Why not consensus reading with multiple readers for evaluating a new (test) device?

Bipasa Biswas
CDRH, FDA, 10903 New Hampshire Avenue, Silver Spring, MD 20993

Abstract
Diagnostic devices, like imaging devices, often require an interface with a reader, who evaluates and interprets to provide a final diagnosis. The diagnosis is a qualitative assessment and thus the reader's responses are yes/no or present/absent. Examples include Colon Capsule Endoscopy (CCE) with cameras that image the colon to detect polyps. This presentation will discuss why a consensus reading of the subject device under evaluation, is not appropriate, when in practice only a single reader is involved in evaluating the images. Since both the subject device and the comparator are evaluated on the same subject, the design is paired and the readers, all read the same images, thus binary data from such a design is correlated. Data from correlated binary data were simulated to assess the impact of consensus readings of the test device output in evaluating the performance measures.

Key Words: Correlated binary data, sensitivity, specificity, bias.

1. Introduction

Diagnostic devices often involve imaging devices that require interpretation by a reader to evaluate the image and often to detect conditions of interest in the image. The focus of this paper is a qualitative assessment by readers of an image where the responses are yes/no or present/absent. In addition, for this paper, each subject provides only one detectable condition (if present) and the current manuscript does not involve multiple conditions per subject/patient. A common example is the Colon Capsule Endoscopy (CCE) where the patient ingests a capsule, which has a camera inside, and the camera takes video images of the colon. The purpose is to detect polyps in the colon. The evaluations and final diagnosis involves a reader or an expert to assess the images and often the assessments are subject to reader variability.

A consensus reading in evaluating findings from an image involves an assessment based on majority of the reader's assessments. In general, three or more readers participate in a consensus reading. For example, if there are three readers then majority evaluation and finding by the three readers is considered as an outcome. (for e.g. if there are 3 readers evaluating the same image and 2 out of the three readers make a positive detection then the consensus assessment will be positive detection and likewise for a negative detection).

Non-invasive CCE with camera is used to detect colon polyps where the detection of polyps is based on reader evaluation of the images of the colon from the camera in the

capsule. The detection/findings are at subject level where at least one polyp >= 6mm or >=3 polyps in the Colon detected is deemed a positive detection for that subject.

Optical Colonoscopy (OC) serves as the clinical reference standard for detecting colon polyps and the CCE with camera is evaluated against OC to determine its accuracy in detecting polyps. Also, in rare instances one CCE with camera is compared against another used as a comparator, which is not a reference standard.

1.1 Reader study framework for imaging devices

In general, diagnostic devices with dichotomous output e.g. presence or absence of the condition of interest, is evaluated against a clinical reference standard used to establish the true condition. The performances are evaluated based on sensitivity and specificity when compared against a clinical reference standard (e.g. optical colonoscopy for colon polyp detection). Thus, if T is the outcome from CCE and R is the outcome from OC; 1 is a positive detection of polyp and 0 no detection of polyps, the performance is evaluated by the following pair:
Sens = P(T=1|R=1)
Spec = P(T=0|R=0)

If the test device is compared and evaluated against an imperfect reference standard, agreement measures-positive percent agreement (PPA) and negative percent agreement (NPA) are reported instead of sensitivity-specificity pair. Thus, if T is the outcome from colon capsules and C is the outcome from comparator; 1 is a positive detection of polyp and 0 no detection of polyps) the performance as PPA and NPA are:
PPA = P(T=1|C=1)
NPA = P(T=0|C=0).

2. Correlated Binary Data

This paper focuses on one finding per subject (where a significant finding per subject is either a polyp >=6 mm or number of Polyps >=3 in the colon). And multiple readers (generally three or more readers) assess the same images. Multiple readers result in correlated data as same readers evaluate all subject images to detect polyps.

The design is based on three readers each reading images from the test device (denoted by T) and the images from the comparator (denoted by C).

2.1 Framework for correlated binary data

Say, T1, T2, T3, C1, C2, C3 are binary random variables with 1= polyp detected; 0=no polyp detected, for observations by the three readers reading from each test and the comparator images. Probabilities are denoted as
p_{Ti}=P(Ti=1); q_{Ti}=P(Ti=0); i=1,2,3
p_{Cj}=P(Cj=1); q_{Cj}=P(Cj=0); j=1,2,3
p_{TiTj}=P(Ti=1, Tj=1); i,j=1,2,3 and i≠j
p_{CiCj}=P(Ci=1, Cj=1); i,j=1,2,3 and i≠j
p_{TiCi}=P(Ti=1, Cj=1); i,j=1,2,3

The correlation and the joint probabilities of two random variables can be written as:

$r_{TiTj}=(p_{TiTj}-p_{Ti}p_{Tj})/sqrt(p_{Ti}q_{Ti}p_{Tj}q_{Tj})$
$p_{TiTj} = p_{Ti}p_{Tj} + r_{TiTj} *sqrt(p_{Ti}q_{Ti}p_{Tj}q_{Tj})$ where i,j=1,2,3 and i≠j;
$r_{TiCj}=(p_{TiCj}-p_{Ti}p_{Cj})/sqrt(p_{Ti}q_{Ti}p_{Cj}q_{Cj})$
$p_{TiCj} = p_{Ti}p_{Cj} + r_{TiCj} *sqrt(p_{Ti}q_{Ti}p_{Cj}q_{Cj})$ where i,j=1,2,3;
$r_{CiCj}=(p_{CiCj}-p_{Ci}p_{Cj})/sqrt(p_{Ci}q_{Ci}p_{Cj}q_{Cj})$
$p_{CiCj} = p_{Ci}p_{Cj} + r_{CiCj} *sqrt(p_{Ci}q_{Ci}p_{Cj}q_{Cj})$ where i,j=1,2,3 and i≠j;

Some necessary conditions hold as folows:
$0 \le p_{Ti}, p_{Ci} \le 1$; i= 1,2,3
$\max (p_{Ti} + p_{Tj} -1,0) \le p_{TiTj} \le \min (p_{Ti} , p_{Tj})$;i,j=1,2,3 and i≠j
$\max (p_{Ci} + p_{Cj} -1,0) \le p_{CiCj} \le \min (p_{Ci} , p_{Cj})$; i,j=1,2,3 and i≠j
$\max (p_{Ti} + p_{Cj} -1,0) \le p_{TiCj} \le \min (p_{TiCj} , p_{TiCj})$; i,j=1,2,3
$p_{T1} +p_{T2} +p_{T3} +p_{C1} +p_{C2} +p_{C3} –(p_{T1T2} +p_{T1T3} + p_{T2T3} +p_{C1C2} +p_{C1C3} + p_{C2C3} + p_{T1C1} +p_{T1C2} + p_{T1C3} +p_{T2C1}+p_{T2C2}+p_{T2C3}+p_{T3C1}+p_{T3C2}+p_{T3C3}) \le 1$

Thus, the six correlated binary variables have the following correlation structure

$$\begin{bmatrix} 1 & r_{T1T2} & r_{T1T3}r_{T1C1} & r_{T1C2} & r_{T1C3} \\ r_{T2T1} & 1 & r_{T2T3}r_{T2C1} & r_{T2C2} & r_{T2C3} \\ r_{T3T1} & r_{T3T2} & 1 \quad r_{T3C1} & r_{T3C2} & r_{T3C3} \\ r_{C1T1} & r_{C1T2} & r_{C1T3} \quad 1 & r_{C1C2} & r_{C1C3} \\ r_{C2T1} & r_{C2T2} & r_{C2T3}r_{C2C1} & 1 & r_{C2C3} \\ r_{C3T1} & r_{C3T2} & r_{C3T3}r_{C3C1} & r_{C3C2} & 1 \end{bmatrix}$$

Now, if the evaluation and comparison is against a clinical reference standard (e.g. OC). Say, the three readers are T1, T2 and T3. The following are some necessary conditions:
$0 \le p_{Ti} \le 1$; i= 1,2,3
$\max (p_{Ti} + p_{Tj} -1,0) \le p_{TiTj} \le \min (p_{Ti} , p_{Tj})$;i,j=1,2,3 and i≠j
$p_{T1} +p_{T2} +p_{T3} –(p_{T1T2} +p_{T1T3} + p_{T2T3}) \le 1$

The correlation matrix is given by

$$\begin{pmatrix} \mathbf{1} & \boldsymbol{r_{T1T2}} & \boldsymbol{r_{T1T3}} \\ \boldsymbol{r_{T2T1}} & \mathbf{1} & \boldsymbol{r_{T2T3}} \\ \boldsymbol{r_{T3T1}} & \boldsymbol{r_{T3T2}} & \mathbf{1} \end{pmatrix}$$

2.2 Simulations

Correlated binary data were simulated using the same correlation structure between pairs of the six-correlated binary random variables. The correlations used were 0.0, 0.25,0.50, and 0.75. The same probability for positive detection of polyps by each reader and the modality was used. The probabilities used are pTi, pCi, = 0.1, 0.2, and 0.3(i=1,2,3). The consensus was based on 2 out of 3 giving the same assessment. A sample size N=750 was used in the simulations.

The tables 1,2 and 3 denote the estimates of positive percent agreement (PPA) and negative percent agreement (NPA) for each individual reader (Readers 1, 2, and 3), the average for the three readers and the PPA and NPA when consensus of the three readers reading the images from the test device for p_{Ti}, p_{Ci}, = 0.1, 0.2, and 0.3 respectively.

Table 1: p_{Ti}=0.1.

Correlation	Performance Measures	Reader 1	Reader 2	Reader 3		Consensus of the three readers with Test Device
0.0	PPA (%)	21.1	21.1	47.4	29.8	0.0
	NPA (%)	83.7	82.5	83.3	83.1	97.0
0.25	PPA(%)	45.7	40.0	62.9	49.5	37.1
	NPA(%)	88.0	88.0	87.3	87.7	94.8
0.50	PPA(%)	62.5	44.6	69.6	58.9	51.7
	NPA(%)	93.2	94.1	92.2	93.2	96.0
0.75	PPA(%)	79.4	69.1	79.4	76.0	80.9
	NPA(%)	97.7	97.3	96.8	97.1	98.2

Table 2: p_{Ti}=0.2.

Correlation	Performance Measures	Reader 1	Reader 2	Reader 3	Average of the 3 readers	Consensus of the three readers with Test Device
0.0	PPA (%)	36.2	37.7	47.8	40.6	1.5
	NPA (%)	72.4	68.3	72.2	71.0	89.3
0.25	PPA(%)	46.9	44.2	51.3	47.5	36.3
	NPA(%)	80.5	78.0	82.3	80.3	90.9
0.50	PPA(%)	61.6	62.4	63.9	62.7	63.9
	NPA(%)	88.8	86.4	87.5	87.6	94.0
0.75	PPA(%)	77.7	83.1	73.0	77.9	85.1
	NPA(%)	95.5	94.7	95.3	95.2	97.5

Table 3: p_{Ti}=0.3.

Correlation	Performance Measures	Reader 1	Reader 2	Reader 3	Average of the 3 readers	Consensus of the three readers with Test Device
0.0	PPA (%)	36.6	39.2	46.4	40.7	13.7
	NPA (%)	62.1	60.0	61.1	61.0	78.6
0.25	PPA(%)	50.5	51.5	55.6	57.5	50.0
	NPA(%)	74.6	69.7	70.8	71.7	82.1
0.50	PPA(%)	62.8	65.1	64.7	64.2	71.2
	NPA(%)	83.2	79.8	81.9	81.6	88.8
0.75	PPA(%)	81.6	83.4	86.6	83.9	85.7
	NPA(%)	92.5	90.2	89.9	90.9	93.6

If the intent of use involves a single reader evaluating the images from a CCE which is under evaluation (or is the test device) and if instead a consensus (majority of 2 out of 3 evaluation by readers evaluating the test device) is used, the PPA and NPA could be very different from the PPA and NPA observed for individual readers. Best to report PPA and NPA for each reader separately and assess the reader variability.

Now if the CCE is compared against a clinical reference standard, OC, then the device and reader performance can be evaluated by the sensitivity and specificity pair. Three binary correlated data was simulated using a sample size of 750 with number of subject with positive detections chosen as 75 and subjects with negative detections being 675. The results in table 4 indicate that both sensitivity and specificity could be biased upwards. Thus, best to report sensitivity and specificity for each reader separately and assess the reader variability.

Table 4: Sensitivity and specificity (npos=75, nneg=675)

Correlation	Performance Measures	Reader 1	Reader 2	Reader 3	Average of the 3 readers	Consensus of the three readers with Test Device
0.0	Sens(%)	85.3	73.3	72.0	76.9	84.0
	Spec(%)	91.0	90.0	90.5	90.4	96.4
0.25	Sens(%)	82.7	68.0	72.0	74.2	78.7
	Spec(%)	91.7	91.0	90.4	91.0	95.0
0.50	Sens(%)	77.3	73.3	74.7	75.1	76.0
	Spec(%)	92.4	91.4	90.5	91.5	93.0
0.75	Sens(%)	74.7	70.7	76.0	73.8	72.0
	Spec(%)	91.4	91.0	91.4	91.3	91.4

3. Conclusion

CCE with camera is used to detect colon polyps where the detection of polyps is based on reader evaluation of the images of the colon from the camera in the capsule. Optical Colonoscopy (OC) serves as the clinical reference standard for detecting colon polyps and the CCE with camera is evaluated against OC to determine its accuracy in detecting polyps. Also, in rare instances one CCE with camera is compared against another used as a comparator, which is not a reference standard.

When a new device is compared to an already marketed device, if the device images are intended to be evaluated by a single reader then it is best to report PPA and NPA for each reader separately. A consensus of the reader reading the images from a test device should not be used to evaluate the test imaging device. In addition, the reader variability needs to be evaluated. Likewise, when a new device is evaluated against a clinical reference standard, if the device images are intended to be evaluated by a single reader, then it is best to report both sensitivity and specificity for each reader and the reader variability.

References

(1) Lunn A. D. and Davies S. J. A note on generating correlated binary variables. Biometrika 1998; 85: 487-490.
(2) Leisch F., Weingessel A. and Hornik K. On generation of correlated artificial binary data. Working paper series, SFB "Adaptive Information Systems and Modelling in Economics and Management Science", Vienna University of Economics.
(3) Rokkas T., Papaxoinis K., Ladas S. A meta-analysis evaluating the accuracy of colon capsule endoscopy in detecting colon polyps. : II. Resolving the paradoxes. Gastrointestinal Endoscopy 2010; 71(4) 792-798.
(4) Spada C., Hassan C., Sturniolo G.C., Marmo R., Riccioni, M. E., de Francis R. Van Gossum A. and Costamagna, G. Literature review and recommendations for clinical application of Colon Capsule Endoscopy. Digestive and Liver Disease 2011; 43:251-258.