Investigating semi-automated ki67 scoring efficacy

Tian Yu Liu[a], Peiqi Wang[b], Susan J. Done[c,*]

[a] *Faculty of Music, Univeristy of Toronto, ON, Canada*
[b] *Department of Molecular Genetics and Microbiology, University of Toronto, Canada*
[c] *The Campbell Family Institute for Breast Cancer Research, Canada*

Abstract

300words, work done, result obtained, conclusions drawn. filled in later

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 expressed in virtually cells of every tissue origin and is highly sensitive to cell cycle changes, making it an ideal marker for quantifying uncontroled proliferation, a hallmark of cancer. Unsurprisingly, Ki-67 immunohistochemical (IHC) staining of human neoplasmic cell 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 particuarly in carcinomas of the breast, where a multitude of studies reported the use of Ki-67 labeling index in predicting disease free/overall survival and tumour recurrence [3, 4] as well as in guiding neoadjuvant chemotherapy. [5–7] Practically, Ki-67 labeling index served as a feasible alternative to gene signature based assessments such as OncotypeDx in cancer subtyping when used in conjunction with established breast histopathological markers such as ER, PgR, and HER2. [8] 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- amd interobserver variability. [9, 10] Although recommendations and guidlines exist in an effort to harmonize such variability, [11] the choice of scoring methods and selection of cutoff for Ki-67 positivity remain a subject of debate. One promising approach to the problem utilizes digital image analysis (DIA), which reduces interobserver variability and alleviatas labor intensive work. In some cases, DIA was reported to outperform traditional manual scoring methods. [12]

In this study, we evaluated the efficacy and reproducibility of two digital image analysis methods - Aperio ePathology and Definiens Tissue Studio. Additionally, we measured their agreement to a set of manual scores previously identified to be a predictor of ipsilateral breast relapse in the the Toronto-British Columbia (TBC) trial patient cohort. [13]

[2]

Materials and Methods

Patients and Sample Collection

A subset of patient cohort from the TBC trial were used for this study. 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. [14] 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, where used for subsequent IHC and image analysis.

Ki-67 Immunohistochemistry

TMA blocks were cut in 0.5 μm sections and incubated with monoclonal MIB-1 antibody (Dako) at [time, temperature] and counter-stained with hematoxylin. [more on positive negative control and specifics of staining, are they all from the same staining]

Scoring Methodologies

Manual Assessment

A trained pathologist counted the number of brown staining for at least 100 cells within tumour hot spot, or areas in which Ki-67 staining is predominant, for each core. The total number of nucleus and positively stained nucleus 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.

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

Digital Image Analysis (DIA)

Stained sections were scanned by [detail regarding scaninng process]. To assess reproducibility of DIA, specifically the Aperio system, 2 pathologists independently marked tumour region of interest. Settings for the detection algorithm were adjusted for by another experienced pathologist and used in both set of annotations. Annotated images were analyzed to quantify inter-rater reliability when using a DIA method. To assess the agreement of DIA method to the manual scores, the same set of images were analyzed using the Definiens system in addition to the Aperio system. In this case, a technician marked tumour areas in a few cores, which trained the software to recognize tumour areas for every other cases.

Statistics

Data distribution cross different scoring methods were visulized using boxplot, accompanied by summary statistics. Inter-rater reliability (IRR) was quantified using a two-way mixed, average-measures intraclass correlation coefficient(ICC) to assess the degree that raters provide absolute agrement in their ratings of Ki-67 labeling index using the Aperio system. An ICC close to 1 represents high reliability. Similarly, ICC was used to assess the degree that different scoring methods agree on the same patient cohort. Bland-Altman plot was used to visualize agreement between results from manual assessment and the two DIA methods. [15] 95% confidence interval for the limits of agreement as well as the mean of difference was calculated based on an alpha of 0.05. High agreement between scoring methods equates to a mean of differences centered about zero with a small standard deviation. Fleiss Kappa were calculated based a cutoff for Ki-67 labeling index of 15 to quantify the practicality of correctly classfying a core into clinically relevant groups , namely Ki-67 low and Ki-67 high.

Results

Overall distribution

Boxplot of untransformed Ki-67 labeling index as well as summary statistics of log2-transformed Ki-67 labeling index were presented in Figure 1, and Table 1. The Aperio system tended to overestimate Ki-67 labeling index; whereas the Definiens system corresponded well to manual assessment.

Inter-rater Reliability Using a DIA Method

ICC between two raters using the the Aperio system was 0.675 (95%CI: 0.534-0.768) The resulting ICC could only be considered fair, suggesting that a considerable amount of error was introduced by annotating appropriate tumour areas for analysis. [16] Additionally, the Kappa statistics was fairly low when comparing how two raters agree using the Aperio system alone (0.256).

Agreement of DIA methods to Manual Assessment

Bland-Altman plot for every DIA method with respect to corresponding manual assessment was listed in Figure 2, 3, 4. The limits of agreement using the Aperio system against manual assessment were -53.4(95%CI: -57.2–49.5) to 14.0 (95%CI: 10.1-17.8) with a mean of difference of -19.7 (95%CI: -21.9-17.8) and -51.8 (95%CI: -56.2–12.6) to 21.5 (95%CI: -17.7-25.9) with a mean of -15.1 (95%CI: -47.3-17.0) respectively. It is apparent that the Aperio system systematically overestimated Ki-67 labeling index by a large margin, perhaps a direct result of inaccurate calibration procedure as there is no a priori settings for the algorithm. The limits of agreement using the Definiens system against manual assessment were -10.383 (95%CI: -12.314–8.451) to 10.960 (95%CI: 9.029-12.892) with a mean difference of 0.289 (95%CI: -0.826-1.404). High agreement was observed with minimal bias. This may be attributable to better detection algorithms and design of workflow, where raters would necessarily rely on computer algorithm to detect tumour region. ICC of two raters using the Aperio system when compared directly to the manual assessment was 0.185 (95%CI -0.25-0.475) and 0.333 (95%CI -0.166-0.595) respectively; whereas ICC of rater using the Definiens system when compared to the manual assessment was 0.935 (95%CI 0.903-0.957). Such dichotomizing CCI implicated a substantial difference in the analytical accuracy of different DIA methods. Kappa statistics were -0.142 and 0.127 respectively for using the Aperio system, indicating the inability of the Aperio system to classify cores into clinically relevant groups. For the Definiens system, the Kappa statistics was 0.787, indicating high agreement in making clinically relevant classifications.

Discussions

interpretation of significance of findings, relate to other works, further research directions

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. 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).
6. 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).
7. 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).

8. 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).
9. 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).
10. 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).
11. 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).
12. Stålhammar, G. *et al.* Digital image analysis outperforms manual biomarker assessment in breast cancer. *Modern Pathology* **29,** 318–329. ISSN: 0893-3952 (Apr. 2016).
13. 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).
14. 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).
15. Bland, J. M. & Altman, D. G. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. *Lancet* **i,** 307–310 (1986).
16. 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).

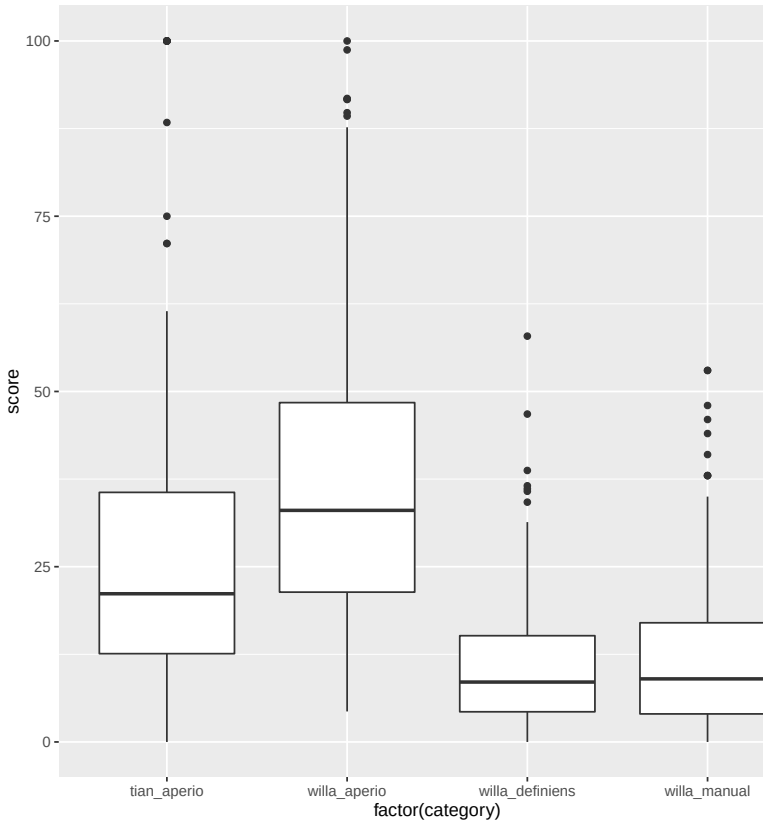

Figure 1: **Summary boxplot of Ki-67 labeling index**

s	tian_aperio	willa_aperio	willa_definiens	willa_manual
Min.	−3.32	2.15	−3.32	−3.32
1st Qu.	3.67	4.42	2.14	2.04
Median	4.41	5.05	3.11	3.19
Mean	4.35	4.95	2.94	2.79
3rd Qu.	5.16	5.60	3.93	4.10
Max.	6.65	6.65	5.86	5.73

Table 1: **Summary statistics for log2-transformed Ki-67 labeling index**

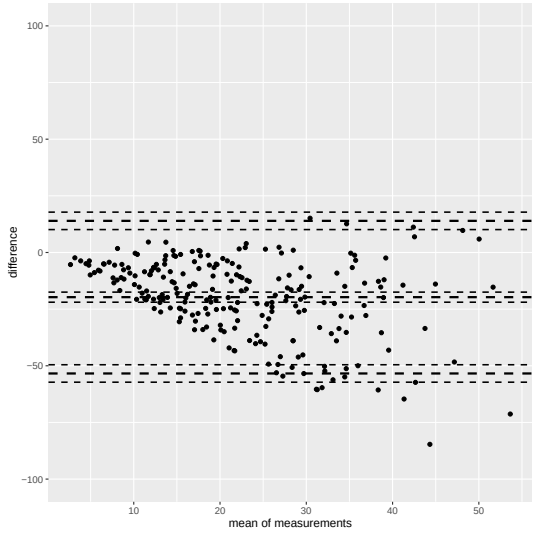

Figure 2: **Bland-Altman Plot for Aperio DIA rater 1 vs. manual assessment**

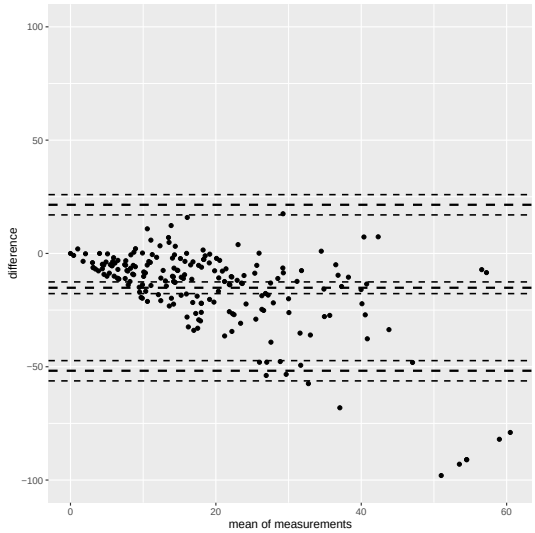

Figure 3: **Bland-Altman Plot for Aperio DIA rater 2 vs. manual assessment**

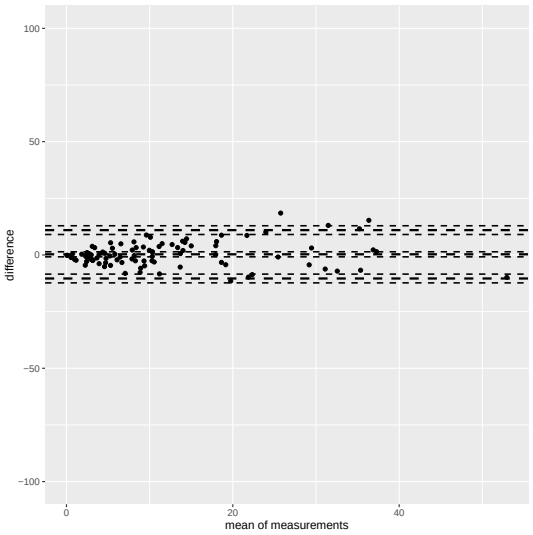

Figure 4: **Bland-Altman Plot for Definiens DIA vs. manual assessment**