Complete manuscript title:

An International Ki67 Reproducibility Study in Adrenal Cortical Carcinoma.

Thomas G. Papathomas, MD [1, 2*], **Eugenio Pucci, MD** [1, 3*], **Thomas J. Giordano, MD, PhD** [4], **Hao Lu, PhD** [5], **Eleonora Duregon, MD** [6], **Marco Volante, MD, PhD** [6], **Mauro Papotti, MD** [6], **Ricardo V. Lloyd, MD, PhD** [7], **Arthur S. Tischler, MD** [8], **Francien H. van Nederveen, MD, PhD** [9], **Vania Nose, MD, PhD** [10], **Lori Erickson, MD** [11], **Ozgur Mete, MD** [12], **Sylvia L. Asa, MD, PhD** [12], **John Turchini, MD** [13], **Anthony J. Gill, MD, FRCPA** [13], **Xavier Matias-Guiu, MD, PhD** [14], **Kassiani Skordilis, MD, FRCPath** [15], **Timothy J. Stephenson, MD, FRCPath** [16], **Frédérique Tissier MD, PhD** [17-18], **Richard A. Feelders, MD, PhD** [19], **Marcel Smid, BSc** [20], **Alex Nigg** [1], **Esther Korpershoek, PhD** [1], **Peter J. van der Spek, PhD** [5], **Winand N.M. Dinjens, PhD** [1], **Andrew P. Stubbs, PhD** [5], **Ronald R. de Krijger, MD, PhD** [1, 21-22]

1. Department of Pathology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, the Netherlands
2. Department of Histopathology, King's College Hospital, London, UK
3. Department of Clinical and Molecular Medicine, Pathology Unit, Sant' Andrea Hospital, Sapienza University, Rome, Italy
4. Department of Pathology, Department of Internal Medicine, University of Michigan Comprehensive Cancer Center, University of Michigan Health System, USA
5. Department of Bioinformatics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

6. Department of Oncology, University of Turin at San Luigi Hospital, Orbassano, Italy
7. Department of Pathology and Laboratory Medicine, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA
8. Department of Pathology and Laboratory Medicine, Tufts Medical Center, Tufts University School of Medicine, Boston, Massachusetts, USA
9. Laboratory for Pathology, Dordrecht, the Netherlands
10. Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA
11. Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA
12. Department of Pathology, University Health Network, University of Toronto, Toronto, Ontario, Canada
13. Department of Anatomical Pathology, Royal North Shore Hospital and University of Sydney, Australia
14. Department of Pathology and Molecular Genetics and Research Laboratory, Hospital Universitari Arnau de Vilanova, IRBLLEIDA, University of Lleida, Lleida, Spain
15. Department of Pathology, University Hospitals Birmingham, Birmingham, UK
16. Department of Histopathology, Royal Hallamshire Hospital, Sheffield, UK
17. Institut National de la Santé et de la Recherche Médicale U1016, Institut Cochin, Centre National de la Recherche Scientifique UMR8104, Université Paris

Descartes, Sorbonne Paris Cité, Rare Adrenal Cancer Network COMETE, Paris, France

18. Department of Pathology, Hôpital Pitié-Salpêtrière, Université Pierre et Marie Curie, Paris, France
19. Department of Internal Medicine, Division of Endocrinology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, the Netherlands
20. Department of Medical Oncology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, the Netherlands
21. Department of Pathology, Reinier de Graaf Hospital, Delft, the Netherlands
22. University Medical Center Utrecht, Princess Maxima Center for Pediatric Oncology, Utrecht, the Netherlands

* equal contribution

Corresponding author:

Thomas G. Papathomas, MD

Department of Histopathology,

King's College Hospital, Denmark Hill,

London, UK

Tel: +44-7437416650

t.papathomas@erasmusmc.nl

thomaspapathomas@nhs.net

Disclosure of funding: This study was supported by the Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 259735 (ENS@T-Cancer). Further support was partially provided by grants from AIRC, Milan no. IG/14820/2013 (to MP).

Abstract

Despite the established role of Ki67 labelling index in prognostic stratification of adrenocortical carcinomas and its recent integration into treatment flow charts, the reproducibility of the assessment method has not been determined. The aim of this study was to investigate inter-observer variability among endocrine pathologists using a web-based virtual microscopy approach. Ki67 stained slides of 76 adrenocortical carcinomas were analyzed independently by 14 observers; each according to their method of preference including eyeballing, formal manual counting and digital image analysis. The inter-observer variation was statistically significant (p<0.001) in the absence of any correlation between the various methods. Subsequently, 61 static images were distributed among fifteen observers who were instructed to follow a category-based scoring approach. Low levels of inter-observer (F=6.99; Fcrit= 1.70; p<0.001) as well as intra-observer concordance (n=11; Cohen's Kappa ranging from -0.057 to 0.361) were detected. To improve harmonization of Ki67 analysis, we tested the utility of an open source Galaxy virtual machine application, namely *Automated Selection of Hotspots*, in 61 virtual slides. The software-provided Ki67 values were validated by digital image analysis in identical images, displaying strong correlation 0.96 (p<0.0001) and dividing the cases into 3 classes (cut-offs of 0%-15%-30% and/or 0%-10%-20%) with significantly different overall survivals (p<0.05). We conclude that current practices in Ki67 scoring assessment vary greatly and inter-observer variation sets particular limitations to its clinical utility, especially around clinically relevant cut-off

values. Novel digital microscopy-enabled methods could provide critical aid in reducing variation, increasing reproducibility and improving reliability in the clinical setting.

Key words or phrases: Ki67 labelling index; proliferation; adrenal cortical carcinoma; interobserver variation; digital pathology

Introduction

Adrenocortical carcinoma (ACC) is a rare endocrine malignancy with a poor overall prognosis and an estimated incidence of 0.7-2 cases per million (**1**). When confronted with this tumor, pathologists are expected to provide the Weiss score, the status of resection margins and prognosticators including the Weiss score, mitotic grade and Ki67 labelling index (LI), and, if diagnostically challenging, confirm its adrenocortical origin on immunohistochemical grounds (**2-3**). It has been shown (**4**) that ACCs can be subdivided using a variety of methods including the mitotic frequency into low-grade (≤20 mitoses/50 HPFs) and high-grade (>20 mitoses/50 HPFs) (**5**), Stereoidogenic Factor-1 immunohistochemistry (**6-7**) and other proliferation-based scoring methods such as phosphohistone H3–specific immunohistochemistry (**8**).

According to recent data generated by the European Network for the Study of Adrenal Tumors (ENS@T) ACC study group (**9-10**), the resection status and the Ki67 labelling index in both localized and advanced ACCs constitute the most relevant prognostic parameters (**2**). In accordance, Duregon *et al*. (**8**) demonstrated that Ki67 LI is the most powerful tool in terms of prognostic stratification. In addition to its emerging value as a critical determinant of prognosis, Ki67 LI has been recently integrated in treatment flow charts for adrenocortical cancer patients suffering from tumors either amenable to radical resection or at advanced presentation. Accordingly, thresholds of 10%, 20%, and 30% seem to be crucial in therapeutic decisions, including adjuvant mitotane, radiotherapy of the tumor bed as well as combination therapy of mitotane and three cycles of cisplatin respectively (**1-2**).

The standardized assessment of Ki67 LI is important and remains a key issue and responsibility of histopathologists. Nevertheless, various factors, such as pre-analytical, analytical, interpretation, scoring, and data analysis, might affect the Ki67 LI (**11**). In particular, lack of uniformity and consistency in quantification (**12**) as well as intratumoral heterogeneity of proliferation (**5, 11, 13-14**) might limit its assessment. In this context, we have implemented an open source toolset, namely Automated Selection of Hotspots (ASH) aiming at improved accuracy and reproducibility of reporting of the Ki67 LI (**15**).

In the present study, we determined the inter-observer variability for Ki67 LI and examined the current practices among expert endocrine pathologists in a multicenter cohort of conventional ACCs using virtual microscopy. The impact of various parameters, i.e. readout technique of preference in diagnostics, selected fields for evaluation and estimated total number of cells, on Ki67 assessment was further investigated. Moreover, we evaluated the variability of Ki67 LI around clinically relevant cut-offs (**1-2**) and validated the efficiency of ASH as compared to the human independent selection of hotspot areas.

Materials and Methods

Case selection and KI67 (MIB1) immunohistochemistry

One hundred and one conventional ACCs were collected from four specialized centres from Europe and United States: (**1**) **San Luigi Gonzaga Hospital** and **University of Turin**, Turin, Italy (25 samples), (**2**) **Erasmus MC Cancer Institute**,

Rotterdam, The Netherlands (12 samples), (**3**) **University of Wisconsin School of Medicine and Public Health** (5 samples) and (**4**) **University of Michigan Health System** (59 samples). Borderline/ atypical adrenocortical neoplasms as well as ACC variants (oncocytic, myxoid and sarcomatoid) were not included in the present study. Each case was thoroughly reviewed and representative unstained glass slide(s) were selected and provided for immunohistochemical analysis within a single center (Department of Pathology, Erasmus MC Cancer Institute, Rotterdam, the Netherlands) with the following protocol. Slides and formalin-fixed paraffin-embedded (FFPE) whole-tissue sections of 4μm thickness were stained with a commercially available antibody: mouse monoclonal MIB1 M7240 antibody (Dako, Glostrup, Denmark; 1:400 dilution) against Ki67 on an automatic Ventana Benchmark Ultra System (Ventana Medical Systems Inc. Tuscon, AZ, USA) using Ultraview DAB detection system preceded by heat-induced epitope retrieval with Ventana Cell Conditioning 1 (pH 8.4) at 97°C for 52 minutes. Diaminobenzidine was used as the chromogen. All cases were assessed anonymously according to the Proper Secondary Use of Human Tissue code established by the Dutch Federation of Medical Scientific Societies (http:// www.federa.org). The Medical Ethical Committee of the Erasmus MC approved the study. Cases displaying artefactual intratumoral variation in labeling were excluded by use of Ki67 labeled mitotic figures as internal positive controls.

Digital pathology application

High-resolution, whole-slide images were acquired from all Ki67 (MIB1) stained slides using a NanoZoomer Digital Pathology (NDP) System (Hamamatsu Photonics

K.K. Japan) working at a resolution of 0.23 µm/pixel. The immunostains were scanned at x40 magnification and automatically digitized in their proprietary NDP Image (NDPI) file format. Between October 2013 and March 2014, digital files were consecutively uploaded in one set to a server at Erasmus MC through the standard File transfer Protocol (FTP) in the DMZ with URL: http://digimic.erasmusmc.nl/; enabling online worldwide viewing through a virtual microscopy interface (NDP.view Viewer Software, Hamamatsu Photonics K.K. Japan).

Participants and interpretation of staining results

In the first round (**Supplemental Digital Content 1**), fourteen observers, among which 11 expert endocrine pathologists (R.V.L., L.E., V.N., O.M., S.L.A., X.G., T.J.S., K.S., F.T., F.H. vanN., R.R.deK.) and 3 residents (T.G.P., E.D. & J.T) received: (**i**) an email detailing the objectives of the project and clearly stating that only nuclear staining (plus mitotic figures which are stained by Ki67) should be incorporated into the Ki67 score defined as the percentage of positively stained cells among the total number of malignant cells scored with staining intensity being of no relevance (**11**), (**ii**) the corresponding link providing access to the virtual slides, and (**iii**) a scoring list to be completed during Ki67 immunohistochemical evaluations.

All virtual slides were distributed online, reviewed by each observer in a blinded fashion without knowledge of the corresponding clinicopathological data or scores assigned by other pathologists. In particular, participants were asked to assess (**i**) the Ki67 LI based on (**ii**) the method of their preference/practice in diagnostics (visual estimation, formal manual count or Digital Image Analysis [DIA]) reporting on (**iii**) the

estimated *total number of cells* and (**iv**) the selected fields for evaluation i.e. hot spot area(s) *or* average score across the section, *or* average score across the section adding hot spot area(s).

Twenty-five cases were excluded from the analysis due to subtoptimal staining, poor scan quality and fixation artifacts. The remaining tumors from 76 patients of mean age 47.6 years (ranging from 8 to 85 years; 1.17 female:male ratio) comprised 62 primary tumors, 6 recurrences and 8 metastases. Thirty-four patients died of the disease, while 42 are alive with or without evidence of disease. The latter are currently in follow-up at various institutions with a mean of 34.27 months (range, 1 week to 169 months).

In the second round of assessment performed 9 months later (**Supplemental Digital Content 1**), 61 static images (.JPG files) were circulated among 15 observers, including 11 expert endocrine pathologists (R.V.L., L.E., O.M., S.L.A., T.J.S., K.S., M.V., A.S.T., A.J.G., F.H. vanN., R.R.deK.) and 4 residents (T.G.P., E.P., E.D. & J.T). These images were selected as the most active areas based on an automated approach (**15**). The participants were instructed to follow a category-based evaluation of the Ki67 LI on the basis of visual estimation without performing formal manual count or DIA.

Software application

Seventy-six virtual slides were assessed with a recently developed open source Galaxy virtual machine application designed for Ki67 hotspot detection in adrenocortical cancer (**Supplemental Digital Content 1**). In brief, ASH comprises three classes: NDPI Segmentation, Adaptive Step Finding and a Reporting Visualization which utilizes the

NDPI splitter to convert the specific NDPI format digital slide into a conventional tiff or jpeg format image for automated segmentation and adaptive step finding hotspots detection algorithm (**15**). Quantitative hotspot ranking is provided by the functionality from the open source application ImmunoRatio (**16**) as part of the ASH protocol. Accordingly, the output is a ranked set of hotspots with concomitant quantitative values based on whole slide ranking.

Statistical analysis

Inter-observer variability using either virtual microscopy (first evaluation) or visual estimation on static images (second evaluation) as well as differences in the type of assessment was assessed with ANOVA single factor. In order to evaluate intra-observer agreement, Cohen's Kappa was performed following conversion of the Ki67 values of the initial numerical assessment into categorical variables. With regard to automatically selected areas, we compared computerized counts based on ImmunoRatio and DIA respectively in identical images using Pearson correlation with "Wessa, P. (2015), Free Statistics Software, Office for Research Development and Education, version 1.1.23-r7, URL *http://www.wessa.net/*". The correlation between human independent selection and software selection of hotspot areas was examined with Spearman rank order correlation. To compare the results of Ki67 assessment with overall survival, Kaplan-Meier curves were plotted and *P* values were calculated using the Log-rank test. The level of significance was set at $P<0.05$. All other statistical analyses were performed using SPSS software (SPSS version 21; SPSS, Inc., Chicago, IL, USA).

Results

Inter-observer variation in KI67 LI assessment

Seventy-six cases were initially analyzed displaying statistically significant variance between 14 observers (ANOVA F= 10.43; Fcrit= 1.73; p<0.001) (**Fig**. **1**). Differences in current practices concerning the Ki67 LI assessment are highlighted in **Figure 2**. Out of 14 observers, eight preferred formal manual counting, four visual estimation and two DIA (ImageJ software, 1.47v, Wayne Rasband, NIH, USA &KS400 image analysis software, version 3.0, Carl Zeiss Vision GmbH). With regard to the residents, two used formal manual count and one DIA (KS400 image analysis software, version 3.0, Carl Zeiss Vision GmbH). The overall agreement was not affected by different levels of experience in endocrine pathology (data not shown). No statistical significance was found between the different methods of assessment (ANOVA p=0.079), except between visual estimation and formal manual count (t-test p=0.014). Kaplan-Meier curves based on overall survival were plotted against 0%-15%-30% cut-offs (**Fig**. **3**).

Impact of visual estimation on variation

Given the large variation observed in the initial Ki67 assessment, we decided to reduce potential complexities by using visual estimation and following a category-based approach in 61 pre-determined images. In this context, the variation remained statistically significant between 15 observers (ANOVA F=6.99; Fcrit= 1.70; p<0.001). In order to evaluate inter-observer concordance, ASH maximum values were utilized as

"gold standard" and transformed into categorical variables. The highest levels of concordance were achieved within the lowest range of Ki67 values i.e. 0-10% (**Fig**. **4**). Likewise, the overall agreement was not affected by different levels of experience in endocrine pathology (data not shown). In order to assess intra-observer concordance, we transformed those numerical values of the initial assessment into categorical variables. A very low degree of concordance was detected for every observer (n=11) (**Supplemental Digital Content 2**) with the majority having a higher score on visual estimation of pre-determined images (**Fig**. **5**).

Automated Selection of Hotspots (ASH): a virtual microscopy-enabled assessment of KI67 LI

Following software assessment, fifteen out of 76 cases were excluded due to artifacts interfering with the analysis. In order to verify its applicability in the remaining 61 cases, we determined the degree of concordance (**i**) between computerized counts as provided by the software (ImmunoRatio) and by DIA (KS400 image analysis software, version 3.0, Carl Zeiss Vision GmbH) in identical images (n=610; 10 images as selected by the ASH per virtual slide); and (**ii**) between computerized counts as provided by the software (ImmunoRatio) and as generated by human independent selection (DIA) in different images displaying the highest Ki67 expression (n=61; 1 image per virtual slide). To this end, an observer (T.G.P.) selected 10 hotspot areas by visual estimation on a virtual microscopy interface and subsequently performed DIA (KS400 image analysis software, version 3.0, Carl Zeiss Vision GmbH). In this setting, strong correlations of 0.96 and 0.84 were detected respectively ($p<0.001$). From a

clinical standpoint, we determined whether software-provided Ki67 values could divide the cases into 3 classes with significantly different overall survivals. In fact, when overall survival Kaplan-Meier curves were plotted against 0%-15%-30% and/or 0%-10%-20% cut-offs (**Fig**. **6**), overall comparisons were statistically significant (p<0.05).

Discussion

Ki67 immunohistochemistry has been integrated in routine pathology practice not only in diagnostics i.e. grading and tumor classification, diagnosis of intraepithelial neoplasia and assessment of malignant potential, but also as a prognostic and predictive biomarker. With regard to adrenocortical carcinomas, it has been proposed in diagnostics (**17-18**), prognostics (**8-10**, **19**) as well as in guiding treatment decisions (**1-2**). The current study highlights the need for standardized use of the Ki67 LI discouraging visual estimation and verifies the applicability of ASH in Ki67 assessment.

A large variation was noted among 14 observers in Ki67 index determination using a virtual microscopy interface. Because of the stringent centralized staining protocol, all participants were seeing the same slides. The variation therefore could not be explained by technical issues and had to be attributed to different practices with respect to interpretation and scoring such as area(s) of slide read, total number of cells in fields of evaluation and methods of assessment (**11**). In support of the last, we still observed significant levels of variation even when reducing complexities by estimating Ki67 LI levels in pre-selected areas and following a category-based approach using visual estimation. This is consistent with studies in breast carcinomas using a TMA platform

(**20**) as well as in gastroenteropancreatic neuroendocrine tumors using pre-determined images (**12**).

Although visual estimation has been suggested as an acceptable method of assessment on expert diagnostic (**13**) and/or research grounds (**21-23**), our findings further reinforce the notion that this readout technique is subjective, inaccurate and thus unreliable (**12**, **20**, **24-25**). Importantly, low levels of concordance were revealed around categorical cut-off values recently proposed in ACCs. This is in keeping with Tang *et al*. (**12**) who reported significant discordance among 18 observers, which was sufficient to alter the final grade of the majority of 45 neuroendocrine tumors. Whether such discordances could be solely ascribed to the method of assessment or partly to parameters residing in the realm of cognitive psychology (**21**) remains uncertain.

The aforementioned data challenge the clinical applicability of clinically relevant cut-offs in ACCs. In accordance with Polley *et al*. (**20**) and Mengel *et al*. (**26**), Beuschlein *et al*. (**9**) suggested that Ki67LI variability is to be expected in ACCs at different clinical centers highlighting the issue of inter-laboratory variation due to pre-analytical and analytical parameters (**20**, **26**). In this setting, rigorous methods in tissue preparation, i.e. fixation, processing and generation of uniform sections, would seem to be important. Inter-laboratory variables at play, e.g. variation affecting controlled conditions, variability in microtomes used as well as differences in the temperature of the FFPE blocks, might have affected the thickness of the immunostained sections in the current study. In addition to the inter-laboratory variation, inter-observer and intra-observer variation (**12**, **21**, **27-28**) and tumor heterogeneity of Ki67 expression levels (**13-14**, **29-30**) seem to add further levels of complexity to the issue of reproducibility, thereby hampering its

clinical utility. This issue was emphasized by the International Ki67 in Breast Cancer Working Group (**11**) that was unable to reach a consensus in the absence of harmonized methodology with respect to ideal thresholds that could be useful in clinical routine practice. Accordingly, they recommended that cut-offs for prognosis, prediction, and monitoring should be applied only if the results from local practice have been validated against the respective ones in studies that have defined these particular cut-offs (**11**, **20**).

Various approaches have been developed to obtain standardized Ki67 scoring. These include efforts to reduce inter-laboratory variation by calibrating to a common scoring method via a web-based tool (**31**), and efforts to reduce inter- and intra-observer variation by either selecting the most representative tumor areas based on an automated approach (**15, 32-33**) or providing a software-automated quantitation of Ki67 LI (**16**, **34**-**36**). In the setting of computerized image analysis, we verified the applicability of a digital microscopy-enabled method for assessment of Ki67 expression in adrenocortical cancer. The novel approach of software-selected areas aims not only to reduce the inter-observer variation, but also to characterize Ki67 levels of heterogeneity in primary tumors, recurrences and metastases.

User interaction is recommended prior to virtual slide analysis in order to ensure that areas leading to miscalculations, i.e. intrinsic as well as extrinsic pigmentation (deposit artifacts), necrotic areas, tissue folds etc, are excluded (**15**). In this series, excluding certain tissue regions was not sufficient to avoid serious miscalculations with regard to fifteen cases (15 out of 76; 20%) that were subsequently excluded from the analysis, calling into question potential clinical actions based on such cases. Future efforts should

focus on software amendments to overcome technical shortcomings in addition to improving methods of scoring.

In conclusion, current practices in Ki67 scoring assessment vary greatly and inter-observer variation sets particular limitations to the clinical utility of Ki67 LI, especially around clinically relevant cut-off values, in adrenocortical cancer. Our results highlight the need for standardization and suggest that visual estimation should be strongly discouraged as a readout technique, while computerized DIA appears to provide a reliable alternative. To drive forward harmonization of Ki67 analysis, we have previously developed and now validated an open source Galaxy virtual machine application, namely Automated Selection of Hotspots. Given certain pre-analytical and analytical concerns, quality assurance schemes i.e. standardized tissue fixation along with fine-tuned immunohistochemical staining protocols are expected to additionally increase reproducibility and reliability of the Ki67 LI in endocrine pathology practice.

Disclosure/conflict of interest

The authors declare no conflict of interest.

References

1. Fassnacht M, Libé R, Kroiss M, et al. Adrenocortical carcinoma: a clinician's update. *Nat Rev Endocrinol*. 2011;7:323--335.

2. Fassnacht M, Kroiss M, Allolio B. Update in adrenocortical carcinoma. *J Clin Endocrinol Metab*. 2013;98:4551--4564.

3. van't Sant HP1, Bouvy ND, Kazemier G, et al. The prognostic value of two different histopathological scoring systems for adrenocortical carcinomas. *Histopathology*. 2007;51:239--245.

4. Mouat IC, Giordano TJ. Assessing Biological Aggression in Adrenocortical Neoplasia. *Surgical Pathology Clinics*. 2014;7:533--541.

5. Giordano TJ. The argument for mitotic rate-based grading for the prognostication of adrenocortical carcinoma. *Am J Surg Pathol*. 2011;35:471--473.

6. Sbiera S, Schmull S, Assie G, et al. High diagnostic and prognostic value of steroidogenic factor-1 expression in adrenal tumors. *J Clin Endocrinol Metab*. 2010;95:E161--171.

7. Duregon E, Volante M, Giorcelli J, et al. Diagnostic and prognostic role of steroidogenic factor 1 in adrenocortical carcinoma: a validation study focusing on clinical and pathologic correlates. *Hum Pathol*. 2013;44:822--828.

8. Duregon E, Molinaro L, Volante M, et al. Comparative diagnostic and prognostic performances of the hematoxylin-eosin and phospho-histone H3 mitotic count and Ki-67 index in adrenocortical carcinoma. *Mod Pathol*. 2014;27:1246--1254.

9. Beuschlein F, Weigel J, Saeger W, et al. Major prognostic role of Ki67 in localized adrenocortical carcinoma after complete resection. *J Clin Endocrinol Metab*. 2015;100:841--849.

10. Libé R, Borget I, Ronchi CL, et al. Prognostic factors in stage III-IV adrenocortical carcinomas (ACC): an European Network for the Study of Adrenal Tumor (ENSAT) study. *Ann Oncol*. 2015;10:2119--2125.

11. Dowsett M, Nielsen TO, A'Hern R, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *J Natl Cancer Inst*. 2011;103:1656--1664.

12. Tang LH, Gonen M, Hedvat C, et al. Objective quantification of the Ki67 proliferative index in neuroendocrine tumors of the gastroenteropancreatic system: a comparison of digital image analysis with manual methods. *Am J Surg Pathol*. 2012;36:1761--7170.

13. Adsay V. Ki67 labeling index in neuroendocrine tumors of the gastrointestinal and pancreatobiliary tract: to count or not to count is not the question, but rather how to count. *Am J Surg Pathol*. 2012;36:1743--1746.

14. Yang Z, Tang LH, Klimstra DS. Effect of tumor heterogeneity on the assessment of Ki67 labeling index in well-differentiated neuroendocrine tumors metastatic to the liver: implications for prognostic stratification. *Am J Surg Pathol*. 2011;35:853--860.

15. Lu H, Papathomas TG, van Zessen D, et al. Automated Selection of Hotspots (ASH): enhanced automated segmentation and adaptive step finding for Ki67 hotspot detection in adrenal cortical cancer. *Diagn Pathol*. 2014;9:216.

16. Tuominen VJ, Ruotoistenmäki S, Viitanen A, et al. ImmunoRatio: a publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67. *Breast Cancer Res*. 2010;12:R56.

17. Schmitt A, Saremaslani P, Schmid S, et al. IGFII and MIB1 immunohistochemistry is helpful for the differentiation of benign from malignant adrenocortical tumours. *Histopathology*. 2006;49:298--307.

18. Soon PS, Gill AJ, Benn DE, et al. Microarray gene expression and immunohistochemistry analyses of adrenocortical tumors identify IGF2 and Ki-67 as useful in differentiating carcinomas from adenomas. *Endocr Relat Cancer*. 2009;16:573--583.

19. Ip JC, Pang TC, Glover AR, et al. Immunohistochemical Validation of Overexpressed Genes Identified by Global Expression Microarrays in Adrenocortical Carcinoma Reveals Potential Predictive and Prognostic Biomarkers. *Oncologist*. 2015;20:247--256.

20. Polley MY, Leung SC, McShane LM, et al. An international Ki67 reproducibility study. *J Natl Cancer Inst*. 2013;105:1897--1906.

21. Varga Z, Diebold J, Dommann-Scherrer C, et al. How reliable is Ki-67 immunohistochemistry in grade 2 breast carcinomas? A QA study of the Swiss Working Group of Breast- and Gynecopathologists. *PLoS One*. 2012;7:e37379.

22. Hida AI, Bando K, Sugita A, et al. Visual assessment of Ki67 using a 5-grade scale (Eye-5) is easy and practical to classify breast cancer subtypes with high reproducibility. *J Clin Pathol*. 2015;68:356--361.

23. Hida AI, Oshiro Y, Inoue H, et al. Visual assessment of Ki67 at a glance is an easy method to exclude many luminal-type breast cancers from counting 1000 cells. *Breast Cancer*. 2015;22:129--134.

24. Reid MD, Bagci P, Ohike N, et al. Calculation of the Ki67 index in pancreatic neuroendocrine tumors: a comparative analysis of four counting methodologies. *Mod Pathol*. 2015;28:686--694.

25. Mikami Y, Ueno T, Yoshimura K, et al. Interobserver concordance of Ki67 labeling index in breast cancer: Japan Breast Cancer Research Group Ki67 ring study. *Cancer Sci*. 2013;104:1539--1543.

26. Mengel M, von Wasielewski R, Wiese B, et al. Inter-laboratory and inter-observer reproducibility of immunohistochemical assessment of the Ki-67 labelling index in a large multi-centre trial. *J Pathol*. 2002;198:292--299.

27. Niikura N, Sakatani T, Arima N, et al. Assessment of the Ki67 labeling index: a Japanese validation ring study. *Breast Cancer*. 2014 May 3 [Epub ahead of print]

28. Gudlaugsson E, Skaland I, Janssen EA, et al. Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer. *Histopathology*. 2012;61:1134--1144.

29. Shi C, Gonzalez RS, Zhao Z, et al. Liver Metastases of Small Intestine Neuroendocrine Tumors: Ki-67 Heterogeneity and World Health Organization Grade Discordance With Primary Tumors. *Am J Clin Pathol*. 2015;143:398--404.

30. Couvelard A, Deschamps L, Ravaud P, et al. Heterogeneity of tumor prognostic markers: a reproducibility study applied to liver metastases of pancreatic endocrine tumors. *Mod Pathol*. 2009;22:273--281.

31. Polley MY, Leung SC, Gao D, *et al*. An international study to increase concordance in Ki67 scoring. *Mod Pathol*. 2015;28:778--786.

32. Lopez XM, Debeir O, Maris C, et al. Clustering methods applied in the detection of Ki67 hot-spots in whole tumor slide images: an efficient way to characterize heterogeneous tissue-based biomarkers. *Cytometry A*. 2012;81:765--775.

33. Elie N, Plancoulaine B, Signolle JP, et al. A simple way of quantifying immunostained cell nuclei on the whole histologic section. *Cytometry A*. 2003;56:37--45.

34. Klauschen F, Wienert S, Schmitt W, et al. Standardized Ki67 diagnostics using automated scoring - clinical validation in the GeparTrio breast cancer study. *Clin Cancer Res*. 2015;21:3651--3657.

35. Samols MA, Smith NE, Gerber JM, et al. Software-automated counting of Ki-67 proliferation index correlates with pathologic grade and disease progression of follicular lymphomas. *Am J Clin Pathol*. 2013;140:579--587.

36. Schaffel R, Hedvat CV, Teruya-Feldstein J, et al. Prognostic impact of proliferative index determined by quantitative image analysis and the International Prognostic Index in patients with mantle cell lymphoma. *Ann Oncol*. 2010;21:133--139.

Figure Legends

Figure 1. Ki67 Labelling Index determined by14 observers on 76 Virtual Slides with various method of assessment. Ki67 was quantified as percentage of positive immunoreactive tumor cells against total tumor cells and was expressed as mean.

Figure 2. Observers' evaluation as referred to the method of assessment, fields of evaluation and total number of cells utilized in the count.

Figure 3. Overall survival for DIA-MC performers (A), best performer (B), eyeballers (C) and all pathologists (D) using 0%-15%-30% as cut-offs.

Figure 4. Levels of concordance between observers following a category-based Ki67 scoring by visual estimation.

Figure 5. Intra-observer concordance of 11 observers participating both in numerical and category-based assessment of the Ki67 Labelling Index (= equal > higher < lower score on visual estimation of pre-determined images).

Figure 6. Overall survival determined by pathologists using 0%-15%-30% (A) and 0%-10%-20% (B) cut-offs compared to the software (ASH) cut-off ranges of 0%-15%-30% (C) and 0%-10%-20% (D) respectively.

Supplemental Digital Content

Supplemental Digital Content 1. Flowchart illustrating various steps of analysis through the study.

Supplemental Digital Content 2. Intra-observer variability as evaluated by Cohen's Kappa for 11 observers participating both in numerical and category-based assessment of the Ki67 Labelling Index (number of observations with = equal and/or > higher and/or < lower score on visual estimation of pre-determined images)