



Verification Bias Correction in Endometrial Abnormalities in Infertile Women Referred to Royan Institute Using Statistical Methods

Fatemeh Niknejad^{1,2}, Firoozeh Ahmadi², Masoud Roudbari^{1*}

Received: 26 Apr 2023

Published: 14 Nov 2023

Abstract

Background: Verification bias is a common bias in the diagnostic accuracy of diagnostic tests and occurs when a number of individuals do not perform the gold standard test. In this study, we review the correcting methods of verification bias.

Methods: In a cross-sectional study in 2020, 567 infertile women who were referred to Royan Research Institute were evaluated. The ultrasound is the performed test and the gold standard are hysteroscopy for some, and pathology for other abnormalities. For correcting verification bias conventional, Begg and Greens, Zhou, and logistic regression methods were used.

Results: In the gold standard hysteroscopy test, the sensitivity (SEN) and specificity (SPEC) obtained in conventional, Begg and Greens, Zhou, and logistics Regression methods were (50%, 90.3%), (48%, 96%), (22%, 77%), (50%, 90%), and (72.8, 77) respectively. Furthermore, the area under the curve (AUC) index and kappa statistics were calculated as 70.2%, and 43.6% respectively. In the pathology gold standard test, the SEN and SPEC for the conventional methods, Begg and Greens, Zhou and logistics regression were (67.7%, 86.7%), (66%, 88%), (29%, 70%), (66.9%, 87.6%), and (73%, 83.9%) respectively. Also, the AUC index and kappa statistics were 77%, and 55% respectively.

Conclusion: In the study on endometrial abnormalities in infertile women, assuming that the missing data mechanism is random, the amount of bias in calculating SEN and SPEC is very low in the diagnostic tests calculated before and after correction, using Begg and Greens and logistic regression method. But Zhou's method gives rather large biased estimates.

Keywords: Verification bias, Ultrasound, Hysteroscopy, Endometrium, Sensitivity, Specificity

Conflicts of Interest: None declared

Funding: None

*This work has been published under CC BY-NC-SA 1.0 license.

Copyright© Iran University of Medical Sciences

Cite this article as: Niknejad F, Ahmadi F, Roudbari M. Verification Bias Correction in Endometrial Abnormalities in Infertile Women Referred to Royan Institute Using Statistical Methods. *Med J Islam Repub Iran*. 2023 (14 Nov);37:122. <https://doi.org/10.47176/mjiri.37.122>

Introduction

Verification bias occurs when only some participants of diagnostic tests perform the gold standard test or when some participants perform one gold standard test and others perform another one (1).

In Diagnostic test studies, accurate and uniform verification of the disease is very important. Using two different gold standard tests leads to different accuracy in verifying

the disease.

Many gold standard tests are invasive and expensive or dangerous (such as angiography, biopsy, and surgery), or in some gold standard, such as ultrasounds, the lesions exist, but it is less important from the clinical point of view than the patients do not undergo the exploratory surgery and the existence or absence of lesions do not confirm. In

Corresponding author: Dr Masoud Roudbari, Roudbari.m@iums.ac.ir

¹ Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran

² Department of Reproductive Imaging, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran

↑What is “already known” in this topic:

In diagnostic experiments, the missing data can affect the values of sensitivity and specificity of the tests. Also, different gold standards have different effects on the value of the above indices. Using different gold standards and correction methods to reduce the verification bias can help the researchers to determine the best bias correction method.

→What this article adds:

The article introduces different methods for correction of the verification bias to find a unique accuracy when the patients do not undergo the gold standard or receive different gold standards. Furthermore, different amount of missing data was considered, and its effect was compared in various correction methods.

this case, the real situation of the patient will be missing. Sometimes in ultrasounds or exploratory surgery gold standard, the lesions are not recognizable, then the real situation of the patients will be unknown and the data in these cases will be missing. Therefore, in many studies, verification bias is inevitable by the researchers.

As a result of the advancement of medical science and technology in diagnostic aids equipment, new methods have been provided to physicians for more accurate diagnosis of the diseases of patients who visit medical centers. Currently, diagnostic tests compare and test new methods with previous common methods using statistical methods. Physician-scientists, to investigate all kinds of diseases, are always looking for diagnostic aids methods that have a faster and more accurate diagnosis and are accompanied by less pain and less intervention (less invasive) for the patient. Moreover, saving money and time for the patient and the physician should also be taken into consideration.

Imaging methods, including types of ultrasounds and radiology, are among these types of diagnostic aid methods, which have significantly helped physicians and patients. Many studies have been conducted on all kinds of diseases and complications using ultrasound. These diagnostic methods, like other methods, depend on the physician's experience and the type and quality of the device.

One of the factors of infertility is the intrauterine environment. Implantation in the uterus during Assisted Reproductive Treatment is influenced by the morphology and thickness of the endometrium (endometrium is an inner epithelial layer along with its mucous membrane. The endometrium is the inner wall of the uterus) and is the uterine cavity. Fibromas, congenital uterine anomalies, endometrial polyps, and uterine synechiae are among the potential causes of infertility. The improper shape of the uterine cavity due to the fibroma or septum can lead to implantation failure and frequent premature abortions (2). Fibroids, in 10% of cases, have a destructive effect on women's fertility (3) and cause an increase in the risk of miscarriage in women who had a natural pregnancy, as well as an increase in miscarriage in half of the pregnancies in IVF (In Vitro Fertilization) cycles (4).

Evaluation of the uterine cavity is a major part of the complete evaluation of an infertile individual. The examination methods may be different, and it is better to be done according to the individual needs of the patient, which include ultrasound or hysteroscopy (5).

The first tool for diagnosing uterine anomalies is ultrasound. Ultrasound is a safe, non-invasive, and almost bloodless method which can currently be used without special equipment in the evaluation of the uterine cavity of infertile women. Ultrasound has been proven to be a method with high reliability in diagnosing endometrial abnormalities. Sonographic examination of the endometrium can show structural abnormalities; however, it is difficult to correctly diagnose the type of lesion and its exact location in the uterine cavity, and some lesions may be overlooked or not detected. Moreover, the conducted studies to examine the diagnostic accuracy of this method have reported different results (2).

Hysteroscopy is a therapeutic-diagnostic method which

enables direct imaging of the uterine cavity and can be used to sample suspicious lesions. Furthermore, as a gold standard method, it is considered to identify intrauterine lesions and congenital uterine anomalies, including arcuate uterus and septum (for adventitious lesions such as a polyp, myoma, synechiae, hyperplasia, retained products of conception, pathology is used as the gold standard Hysteroscopy, as the second step after the ultrasound, is used for screening and distinguishing patients from non-patients. The advantage of hysteroscopy is that it can be used for diagnostic and therapeutic measures at the same time for people (6).

It is worth mentioning that hysteroscopy is an invasive procedure that causes patients to suffer anesthesia. Hysteroscopy is used to examine the cervix and uterine cavity. Hysteroscopy has risks such as uterine perforation, infection, bleeding, and embolism (2).

For example, a patient who does not have any lesions according to the diagnosis of the physician and sonography should be subjected to anesthesia and exploratory surgery to examine the absence of lesions in her, but in dealing with patients, due to the invasiveness of the method, high cost and life risks, the gold standard test is not ethically performed for all people, especially healthy one. This issue seems reasonable in clinical methods, but in the evaluation of diagnostic tests, the impossibility of calculating sensitivity (SEN) and specificity (SPEC) will lead to missing data and then bias in the results (7). Verification bias can also cause researchers to make mistakes in their conclusions. For example, are the results of exploratory surgery more accurate for patients, or is ultrasound sufficient? This can lead to irreparable consequences, especially if diagnostic tests are performed based on incorrect results (8). The purpose of this research is to use verification bias correction in endometrial abnormalities in infertile women referred to Royan Institute, using Begg and Greens, Zhou, and Logistic Regression methods, and to compare these methods.

Methods

Patients' Selection

In this cross-sectional study in 2020, the data of referred patients to Royan Institute were collected. The inclusion criteria were those women who were referred to Royan Institute with the infertility diagnosis. The data was from 576 patients who were checked by a gynecologist with infertility diagnosis. In infertile patients with possible endometrial abnormalities, an ultrasound examination was performed, and patients with lesions were referred for exploratory surgery. It is worth noting that exploratory surgery is performed only to diagnose lesions and diseases, and is not a treatment.

In some lesions, the physician cannot recognize its type just by seeing the lesion, and a portion of the lesion must be removed from the patient's body and send to the pathology laboratory for microscopic examination and the final opinion depends on the laboratory's diagnosis; this operation of removing the lesion is called Biopsy. The patients with adventitious lesions undergo a biopsy test which is determined as the gold standard, and then the calculations of diagnostic tests, including SEN, SPEC, and positive and negative predictive values, are performed. In patients with

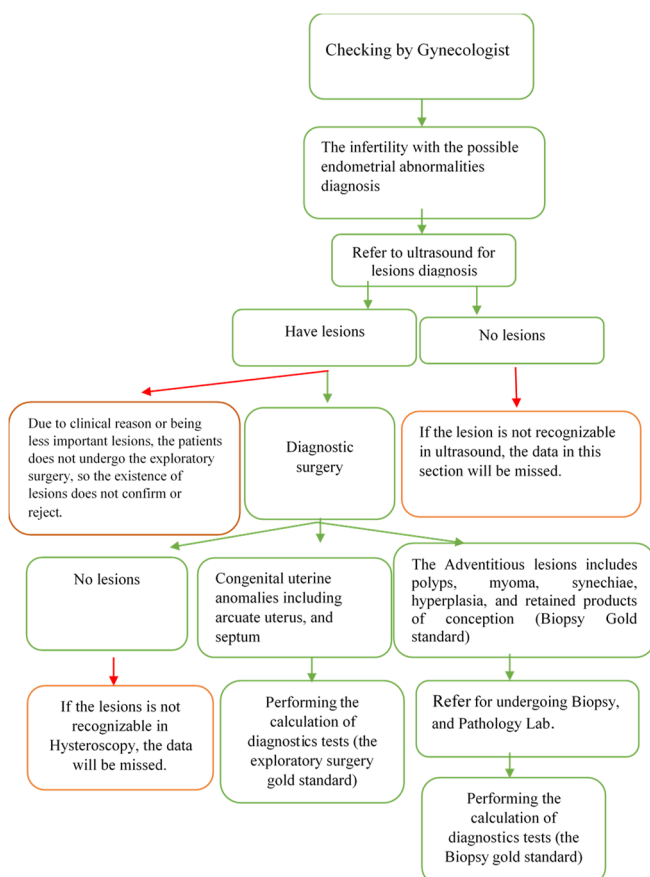


Figure 1. The flowchart of performing the research process

congenital lesions, the above indices were calculated using exploratory surgery as the gold standard. Figure 1 is the flowchart which is introduced for a better understanding of the research process.

Sample size

The data was collected from the patients' files who were referred to Royan Institute. Using the formula for comparison of two diagnosis tests in the same population, we have (9):

$$N_{Disease} = \frac{\{z_{1-\alpha/2} \Lambda + z_{1-\beta} \sqrt{\Lambda^2 - \zeta^2 (3+\Lambda)/4}\}^2}{\Lambda \pi^2}$$

$$N = \frac{N_{Disease}}{\pi_{Disease}}$$

Where $\pi_{Disease}$ is the predicted prevalence of the disease, and also

$$\Lambda = (1 - Sen_1)Sen_2 + (1 - Sen_2)Sen_1 \text{ and } \zeta = (1 - Sen_1)Sen_2 - (1 - Sen_2)Sen_1$$

Where Sen_1 and Sen_2 are the SEN of the first and second diagnosis tests respectively. The prevalence of

Endometrial abnormalities are determined to be at least 13.5%. Also, the SEN of sonography and Hysteroscopy were arranged to be 0.79 and 0.94 respectively. Then using the type one error as 0.05, we have $N=567$.

Accuracy in diagnostic tests

In statistics, SEN and SPEC are two evaluation indices of the result of a two-way test (diagnostic test). When the data can be divided into positive (sick) and negative (healthy) groups, the accuracy of the test results that divide people into these two groups can be measured and described using the SEN and SPEC indices. SEN refers to the proportion of positive cases that the diagnostic test correctly marks them as positive. SPEC refers to the proportion of negative cases that the diagnostic test correctly marks as negative. We also define the following:

True Positive (TP): Frequency of patients in which the patient is correctly diagnosed as sick.

False positive (FP): is the frequency of healthy people in which a healthy person is incorrectly diagnosed as sick.

True Negative (TN): Frequency of healthy people in which a healthy person is correctly diagnosed as healthy.

False Negative (FN): It is the frequency of patients in which the patient is incorrectly diagnosed as healthy.

In other words, SEN is the result of dividing TP cases by the sum of TP and FN cases.

The formulas for calculating SEN and SPEC are as follows:

$$Sensitivity = SEN = TP / (TP + FN) \quad [1]$$

$$Specificity = SPEC = \frac{TN}{TN + FP} \quad [2]$$

Several articles have been published about verification

Table 1. The data contingency table

Observed data	A New Test Result		
	T=1 (positive test)	T=0 (negative test)	
Gold standard test results	D=1 (patient) D=0 (healthy)	s_1 r_1	s_0 r_0
Unconfirmed patients		u_1	u_0
Total		m_1	m_0

bias correction methods, and we will review the most important ones.

Methods of correction of verification bias

Begg and Greens Method: Let's assume the gold standard missing data are random. Then, all patients undergo a diagnostic test, but only some have the gold standard test. This means that not all patients undergo surgery or biopsy.

To correct this bias, as suggested by Begg and Greens, the following maximum likelihood (ML) estimation formulas are used to calculate their SEN and SPEC (10).

$$\hat{sen} = \frac{m_1 s_1 / [N(s_1 + r_1)]}{m_0 s_0 / [N(s_0 + r_0)] + (m_1 s_1 / [N(s_1 + r_1)])} \quad [3]$$

$$\hat{spec} = \frac{\frac{r_0 m_0}{[N(s_0 + r_0)]}}{\frac{m_0 r_0}{[N(s_0 + r_0)]} + \frac{m_1 r_1}{[N(s_1 + r_1)]}} \quad [4]$$

The values in equations (3) and (4) are presented in the Begg and Greens contingency table (Table 1).

Zhou's Maximum Likelihood (ML) Method: The gold standard hypothesis of missing data is not random in this method and occurs when the verification process depends on unobserved data (11). This situation often occurs when there is one of the following:

1. The interval between the initial diagnostic test and the gold standard is long.
2. Different researchers in different centers or laboratories have conducted multiple research projects.
3. The patient population is very heterogeneous.
4. An unknown disease process has been used (10).

In this case, the ML estimation method was suggested by Zhou to calculate the SEN and SPEC. In this method, the log-likelihood function for estimating SEN and SPEC is as follows (11), where t shows the sickness or no sickness of the case in the gold standard test.

$$\sum_{t=0}^1 m_t \log \varphi_{1t} + \sum_{t=0}^1 s_t \log (e_t \lambda_{0t} \varphi_{2t}) + r_t \log [\lambda_{0t} (1 - \varphi_{2t})] + u_t \log [(1 - e_t \lambda_{0t}) \varphi_{2t} + (1 - \lambda_{0t}) (1 - \varphi_{2t})] \quad [5]$$

Where
 $\varphi_{1t} = P(T=t)$ $\varphi_{2t} = P(D=1 | T=t)$ $t=(0,1)$ [6]

Also, we have

$$e_t = \frac{\lambda_{1t}}{\lambda_{0t}} \quad [7]$$

Where:

λ_{10} is the probability of choosing a really sick person with a negative test.

λ_{11} is the probability of choosing a really sick person with a positive test.

λ_{00} is the probability of choosing a really healthy person with a negative test.

λ_{01} is the probability of choosing a really healthy person with a positive test.

e_0 is the probability of choosing a really sick person with a negative test divided by the probability of choosing a really healthy person with a negative test.

e_1 is the probability of choosing a really sick person with a positive test divided by the probability of choosing a really healthy person with a positive test.

Therefore, the ML estimator for SEN and SPEC is as follows (10), where the unknown symbols are presented in Table 1.

$$\hat{sen}(e_0, e_1) = \frac{\frac{s_1 m_1}{s_1 + e_1 r_1}}{s_1 m_1 / (s_1 + e_1 r_1) + (\frac{s_0 m_0}{s_0 + e_0 r_0})} \quad [8]$$

$$\hat{spec}(e_0, e_1) = \frac{e_0 r_0 m_0 / (s_0 + e_0 r_0)}{e_0 r_0 m_0 / (s_0 + e_0 r_0) + e_1 r_1 m_1 / (s_1 + e_1 r_1)} \quad [9]$$

Logistic regression method: Kosinski and Barnhart proposed a likelihood-based regression approach, which can be used according to different types of missing data (12). They assumed that there are p variables for all patients tested in diagnostic tests and its likelihood function was obtained based on the observed data as follows:

$$L_{obs} = \prod_{i=1}^N p(R_i, T_i, D_i | x_i)^{R_i} P(R_i, T_i | x_i)^{1-R_i} = \prod_{i=1}^N p(R_i, T_i, D_i | x_i)^{R_i} \left\{ \sum_{d=0}^1 p(R_i, T_i, D_i = d | x_i) \right\}^{1-R_i} \quad [10]$$

Where i index is for the patients. Also, we have:

R_i is the sickness situation of i^{th} patients according to the gold standard.

T_i is the diagnostic test situation of i^{th} patients.

D_i is the sickness situation of i^{th} patients.

X_i is a binary variable for i^{th} patients.

To write P(R,T, D) as the multiplication of some conditional probability, Baker (12) introduced the following way:

$$p(R, T, D) = P(T)P(D|T)P(R|T, D) \quad [11]$$

Then, according to the Baker method, we have

$$P(R, T, D|x) = P(D|x) \times P(T|D, x)P(R|T, D, x) \quad [12]$$

Then we can calculate SEN and SPEC from $P(T|D, x)$, so the possibility of modeling for $P(R|T, D, x)$, which is the same missing data mechanism, is feasible. Therefore, with the above component, we can write the logistic model for disease, diagnostic tests, and missing data mechanism components (13).

Other calculated indices to examine and compare the above methods are as follows:

Receiver Operating Characteristic (ROC): ROC is one of the methods of analyzing and evaluating the function of binary classification. ROC is considered a binary classification to show the evaluation ability of a system, and its detection threshold is also variable. The axes are calibrated based on SEN and SPEC complement (1-SPEC) (14).

Area under Curve (AUC): This is the optimal area of the curve above the bisector, and we get the best results when the SEN and the complement of SPEC are at their highest and lowest values, respectively. The area under the ROC curve is called the AUC (14).

Cohen's kappa coefficient: In statistical inference, there is a concept called a measurement of agreement, which examines and evaluates the relationship between two quantities. The difference between this concept and other statistical concepts is the separate measurement of these two quantities by two people, phenomena, or two decision-making sources; the difference between the Kappa coefficient and the percentage of simple agreement is in the elimination of random agreements (15).

This research was performed at Iran University of Medical Sciences in 2020 and was registered with the code IR.IUMs.REC.1299.253 in the National System of Ethics in Biomedical Research.

Results

According to the research results, the average age of the subjects in this study is 33.1 ± 5.64 years. The duration of their infertility was also investigated, and the average duration of infertility and its standard deviation is 7.4 years and 5.14, respectively.

In addition, among these subjects, 338 subjects (59.6%) had primary infertility, 172 subjects (30.3%) had secondary infertility, 2 subjects (0.4%) did not have infertility, and the information of 55 subjects (9.7%) was not available.

Medical ultrasound was performed for all 567 patients, and they were evaluated for endometrial abnormalities with hysteroscopy gold standard. From the above patients, 523 patients also underwent hysteroscopy, and 44 patients are part of the missing data; the information of the patients' Contingency tables is presented in Table 2 based on a positive or negative result.

Moreover, all 567 patients were examined for endometrial abnormalities with the gold standard of pathology by ultrasound, of which 380 patients had known data and had been divided into positive or negative results, and 187 patients were part of the missing data. The information of their Contingency tables is presented in Table 3 based on a positive or negative result.

According to the obtained Contingency tables, the SEN and SPEC were calculated by the mentioned methods, and also, the AUC, the Kappa, absolute value Bias (16), and independent chi-square statistics were calculated. The results

Table 2. Contingency table between ultrasound and hysteroscopy results in the gold standard of the hysteroscopy group

Medical ultrasound	Hysteroscopy (gold standard)			
	Negative	Positive	Missing data	Total
Negative	308	91	40	439
Positive	33	91	4	128
Missing data	0	0	0	0
Total	341	182	44	567

Table 3. Contingency table between ultrasound and pathology results in the gold standard pathology group

Medical ultrasound	Pathology (gold standard)			
	Negative	Positive	Missing data	Total
Negative	222	40	139	401
Positive	34	84	48	166
Missing data	0	0	0	0
Total	256	124	187	567

Table 4. Comparing the diagnostic accuracy of ultrasound and hysteroscopy, in the gold standard group of Hysteroscopy, by different methods

Method	SEN (%)	Bias	SPEC (%)	
Standard method	50	0	90.3	
Correction Methods	Begg and Greens' method	48	2	96
	Zhou method	22	28	77
	Logistic regression method of predictive power of ultrasound	50	0	90
	Logistic regression method of predictive power of hysteroscopy	72.8	22.8	77
AUC	0.702 ($P < 0.001$)			
Kappa statistic	0.436 ($P < 0.001$)			
Chi-square statistic	106.667 ($P < 0.001$)			

Table 5. Comparing the diagnostic accuracy of ultrasound and pathology in the gold standard pathology group by different methods

Method	Sensitivity (%)	Bias	Specificity (%)	
Standard method	67.7	0	86.7	
Correction Methods	Begg and Greens' method	66	1.7	88
	Zhou method	29	38.7	70
	Logistic regression method of predictive power of ultrasound	66.9	0.8	87.6
	Logistic regression method of predictive power of hysteroscopy	73	5.3	83.9
AUC	0.772 ($P < 0.001$)			
Kappa statistic	0.551 ($P < 0.001$)			
Chi-square statistic	115.725 ($P < 0.001$)			

are shown in Tables 4 and 5 based on the gold standard type.

Also, using the Homogeneity chi-square test in Table 4, the SEN in different methods are significantly different ($P < 0.001$)

Furthermore, the SEN in different methods is significantly different ($P = 0.000$) using the Homogeneity chi-square test in Table 5.

It is essential to add that the two AUCs are almost the same, and there is no significant difference between them using the Delong test.

Figures 2 and 3 also show the calculated ROC curve for each gold standard. The optimal area of the diagram is above the bisector, and we get the best results when the SEN and the SPEC complement are at their highest and lowest values, respectively.

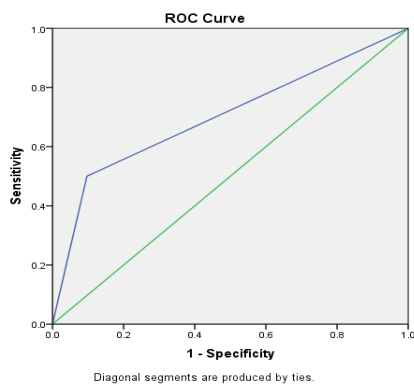


Figure 2. ROC curve of comparing ultrasound and hysteroscopy (gold standard)

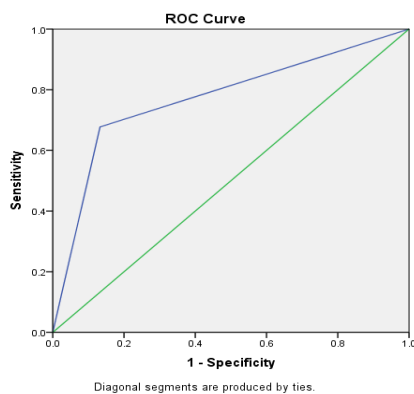


Figure 3. ROC curve comparing ultrasound and pathology (gold standard)

Discussion

It was shown that in the hysteroscopy gold standard group, the SEN and SPEC were 50% and 90.3%, respectively, according to the standard method, 48% and 96% according to Begg and Green's method, 22% and 77% according to the Zhou method, and according to the logistic regression method to analyze the predictive power of ultrasound, it was 50% and 90%. Using the logistic regression method to analyze the predictive power of pathology gold standard, they were calculated to be 72.8% and 77%

In the gold standard pathology group, the SEN and SPEC were 67.7% and 86.7%, respectively, by the standard method, 66% and 88%, by Begg and Green's method, 29% and 70%, by the Zhou method, and by logistic regression method to analyze the predictive power of ultrasound it was 66.9% and 87.6%, and also using the logistic regression method, to analyze the predictive power of pathology, they were calculated to be 73% and 83.9%.

In a similar study on uterine anomalies in 2019, 1141 people were subjected to ultrasound, 180 and 961 of whom had a positive and a negative ultrasound test, respectively. Out of 1141 subjects, 351 cases showed uterine abnormalities in hysteroscopy, of which 205 had negative ultrasounds (17). This shows that ultrasound in the present study has performed better than similar articles. Therefore, to determine patients with uterine abnormalities, ultrasound is a non-invasive method for diagnosis of intrauterine abnormalities. It is a valuable equivalent to hysteroscopy with high accuracy for the determination and characterization of uterine abnormalities. This may lead to a more precise surgery plan and performance.

Zhou also presented a new method based on the ML estimation for SEN and SPEC, which adjusted the effect of this bias and presented some new correction formulas based on it. The distinguishing point of Zhou's theory is in the assumption that missing data are not at random. He used the results of this study to examine the diagnostic accuracy of the liver scan to diagnose liver patients, and out of 650 studied patients, 306 patients were not confirmed due to not performing gold standard. According to the results, the SEN and SPEC estimators are 63% and 90%, respectively, and the ML estimator by Zhou's method is between 68% and 95% for SEN and between 74% and 84% for SPEC (11). To compare the results of ultrasound and hysteroscopy, in the evaluation of anomalies septum, arcuate and uterine synechiae, the calculated SEN and SPEC by the standard method were 50% and 90.3%, and by the Zhou

correction method were calculated at 22% and 77%. Therefore, the obtained results in the presented research are similar to Zhou's method (11), and to determine the liver lesions in patients, the Zhou method and the presented research perform the same, so there are not many differences between the SEN and SPEC indices between the two researches.

Kosinski and Barnhart presented a complementary method to previous studies, assuming missing data are not at random. This method can repeatedly use a logistic regression module and is based on the ML method, it has been tested on the data of 2688 cardiac patients, of which 2217 patients did not perform the gold standard (coronary angiography) test, and only 471 people have a confirmed condition in terms of the disease; that means there were 82.5% missing data. The SEN and SPEC of the standard method using the Kosinski and Barnhart method were 98% and 14%; Then, it was modeled using different assumptions, and the obtained SEN in the case of the highest and the lowest values were 81% and 66%, respectively. Also, the SPEC was between 59% and 65% (12) using the Kosinski and Barnhart method. To compare the results of ultrasound and pathology to evaluate abnormalities of polyps, fibroids, hyperplasia, and retained products of conception, the calculated SEN and SPEC by the standard method are 67.7% and 86.7%, respectively, and by the logistic regression correction method, was 67.7% and 86.7%. Since the data of Kosinski and Barnhart's study has 32% missing data, the results obtained in the above research are similar to the results of the present research. Therefore, to determine the situation of coronary disease patients with a high percentage of missing data, the results are similar to the presented research.

Ünal and Burgut have used a set of available methods, including Begg and Green's method, Zhou's ML method, and the logistic regression method in the correction of verification bias. These methods have been used to calculate the diagnostic accuracy of dual-phase MIBI parathyroid in the diagnosis of primary and secondary hyperparathyroidism, in which out of 69 patients, 48 patients did not have histopathology (gold standard) results. Their studies on real data have shown that verification bias should not be ignored. Otherwise, the diagnostic accuracy of the test will be incorrect, and as a result, the diagnostic accuracy will be underestimated or overestimated. Assuming the randomness of the data, Begg and Green's method has better performance in this study, and Zhou's method may not be suitable for correction due to high bias. The values of SEN and SPEC obtained in this study were 72% and 93% by the standard method, 75% and 90% by Begg and Green's method, at least 56% and 35% by the Zhou method, and at least 70% and 90% by the logistic regression method (10).

According to different methods, the values of SEN and SPEC in the present study are very similar, but these values in the Zhou method are very different from other methods and have the most biased values. At the same time, the results of the present study on endometrial abnormalities are similar to the study of Ünal and Burgut for all described methods, but, in Zhou method in Ünal and Burgut study the

values of SEN and SPEC had some bias similar to the present study. Therefore, both studies had the same result for SEN and SPEC, except for the Zhou method, which is an estimation for both studies and both indices are different from all the above methods. Therefore, to determine the situation of hyperparathyroidism patients, with the gold standard of histopathology, the results are similar to the presented research, so more bias in the Zhou method and fair bias in other methods.

A systematic review research studied 793 articles over the past ten years and examined the diagnostic accuracy of one type of test in diagnosing celiac disease. In these articles, the SEN fluctuated and decreased from 92% to 57%. Moreover, SPEC has increased from 97% to 99%. As a result, if verification bias correction methods are not used, the SEN may be estimated to be much higher than reality in the studies, and this highlights the effect of verification bias on the estimation of diagnostic accuracy. In addition, for systematic review studies, the author has suggested excluding biased studies from the systematic review (18). However, in the current study on endometrial abnormalities, the amount of bias was very low, and there was not much difference in the calculated SEN before and after the verification bias correction, which indicates the good accuracy of ultrasound diagnosis in this study. Therefore, to investigate Celiac disease to determine endometrial abnormalities, the bias was low so SEN before and after correction are the same, which is different from the current research.

Conclusion

Following examination of different methods of verification bias, the best methods were introduced. In the studies with missing data that focus on the SEN and SPEC of a new method for diagnosing diseases, the methods used in the present study help to correct calculations; therefore, researchers who will conduct research in this field in the future are recommended to use available methods to correct the bias and calculate the real accuracy of the tests.

It is suggested that if other variables are also effective in the diagnosis of the disease, the researcher should collect relevant data to increase the accuracy of prediction in future research.

For research projects in which the purpose of the study is to calculate the diagnostic accuracy of the tests in studying alternative diagnostic methods, the use of verification bias correction methods is recommended since they play an important role in reducing the effect of missing data in the final results of the study.

Moreover, in similar articles, there are other methods for the correction of verification bias, which are worth investigating and studying. Among these methods are Artificial Neural Networks and the data simulation method.

Acknowledgment

We are very grateful to Dr. Firozeh Ghaffari, surgeon and obstetrician-gynecologist, faculty member of Royan Institute.

Ethical consideration

This research was performed at Iran University of Medical Sciences in 2020 and was registered with the code IR.IUMs.REC.1299.253 in the National System of Ethics in Biomedical Research.

Authors contribution

Fatemeh Niknejad collection of the data and data analysis. Firoozeh Ahmadi: Conception of the research and checking the final draft. Masoud Roudbari: Data analysis and drafting of the work.

Conflict of Interests

The authors declare that they have no competing interests.

References

1. O'Sullivan JW, Banerjee A, Heneghan C, Pluddemann A. Verification bias. *BMJ Evid Based Med*. 2018, Apr;23(2):54-5.
2. El Huseiny AM, Soliman BS. Hysteroscopic findings in infertile women: a retrospective study. *Middle East Fertil Soc J*. 2013;18(3):154-8.
3. Hart R, Khalaf Y, Yeong CT, Seed P, Taylor A, Braude P. prospective controlled study of the effect of intramural uterine fibroids on the outcome of assisted conception. *Hum Reprod*. 2001;16(11):2411-7.
4. Guven MA, Bese T, Demirkiran F, Idil M, Mgoyi L. Hydrosonegography in screening for intracavitary pathology in infertile women. *Int J Gynecol Obstet*. 2004;86(3):377-383.
5. Hajishiha M, Ghasemi-rad M, Karimpour N, Mladkova N, Boromand F. Transvaginal sonographic evaluation at different menstrual cycle phases in diagnosis of uterine lesions. *Int J Women's Health*. 2011;3:353-7.
6. Balić D, Balić A. Office hysteroscopy, transvaginal ultrasound and endometrial histology: A comparison in infertile patients. *Acta Med*. 2011;40(1):2011:34.
7. Cheharazi M, Shamsipour M, Norouzi M, Jafari F, Ramazan Ali F. A New Method for Correcting Verification Bias in Diagnostic Accuracy Studies Using A Bayesian Approach. *Iran J Epidemiol*. 2012;8(2):20-28
8. Alonzo TA. Verification bias-impact and methods for correction when assessing accuracy of diagnostic tests. *Rev Stat*. 2014;12(1):67-83.
- 9- Motamed N, Zamani F. Sample size in medical research: with a practical approach. Tehran: Asre Roshanbini, 2017.
10. Ünal I, Burgut R. Verification bias on sensitivity and specificity measurements in diagnostic medicine: a comparison of some approaches used for correction. *J Appl Stat*. 2014;41(5):1091-1104.
11. Zhou XH. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Commun Stat Theory Methods*. 1993;22(11):3177-3198.
12. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics*. 2003;59(1):163-71.
13. Baker SG. Evaluating multiple diagnostic tests with partial verification. *Biometrics*. 1995;51(1):330-7.
14. Rosner B. *Fundamentals of Biostatistics*. Boston: Brooks/Cole, Cengage Learning, 2011.
15. McHugh ML. Interrater reliability: The kappa statistic. *Biochem Med (Zagreb)*. 2012;2(3):276-282.
16. Arifin WN, Yusof UK. Partial Verification Bias Correction Using Inverse Probability Bootstrap Sampling for Binary Diagnostic Tests. *Diagnostics (Basel, Switzerland)*. 2022 Nov;12(11):2839.
17. Monteiro CS, Cavalu LK, Dias JA, Pereira FAN, Reis FM. Uterine alterations in women undergoing routine hysteroscopy before in vitro fertilization: high prevalence of unsuspected lesions. *JBRA Assist Reprod*. 2019;23(4):396-401.
18. Hujjoel IA, Jansson-Knodell CL, Hujjoel PP, Hujjoel MLA, Choung RS, Murray JA, et al. Estimating the Impact of Verification Bias on Celiac Disease Testing. *J Clin Gastroenterol*. 2020;55(4):327-334.