The performance of testing procedures for clinical trials with multiple binary endpoints

Kouji Yamamoto[*] and Kanae Takahashi[†]

Abstract — **In medicine, screening tests or diagnostic tests are important for early detection and treatment of disease. The sensitivity (SE) and specificity (SP) are freqently used to refer the ability of a diagnostic test. The SE is the probability that a diseased individual has a positive result, and the SP is the probability that a non-diseased individual has a negative result. On the other hand, the positive predictive value (PPV) and the negative predictive value (NPV) describe the performance of a diagnostic test. The PPV is the probability of disease when the diagnostic test result is positive, and the NPV is the probability of no disease when the diagnostic test result is negative. The SE, SP, PPV and NPV are useful clinically, and may influence the treatment decision. There are several methods to compare these measures of two diagnostic tests separately. Practically, however, the performance of a diagnostic test may be assessed comprehensively with multiple measures, and it may also be insufficient to show the speriority of only one measure. Therefore, this talk deals with a case that the effectiveness of a new dignostic test is confirmed only when the superiority of the new test to the other test is shown in at least one measure and non-inferiority is shown in the other measures. There are some testing procedures for such a situation. In this talk, we perform simulation studies to investigate the performance of these procedures for evaluating diagnostic tests. We assume multiple primary binary endpoints such as SE, SP, or as PPV and NPV.**

Keyword: *Bootstrap test; gatekeeping procedure; joint hypothesis.*

1 Introduction and methods

The sensitivity (SE) and specificity (SP) are freqently used to refer the ability of a diagnostic test. The SE is the probability that a diseased individual has a positive result, and the SP is the probability that a non-diseased individual has a negative result. On the other hand, the positive predictive value (PPV) and the negative predictive value (NPV) describe the performance of a diagnostic test. The PPV is the probability of disease when the diagnostic test result is positive, and the NPV is the probability of no disease when the diagnostic test result is negative. The SE, SP, PPV and NPV are useful clinically, and may influence the treatment decision. There are several methods to compare these measures of two diagnostic tests separately. Generally, we use McNemar test for comparing two SEs or SPs. On the other hand, there are a few research for testing the equality of predictive values (Kosinski, 2013). We can use the method described above for clinical trials with only one endpoint.

In practice, however, the performance of a diagnostic test may be assessed comprehensively with multiple measures, and there may be some cases that it is insufficient to show the superiority of only one specified measure. So we now consider the case that the effectiveness of a new dignostic test is confirmed only when the superiority of the new test to the other test is shown in at least one measure and non-inferiority is shown in the other measures.

For such a situation, Bloch *et al.* (2007) considered a combined superiority and non-inferiority approach for multiple binary endpoints based on a bootstrap test, and Mascha and Turan (2012) proposed a framework for joint hypothesis testing based on gatekeeping procedures.

However, these researches did not mention the details much for binay endpoints.

Therefore, in this talk, we investigate the performance of these testing procedures with respect to the empirical type I error rate and empirical power via simulation studies.

[*] Department of Clinical Epidemiology and Biostatistics, Graduate School of Medicine, Osaka University, 2-2 Yamadaoka, Suita, Osaka, 565-0871, Japan, E-mail: yamamoto-k@stat.med.osaka-u.ac.jp, Tel: +81-6-6210-8373

[†] Department of Medical Innovation, Osaka University Hospital, 2-2 Yamadaoka, Suita, Osaka, 565-0871, Japan

2 Simulation studies

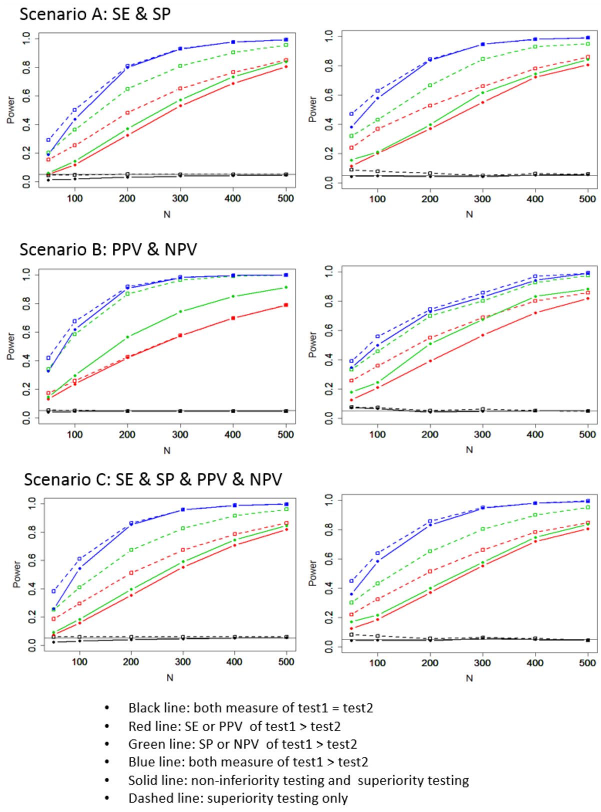
We generate datasets based on Kosinski (2013) method with several scenarios shown in Table 1, and assume non-inferiority margin is 0.1, and nominal type I error rate is 0.05.

Figure 1 shows the results for empirical type I error and power using Bloch *et al.* (2007) and Mascha and Turan (2012) methods. We can see from Figure 1 that (1) in views of type I error rate and statistical power, the two methods may not be much different, and (2) dropping non-inferiority for all endpoints does not gain much in power for the method. This means the advantages of combined superiority and non-inferiority approach to multiple endpoints.

Table 1: Simulation settings

Scenarios	SE1	SE2	SP1	SP2	PPV1	PPV2	NPV1	NPV2
A	0.75 0.85	0.75	0.80 0.90	0.80	-	-	-	-
B	-	-	-	-	0.75 0.85	0.75	0.80 0.90	0.80
C	0.75 0.85	0.75	0.80 0.90	0.80	Derived from SEs and SPs			

SE: sensitivity, SP: specificity, PPV: positive predictive value, NPV: negative predictive value.


Scenario A: SE & SP

Scenario B: PPV & NPV

Scenario C: SE & SP & PPV & NPV

- Black line: both measure of test1 = test2
- Red line: SE or PPV of test1 > test2
- Green line: SP or NPV of test1 > test2
- Blue line: both measure of test1 > test2
- Solid line: non-inferiority testing and superiority testing
- Dashed line: superiority testing only

Figure 1: Empirical power for several scenarios

References

[1] Bloch, D. A. et al. (2007). A combined superiority and non-inferiority approach to multiple endpoints in clinical trials. *Statistics in medicine*, **26**, 11931207.

[2] Kosinski, A. S. (2013). A weighted generalized score statistic for comparison of predictive values of diagnostic tests. *Statistics in medicine*, **32**, 96477.

[3] Mascha, E. J. and Turan, A. (2012). Joint hypothesis testing and gatekeeping procedures. *Anesthesia and Analgesia*, **114**, 13041317.