World Journal of Research and Review (WJRR)
ISSN:2455-3956, Volume-4, Issue-4, April 2017 Pages 58-60

Accuracy and Precision in Medical Researches; Common Mistakes and Misinterpretations

Siamak Sabour*, Ommolbanin Abbasnezhad, Samaneh Mozaffarian, Hajar Nazari Kangavari

***Abstract*- Background: When we use a single test in clinical care, for appropriate management of patients and correct diagnosis in clinical care the validity and reliability of that single test is important. The purpose of this study was to evaluate the statistical issues about validity and reliability used in medical papers focusing on common mistakes and misinterpretations.**

METHODS: The articles about validity and reliability published in PubMed in 2012- 2015, were searched using MESH term. 200 most relevant papers with our topic were reviewed for assessing the correctness of methodology and statistical tests used to assess validity and reliability.

Results: our study showed that the clinical researchers make many mistakes in assessing of validity and reliability of a single test. In more than half of the papers the methodology and statistical tests used for evaluating of validity and reliability of a single test were incorrect or incomplete.

Conclusion: In analysis of validity and reliability of a test in published papers, there are many mistakes and clinical researchers need to gain more knowledge about that.

***Index Terms*— Validity, reliability, clinical research, mistake**

I. Introduction

Reliability (precision) and validity (accuracy) are two important methodological issues in all fields of researches. For reliability and validity analysis, appropriate tests should be applied by clinical researchers. Misdiagnosis and mismanagement of the patients in routine clinical care cannot be avoided using inappropriate tests to assess reliability and validity. Whenever a test or other measuring device is used as part of the data collection process, the validity and reliability of that test is important.

Validity refers to the degree in which our test or other measuring device is truly measuring what we intend to measure. In contrast, reliability assesses the extent to which results agree when obtained by different approaches- that is, different observers, study instruments, or procedures- or by the same approach at different point in time. [1]

Reliability (repeatability or reproducibility) and validity (accuracy) is being assessed by different statistical tests. In many papers, the researchers use wrong statistical tests. *The aim of this study was to examine the correctness of methodology and statistical tests used for assessing reliability and validity of a diagnostic test in medicine in 2012-2015.*

Material and Methods

In this study, 200 published articles were assessed in PubMed search engine using MESH term about Validity and Reliability of diagnostic tests. Validity, accuracy, reliability, precision, reproducibility and repeatability were used as key words. These articles were reviewed and their statistical tests were checked and then assessed in term of correct, incorrect and incomplete methodology and statistical tests.

For reliability (precision), depending on quantitative or qualitative type of the variable, Intra Class Correlation Coefficient (ICC), Bland Altman or even coefficient of variance (CV) and weighted kappa should be used. These tests utilize as correct tests for assessing reliability.

The use of simple kappa (kappa Cohen) test for evaluating of reliability is not completely correct because it not only depends upon the prevalence in each category but also depends upon the number of categories of the variable.

Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), likelihood ratio positive and likelihood ratio negative as well as diagnostic accuracy and odds ratio are among the tests to evaluate the validity (accuracy) of a test. These tests for qualitative and binomial variables are considered as correct tests. Moreover, for quantitative variables depending on its distribution, Pearson and spearman can be used to evaluate validity. Most of papers use only two or three tests mentioned above for evaluation of validity. So, they considered as incomplete evaluations.

RESULTS AND DISCUSSION

Of the most relevant published studies in field of medicine from 2012 to 2015, 200 papers were reviewed. Among those, surprisingly only in 33(40. 7%) articles correct methodology and statistical tests, in 14(17. 2%) articles incorrect tests and in 34(41.9%) articles incomplete tests were used for assessing of validity. For evaluating of reliability 119 papers were

Siamak Sabour, MD, PhD, Department of Epidemiology, School of Public Health, Shaheed Beheshti University of Medical Sciences, Tehran, IRAN

Ommolbanin Abbasnezhad, MSc of Epidemiology, Department of Epidemiology, School of Public Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Samaneh Mozaffarian, MSc of Epidemiology, Department of Epidemiology, School of Public Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Hajar Nazari, MSc of Epidemiology, Department of Epidemiology, School of Public Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran

investigated. Among these the researchers used correct tests in 65(54. 6%) papers, incorrect tests in 45(37%) papers and incomplete tests in 9 (7.5%) papers. The main findings were shown in Fig 1.

Whenever a test or other measuring device is used as part of the data collection process, the validity and reliability of that test is important. Just as we would not use a math test to assess verbal skills, we would not want to use a measuring device for research that was not truly measuring what we proposed it to measure. After all, we are relying on the results to show support or a lack of support for our theory and if the data collection methods are erroneous, the data we analyze will also be erroneous.[1]

Reliability (repeatability or reproducibility) is being assessed by different statistical tests such as Pearson r, least square and paired t.test which all of them are among common mistakes in reliability analysis [1] and is being published by high impact journals. [2-8].

Briefly, for quantitative variable Intra Class Correlation Coefficient (ICC), Bland Altman or even coefficient of variance (CV) and for qualitative variables weighted kappa should be used with caution because kappa has its own limitation too. [9-39].

It is crucial to know that there is no value of kappa that can be regarded universally as indication good agreement. Two important weaknesses of simple kappa value to assess agreement of a qualitative variable are as follow: It depends upon the prevalence in each category which means *it can be possible to have different kappa value having the same percentage for both concordant and discordant cells!* Fig 2 shows that in both (a) and (b) situations the prevalence of concordant cells are 80% and discordant cells are 20%, however, we get different kappa value (0.38 and 0.60) respectively. Kappa value also depends upon the number of categories [9-39].

Sensitivity (Percent with the disease who test positive, True Positives / (True Positives + False Negative)), specificity (Percent healthy who test negative, True Negatives / (True Negatives + False Positive)) positive predictive value (PPV), (Percent of positive tests who actually are diseased, True Positives / (True Positives + False Positive)), negative predictive value (NPV) (Percent of negative tests who are healthy, True Negatives / (True Negatives + False Negative)), likelihood ratio positive and likelihood ratio negative as well as diagnostic accuracy [(both true positive and true negative results / total)* 100] and odds ratio (true results / false results) preferably more than 50, are among the tests to evaluate the validity of a single test compared to a gold standard. [9-39].

As a take home message, the present methodological and statistical situation of publications in the field of medicine in purpose of reliability and validity analysis, is not acceptable. Therefore, misdiagnosis and mismanagement of the patients in medicine cannot be avoided.

CONCLUSION

As a take home message, the present methodological and statistical situation of publications in the field of medicine in purpose of reliability and validity analysis, is not acceptable. Therefore, misdiagnosis and mismanagement of the patients in clinical care cannot be avoided. So, the clinical researchers must enhance their knowledge about these topics and improve the situation of analysis of validity and reliability in their researches and help the authorities and clinical practitioners to make appropriate decisions in management of the patients.

REFERENCES

[1] I. Lawrence , K. Lin, A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989:255-68.
[2] V. Jain, J. Duda, B. Avants, M. Giannetta, S.X. Xie, T. Roberts, et al, Longitudinal reproducibility and accuracy of pseudo-continuous arterial spin–labeled perfusion MR imaging in typically developing children. Radiology. 2012;263(2):527-36.
[3] S. AlBarakati, K. Kula, A, Ghoneima. The reliability and reproducibility of cephalometric measurements: a comparison of conventional and digital methods. Dentomaxillofacial Radiology. 2014.
[4] V. Lai, W.K. Tsang, W.C. Chan, T.W. Yeung, Diagnostic accuracy of mediastinal width measurement on posteroanterior and anteroposterior chest radiographs in the depiction of acute nontraumatic thoracic aortic dissection. Emergency radiology. 2012;19(4):309-15.
[5] L.F. Ling, N.A. Obuchowski, L. Rodriguez, Z. Popovic, D. Kwon, T.H. Marwick, Accuracy and interobserver concordance of echocardiographic assessment of right ventricular size and systolic function: a quality control exercise. Journal of the American Society of Echocardiography. 2012;25(7):709-13.
[6] T. Sato, I. Tsujino, H. Ohira, N. Oyama-Manabe, A. Yamada, Y.M. Ito, et al, Validation study on the accuracy of echocardiographic measurements of right ventricular systolic function in pulmonary hypertension. Journal of the American Society of Echocardiography. 2012;25(3):280-6.
[7] G. Maislin, M.M. Ahmed, N. Gooneratne, M. Thorne-Fitzgerald, C. Kim, K. Teff, et al, Single slice vs. volumetric MR assessment of visceral adipose tissue: reliability and validity among the overweight and obese. Obesity. 2012;20(10):2124-32./www.wiley.com/
[8] V. Soviero, S. Leal, R. Silva, R. Azevedo, Validity of MicroCT for in vitro detection of proximal carious lesions in primary molars. Journal of dentistry. 2012;40(1):35-40./www.researchgate.net/
[9] K.J. Rothman, S. Greenland, T.L. Lash, Modern Epidemiology, 4th edition. Lippincott Williams & Wilkins, Baltimore, United States. 2010.
[10] M. Szklo, F.J. Nieto, Epidemiology beyond the basics, 2 nd edition, . Manhattan, new York, United State: Jones and Bartlett Publisher; 2007.
[11] S. Sabour, Interlaboratory and interstudy reproducibility of a novel lateral-flow device: a statistical issue. Journal of clinical microbiology. 2013;51(5):1652-.
[12] S. Sabour, F. Ghassemi, Interrater reliability of intensive care unit electroencephalogram revised terminology: pitfalls and challenges of using kappa value. Journal of Clinical Neurophysiology. 2013;30(2):210.
[13] S. Sabour, A quantitative assessment of the accuracy and reliability of O-arm images for deep brain stimulation surgery. Neurosurgery. 2013;72(4):E696.
[14] S. Sabour, Single slice vs. volumetric MR assessment of visceral adipose tissue: reliability and validity among the overweight and obese. Obesity. 2013;21(1):6-7.
[15] S. Sabour, E.V. Dastjerdi, Validity of a visual scoring method for masticatory ability using test gummy jelly: methodological mistake. Gerodontology. 2013;30(1):85-.
[16] S. Sabour, E.V. Dastjerdi, Reliability of assessment of nasal flow rate for nostril selection during nasotracheal intubation: common mistakes in reliability analysis. Journal of clinical anesthesia. 2013;2(25):162.
[17] S. Sabour, F. Ghassemi, Ocular duction studies: statistical issues. Ophthalmology. 2013;1(120):222-3.
[18] S. Sabour, Recovery, dependence or death after discharge. Journal of general internal medicine. 2013;28(3):342-.
[19] S. Sabour, Reliability and repeatability of toe pressures measured with laser Doppler and portable and stationary photoplethysmography devices. Annals of vascular surgery. 2012;26(8):1167.
[20] S. Sabour, Re: Validity of chronic obstructive pulmonary disease diagnoses in a large administrative database: The rule of thumb in validity

analysis of a test. Canadian respiratory journal : journal of the Canadian Thoracic Society. 2012;19(5):331.
[21] S. Sabour, F. Ghassemi, The reproducibility of measurements of differential renal function in paediatric 99mTc-MAG3 renography: is this correct? Nuclear medicine communications. 2012;33(12):1311.
[22] S. Sabour, F. Ghassemi, Accuracy, Validity, and Reliability of the Infrared Optical Head Tracker (IOHT) Letters. Investigative ophthalmology & visual science. 2012;53(8):4776-.
[23] S. Sabour, F. Ghassemi, Validity of self-reported eye disease and treatment: is that correct?! British Journal of Ophthalmology. 2012:bjophthalmol-2012-302205.
[24] S. Sabour, F. Ghassemi, Reliability and validity of conjunctival ultraviolet autofluorescence measurement. The British journal of ophthalmology. 2012;96(9):1271.

[25] S. Sabour, Reliability of SleepStrip as a screening test in obstructive sleep apnea patients: methodological issues to avoid misinterpretation. Eur Arch Otorhinolaryngol. 2015 May;272(5):1299-300.
[26] S. Sabour, Validity and reliability of the robotic objective structured assessment of technical skills. Obstet Gynecol. 2014 Oct;124(4):839.
[27] S. Sabour, Methodologic concerns in reliability of noncalcified coronary artery plaque burden quantification. AJR Am J Roentgenol. 2014 Sep;203(3):W343.
[28] S. Sabour, Reliability of categorical loudness scaling in the electrical domain: common mistakes. Int J Audiol. 2014 Nov;53(11):836-7.
[29] S. Sabour, The reliability of routine clinical post-processing software in assessing potential diffusion-weighted MRI "biomarkers" in brain metastases, common mistake. Magn Reson Imaging. 2014 Nov;32(9):1162.
[30] S. Sabour, Reliability and benefits of medical student peers in rating complex clinical skills; common mistake.Med Teach. 2014 Nov;36(11):1007-8.
[31] S. Sabour, F. Ghassemi, Reliability of on-call radiology residents' interpretation of 64-slice CT pulmonary angiography for the detection of pulmonary embolism: methodological error. Acta Radiol. 2014 May;55(4):427.
[32] S. Sabour, Interrater reliability of assessing levator ani deficiency with 360° 3D endovaginal ultrasound: mistake and misinterpretation in reliability analysis. Int Urogynecol J. 2014 May;25(5):707.
[33] S. Sabour, Intraobserver and interobserver agreement in visual inspection for xanthochromia: implications for subarachnoid hemorrhage diagnosis, computed tomography validation studies, and the Walton rule: methodological mistake. Neurosurgery. 2014 Jun;74(6):E702.
[34] S. Sabour, Bedside ultrasonography as a diagnostic tool for the fifth metatarsal fractures: methodological concern in reliability analysis. Am J Emerg Med. 2014 May;32(5):470.
[35] S. Sabour, Reliability and accuracy of skeletal muscle imaging in limb-girdle muscular dystrophies. Neurology. 2013 Jun 11;80(24):2275.
[36] S. Sabour, Interlaboratory and interstudy reproducibility of a novel lateral-flow device: a statistical issue. J Clin Microbiol. 2013 May;51(5):1652.
[37] S. Sabour, F. Ghassemi, Interrater reliability of intensive care unit electroencephalogram revised terminology: pitfalls and challenges of using kappa value. J Clin Neurophysiol. 2013 Apr;30(2):210.
[38] S. Sabour, A quantitative assessment of the accuracy and reliability of O-arm images for deep brain stimulation surgery. Neurosurgery. 2013 Apr;72(4):E696.
[39] S. Sabour, Single slice vs. volumetric MR assessment of visceral adipose tissue: reliability and validity among the overweight and obese. Obesity (Silver Spring). 2013 Jan;21(1):6-7.

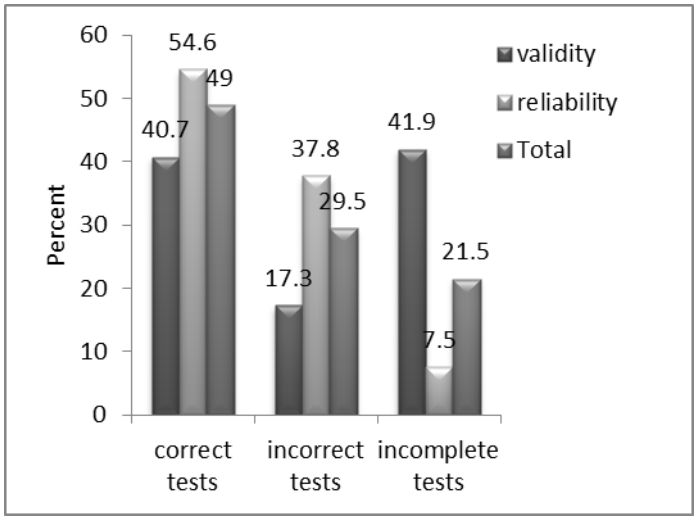

Figure 1. Percentage of incorrect and incomplete tests regarding validity and reliability used in published medical papers from 2012 to 2015

(a)

		Observer 1		
		Positive	Negative	Total
Observer 2	Positive	70	10	80
	Negative	10	10	20
	Total	80	20	100

K=0.38

(b)

		Observer 1		
		Positive	Negative	Total
Observer 2	Positive	40	10	50
	Negative	10	40	50
	Total	50	50	100

K=0.60

Figure 2. Comparison of two observers' diagnosis with different prevalence in the two categories