1) have been investigated for inter-observer agreement⁷. These cut-offs for HER2, PR and ER are clinically important and are used in clinical practice by pathologists.

Clinically relevant cut-offs are important for biomarker evaluation. We sought to investigate three cut-offs i.e. 10%, 20% and +2/+3 with the hope that if evidence of their scoring reproducibility is provided, they could potentially help the clinical translation of IHC biomarkers. Interestingly, the purpose of cut-offs differ for different biomarkers but these three cut-offs have been used for diagnostic (S100P, pVHL, KIT, HMG1(Y), CK20, P53, Ki-67)^{17,26,27,33,34}, prognostic (Ki-67, p53) and predictive (APAF-1, EGFR) biomarkers³⁵⁻³⁸. Therefore, investigating the strength of agreement between pathologists for these three cut-offs has significant clinical importance.

Inter-observer agreement between pathologists was used to elucidate the reliability of cut-offs. A 'substantial' agreement was observed with overall mean κ scores of 0.64 and 0.62 for 10% and +2/+3 cut-offs respectively, whereas 'moderate' agreement with a κ score of 0.59 was observed for 20% cut-off. In a study comparing the 10% positivity with 30% positivity for HER2, the mean κ scores for inter-observer agreement were 0.49 for 10% positive cells and 0.54 for 30% positive cells⁸. Clearly, the κ scores generated for the three cut-offs under investigation in our project are comparable to the κ scores for HER2 which is

already in clinical practice as a predictive biomarker. Moreover, studies looking at the inter-observer reproducibility in histopathology and the IHC literature have shown that κ scores more than 0.60 (substantial agreement) are regarded as a good level of agreement. In comparison, κ scores less than 0.40 (fair agreement) are regarded as an unacceptably low level of agreements for diagnostic purposes³⁹⁻⁴³.

Intra-observer agreement of the scoring and then re-scoring of the same image was used to assess reproducibility of the cut-offs. Again a pattern similar to inter-observer agreement emerged with 'substantial' agreements for the 10% and +2/+3 cut-offs and 'moderate' agreement for 20% cut-off. However, the intra-observer agreements (0.71, 0.60 and 0.73) in the present study are higher than inter-observer agreements (0.64, 0.59 and 0.62) for the three cut-offs. This finding agrees with the previous literature that the intra-observer agreement is more than the inter-observer agreement. For example the intra-observer agreement ($\kappa=0.85$) is better than the inter-observer agreement ($\kappa=0.80$) for PDX-1 IHC staining intensity in prostate cancer⁴⁴. In addition, the intra-observer agreement ($\kappa=0.78$) is better than the inter-observer agreement ($\kappa=0.65$) for evaluation of focal cortical dysplasia categories⁴⁵.

Taking 10% positive cells as a cut-off has been used for a variety of IHC biomarkers in different cancer types. These include S100P and XIAP in the differentiation of pancreatic cancer from non-neoplastic pancreatic tissue, and for a panel of napsin A, TTF 1, CK 5, and P63 in differentiating adenocarcinoma from squamous cell carcinoma of the lung^{19,46}. Moreover, 10% cut-off is prognostic in breast cancer for a panel of Ki67 and p53, predictive of event-free

survival in stage II colon cancer for VEGF and is predictive in rectal tumours treated with preoperative, high-dose-rate brachytherapy for APAF-1^{36,38,47}. The use of a 10% cut-off in other areas of pathology means that the more experienced pathologists in the present study will have already had experience in applying this cut-off, which is a possible explanation for why they have a higher agreement than junior pathologists. Studies have attempted to show the reproducibility of the 10% cut-off and the κ scores achieved in the current study (0.64, substantial agreement) is similar to the κ scores (0.57-0.77, moderate to substantial) in the reported literature^{36,48-50}.

The 20% positive staining cut-off has also been used for a variety of IHC biomarkers. These include, Ki-67 as a prognostic biomarker in breast carcinoma⁵¹ and NF-E2 in the differentiation of essential thrombocythemia from primary myelofibrosis⁵². However, studies investigating the variation in interpretation of this cut-off between pathologists are very limited. The current study investigated the 20% cut-off for observer agreement and our results suggested a good level of agreement.

Moderate to strong staining intensity and any percentage of positive cells (+2/+3) as a cut-off has also been used for IHC biomarkers. These include CK20, P53, CK5/6, CD138, and Her2/Neu in the diagnosis of urothelial carcinoma in situ and the use of claudin-4 to distinguish adenocarcinoma from malignant mesothelioma in effusion cytology^{26,53,54}. However, once again studies observing the variation in interpretation of this cut-off between pathologists are very limited. Our results demonstrate that this cut-off is also reliable, reproducible

and easy to score and it can be ranked second to the 10% cut-off from the current study.

The observer agreement was also assessed using staining in different cellular compartments. Staining in only the cytoplasmic compartment achieved higher agreements than other staining patterns.

The interpretation of membranous staining for HER2 in breast cancer is used in routine clinical practice. Hameed et al ⁸ investigated inter-observer agreement using 10% positive cells with membranous staining for HER2 in breast cancer. The authors found a mean inter-observer agreement κ score of 0.49 ⁸. We also investigated inter-observer agreement using 10% positive cells with membranous staining for mesothelin in pancreatic cancer and observed κ score agreement of 0.62. Thus 10% cut-off and membranous staining achieve reasonable observer agreement not only for HER2 in breast cancer but for other biomarker in a different cancer and warrants further investigations.

The sample size was good and seven pathologists participated in the present study. This number is comparable to the IHC biomarker and histopathology literature (4 to 7 participants) where observer agreement was investigated ^{44,55-57}. In addition, the participants in the current study were practising pathologists with variable levels of experience as compared to studies where either physicians (with no formal pathology experience) ⁴⁴ or researchers with experience in IHC were recruited ⁵⁶. Thus the results of this study provide good evidence on the use of cut-offs for IHC biomarkers.

The limitations include: the relatively few number of images due to the time constraints imposed by the clinical work of the pathologists; and all pathologists were from the same institution; the aim was to carry out the study with all of the pathologists present at one session and this was achieved for the inter-observer part but for the intra-observer part we had to arrange an extra session. An important limitation results from the fact that images were shown as a PowerPoint presentation on screen rather than using a standard microscope.

Conclusions

In a day-to-day clinical practice pathologists need scoring systems and cut-offs that are reproducible and easy to use^{1,8}. A wide range of cut-offs have been used for IHC biomarkers. We selected 10%, 20% and +2/+3 cut-offs that have been utilised previously in clinical practice. These three cut-offs are reliable and reproducible achieving a reasonably good agreement level between pathologists (when compared with the literature). They could facilitate translational biomarker studies and could potentially be used by scientists who are not trained pathologists but are involved in investigating IHC biomarkers. A biomarker achieving diagnostic, prognostic and/or predictive significance with any of the three cut-offs may have translational potential. Further studies are required to assess these cut-offs with pathologists from different institutions and using a larger sample of images.

Abbreviations

IHC: Immunohistochemistry

10% cut-off: 10% positive epithelial cells

20% cut-off: 20% positive epithelial cells

+2/+3 cut-off: moderate to strong staining intensity with any proportion of positive cells

K: Kappa

ROC: Receiver operating characteristic curve

HER2: Human epidermal growth factor receptor 2

ER: Oestrogen receptor

PR: Progesterone receptor

H-score: Histoscore

CN: Cytoplasmic and/or nuclear staining

CM: Cytoplasmic and/or membranous staining

CK7: Cytokeratin 7

CK20: Cytokeratin 20

TTF1: Thyroid transcription factor 1

HMGI: High-mobility group protein I

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AA, KAO and FRD designed the study, performed the statistical analyses and drafted the manuscript. MHD, SB and AB provided assistance for the statistical analyses and helped draft the manuscript. SB, JS, VL, ME, KAO and FRD evaluated the immunostaining and helped draft the manuscript. NBJ, DC, VB, SD, CO, CM, RC provided samples, performed immunostaining and helped draft the manuscript.

Tables

Table 1: Pairwise k scores of inter-observer agreements between pathologists for the three cut-offs.

10% Cut-Off								
Observers								
Observers		A	B	C	D	E	F	G
	A		0.8	0.82	0.62	0.47	0.72	0.89
	B			0.82	0.62	0.47	0.53	0.89
	C				0.48	0.36	0.58	0.72
	D					0.48	0.55	0.72
	E						0.74	0.54
	F							0.6
	G							
Mean k score				0.64 (95% CI, 0.57-0.70)				
20% Cut-Off								
Observers								
Observers		A	B	C	D	E	F	G
	A		0.85	0.57	0.64	0.54	0.49	0.92
	B			0.7	0.53	0.43	0.38	0.77
	C				0.64	0.42	0.48	0.51
	D					0.62	0.82	0.56
	E						0.71	0.46
	F							0.42
	G							
Mean k score				0.59 (95% CI, 0.52-0.66)				
+2/+3 Cut-Off								
Observers								
Observers		A	B	C	D	E	F	G
	A		0.75	0.42	0.6	0.55	0.46	0.65
	B			0.48	0.54	0.49	0.52	0.59
	C				0.55	0.61	0.72	0.61
	D					0.58	0.6	0.7
	E						0.88	0.88
	F							0.77
	G							
Mean k score				0.62 (95% CI, 0.56-0.67)				

Note: Comparison of pairwise k scores with colour codes between pathologists (A-G) in the evaluation of immunohistochemistry using 10%, 20% and +2/+3 cut-offs are shown with mean k score and 95% CI separately for each cut-off.

Table 2: Pairwise k scores of inter-observer agreements between experienced and junior pathologists for the three cut-offs.

A				
Experienced pathologists				
10% Cut-Off				
Observers				
Observers		A	B	C
	A		0.80	0.82
	B			0.82
Mean K score		0.81 (Range: 0.80-0.82)		
20% Cut-Off				
Observers				
Observers		A	B	C
	A		0.85	0.57
	B			0.70
Mean K score		0.71 (Range: 0.57-0.85)		
+2/+3 Cut-Off				
Observers				
Observers		A	B	C
	A		0.75	0.42
	B			0.48
Mean K score		0.55 (Range: 0.42-0.75)		

B					
Junior pathologists					
10% Cut-Off					
Observers					
Observers		D	E	F	G
	D		0.48	0.55	0.72
	E			0.74	0.54
	F				0.60
	G				
Mean k score		0.61 (Range: 0.48-0.74)			
20% Cut-Off					
Observers					
Observers		D	E	F	G
	D		0.62	0.82	0.56
	E			0.71	0.46
	F				0.42
	G				
Mean k score		0.60 (Range: 0.42-0.82)			
+2/+3 Cut-Off					
Observers					
Observers		D	E	F	G
	D		0.58	0.60	0.7
	E			0.88	0.88
	F				0.77
	G				
Mean k score		0.73 (Range: 0.58-0.88)			

A, Comparison of pairwise K scores with colour codes between experienced pathologists (A-C) in the evaluation of immunohistochemistry using 10%, 20% and +2/+3 cut-offs shown with mean k score and 95% CI separately for each cut-off.

B, Comparison of pairwise K scores with colour codes between junior pathologists (D-G) in the evaluation of immunohistochemistry using 10%, 20% and +2/+3 cut-offs shown with mean k score and 95% CI separately for each cut-off.

Table 3: Mean k scores with p values for staining of different cellular compartments.

Cut-Offs	C vs. CN		C vs. CM		CN vs. CM	
	mean	p value*	mean	p value	mean	p value
10%	0.77 vs. 0.71	0.380	0.77 vs. 0.64	0.150	0.71 vs. 0.64	0.500
20%	0.75 vs. 0.63	0.100	0.75 vs. 0.40	<0.001	0.63 vs. 0.40	0.009
+2/+3	0.81 vs. 0.58	0.001	0.81 vs. 0.40	<0.001	0.58 vs. 0.40	0.010

Note: *Paired sample t test

Abbreviations: C (cytoplasmic staining), CM (cytoplasmic and membranous staining) and CN (cytoplasmic and nuclear staining).

Table 4: Multivariable linear regression for 10%, 20% and +2/+3 cut-offs as predictor variables and perceived ease of interpretation as dependent variable.

Variables	Unstandardized Coefficients		Standardized Coefficients	P value
	B	Std. Error	Beta	
(Constant)	-5.67	1.07		<0.001
10% Cut-off	0.71	0.16	0.41	<0.001
+2/+3 Cut-off	0.46	0.12	0.38	0.001
20% Cut-off	0.51	0.16	0.34	0.004

Note: The 10%, 20% and +2/+3 are predictor variables i.e. they are variables that are predicting an outcome (the ease of interpretation). In this regression model ease of interpretation is a dependent variable i.e. a variable which “depends” on the predictor variable. The standardised beta coefficients were used as an estimate of association between predictor and dependent variable. The higher the beta coefficient the higher is the p-value significance and the stronger is the association between predictor and dependent variables. Beta coefficient in this model is highest (0.41) for 10% cut-off, followed by +2/+3 (0.38) and 20% (0.34). However, the p-value for all three cut-offs is significant showing a positive association with ease of interpretation.

Table 5: Pairwise k scores of intra-observer agreements for pathologists for the three cut-offs.

K Scores						
Codes	10%	P value	20%	P value	+2/+3	P value
A-A	0.76	<0.001	0.74	<0.001	0.80	<0.001
B-B	0.89	<0.001	0.68	<0.001	0.72	<0.001
C-C	0.84	<0.001	0.59	<0.001	0.50	0.003
D-D	0.43	0.002	0.55	0.001	0.59	<0.001
E-E	0.68	<0.001	0.59	<0.001	0.88	<0.001
F-F	0.47	0.003	0.51	0.002	0.94	<0.001
G-G	0.87	<0.001	0.53	0.001	0.68	<0.001
Mean (95% CI)	0.71 (0.53-0.88)		0.60 (0.52-0.68)		0.73 (0.59-0.87)	

Note: Pairwise k scores showing intra-observer reproducibility from scoring and re-scoring (for example A-A) of all seven pathologists (A-G) in the evaluation of Immunohistochemistry using 10%, 20% and +2/+3 cut-offs.

Figure Legends:

Figure 1: Representative images of high and low inter-observer agreement between all pathologists for 10%, 20% and +2/+3 cut-offs.

Figure 1 legend: The high (left column grid) and low (right column grid) inter-observer agreement of IHC interpretation is shown for the three cut-offs. The staining for the three cut-offs was recorded only in the tumour epithelium. The high level agreement is attributed to the strong staining intensity and higher proportion of positive cells as illustrated in left column grid. All pathologists agreed on the images in the left column grid for all three cut-offs. However, there were differences in the number of pathologists agreeing on the images in the right column grid. For 10% cut-off (right upper image) 4/7 pathologists agreed, for 20% cut-off (right middle image) 5/7 pathologists agreed and for the +2/+3 cut-off (right lower image) 3/7 pathologists agreed.

References

1. Zlobec I, Steele R, Terracciano L, Jass JR, Lugli A. Selecting immunohistochemical cut-off scores for novel biomarkers of progression and survival in colorectal cancer. *Journal of clinical pathology*. Oct 2007;60(10):1112-1116.
2. Budczies J, Klauschen F, Sinn BV, et al. Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. *PloS one*. 2012;7(12):e51862.
3. Mazumdar M, Glassman JR. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Statistics in medicine*. Jan 15 2000;19(1):113-132.
4. Harvey JM, Clark GM, Osborne CK, Allred DC. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. May 1999;17(5):1474-1481.
5. Weiss HL, Niwas S, Grizzle WE, Piyathilake C. Receiver operating characteristic (ROC) to determine cut-off points of biomarkers in lung cancer patients. *Disease markers*. 2003;19(6):273-278.
6. Borlot VF, Biasoli I, Schaffel R, et al. Evaluation of intra- and interobserver agreement and its clinical significance for scoring bcl-2 immunohistochemical expression in diffuse large B-cell lymphoma. *Pathology international*. Sep 2008;58(9):596-600.
7. Cohen DA, Dabbs DJ, Cooper KL, et al. Interobserver agreement among pathologists for semiquantitative hormone receptor scoring in breast carcinoma. *American journal of clinical pathology*. Dec 2012;138(6):796-802.
8. Hameed O, Adams AL, Baker AC, et al. Using a higher cutoff for the percentage of HER2+ cells decreases interobserver variability in the interpretation of HER2 immunohistochemical analysis. *American journal of clinical pathology*. Sep 2008;130(3):425-427.
9. Kim MJ, Shin HC, Shin KC, Ro JY. Best immunohistochemical panel in distinguishing adenocarcinoma from squamous cell carcinoma of lung: tissue microarray assay in resected lung cancer specimens. *Annals of diagnostic pathology*. Feb 2013;17(1):85-90.
10. Galgano MT, Castle PE, Atkins KA, Brix WK, Nassau SR, Stoler MH. Using biomarkers as objective standards in the diagnosis of cervical biopsies. *The American journal of surgical pathology*. Aug 2010;34(8):1077-1087.
11. Boltze C, Schneider-Stock R, Aust G, et al. CD97, CD95 and Fas-L clearly discriminate between chronic pancreatitis and pancreatic ductal adenocarcinoma in perioperative evaluation of cryocut sections. *Pathology international*. Feb 2002;52(2):83-88.
12. Oh YL, Song SY, Ahn G. Expression of maspin in pancreatic neoplasms: application of maspin immunohistochemistry to the differential diagnosis. *Appl Immunohistochem Mol Morphol*. Mar 2002;10(1):62-66.
13. Cheuk W, Wong KO, Wong CS, Dinkel JE, Ben-Dor D, Chan JK. Immunostaining for human herpesvirus 8 latent nuclear antigen-1 helps

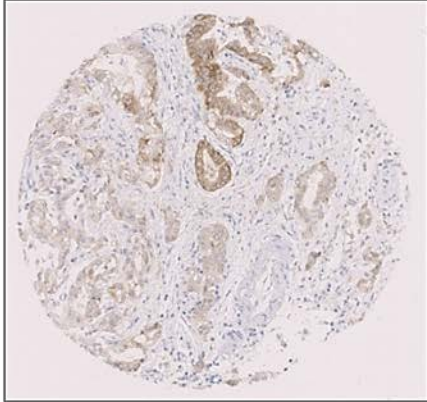
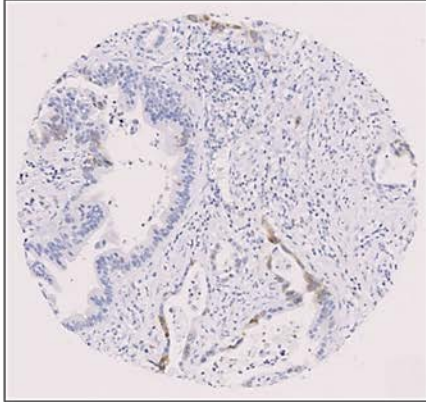
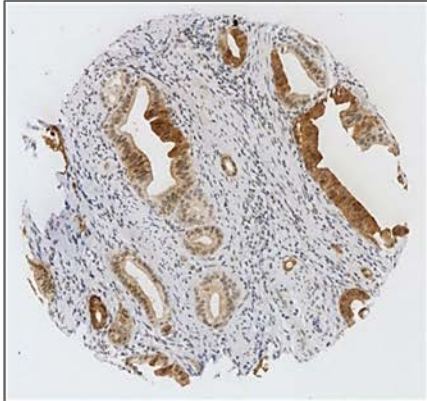
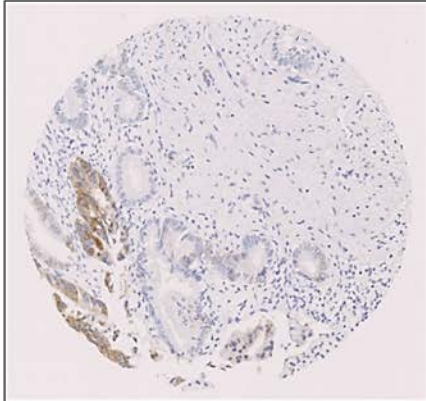
- distinguish Kaposi sarcoma from its mimickers. *American journal of clinical pathology*. Mar 2004;121(3):335-342.
14. Lin F, Shi J, Liu H, et al. Diagnostic utility of S100P and von Hippel-Lindau gene product (pVHL) in pancreatic adenocarcinoma-with implication of their roles in early tumorigenesis. *The American journal of surgical pathology*. Jan 2008;32(1):78-91.
 15. Maass N, Hojo T, Ueding M, et al. Expression of the tumor suppressor gene Maspin in human pancreatic cancers. *Clinical cancer research : an official journal of the American Association for Cancer Research*. Apr 2001;7(4):812-817.
 16. Strickland LA, Ross J, Williams S, et al. Preclinical evaluation of carcinoembryonic cell adhesion molecule (CEACAM) 6 as potential therapy target for pancreatic adenocarcinoma. *The Journal of pathology*. Jul 2009;218(3):380-390.
 17. Yamaguchi U, Hasegawa T, Sakurai S, et al. Interobserver variability in histologic recognition, interpretation of KIT immunostaining, and determining MIB-1 labeling indices in gastrointestinal stromal tumors and other spindle cell tumors of the gastrointestinal tract. *Applied immunohistochemistry & molecular morphology : AIMM / official publication of the Society for Applied Immunohistochemistry*. Mar 2006;14(1):46-51.
 18. Wentzensen N, Schwartz L, Zuna RE, et al. Performance of p16/Ki-67 immunostaining to detect cervical cancer precursors in a colposcopy referral population. *Clinical cancer research : an official journal of the American Association for Cancer Research*. Aug 1 2012;18(15):4154-4162.
 19. Brunnstrom H, Johansson L, Jirstrom K, Jonsson M, Jonsson P, Planck M. Immunohistochemistry in the differential diagnostics of primary lung cancer: an investigation within the Southern Swedish Lung Cancer Study. *American journal of clinical pathology*. Jul 2013;140(1):37-46.
 20. Wolff AC, Hammond ME, Schwartz JN, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. Jan 1 2007;25(1):118-145.
 21. Chu P, Wu E, Weiss LM. Cytokeratin 7 and Cytokeratin 20 Expression in Epithelial Neoplasms: A Survey of 435 Cases. *Mod Pathol*. //print 0000;13(9):962-972.
 22. Polley MY, Leung SC, McShane LM, et al. An international Ki67 reproducibility study. *Journal of the National Cancer Institute*. Dec 18 2013;105(24):1897-1906.
 23. Ali A, Brown V, Denley S, et al. Expression of KOC, S100P, mesothelin and MUC1 in pancreatico-biliary adenocarcinomas: development and utility of a potential diagnostic immunohistochemistry panel. *BMC Clinical Pathology*. 2014;14(1):35.
 24. Abe N, Watanabe T, Masaki T, et al. Pancreatic duct cell carcinomas express high levels of high mobility group I(Y) proteins. *Cancer research*. Jun 15 2000;60(12):3117-3122.
 25. Crocetti E, Caldarella A, Ferretti S, et al. Consistency and inconsistency in testing biomarkers in breast cancer. A GRELL study in cut-off variability in the Romance language countries. *Breast (Edinburgh, Scotland)*. Aug 2013;22(4):476-481.
 26. Jung S, Wu C, Eslami Z, et al. The role of immunohistochemistry in the diagnosis of flat urothelial lesions: a study using CK20, CK5/6, P53, Cd138, and Her2/Neu. *Annals of diagnostic pathology*. Feb 2014;18(1):27-32.

27. Mallofre C, Castillo M, Morente V, Sole M. Immunohistochemical expression of CK20, p53, and Ki-67 as objective markers of urothelial dysplasia. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc.* Mar 2003;16(3):187-191.
28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* Mar 1977;33(1):159-174.
29. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American journal of epidemiology.* Apr 1 2006;163(7):670-675.
30. Goshu M, Nagashima K, Sato Y. Study designs and statistical analyses for biomarker research. *Sensors (Basel, Switzerland).* 2012;12(7):8966-8986.
31. Tzankov A, Zlobec I, Went P, Robl H, Hoeller S, Dirnhofer S. Prognostic immunophenotypic biomarker studies in diffuse large B cell lymphoma with special emphasis on rational determination of cut-off scores. *Leukemia & lymphoma.* Feb 2010;51(2):199-212.
32. Simon R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. 20110929 (1741-0541 (Print)).
33. Lin F, Shi J, Liu H, et al. Diagnostic utility of S100P and von Hippel-Lindau gene product (pVHL) in pancreatic adenocarcinoma - With implication of their roles in early tumorigenesis. *American Journal of Surgical Pathology.* January 2008;32(1):78-91.
34. Abe N, Watanabe T, Masaki T, et al. Pancreatic duct cell carcinomas express high levels of high mobility group I(Y) proteins. *Cancer Research.* 15 Jun 2000;60(12):3117-3122.
35. Fisher G, Yang ZH, Kudahetti S, et al. Prognostic value of Ki-67 for prostate cancer death in a conservatively managed cohort. *British journal of cancer.* Feb 5 2013;108(2):271-277.
36. Zlobec I, Vuong T, Compton CC. The predictive value of apoptosis protease-activating factor 1 in rectal tumors treated with preoperative, high-dose-rate brachytherapy. *Cancer.* Jan 15 2006;106(2):284-286.
37. Hirsch FR, Dziadziuszko R, Thatcher N, et al. Epidermal growth factor receptor immunohistochemistry: comparison of antibodies and cutoff points to predict benefit from gefitinib in a phase 3 placebo-controlled study in advanced nonsmall-cell lung cancer. *Cancer.* Mar 1 2008;112(5):1114-1121.
38. Kobayashi T, Iwaya K, Moriya T, et al. A simple immunohistochemical panel comprising 2 conventional markers, Ki67 and p53, is a powerful tool for predicting patient outcome in luminal-type breast cancer. *BMC Clinical Pathology.* 2013;13(1):5.
39. Kerkhof M, van Dekken H, Steyerberg EW, et al. Grading of dysplasia in Barrett's oesophagus: substantial interobserver variation between general and gastrointestinal pathologists. *Histopathology.* Jun 2007;50(7):920-927.
40. Lorinc E, Jakobsson B, Landberg G, Veress B. Ki67 and p53 immunohistochemistry reduces interobserver variation in assessment of Barrett's oesophagus. *Histopathology.* Jun 2005;46(6):642-648.
41. Turner JK, Williams GT, Morgan M, Wright M, Dolwani S. Interobserver agreement in the reporting of colorectal polyp pathology among bowel cancer screening pathologists in Wales. *Histopathology.* May 2013;62(6):916-924.
42. von Wasielewski R, Mengel M, Wiese B, Rudiger T, Muller-Hermelink HK, Kreipe H. Tissue array technology for testing interlaboratory and interobserver reproducibility of immunohistochemical estrogen receptor

- analysis in a large multicenter trial. *American journal of clinical pathology*. Nov 2002;118(5):675-682.
43. Cross SS, Betmouni S, Burton JL, et al. What levels of agreement can be expected between histopathologists assigning cases to discrete nominal categories? A study of the diagnosis of hyperplastic and adenomatous colorectal polyps. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*. Sep 2000;13(9):941-944.
 44. Jaraj SJ, Camparo P, Boyle H, et al. Intra- and interobserver reproducibility of interpretation of immunohistochemical stains of prostate cancer. *Virchows Arch*. Oct 2009;455(4):375-381.
 45. Coras R, de Boer OJ, Armstrong D, et al. Good interobserver and intraobserver agreement in the evaluation of the new ILAE classification of focal cortical dysplasias. *Epilepsia*. Aug 2012;53(8):1341-1348.
 46. Kosarac O, Takei H, Zhai QJ, Schwartz MR, Mody DR. S100P and XIAP expression in pancreatic ductal adenocarcinoma: potential novel biomarkers as a diagnostic adjunct to fine needle aspiration cytology. *Acta cytologica*. 2011;55(2):142-148.
 47. Cascinu S, Staccioli MP, Gasparini G, et al. Expression of vascular endothelial growth factor can predict event-free survival in stage II colon cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*. Jul 2000;6(7):2803-2807.
 48. Hoang MP, Sahin AA, Ordonez NG, Sneige N. HER-2/neu gene amplification compared with HER-2/neu protein overexpression and interobserver reproducibility in invasive breast carcinoma. *American journal of clinical pathology*. Jun 2000;113(6):852-859.
 49. Ogino J, Asanuma H, Hatanaka Y, et al. Validity and reproducibility of Ki-67 assessment in gastrointestinal stromal tumors and leiomyosarcomas. *Pathology international*. Feb 2013;63(2):102-107.
 50. Thomson TA, Hayes MM, Spinelli JJ, et al. HER-2/neu in breast cancer: interobserver variability and performance of immunohistochemistry with 4 antibodies compared with fluorescent in situ hybridization. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*. Nov 2001;14(11):1079-1086.
 51. Park D, Kåresen R, Noren T, Sauer T. Ki-67 expression in primary breast carcinomas and their axillary lymph node metastases: clinical implications. *Virchows Archiv*. 2007/07/01 2007;451(1):11-18.
 52. Aumann K, Frey AV, May AM, et al. Subcellular mislocalization of the transcription factor NF-E2 in erythroid cells discriminates prefibrotic primary myelofibrosis from essential thrombocythemia. *Blood*. Jul 4 2013;122(1):93-99.
 53. Trivedi A, Cartun RW, Ligato S. Role of lymphovascular invasion and immunohistochemical expression of IMP3 in the risk stratification of superficially invasive pT1 esophageal adenocarcinoma. *Pathology, research and practice*. Mar 13 2014.
 54. Jo VY, Cibas ES, Pinkus GS. Claudin-4 immunohistochemistry is highly effective in distinguishing adenocarcinoma from malignant mesothelioma in effusion cytology. *Cancer cytopathology*. Apr 2014;122(4):299-306.
 55. Betta PG, Andrion A, Donna A, et al. Malignant mesothelioma of the pleura. The reproducibility of the immunohistological diagnosis. *Pathol Res Pract*. 1997;193(11-12):759-765.
 56. Kirkegaard T, Edwards J, Tovey S, et al. Observer variation in immunohistochemical analysis of protein expression, time for a change? *Histopathology*. Jun 2006;48(7):787-794.

57. Atkinson R, Mollerup J, Laenkholm AV, et al. Effects of the change in cutoff values for human epidermal growth factor receptor 2 status by immunohistochemistry and fluorescence in situ hybridization: a study comparing conventional brightfield microscopy, image analysis-assisted microscopy, and interobserver variation. *Arch Pathol Lab Med.* Aug 2011;135(8):1010-1016.

Figure 1: Representative images of high and low inter-observer agreement between all pathologists for 10%, 20% and +2/+3 cut-offs.

Cut-Offs	High Interobserver agreement	Low Interobserver agreement
10%		
20%		
+2/+3	