Ki-67 Digital Image Analysis: Reliability and Variability

Peiqi Wang[a], Tian Yu Liu[b], Willa Shi[c], Sehrish Butt[d], Trevor McKee[d], Fei-Fei Liu[a,c], Naomi A. Miller[c], Adewunmi Adeoye[c], David McCready[c], Anthony Fyles[c], Susan J. Done[a,c,*]

[a]*Department of Medical Biophysics, University of Toronto, Canada*
[b]*Faculty of Music, University of Toronto, ON, Canada*
[c]*Princess Margaret Cancer Centre, Canada*
[d]*STTARR Medical Diagnostics Imaging Center*

Abstract

Ki-67 is a nuclear protein reflective of tumour proliferative state. The usage of Ki-67 labeling index as a prognostic and predictive biomarker in breast cancer is encouraging. Nonetheless, it is not recommended for routine clinical usage because of a lack of standardization. Digital image analysis (DIA) holds good potential; however, investigation on the reliability as well as intra- and inter-algorithmic variability of different DIA methods is lacking. In this study, we scored a set of cases (n=278) utilizing manual hotspot counting and two digital image analysis (DIA) methods, Aperio ePathology and Definiens Tissue Studio. The Definiens system achieved high agreement (ICC: 0.892) with manual score reference and classified cases accurately (κ: 0.67) based on a cut-off of 14%. DIA is able to achieve comparable results with manual hotspot counting. The Aperio system was run twice on two sets of independently segmented images. Agreement to the manual score reference was at best moderate (ICC: 0.173 and 0.439, κ: 0.155 and 0.233) due to over-estimation bias of the system. Considerable intra- and inter-algorithmic variability was observed. In addition to heterogeneous tumour biology and varying algorithm implementation, settings assignments and image segmentation are main contributors to such variability. Calibrating analytical settings and novel designs of automatic image segmentation are crucial toward harmonizing DIA.

Keywords: ki67, breast cancer

Introduction

Ki-67 is a human nuclear protein detected exclusively in the active phases of the cell cycle, namely G_1, S, G_2, and mitosis, while absent in the resting G_0 phase.[1] It is highly sensitive to cell cycle changes, making it an ideal marker for quantifying uncontrolled proliferation, a hallmark of cancer. Unsurprisingly, Ki-67 immunohistochemical (IHC) staining of human neoplsmic cells has emerged as a rapid and cost-effective analytics capable of determining the growth fraction of tumour cell populations, [2] The use of Ki-67 labelling index, or the percentage of Ki-67-positive cells, has great prognostic potential particularly in carcinomas of the breast, where a multitude of studies report the use of Ki-67 labeling index in predicting disease free/overall survival and tumour recurrence [3–5] as well as in guiding neoadjuvant chemotherapy. [6–8] Practically, Ki-67 labeling index may contribute to improved tumour grading, where proliferation is routinely assessed using mitotic count. [9] Additionally, Ki-67 labeling index used in conjunction with established breast histopathological markers may serve as a feasible and cost-effective alternative to gene signature based assessments such as OncotypeDx in cancer subtyping. [10]

Despite its apparent value in cancer prognosis, widespread use of Ki-67 labeling index in clinical pathology is hampered by the lack of standardization and suffers from substantial intra- and interobserver variability. [11, 12] Although recommendations and guidelines exist in an effort to harmonize such variability, [13] the choice of scoring methods and selection of cut-off for Ki-67 positivity remain a subject of debate. One promising approach to the problem utilizes digital image analysis (DIA), which ensures automaticity, repeatability and reproducibility. However, aforementioned characteristics do not guarantee objectivity; Differences in image segmentation and algorithm used could still give rise to variability. [14] Some DIA methods were reported to agree comparably with [15, 16] or even outperform visual assessments; [17, 18] Others suggested that DIA methods were less reliable and prognostic. [19] It is apparent that inter-algorithmic variability is high and performance is context dependent. Therefore, there is a great need to validate the reliability of existent Ki-67 DIA methods so as to identify major sources of variability and potential solutions.

In this study, we evaluated two digital image analysis methods - Aperio ePathology and Definiens Tissue Studio. The former requires explicit manual image segmentation; while the latter semi-automatically segments the images by first calibrating against a few test cases. We assessed the reliability of the two DIA methods by reporting their agreement to a set of manual scores previously identified to be a predictor of ipsilateral breast cancer relapse in the the Toronto-British Columbia (TBC) trial patient cohort. [20] We compared agreements within and be-

*Principal corresponding author
Email address: Susan.Done@uhn.ca (Susan J. Done)

tween the DIA methods so as to evaluate intra- and inter-algorithmic variability. We also discussed potential sources of variability and ways to mitigate them.

Materials and Methods

Sample Collection

A subset of patient cohort from the TBC trial were used for this study. [20] The TBC trial consists of node-negative patients who were older than 50 years of age randomly assigned to receive tamoxifen alone or tamoxifen and breast radiotherapy after breast-conserving surgery. [21] Tissue microarrays were constructed using a triplicate of 0.6 mm tumour cores from formalin-fixed, paraffin-embedded blocks. A total of 6 TMA blocks, amounting to 278 cases, were taken directly from a previous study and used for subsequent image analysis. [20] TMA blocks were cut in 0.5 µm sections, stained with 1:500 dilution SP6 (NeoMarker) and counter-stained with hematoxylin.

Scoring Methodologies

Manual Assessment

A trained individual, assigned as rater 1, counted at least 200 cells within tumour hot spot, or areas in which Ki-67 most frequently expressed, for each core. The total number of nuclei and positively stained nuclei over the span of three cores were summed and the Ki-67 labeling index was calculated for each case. 10% of the samples were randomly chosen and rescored for quality assurance by experienced pathologists (NM, SD). As the scores resulting from this set of manual assessment was clinically significant in predicting ipsilateral breast cancer relapse, they were used as a reference to be compared with results from other scoring methods on the same set of stained slides.

Digital Image Analysis (DIA)

To assess intra-algorithmic variability of the DIA methods, specifically the Aperio system, 2 trained individuals, assigned as rater 1 and rater 2, independently marked tumour region of interest (ROI) for proper image segmentation. Analytical Settings, such as minimum nucleus radius and staining intensity threshold, for the algorithm were subjectively adjusted for by another experienced pathologist and used in both set of images. The same set of images were analyzed using the Definiens system in addition to the Aperio system. In this case, a technician, assigned as rater 3, segmented images in a few cases, which calibrated the software to perform semi-automatic segmentation. Minor adjustments were made to correct for faulty segmentation.

Statistics

Data distribution for different scoring methods were visualized using boxplot, accompanied by summary statistics. Scatter plot as well as Bland-Altman plot were used to visualize agreements between results from two DIA methods in relation to manual score reference. [22] 95% confidence interval for the limits of agreement as well as the mean difference was calculated based on an alpha of 0.05. Two methods were considered unbiased if the mean difference centered about zero. [23] To correct for positive skewness, Ki-67 labeling indices were log base 2 transformed after incrementing by 1% for subsequent statistical calculation. Concordance between methods was quantified using a two-way mixed, average-measures intraclass correlation coefficient (ICC) to assess the degree that raters provide absolute agreement in their ratings of Ki-67 labeling index. [24] An ICC close to 1 represents high concordance. Conger generalized Kappa (κ) were calculated based on a set of commonly used cut-offs for Ki-67 positivity to evaluate the practicality of consistent classification using results from methods tested. [25] Agreement of the two DIA methods to the manual score reference was computed using ICC and κ. intra-algorithmic variability was evaluated by comparing ICC and κ between two scoring instances for the Aperio system. inter-algorithmic variability was evaluated by comparing the agreements of two DIA methods with each other. R (version 3.2.4) was used generate all statistics and graphs.

Results

Overall Distribution of different scoring methods

Boxplot of untransformed Ki-67 labeling index as well as summary statistics presented in Figure 1. The Aperio system tended to overestimate Ki-67 labeling index; whereas the Definiens system showed a similar distribution to manual score reference.

The Definiens system agreed well with manual score reference

Scatter plot comparing DIA methods to manual assessment was presented in Figure 2. Bland-Altman plot for every DIA method compared to manual score reference and relevant statistics were presented in Figure 3. It was apparent that the Aperio system systematically overestimated Ki-67 labeling index by a large margin in both scoring instances. The Definiens system faired better in introducing minimal bias, but still exhibited non-negligible variability.

ICCs comparing two rating instances using the Aperio system and the manual score reference were 0.173 (95%CI -0.245 ~ 0.459) and 0.439 (95%CI -0.258 ~ 0.72) respectively, representing poor to moderate agreements. ICC comparing the Definiens system and the manual score reference was 0.892 (95%CI 0.841 ~ 0.924), indicating high degree of agreement. Ki-67 labeling index was scored similarly using manual assessment and the Definiens system.

It may be misleading to solely measure absolute agreement, as ultimately cases would be classified into clinically relevant groups based on the Ki-67 labeling index. Kappa statistics calculated using cut-offs from a meta-analysis study were listed in Figure 4. [5] With a 14% cut-off used to distinguish luminal B from luminal A tumours, [26] κ obtained using the Definiens system was 0.67, suggesting a substantial agreement in making clinically relevant classifications. [27] With a hypothetical 25% cut-off used to distinguish 'luminal B-like' tumours proposed in the recent St. Gallen Breast Cancer Conference, [28]

the two DIA methods achieved fair to moderate agreement with κ of 0.35 and 0.57 respectively.

The discrepancy in agreements of the two DIA methods to manual score reference could be largely attributable to subjective assignments of analytical settings in addition to varying algorithm implementation.

High Intra-algorithmic Variability within the Aperio system

Aforementioned ICC comparing the Aperio system and the manual score reference varied (ICC: 0.173 and 0.439). ICC between two rating instances when using the Aperio system was 0.538 (95%CI: 0.31 ~ 0.68), which represented moderate agreement. Additionally, κ comparing two rating instances using the Aperio system indicated slight to fair agreement as presented in Figure 4. [27] As the same analytical settings were used for both rating instances, manual image segmentation was responsible for the differences in Ki-67 labeling index generated. Both ICC and κ suggested that a substantial amount of variability was introduced in the process of image segmentation. [29] The systematic bias observed in the Bland-Altman plot may also be responsible for such disagreement.

High Inter-algorithmic Variability between the two DIA methods

ICC comparing two rating instances using the Aperio system and the Definiens system were 0.351 (95%CI: -0.315 ~ 0.678) and 0.596 (95%CI: -0.182 ~ 0.823). κ comparing the two DIA methods was slight to fair as presented in Figure 4. [27] As a result of ICC and κ, high inter-algorithmic variability was observed between the two DIA methods. It was difficult to control for differences when using the two DIA methods. However, algorithm implementation, segmentation procedure, and rater bias all contribute to perceived high inter-algorithmic variability.

Discussions

There has been developing interest in using Ki-67 labeling index to quantify proliferation levels in cancer research, for cancer subtyping, prognosis, and deciding treatments. However, only a highly reproducible and accurate procedure could be used routinely and reliably in the clinics. Comparison of counting methodologies yield varying results. [30, 31] Although there are promising efforts to standardize manual counting methodologies [12, 13], inherent limitations, such as poor scalability for high throughput assays, are often left unconsidered. Digital image analysis (DIA) offers a viable alternative that is more efficient, repeatable, and scalable. Systematic evaluation of intra and inter-algorithmic variability is warranted but rarely done.

In this study, we assessed agreements of results from two DIA methods to a set of manual score reference (n=278) that is prognostically relevant. The Definiens system was observed to agree well with the manual score reference, both in absolute value of Ki-67 labeling index and in its ability to segregate cases into clinically relevant groups. Although DIA can be highly accurate, not all DIA methods have a good performance. High intra- and inter-algorithmic variability was observed when the two DIA methods were compared within and between themselves. We identified image segmentation as a hugely important contributor to high intra-algorithmic variability in the Aperio system, when all else is kept consistent. Additionally, systematic bias was detrimental toward achieving high agreement. In our study, we identified settings assignments as potential major source of such discrepancy, in addition to different algorithm implementations.

A study reported using a test validation, calibration and measurement error correction methodology to fine-tune settings for accurate Ki-67 labeling index estimation, achieving a 2X reduction in misclassification rates. [17] Such approach could be readily adapted to other DIA methods and reduce variability arisen from subject setting assignments, a predominant source of variability found in this study. However, test validation in the form of stereological test grid counting necessitates manual effort, which goes against the very idea of automation. Deciding the optimal settings for a given set of cases requires further investigation on developing automated calibration methodologies.

Image segmentation has long been a computer vision problem. However, accurate segmentation regardless of the subject and quality of scanned images is particularly challenging. Varied level of human subjectivity is introduced so as to simplify the problem. [14] Improving segmentation algorithms is perhaps the fasted way to reduce intra-algorithmic variability.

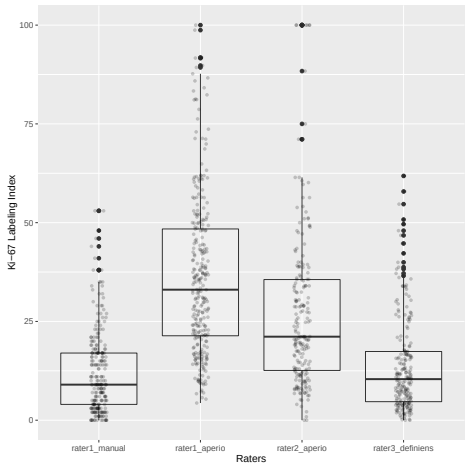
One caveat of this study lies in comparing scores from DIA methods to a set of manual 'gold standards'. Even though the manual score reference was a predictor of clinical outcome, the comparison was nonetheless an indirect one. Instead of pursuing agreements in Ki-67 labeling index, one can assess statistical association of Ki-67 labeling index generated using DIA methods with clinical outcome directly and validate its significance with that of manual assessment. Other studies have explored such idea and found that results from DIA may be a superior prognostic factor. [18]

In conclusion, DIA method can perform comparably with traditional manual assessment methods. Intra- and inter-algorithmic variability is considerable amongst the two DIA methods tested and can be a prevalent phenomenon, hindering valid comparison across different DIA platforms. Settings assignments and image segmentation are major sources of such variability. Novel algorithms on calibration and segmentation are needed toward standardized DIA procedure.

References

1. Gerdes, J. *et al.* Cell cycle analysis of a cell proliferation-associated human nuclear antigen defined by the monoclonal antibody Ki-67. *The Journal of Immunology* **133,** 1710–1715. ISSN: 0022-1767 (1984).
2. Scholzen, T. & Gerdes, J. The Ki-67 protein: From the known and the unknown. *Journal of Cellular Physiology* **182,** 311–322. ISSN: 0021-9541 (Mar. 2000).
3. Stuart-Harris, R. *et al.* Proliferation markers and survival in early breast cancer: A systematic review and meta-analysis of 85 studies in 32,825 patients. *The Breast* **17,** 323–334 (2005).

4. De Azambuja, E. *et al.* Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12,155 patients. *British journal of cancer* **96,** 1504–13. ISSN: 0007-0920 (May 2007).
5. Petrelli, F., Viale, G., Cabiddu, M. & Barni, S. Prognostic value of different cut-off levels of Ki-67 in breast cancer: a systematic review and meta-analysis of 64,196 patients. *Breast Cancer Research and Treatment* **153,** 477–491. ISSN: 0167-6806 (Oct. 2015).
6. Jones, R. L. *et al.* The prognostic significance of Ki67 before and after neoadjuvant chemotherapy in breast cancer. *Breast Cancer Research and Treatment* **116,** 53–68. ISSN: 0167-6806 (July 2009).
7. Nishimura, R., Osako, T., Okumura, Y., Hayashi, M. & Arima, N. Clinical significance of Ki-67 in neoadjuvant chemotherapy for primary breast cancer as a predictor for chemosensitivity and for prognosis. *Breast Cancer* **17,** 269–275. ISSN: 1340-6868 (Oct. 2010).
8. Fasching, P. A. *et al.* Ki67, chemotherapy response, and prognosis in breast cancer patients receiving neoadjuvant treatment. *BMC Cancer* **11,** 486. ISSN: 1471-2407 (Dec. 2011).
9. Van Diest, P. J., van der Wall, E. & Baak, J. P. A. Prognostic value of proliferation in invasive breast cancer: a review. *Journal of clinical pathology* **57,** 675–81. ISSN: 0021-9746 (July 2004).
10. Cuzick, J. *et al.* Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the Genomic Health recurrence score in early breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **29,** 4273–8. ISSN: 1527-7755 (Nov. 2011).
11. Dowsett, M. *et al.* Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *Journal of the National Cancer Institute* **103,** 1656–64. ISSN: 1460-2105 (Nov. 2011).
12. Polley, M.-Y. C. *et al.* An international Ki67 reproducibility study. *Journal of the National Cancer Institute* **105,** 1897–906. ISSN: 1460-2105 (Dec. 2013).
13. Polley, M.-Y. C. *et al.* An international study to increase concordance in Ki67 scoring. *Modern Pathology* **28,** 778–786. ISSN: 0893-3952 (June 2015).
14. Tadrous, P. J. On the concept of objectivity in digital image analysis in pathology. *Pathology* **42,** 207–11. ISSN: 1465-3931 (Apr. 2010).
15. Mohammed, Z. M. A. *et al.* Comparison of visual and automated assessment of Ki-67 proliferative activity and their impact on outcome in primary operable invasive ductal breast cancer. *British journal of cancer* **106,** 383–8. ISSN: 1532-1827 (Jan. 2012).
16. Tang, L. H., Gonen, M., Hedvat, C., Modlin, I. M. & Klimstra, D. S. Objective quantification of the Ki67 proliferative index in neuroendocrine tumors of the gastroenteropancreatic system: a comparison of digital image analysis with manual methods. *The American journal of surgical pathology* **36,** 1761–70. ISSN: 1532-0979 (Dec. 2012).
17. Laurinavicius, A. *et al.* A methodology to ensure and improve accuracy of Ki67 labelling index estimation by automated digital image analysis in breast cancer tissue. *Breast cancer research : BCR* **16,** R35. ISSN: 1465-542X (2014).
18. Stålhammar, G. *et al.* Digital image analysis outperforms manual biomarker assessment in breast cancer. *Modern Pathology* **29,** 318–329. ISSN: 0893-3952 (Apr. 2016).
19. Chabot-Richards, D. S., Martin, D. R., Myers, O. B., Czuchlewski, D. R. & Hunt, K. E. Quantitative image analysis in the assessment of diffuse large B-cell lymphoma. *Modern Pathology* **24,** 1598–1605. ISSN: 0893-3952 (Dec. 2011).
20. Liu, F.-F. *et al.* Identification of a Low-Risk Luminal A Breast Cancer Cohort That May Not Benefit From Breast Radiotherapy. *Journal of Clinical Oncology* **33,** 2035–2040. ISSN: 0732-183X (June 2015).
21. Fyles, A. W. *et al.* Tamoxifen with or without Breast Irradiation in Women 50 Years of Age or Older with Early Breast Cancer. *http://dx.doi.org/10.1056/NEJMoa040595* (2009).
22. Bland, J. M. & Altman, D. G. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. *Lancet* **i,** 307–310 (1986).
23. Hanneman, S. K. Design, analysis, and interpretation of method-comparison studies. *AACN advanced critical care* **19,** 223–34. ISSN: 1559-7768 (2008).
24. Shrout, P. E. & Fleiss, J. L. Intraclass Correlations : Uses in Assessing Rater Reliability. *Psychological Bulletin* **86,** 420–428 (1979).
25. Conger, A. J. Integration and generalization of kappas for multiple raters. *Psychological Bulletin* **88,** 322–328. ISSN: 0033-2909 (1980).
26. Cheang, M. C. U. *et al.* Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *Journal of the National Cancer Institute* **101,** 736–50. ISSN: 1460-2105 (May 2009).
27. Landis, J. R. & Koch, G. G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **33,** 159. ISSN: 0006341X (Mar. 1977).
28. Coates, A. S. *et al.* Tailoring therapies–improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO* **26,** 1533–46. ISSN: 1569-8041 (Aug. 2015).
29. Cicchetti, D. V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* **6,** 284–290. ISSN: 1939-134X (1994).
30. Mikami, Y. *et al.* Interobserver concordance of Ki67 labeling index in breast cancer: Japan Breast Cancer Research Group Ki67 ring study. *Cancer science* **104,** 1539–43. ISSN: 1349-7006 (Nov. 2013).
31. Reid, M. D. *et al.* Calculation of the Ki67 index in pancreatic neuroendocrine tumors: a comparative analysis of four counting methodologies. *Modern Pathology* **28,** 686–694. ISSN: 0893-3952 (May 2015).



	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Rater 1 Manual	0	4	9	12.3	17	53
Rater 1 Aperio	4.35	21.4	33	36.4	48.4	100
Rater 2 Aperio	0	12.6	21.1	27.4	35.6	100
Rater 3 Definiens	0	4.69	10.4	13.8	17.4	61.8

Figure 1: **Boxplot and summary statistics** Distribution of Ki-67 labeling index generated using manual assessment and DIA methods. Outliers are represented as darkened circles. Corresponding summary statistics quantitatively describes the boxplot.

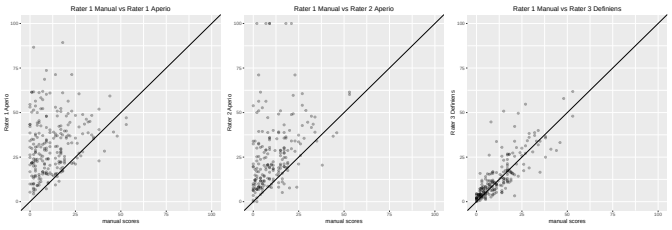
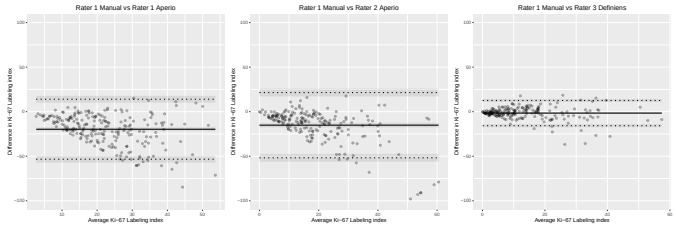


Figure 2: **Scatter plot** Scatter plot of Ki-67 labeling index generated using manual assessment as well as DIA methods. The line of perfect agreement where scores from two methods in comparison are identical is also drawn.



	Lower Limit (95% CI)	Mean Difference (95% CI)	Upper Limit (95% CI)
Rater 1 Manual vs. Rater 1 Aperio	−53.39 (−57.239 ~ −49.54)	−19.715 (−21.938 ~ −17.493)	13.959 (10.109 ~ 17.808)
Rater 1 Manual vs. Rater 2 Aperio	−51.771 (−56.239 ~ −47.303)	−15.147 (−17.726 ~ −12.567)	21.478 (17.01 ~ 25.945)
Rater 1 Manual vs. Rater 3 Definiens	−15.757 (−17.469 ~ −14.046)	−1.592 (−2.58 ~ −0.604)	12.573 (10.862 ~ 14.284)

Figure 3: **Bland-Altman Plot** Bias and agreement interval of manual reference score compared to results from DIA methods. Bland-Altman plot consists of a scatterplot, with each data point representing paired Ki-67 labeling index generated using methods in comparison. X axis is the average of paired measurement while the Y axis is the difference of paired measurement. Data points are flanked by dashed lines, which represent limits of agreement, within which 95% of differences fall. Confidence intervals of mean difference as well as upper and lower limit of agreement are shown as grey area surrounding them. Statistics relevant to the plot are tabulated in the accompanying table.

	5	10	14	20	25	30
Rater 1 Manual vs. Rater 1 Aperio	−0.175	−0.161	−0.17	−0.151	−0.101	−0.086
Rater 1 Manual vs. Rater 2 Aperio	−0.057	0.085	0.155	0.12	0.101	0.067
Rater 1 Aperio vs. Rater 2 Aperio	−0.018	0.154	0.233	0.24	0.348	0.306
Rater 1 Manual vs. Rater 3 Definiens	0.645	0.654	0.67	0.689	0.568	0.507
Rater 1 Aperio vs. Rater 3 Definiens	0.002	0.213	0.248	0.239	0.333	0.362
Rater 2 Aperio vs. Rater 3 Definiens	−0.112	−0.062	−0.146	−0.068	0.096	0.055

Figure 4: **Kappa statistics** κ calculated is a measurement of the extent that different rating instances correctly classify Ki-67 labeling index into Ki-67 low and Ki-67 high based on a selection of cut-offs, namely 5%, 10%, 14%, 20%, 25%, 30%.