Scientific
Research
Publishing

Automatic and Manual Proliferation Rate Estimation from Digital Pathology Images

Lama Rajab[1], Heba Z. Al-Lahham[1], Raja S. Alomari[2], Fatima Obaidat[1], Vipin Chaudhary[2]

[1]The University of Jordan, Amman, Jordan
[2]The University at Buffalo, Buffalo, USA
Email: Lama.rajab@ju.edu.jo, ralomari@buffalo.edu

Abstract

Digital pathology is a major revolution in pathology and is changing the clinical routine for pathologists. We work on providing a computer aided diagnosis system that automatically and robustly provides the pathologist with a second opinion for many diagnosis tasks. However, interobserver variability prevents thorough validation of any proposed technique for any specific problems. In this work, we study the variability and reliability of proliferation rate estimation from digital pathology images for breast cancer proliferation rate estimation. We also study the robustness of our recently proposed method CAD system for PRE estimation. Three statistical significance tests showed that our automated CAD system was as reliable as the expert pathologist in both brown and blue nuclei estimation on a dataset of 100 images.

Keywords

1. Introduction

The development and continued growth of cancerous cells involve various changes at both macro and micro levels of the body. Cell proliferation is usually among the major indicators for proliferation of cancerous cells. Specifically, breast cancer proliferation rate estimation (PRE) is a crucial step for determining the cancer level and is used as a prognostic indicator [1]. In conjunction with tumor size and grade, lymph node status and histological grade, PRE is an indicator for the aggressiveness of individual cancers and helps setting the treatment plan [2].

Traditionally, pathologists perform proliferation rate estimation for breast cancer by examining the whole slides via a microscope. Over the past two decades, digital pathology enabled the usage of high resolution digi-

tizers to provide high resolution images that replace the microscope as shown in our previous work [3].

There are many clinically approved techniques to estimate the PRE including: mitotic index, S-phase fraction, nuclear antigen ImmunoHistoChemistry (IHC) including KI-67 and PCNA-staining Cyclins and PET [4] [5]. Each one of these methods has its advantages or disadvantages based on the clinical settings.

In our work, we use Ki-67-stained biopsy images for PRE. In this technique, PRE is estimated by counting the number of brown nuclei and the number of blue nuclei as shown in **Figure 1**. Stromal areas are clinically excluded from counting because stromal area does not become cancerous. In our previous work [6], we performed digital stromal area removal to eliminate this ambiguous area for both junior pathologists and automated PRE systems.

Manual PRE is time-consuming and laborious for pathologists. An average of six minutes per image is required for PRE by an expert pathologist. Our expert pathologist requires over 10 hours estimating the proliferation rate for our dataset containing 100 images. Many authors target automation of PRE including our recent work [7]. However, one major concern was not investigated in all these efforts which was the inter-variability between the expert pathologists [8].

In this paper, we study the statistical inter-pathologist variability for the various manual PRE we have between four expert pathologists. Moreover, we investigate the reliability of our proposed automated PRE compared to the four pathologist opinions for the 100 images in our dataset.

2. Materials and Methods

Manual ground truth estimation is a major area of interest due to the various human factors that influence the experts. Specifically for breast cancer PRE [9] [10], we find that pathologists provide variable ground truth estimations which make it hard to evaluate any automated PRE estimated technique. Many automated PRE techniques have been proposed in the literature and we recently proposed our technique in [7], an exhaustive review of the techniques as well as a detailed description of our techniques are presented in [7]. In this paper, we provide the necessary statistical study for the inter pathologists variability. Furthermore, we study the statistical variability between the four manual ground truth and our automated technique. In [7], we compared the automated results with one expert pathologist and a student trained by a pathologist. In this paper, we run our statistical study to include for expert pathologists and one automated technique.

We study three statistical significance tests to show the inter-observer variability. Moreover, we study the manual vs automated [7] PRE variability. We study three statistical significance measures: correlation coefficient, T-Test, and Ch-Square test. We briefly describe them due to space limitations.

2.1. Correlation Coefficient

The value of the correlation coefficient $\rho_{x,y}$ gives an indication about the strength and the nature of the relationship between two random variables x and y. It ranges between −1 and +1. A value of +1 means perfect correlation between these two random variables while a −1 values indicates maximal uncorrelated variables [11] the equation for the correlation is shown in Equation (1):
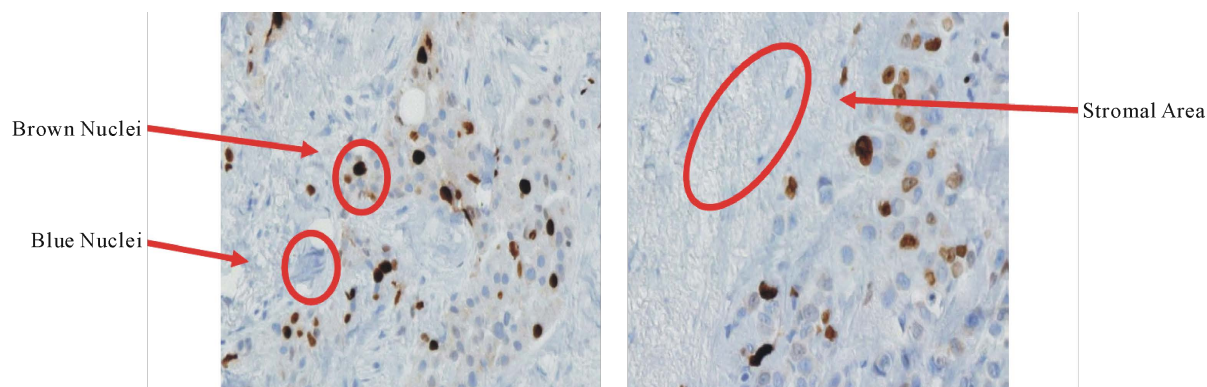
Figure 1. The sample images of Ki-67 stained pathology images showing sample blue nuclei, and stromal areas.

$$\rho_{x,y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\left(\sum (x - \bar{x})^2 \sum (y - \bar{y})^2 \right)}} \tag{1}$$

where x and y are two random variables, $\bar{x}$ and $\bar{y}$ are the corresponding mean values for each sample.

2.2. Student T-Test

Student T-Test (or t test for short) is one of a number of hypothesis tests. The t-test looks at the t statistic, t distribution and degrees of freedom to determine a t value (probability) that can be used to determine whether the two underlying distributions of the two random variables are different as shown in Equation (2):

$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\dfrac{\sigma_r^2}{n_T} + \dfrac{\sigma_c^2}{n_C}}} \tag{2}$$

where $\bar{x}_T$ and $\bar{x}_C$ are the two mean values for the corresponding two data samples r and c, σ_r^2 and σ_c^2 are the corresponding variances for the two data samples r and c, n_T and n_C are the number of the corresponding samples. Moreover, the degrees of freedom (df) for the test should be determined. In the t-test, the degree of freedom is the sum of the persons in both groups minus 2. Given the alpha level, the df, and the t value, you can look the t value up in a standard table of significance Typically, when $t > 0.05$, the two random variables (two underlying data samples) are said to be statistically insignificance, *i.e.*, highly correlated [12].

2.3. Chi-Square

Chi square X^2 is a statistical test commonly used to compare observed with unobserved data upon a specific hypothesis as in Equation (3):

$$X^2 = \frac{\sum_{i=1}^{r} \sum_{j=1}^{c} (Oij - Eij)^2}{Eij} \tag{3}$$

where Oij is the observed frequency in the ith row and j^{th} column, Eij is the expected frequency in the ith row and j^{th} column, r is the number of rows and c is the number of columns. The appropriate number of degrees of freedom (df) is calculated as the number of rows-1 multiplied by the number of columns. If X^2 is greater than what is known as the critical value, then the two samples are dependent.

3. Experimental Results and Analysis

Our data set contain 100 Ki-67-stained histopathology digital images for breast cancer. The blue nuclei are negative positive cells while the brown nuclei are the positive ones. Our collaborating pathologists provided us with the ground truth from four different pathologists including herself as the most senior pathologist. We provided each pathologist with anonymized images labeled in sequence along with an sheet to score for the blue and brown nuclei. None of the four pathologists knew about the other and they were scoring independently. Our most senior pathologist (coauthor) spent over 10 hours for scoring the 100 cases which means an average of 6 minutes per case. Moreover, we run our proposed automated PRE system proposed in [7] over the same 100 images and recorded the automated scoring for both the blue and the brown nuclei.
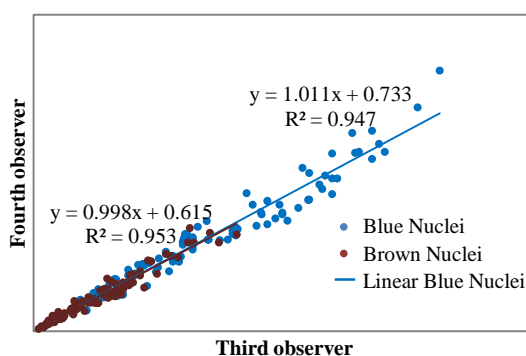
3.1. Correlation Coefficient

The inter-observer reproducibility is first measured by using the correlation coefficient [13] [14]. Overall, there is a higher correlation between pathologists in brown nuclei estimation than blue nuclei estimation. Moreover, our automated CAD system has also a higher correlation coefficient for brown nuclei compared to blue ones. **Table 1** summarizes the inter pathologists correlation coefficient values and manual vs automated correlation coefficient values.

From **Table 1**, we note that the correlation coefficient indicates a very high correlation between the four observers on the brown nuclei counting. However, the correlation is highly variable for blue nuclei counting from an upper value of 0.73 down to 0.768. **Figure 2** and **Figure 3** show the relationship for the manual PRE for

Table 1. Significance values of correlation coefficients.

Expert observers	Correlation coefficient	
	Brown count	Blue count
Observer 1 vs Observer 2	0.987	0.969
Observer 1 vs Observer 3	0.953	0.807
Observer 1 vs Observer 4	0.965	0.806
Observer 2 vs Observer 3	0.959	0.968
Observer 2 vs Observer 4	0.965	0.806
Observer 3 vs Observer 4	0.977	0.973
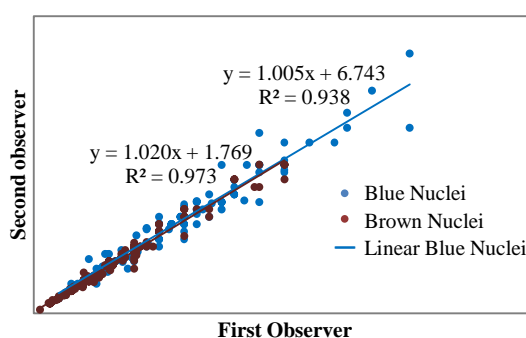
Figure 2. Relationship between first and second observers' nuclei count estimates.

Figure 3. Relationship between third and fourth observers' nuclei count estimates.

observer 1 vs observer 2 and observer 3 vs observer 4, respectively.

On the other hand, we study the correlation coefficients between the manual of each of the four experts and our proposed automated system as shown in Table 2. As we examine this table, the brown nuclei counting is highly correlated to the various observers which indicates an almost perfect reliability of our proposed automated system for brown nuclei estimation. Furthermore, the blue nuclei counting are comparable to the correlation between the manual observers. In other words, our automated blue nuclei estimation is as good as the manual estimation which proves its clinical reliability.

3.2. T-Test

We performed a two-tailed paired T-Test on all the pairs between the four observers and the automated system.

Our Null Hypothesis is that there is a difference between the observers in one hand and the automated system on the other hand. All of the reported significance probability values in **Table 3** shows insignificant statistical difference between the manual expert estimations themselves on one hand and between both the 3rd and 4th observers with the automated system on the other hand. In other words according to **Table 4** which shows the interpretation for the p-value. As you see the p value is less than 0.01 which means that we have a strong evidence to reject the hypothesis that says that there is no relationship (there is a difference) between observers on one hand and the automated system in the other hand in both Brown and Blue nuclei counts estimation.

3.3. Chi Square

We computed Chi-square test all pairs, and it compared with the critical chi square value with $df = 1$, confidence level 99% (probability $= 1 − 0.99 = 0.01$). In all pairs (including inter-observer and our automated method), the calculated chi square value is greater than the critical value, which means that each pair of samples are dependent. In other words, it is statistically reliable to consider any of the expert scoring or the automated scoring values. **Figure 4** and **Figure 5** show two samples images where we high agreement between observes, and a low agreement between observers, respectively.

**Table 2.** Manual vs automated significance values of correlation coefficients.

Expert observers	Correlation coefficient	
	Brown count	Blue count
Observer 1 vs automated	0.974	0.847
Observer 2 vs automated	0.984	0.886
Observer 3 vs automated	0.959	0.661
Observer 4 vs automated	0.963	0.686

**Table 3.** Significance values resulting from paired T-Test.

Expert observers	p-value	
	p (brown count)	p (blue count)
Observer 1 vs observer 2	8.14×10^{-5}	7.2×10^{-6}
Observer 1 vs observer 3	1.5×10^{-4}	1.5×10^{-15}
Observer 1 vs observer 4	8.7×10^{-5}	2.2×10^{-15}
Observer 2 vs observer 3	8.9×10^{-9}	1.4×10^{-11}
Observer 2 vs observer 4	4.3×10^{-9}	7.1×10^{-12}
Observer 3 vs automated	2.9×10^{-7}	6.9×10^{-11}
Observer 4 vs automated	4.0×10^{-7}	1.5×10^{-11}

Table 4. Interpreting significance probability (p-value).

Significance probability p-value	Interpretation
$p < 0.01$	Strong evidence to reject H_0
$0.01 < p \leq 0.05$	Significant EVIDENCE TO REject H_0
$0.05 < p \leq 0.10$	Weak evidence against H_0
$p > 0.10$	Insignificant evidence to reject H_0
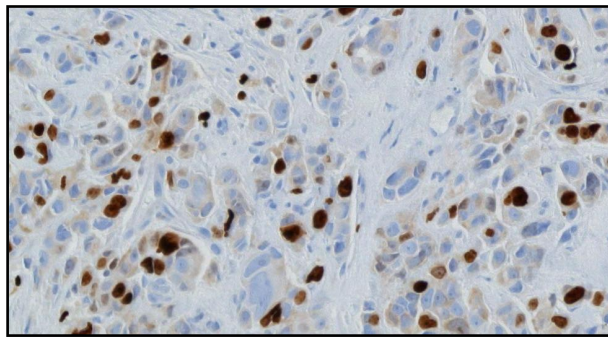
Figure 4. Example of an image has the same value for the brown nuclei in all observers.
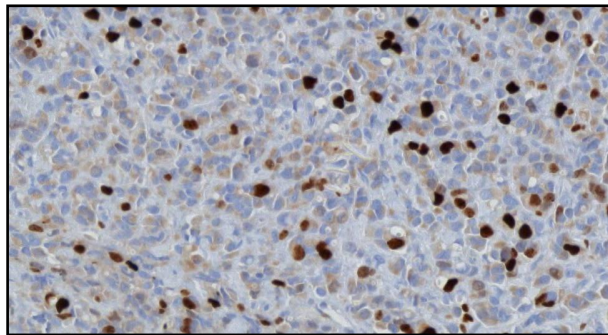
Figure 5. Example of an image where the observers results are completely different.

4. Conclusion

We proposed a detailed statistical study for breast cancer proliferation rate estimation. We studied the inter-observer variability between four expert pathologists on a set of 100 cases. We also studied the reliability of our recently proposed automated PRE system. On the 100 cases, we found that the variability of brown nuclei estimation was statistically insignificant between various pathologists. We also found that our proposed system brown nuclei estimation was statistically reliable. On the other hand, our three statistical significance tests showed fairly high reliability between pathologists for blue nuclei estimation. The same conclusion applies for our proposed automated blue nuclei system.

References

[1] Lord, S.J., Lei, W., Craft, P., *et al.* (2007) A Systematic Review of the Effectiveness of Magnetic Resonance Imaging (MRI) as an Addition to Mammography and Ultrasound in Screening Young Women at High Risk of Breast Cancer. *European Journal of Cancer*, **43**, 1905-1917. http://dx.doi.org/10.1016/j.ejca.2007.06.007

[2] Rakha, E.A., Reis-Filho, J.S., Baehner, F., Dabbs, D.J., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Jacquemier, J., Lakhani, S.R., Palacios, J., Richardson, A.L., Schnitt, S.J., Schmitt, F.C., Tan, P.H., Tse, G.M., Badve, S. and Ellis, I.O. (2010) Breast Cancer Prognostic Classification in the Molecular Era: The Role of Histological Grade. *Breast Cancer Research*, **12**, 207.

[3] Alomari, R.S., Allen, R., Sabata, B. and Chaudhary, V. (2009) Localization of Tissues in High-Resolution Digital Anatomic Pathology Images. *Proceedings of SPIE*, *Medical Imaging*: *Computer-Aided Diagnosis*, **7260**, Article ID: 726016.

[4] Beresford, M.J., Wilson, G.D. and Makris, A. (2006) Measuring Proliferation in Breast Cancer: Practicalities and Applications. *Breast Cancer Research*, **8**, 216. http://dx.doi.org/10.1186/bcr1618

[5] Urruticoechea, S.A., Lan, E. and Dowsett, M. (2005) Proliferation Marker Ki-67 in Early Breast Cancer. *Journal of Clinical Oncology*, **23**, 7212-7220. http://dx.doi.org/10.1200/JCO.2005.07.501

[6] Alomari, R., Ghosh, S., Chaudhary, V. and Al-Kadi, O. (2012) Local Binary Patterns for Stromal Area Removal in

Histology Images. *Proceedings of the SPIE*, *Medical Imaging*: *Computer Aided Diagnosis*, **8315**, Article ID: 831524.

[7] Al-Lahham, H.Z., Alomari, R.S., Hiary, H. and Chaudhary, V. (2012) Automation Proliferation Rate Estimation from Breast Cancer Ki-67 Histology Images. *Proceedings of the SPIE*, *Medical Imaging*: *Computer-Aided Diagnosis*, **8315**, 83152A.

[8] Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M. and Yener, B. (2009) Histopathological Image Analysis: A Review. *IEEE Reviews in Biomedical Engineering*, **2**, 147-171.

[9] Cheng, H.D., Shan, J., Ju, W., Guo, Y. and Zhang, L. (2010) Automated Breast Cancer Detection and Classification Using Ultrasound Images: A Survey. *Pattern Recognition*, **43**, 299-317. http://dx.doi.org/10.1016/j.patcog.2009.05.012

[10] Phukpattaranont, P., Limsiroratana, S. and Boonyaphiphat, P. (2009) Computer-Aided System for Microscopic Images: Application to Breast Cancer Nuclei Counting. *International Journal of Applied Biomedical Engineering*, **2**, 69-74.

[11] Shao, J. and Wang, H.S. (2002) Sample Correlation Coefficients Based on Survey Data Under Regression Imputation. *Journal of the American Statistical Association*, **79**, 544-552.

[12] Cann, J., Ellin, J., Kawano, Y., Knight, B., Long, R.E., Sam, A., Machotka, V. and Smith, A. (2013) Validation of Digital Pathology Systems in the Regulated Nonclinical Environment. Digital Pathology Association, Madison.

[13] Watkins, M.W. and Pacheco, M. (2001) Interobserver Agreement in Behavioral Research: Importance and Calculation. *Journal of Behavioral Education*, **10**, 205-212. http://dx.doi.org/10.1023/A:1012295615144

[14] Yelton, A.R., Wildman, B.G. and Erickson, M.M.T. (1977) A Probability-Based Formula for Calculating Interobserver Agreement. *Journal of applied behavior Analysis*, **10**, 123-131. http://dx.doi.org/10.1901/jaba.1977.10-127