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 binominal 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
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

REFERENCES


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

(a) Observer 1
Observer 2 Positive 70 10 80
Negative 10 10 20
Total 80 20 100
K=0.38

(b) Observer 1
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