ORIGINAL ARTICLE

The Effect on Absence of Clinical History and Demographic Data in Diagnostic Accuracy of Genitourinary Cytopathological Cases Among Undergraduate Medical Laboratory Technology (MLT) Students

Siti Norbaya Mohamad, Mohd Nazri Abu, Najwa Nadeera Roslan, Nur Nadirah Abd Malek, Nur Adlina Alihad

Centre for Medical Laboratory Technology Studies, Faculty of Health Sciences, Universiti Teknologi MARA (UiTM) Selangor, Puncak Alam Campus, 42300 Selangor.

ABSTRACT

Introduction: Genitourinary cytology is a cytomorphological study of benign and malignant urinary cells under microscopic observation. Slide observers were presented with glass slides devoid of demographic information and clinical history in this research. The aims are to evaluate inter- and intra-observer reliability and diagnostic accuracy in genitourinary cytopathology patients without relying on the clinical history and demographic information. **Methods:** A correlational investigation was conducted at the cytology laboratory, Centre for Medical Laboratory Technology (MLT) Studies, Faculty of Health Sciences, Universiti Teknologi MARA (UiTM), Puncak Alam Selangor. Five undergraduate students were recruited as slide observers to screen 26 genitourinary cases using a light microscope. The Fleiss' and Cohen's kappa values were used to assess inter- and intra-observer reliability, respectively, and the receiver operating characteristic (ROC) curve was employed to assess diagnostic accuracy in the absence of clinical history. All collected data were analysed using SPSS software. **Results:** Inter- and intra-observer reliability were interpreted as 'fair agreement' with an average sensitivity of 100%, specificity of 13.16%, Positive Predictive Value (PPV) of 70.4 percent, Negative Predictive Value (NPV) of 100%, and a Likelihood Ratio (LR) of 2.454. The diagnosis accuracy of these genitourinary cases is 70.5%. **Conclusion:** Undergraduate students in MLT, UiTM are sufficiently competent to identify and diagnose genitourinary cytology slides based on cell's morphological characteristics without the assistance of demographic data or patient history.
Malaysian Journal of Medicine and Health Sciences (2022) 18(SUPP15): 264-268. doi:10.47836/mjmhs18.s15.37

Keywords: Clinical history, Diagnostic accuracy, Reliability test, Inter and intra-reliability, Genitourinary

Corresponding Author:
Mohd Nazri Abu, PhD
Email: nazri669@uitm.edu.my
Tel: +603-3258 4433

INTRODUCTION

In the cytology laboratory, all specimens are processed, assessed, diagnosed, interpreted by cytologists or slides observers, and subsequently re-examined and verified by pathologists for final diagnosis. Each sample must accompany by a request form containing all the necessary demographic information including clinical history for reference purposes. An incomplete or incorrect laboratory request form might negatively affect test findings, patient care, and safety. Cytology laboratory play a crucial role in delivering a better prognosis and accurate diagnosis for patients. Louise Nutt et al. (2008) proved that laboratory data affected 70% of medical diagnoses. A present of clinical history aids in diagnostic accuracy improvement (1, 14). According to Raab et al. (2000), he stated that the absence of clinical history contributed to diagnostic inaccuracy since pathologists reported a reactive result rather than malignancy. This constraint will affect the inter- and intra-observer reliability in genitourinary case diagnosis. The kappa value was used to conduct reliability tests, which included inter-observer (agreement between two observers & agreement between separate observers) and intra-observer (agreement between one observer experiences when observing the same case or more than once) (2, 3, 15). It might be difficult for the observer to make an accurate diagnosis in the absence of clinical history, thus reduce the reliability of slide observers to report cases (4, 16). When clinical findings are missing from a slide, the observer will doubt a mistake (5, 16). Unavoidable variations due to inter- and intra-observer variability may be considered an underlying feature

of the discrepancy in reporting system. Failure to correctly assess the variance seen in a research, resulted in incorrect interpretation, which was influenced by the performance quality and patient management of the reporting laboratory (5). As a result, it is critical to evaluate the slide observer's consistency in making the same measurements under the same diagnosis even with absence of clinical history. Therefore the objectives of this study are to evaluate the new screener's performance using blinded genitourinary screening cases, exhibiting dependability and diagnostic accuracy among observers based on their knowledge, skill and abilities without relying on the clinical history and demographic information.

MATERTIALS AND METHODS

Correlational research was performed in this study to determine the relationship between two or more variables using statistical analysis. The association to be evaluated was the lack of clinical history in interpreting genitourinary cases between new screeners, as determined by the reliability test agreement and overall diagnostic accuracy. Twenty-six slides of genitourinary patient's cases were selected from the total number of slides accessible in the cytology laboratory of the Centre for Medical Laboratory Technology Studies, Faculty of Health Sciences, Universiti Teknologi MARA, Malaysia.

Raosoft software was used to determine the sample size. Five new screeners in this trial blindly assessed all 26 cases without considering the clinical history. All screeners were participated voluntarily and chosen based on the inclusion criteria. The study was approved by the Research Ethics Committee, Institute of Research Management & Innovation (RMI), UiTM (reference no: 600-RMI (5/1/16)). All participants who recognised genitourinary cell morphology had at least one year of experience screening cytological slides, operating light microscopes, and performing routine maintenance. The majority of participants were fourth- and fifth-year MLT students who met all inclusion criteria.

The screening session was divided into two parts: the first part (A) and the second part (B), during which the same cases were monitored at different intervals. A month passed between the first and second screening sessions. The first screening session was conducted to establish the agreement between new screeners in preparation for an inter-observer reliability test. In contrast, the second screening session was conducted to obtain results from the same new screeners at different times in preparation for an intra-observer reliability test. Each instance evaluated resulted in a diagnosis by the participants. The diagnostic response form was distributed, and participants were only permitted to choose a final diagnosis following screening and use the standard reporting system.

The inter- and intra-observer reliability agreement was calculated using the kappa coefficient value. The Kappa number reflects the degree of agreement between observers, whether positive or negative. For the remaining parameters, data were tabulated into a contingency table to determine the true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) rates for the new screeners. The sensitivity (%), specificity (%), positive predictive value (PPV), negative predictive value (NPV), and likelihood ratio were used to quantify the data (7, 8). The receiver operating curve (ROC) was plotted using the sensitivity and specificity values. Thus, the diagnostic accuracy is determined by the area under the receiver operating characteristic (ROC) curve. SPSS software was used to analyse the data collected (Statistical Package for Social Science for Windows version 25.0, IBM Corp, Armonk, New York, USA).

RESULTS

Inter-observer reliability test results as measured by Fleiss' Kappa in the first screening session (A) and second screening session (B) can be observed in Table I. The first screening session shows fair agreement with a 0.295 kappa value. The p-value is <0.005, which indicates a statistically significant result. The marginal distribution for the first screening session (A) is 0.186 and 0.403. The second screening session (B) shows fair agreement with a 0.347 kappa value. The marginal distribution for the second screening session (B) is 0.244 and 0.450. The majority of the slide observers classify malignant cases with the 'poor agreement' (0.097), followed by atypical cases with the 'fair agreement' (0.223) and the 'fair agreement' for benign cases (0.391). While, in the second screening session (B), most of the slide observers classified malignant cases with the 'moderate agreement' (0.581), followed by benign and atypical cases with the 'fair agreement,' 0.375 and 0.239, respectively.

The first and second screening sessions of intra-observer reliability were analysed and represented by Cohen's kappa value (Table II). All the slide observers had the 'fair agreement' after comparing the diagnosis between the first and second screening sessions except for slide observers 1 (SO1) and SO3, which showed 'no agreement' with -0.430 and -0.156.

Operating parameters to evaluate overall diagnostic accuracy for each slide observer without providing clinical history were tabulated in Table III. All the slide observers have correctly classified the malignant cases. One malignant case and 11 benign cases have been correctly identified. The other 14 remaining cases had been misclassified as false positives. Slide observers succeeded in classifying malignant cell seen cases and recorded no false-negative cases. Meanwhile, SO4 and SO5 recorded the highest classification of true negative

Table I: Inter-Observer Reliability as Measured by Fleiss' Kappa.

Rating Category	Kappa value	95% Confidence Interval (CI)		Kappa value	95% Confidence Interval (CI)	
		Lower	Upper		Lower	Upper
Screening Session	First Screening Session (A)			Second Screening Session (B)		
Overall results	**0.295**	**0.186**	**0.403**	**0.347**	**0.244**	**0.450**
Benign	0.391	0.269	0.512	0.375	0.251	0.499
Atypical	0.223	0.102	0.345	0.239	0.115	0.363
Malignant	0.097	0.024	0.219	0.581	0.457	0.705

The first (A) and second (B) screening sessions of Fleiss' Kappa value is interpreted as follows: κ value ≤ 0: 'No agreement'; 0.01 – 0.20: 'None to slight agreement'; 0.21 – 0.40: 'Fair agreement'; 0.41 – 0.60: 'Moderate agreement; 0.61 – 0.80: 'Substantial agreement' and 0.81 – 1.00: 'Almost perfect agreement' .

Table II: Intra-observer Reliability as Measured by Cohen's Kappa.

Slide observer	Cohen's Kappa value	Strength of agreement
SO1A, SO1B	-0.430	No agreement
SO2A, SO2B	0.314	Fair agreement
SO3A, SO3B	-0.156	No agreement
SO4A, SO4B	0.217	Fair agreement
SO5A, SO5B	0.262	Fair agreement

SO are the slide observers. The Cohen's Kappa outcome is interpreted as follows, 0.21–0.40 as 'fair agreement,' 0.41–0.60 as 'moderate agreement,' 0.61–0.80 as 'significant agreement,' and 0.81–0.00 as practically 'perfect agreement' values — 0.1–0.20 as 'small agreement' and 0.01–0.20 as 'minor agreement.'

cases, with 21 and 24 out of 26 true negative cases being correct.

Diagnostic accuracy for each slide observer showed the same sensitivity value, which is 100% was recorded for all slide observers. The highest specificity recorded was 20.0% by SO4, while the most minor recorded was 9.1% for SO2. The highest positive predictive value was recorded for SO4 with 80.0%, while the least with 60.0% for SO2. Next, all Negative Predictive Values (NPV) recorded the highest value of up to 100% for all slide observers. The likelihood ratio (LR) value was highest in SO4 with 3.473, while the least recorded by SO2 with 1.775. Therefore, the average value of each operating parameter showed sensitivity is 100%, specificity is 13.16%, Positive Predictive Value (PPV) is 70.4%, Negative Predictive Value (NPV) is 100%, LR is 2.454, and the diagnostic accuracy is 71.54%.

DISCUSSION

Inter-observer Agreement (Fleiss' Kappa)

Fleiss' Kappa was performed to determine the level of agreement between the five slide observers in diagnosing 26 genitourinary slide cases with no clinical history in screening session (A) (Table I). The variables analysed were on malignant, atypical, and benign scales. The value range was read as -1 to 1. A score of -1 indicates complete disagreement, and a value of 1 indicates complete agreement. (9, 10). Inter-observer reliability is used to assess how to separate slide observers make similar diagnoses in identical genitourinary cases without demographic information and clinical history. McHugh (2012) defined k value 0.20 as 'Poor' agreement, 0.21-0.40 as 'Fair' agreement, 0.41-0.60 as 'Moderate' agreement, 0.61-0.80 as 'Good' agreement, and 0.81-1.00 as 'Very good' agreement. (11, 17). Individual kappa coefficients were calculated for each variable (malignant, atypical, and benign). According to the data in Table I, the overall kappa value was 0.295 for the first screening session (A) and 0.347 for the second screening session (B). The 95% confidence interval (CI) suggested that the Fleiss kappa value was true or valid between 0.186 and 0.403 of the marginal distribution for the first screening session (A) and between 0.244 and 0.450 for the second screening session (B). Due to a statistical analysis constraint, the average value for both screening sessions could not be determined due to marginal distribution discrepancies between the first and second sessions. Despite the absence of demographic or clinical information during each screening session,

Table III: Operating Parameters to Evaluate Overall Diagnostic Accuracy for Each Slide Observers Without Providing Clinical History

SO	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Likelihood ratio (LR)	Diagnostic accuracy (%)	Area under curve (AUC)
SO1	100	16.7	80	100	3.070	80.8	0.900
SO2	100	9.1	60	100	1.775	61.5	0.800
SO3	100	10.0	64	100	1.976	65.4	0.820
SO4	100	20.0	84	100	3.473	84.6	0.710
SO5	100	10.0	64	100	1.976	65.4	0.820
Average	**100**	**13.2**	**71**	**100**	**2.454**	**71.5**	**0.810**

SO: Slide Observer; PPV: Positive Predictive Value; NPV: Negative Predictive Value; LR: Likelihood Ratio.

most diagnoses made were in reasonable agreement. On the other hand, the slide observers had difficulties diagnosing benign patients due to their incorrect classification as unusual, suspicious, or malignant. After all, slide observers correctly identified most malignant and benign cases.

Intra Observer Agreement (Cohen Kappa)

For intra-observer reliability tests, the data collected during the first (A) and second (B) screening sessions were analysed and shown as Cohen's kappa value (Table II). Intra-observer reliability was assessed to determine whether there was an agreement between various slide observers selected based on inclusion and exclusion criteria. The value acquired by each slide observer during the first and second screening sessions was calculated using Cohen's Kappa. The same genitourinary cases were detected twice within a month of a gap by the same slide observers.

The consistency of an observer's diagnosis was examined at least twice using the same genitourinary instances to verify the observers' reliability. According to Mchugh (2012), the Kappa result is as follows: 0 indicates no agreement, 0.01–0.20 indicates little to no agreement, 0.21–0.40 indicates reasonable agreement, 0.41–0.60 indicates moderate agreement, 0.61–0.80 indicates substantial agreement, and 0.81–1.00 indicates nearly perfect agreement. According to McHugh, 2012, Kappa values of 0.21–0.40 indicate fair agreement, 0.41–0.60 indicate moderate agreement, 0.61–0.80 indicate significant agreement and 0.81–0.00 indicate almost perfect agreement. 0.1–0.20 as no significant agreement and 0.01–0.20 as no significant agreement (11).

Diagnostic Accuracy

The intra-observer reliability was determined to ascertain agreement between multiple slide observers who were chosen based on inclusion and exclusion criteria. Cohen's Kappa was used to compute the value acquired by each slide observer throughout the first and second screening sessions. The same slide observers recognised the identical genitourinary cases twice within a month of a hiatus. The average value for each operating parameter was obtained by summing up each measured value and calculating the average value among the five new screeners (Table III). It showed a significant value of sensitivity (100%), specificity (13.2%), PPV (71%), NPV (100%), LR (2.454) and diagnostic accuracy (71.5%). This result indicated a significant value in diagnosing genitourinary cases without clinical history.

CONCLUSION

In conclusion, this study discovered that slide observers accurately diagnosed 71.5% of genitourinary cases and had only a decent agreement regarding inter and intra-reliability when diagnosing genitourinary cases. Without demographic or clinical information, the slide observers diagnosed almost all instances correctly. This indicated that most slide observers have sufficient knowledge, experience, and skills for genitourinary case screening. The presence of demographic information and clinical history may assist slide observers in making correct genitourinary diagnoses.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to all staff and lecturers at the Centre for Medical Laboratory Technology Studies.

REFERENCES

1. Stephen S. Raab, MD, Thaira Oweity, MD, Jonathan H. Hughes, MD, Diva R. Salomao, MD, Carolyn M. Kelley, MD, Christopher M. Flynn, MD, Joyce A. D'Antonio, PhD, Michael B. Cohen, MD, Effect of Clinical History on Diagnostic Accuracy in the Cytologic Interpretation of Bronchial Brush Specimens, American Journal of Clinical Pathology, Volume 114, Issue 1, July 2000, Pages 78–83, doi:10.1309/4099-QALD-NVGF-TM4G.
2. Zee M. J. M., Sulaihem R. A., Diercks R. L. et al. Intra- and Interobserver Reliability of Determining the Femoral Footprint of the Torn Anterior Cruciate Ligament on MRI Scans. BMC Musculoskelet Disord. 2022; 22, 493. doi:10.1186/s12891-021-04376-5.
3. Adamu S, Mohammed A, El-Bashir J, Abubakar J, Mshelia D. Incomplete patient data on chemical pathology laboratory forms in a Tertiary Hospital in Nigeria. Annals of Tropical Pathology. 2018;9(1):47-9. doi: 10.4103/atp.atp_44_17
4. Plebani M. Diagnostic Errors and Laboratory Medicine - Causes and Strategies. EJIFCC. 2015;26(1):7-14.. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975219/.
5. Liu, K., Layfield, L. J., Coogan, A. C., Ballo, M. S., Bentz, J. S., & Dodge, R. K. Diagnostic Accuracy in Fine-needle Aspiration of Soft Tissue and Bone Lesions: Influence of Clinical History and Experience. American Journal of Clinical Pathology, 1999; 111(5), 632–640. doi: 10.1093/ajcp/111.5.632.
6. Daniel W.W. Biostatistics: A Foundation for Analysis in the Health Sciences. New York: John Wiley & Sons, 1995; 7(4). ISBN:141-142, 1119282373, 9781119282372. Available from: https://www.ege.fcen.uba.ar/wp-content/uploads/2014/05/Daniel-1995-Biostatistics.pdf.
7. Sim, J., & Wright, C.C. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy. 2005; 85(3), 257-268. Available from: https://pubmed.ncbi.nlm.nih.gov/15733050/.
8. Bujang, M.A. & Baharum, N. Guidelines of the Minimum Sample Size Requirements for

Cohen's Kappa. Epidemiology Biostatistics and Public Health. 2017; 14(2), 1-10. Available from: https://riviste.unimi.it/index.php/ebph/article/view/17614.

9. Altman, Douglas G. 1999. Practical Statistics for Medical Research. Chapman; Hall/CRC Press Available from: https://www.researchgate.net/publication/24917811_Practical_Statistics_for_Medical_Research.
10. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. 1 (33). Biometrics. 1977;159–74. Available from:https://dionysus.psych.wisc.edu/iaml/pdfs/landis_1977_kappa.pdf
11. McHugh ML. Interrater Reliability: The Kappa Statistic. Biochem Med (Zagreb). 2012;22(3):276-282. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/
12. Jones BA, Davey DD. Quality Management in Gynecologic Cytology Using Interlaboratory Comparison. Arch Pathol Lab Med. 2000 May;124(5):672-81. doi: 10.5858/2000-124-0672-QMIGCU.
13. Trevethan R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. Front Public Health. 2017 Nov 20;5:307. doi: 10.3389/fpubh.2017.00307.
14. Hawkins, C. M., et al. "Improving the Availability of Clinical History Accompanying Radiographic Examinations in a Large Pediatric Radiology Department." American Journal of Roentgenology (2014). 202(4): 790-796. doi: 10.2214/AJR.13.11273
15. Belur, Jyoti, Lisa Tompson, Amy Thornton, and Miranda Simon. "Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making." Sociological Methods & Research 50, 2 (2021): 837–65. doi:10.1177/0049124118799372.
16. Ngo A, Gandhi P, Miller WG. Frequency that Laboratory Tests Influence Medical Decisions. J Appl Lab Med. 2017 Jan 1;1(4):410-414. doi: 10.1373/jalm.2016.021634.
17. Roslan, N. N., Abu, M. N., Abd Malek, N,. N., Roslan, N. A., Alihad, N. A., Mohamad, S. N., Md. Isa, K.A. "Effect of Absence Clinical History in Diagnostic Accuracy of Thyroid Fine Needle Aspiration Cytology." Mal J Med Health Sci (2021) 17(SUPP3): 162-167 (eISSN 2636-9346). Available from: https://medic.upm.edu.my/upload/dokumen/2021061417205723_2020_1234.pdf