Statistical Issues Associated with Prognostic Biomarker Evaluation

Kyunghee K. Song
FDA/CDRH, 10903 New Hampshire Ave, Silver Spring, MD 20993

Abstract

In general, the evaluation of diagnostic test using biomarkers with prognostic claim should include the clinical outcome studies. The clinical outcomes are observed in a follow-up fashion and the clinical performance to support the clinical utility of the device depends on the analysis of these clinical outcomes. However, before making any clinical interpretation, it should be noted that the determination of the biomarker status depends on the diagnostic test, and there are likely to be false positives and false negatives when imperfect diagnostic tests are used. Therefore, analyzing and interpreting the survival outcomes without considering these false test outcomes introduce bias in the clinical outcome study. In this presentation, I have conducted a series of simulation study to address these issues and also discuss some regulatory concerns.

Key Words: diagnostic test, prognostic biomarkers, clinical outcomes, log-rank test, simulation study

1. Introduction

Biomarkers are defined as characteristics that is objectively measured and evaluated as an indicator for normal biological processes, pathogenic processes or pharmacologic response to a therapeutic intervention (Biomarkers Definitions Working Group, 2001). Diagnostic test is defined as a measurement or examination used to classify patients into a particular class or clinical state (CLSI EP24) and the diagnostic test detecting/measuring biomarkers are used in clinical practice for the diagnosis and prognosis of clinical conditions of interest as well as screening and monitoring for those conditions. Also, a lot of recent studies investigate whether the use of biomarkers can lead to the specific intervention procedures (such as treatment).

This presentation is focused on the prognostic biomarkers in the regulatory settings.

2. Evaluation of Diagnostic Biomarker

At FDA/Industry meeting 2006, Sargent and Mandrekar have stated that the prognostic marker is a single trait or signature of traits that separates different populations with respect to the risk of an outcome of interest in absence of treatment or despite non targeted “standard” treatment. The “outcome of interest” is specified in McShane et al (2005) as a time-to-event outcome (such as overall survival, recurrence-free survival).

In general, the evaluation of diagnostic test using biomarker is conducted in the analytical study (pre-clinical study) and the clinical study. Analytical study consists of precision, reproducibility, limit of detection, linearity,.. , etc. Clinical study is based on testing clinical samples and comparing the results to the reference method or some proper comparator method. The typical performance measures for the clinical comparison study are sensitivity|specificity or positive|negative percent agreement (also positive|negative likelihood ratio, and other similar measures are also considered depending on the study). In addition, the evaluation of clinical outcome study is required for the diagnostic biomarker with prognostic claim. Thus, in order to evaluate diagnostic biomarkers with prognostic claim, the performance of the test as well as the clinical utility of the test has to be examined subsequently.

3. Evaluation of Diagnostic Test with Prognostic Biomarker with clinical outcomes

MammaPrint is a diagnostic test with the prognostic claim. The intended use says "MammaPrint is a qualitative in vitro diagnostic test service, performed in a single laboratory, using the gene expression profile of fresh frozen breast cancer tissue samples to assess a patient's risk for distant metastasis. The test is performed for breast cancer patients who are less than 61 years old, with Stage I or II disease, with tumor size ≤ 5 cm and who are lymph node negative. The MammaPrint result is indicated for use by physicians as a prognostic marker only, along with other clinicopathological factors."

The clinical study has reported the predictive accuracy of MammaPrint for distant metastasis in 5 years in terms of positive predictive value (PPV) and negative predictive value (NPV), where PPV is defined as Pr(distant metastasis | high risk by gene signature) and NPV as Pr(no distant metastasis | low risk by gene signature). The clinical study has also reported Kaplan-Meier product limit estimates for 5 year and 10 year distant metastasis-free survivals. Both unadjusted and adjusted hazards ratios are reported based on risk classification.

TOP2A FISH pharmDx is also a diagnostic test with the prognostic claim. The intended use says " TOP2A FISH pharmDx Kit is designed to detect amplifications and deletions (copy number changes) of the TOP2A gene using fluorescence in situ hybridization (FISH) technique on formalin-fixed, paraffin-embedded human breast cancer tissue specimens. Deletions and amplifications of the TOP2A gene serves as a marker for poor prognosis in high risk breast cancer patients."

The clinical study has reported Kaplan-Meier product limit estimates for 5 year recurrence-free survival and 5 year overall survival for each treatment arm. The hazard ratio and the 95% confidence limits from multivariate analysis using Cox proportional hazards model are also reported.

Both MammaPrint and TOP2A FISH tests are used to determine the patients into different risk groups and the prognostic claims are evaluated with clinical outcome studies. The interpretation of the clinical outcome study depends on the test results. Therefore, the accuracy of diagnostic test needs to be carefully examined before conducting clinical study as well as analyzing the clinical outcome studies.

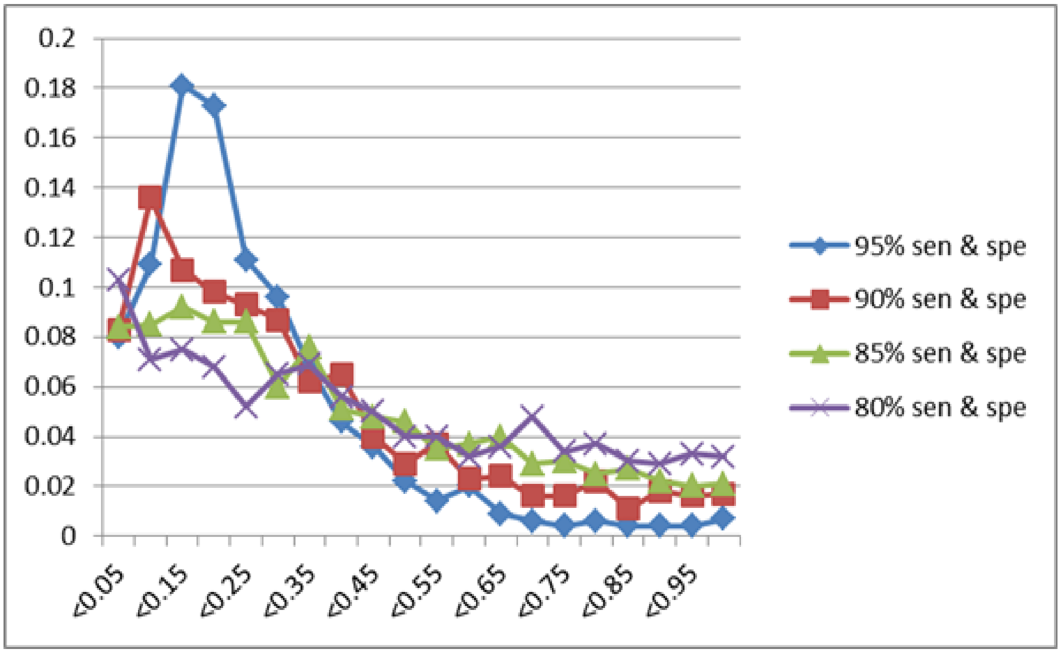
A series of simulation studies are conducted to investigate how the performance of the diagnostic test would affect the analysis of clinical outcome study. It is possible that the patients are grouped into a wrong risk group when a diagnostic test which is not perfect is used to group the patients. The simulation study is designed to examine how this would affect statistical results derived from the clinical outcome study.

4. Simulation Study

Two different failure time distributions ($N_1 = N_2 =50$) are generated from the two exponential distributions with parameters, λ=1.0 for group 1 and λ=1.4 for group 2. These two failure times are compared using the log-rank test. Suppose a diagnostic test is performed to stratify samples into two different groups, such as marker positive/negative or high/low risk. With a perfect diagnostic test (sensitivity = specificity = 1.0), it is assumed that the samples are grouped into the right risk group. However, with an imperfect diagnostic test, there should be false positives and false negatives, and this imperfect test assigns the samples into a wrong group. With these simulated datasets, the log-rank p-value is 0.1563 assuming a perfect diagnostic test.

For the simulation, the following imperfect diagnostic test conditions are considered: 95%, 90%, 85 and 80% for both sensitivity and specificity. A thousand datasets are generated and compared using a log-rank test. The log-rank test p-values are plotted in Figure 1(x-axis for the obtained p-values (binned), y-axis for the proportion out of 1000 generated p-values).

Figure 1



Simulation results show that the accuracy of the diagnostic test greatly affects the statistical interpretation of the clinical outcome study. With 95% sensitivity and specificity, the obtained p-values are mostly around the true p-value, 0.1563, showing a peak between 0.15 and 0.20. However, with 80% sensitivity and specificity, the resulting p-values are widely spread.

Since the diagnostic test accuracy (sensitivity & specificity) greatly affects the statistical inference of clinical outcome study, it is critical to use the diagnostic test with high accuracy.

5. Summary

The unique feature of prognostic biomarkers is that the clinical performance to support the clinical use for the diagnostic device is observed in a follow-up fashion and the clinical utility of the device depends on the analysis of these clinical outcomes. Therefore, the evaluation of diagnostic test using biomarkers with the prognostic claim should include the clinical outcome studies to observe follow-up events.

However, it should be noted that the determination of each group status, such as marker status or gene feature status, might not be perfect. In other words, there might be errors in determining the risk status or genetic aberration status in diagnoses, resulting false positives and false negatives and consequently putting the subjects into a wrong risk group. Therefore, analyzing and interpreting the survival outcomes without considering this possibility introduce bias in the interpretation of the clinical outcome study.

In this presentation, it is shown that the use of the diagnostic test with high accuracy (high sensitivity & high specificity) is critical to examine the statistical inference of clinical outcome study through simulation studies. Therefore, in the prognostic biomarker studies, the efforts should be made to use the best available diagnostic test to determine the risk group (the status of biomarker) to avoid any errors. In some cases, the sensitivity analysis can be considered with not-perfect test. The sensitivity analysis could provide information how the clinical outcome study results would be affected by false results.

References

Biomarkers Definitions Working Group (2001). Clinical Pharmacology and Therapeutics, 69, p.89-95. as referenced in CDER's Draft guidance: Qualification Process for Drug Developmental tools

CLSI EP24-A2 Assessment of the Diagnostic Accuracy of Laboratory Tests Using Receiver Operating Characteristics Curves; Approved Guideline – Second Edition.

McShane LM, Altman DG, Sauerbrei Wtaube SE, Gion M and Clark GM (2005) Reporting recommendations for tumor marker prognostic studies. J Clin Oncol 23:9067-9072

MammaPrint, http://www.accessdata.fda.gov/cdrh_docs/reviews/K062694.pdf

TOP2A, http://www.accessdata.fda.gov/cdrh_docs/pdf5/P050045b.pdf