Journal of Biopharmaceutical Statistics, 18: 167–182, 2008

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400701668316

Taylor & Francis
Taylor & Francis Group

STATISTICAL ISSUES ON THE DIAGNOSTIC MULTIVARIATE INDEX ASSAY FOR TARGETED CLINICAL TRIALS

Jen-pei Liu
Division of Biometry, Department of Agronomy, National Taiwan University and Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taipei, Taiwan

Shein-Chung Chow
Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, USA

In the past decade, pharmacogenomics and microarrays are considered two of the most important scientific breakthroughs for detection and treatment of diseases with many other applications. After completion of the Human Genome Project (HGP), the importance of diagnostic tests for identification of molecular targets increases as more targeted clinical trials are conducted for the individualized treatment of patients in the post-genomic era. As a result, the co-development of drug-device has become the foundation for achieving the ultimate goal of personalized medicine. One of the diagnostic devices for detection of molecular targets is the in vitro diagnostic multivariate index assays (IVDMIA) based on the genomic composite biomarker (GCB) classifiers. Thus, the quality of the IVDMIA, innovative designs, and the evaluation of efficacy for targeted clinical trials are vital for achieving this goal. However, before personalized medicine becomes a reality, many challenges for assurance of the accuracy and precision of the IVDMIA and the estimates of the treatment effect in the population with the molecular targets are to be resolved. In this paper, we identify the following issues on the IVDMIA and targeted clinical trials:* **(i)** *the selection of the differentially expressed genes,* **(ii)** *the optimal representation and algorithm for the genomic composite biomarker (GCB) classifier for the best diagnostic accuracy of the moleculdar target,* **(iii)** *the validation of the IVDMIA, and* **(iv)** *the evaluation of effectiveness and sample size estimation for targeted clinical trials. For each issue, the problem and possible resolutions are discussed. An overall assessment and some concluding remarks are also provided.

Key Words: Genomic composite biomarker classifier; Multivariate index assay; Pharmacogenomics; Targeted clinical trials.

1. INTRODUCTION

As indicated by many researchers (e.g., Casciano and Woodcock, 2006; Dalton and Friend, 2006; Maitournam and Simon, 2005; Simon and Maitournam, 2004;

Received December 1, 2006; Accepted June 20, 2007
Address correspondence to Jen-pei Liu, Division of Biometry, Department of Agronomy, National Taiwan University, Taipei, Taiwan; E-mail: jpliu@ntu.edu.tw

Varmus, 2006), the disease targets at the molecular level can be identified after completion of the Human Genome Project (HGP). As a result, the importance of diagnostic tests for identification of molecular targets increases as more targeted clinical trials will be conducted for the individualized treatment of patients. For example, based on the risk of distant recurrence determined by a 21-gene Oncotype DX® breast cancer assay, patients with a recurrence score of 11 to 25 in the TAILORx (Trial Assigning Individualized Options for Treatment) trial sponsored by the United States National Cancer Institute (NCI) are randomly assigned to receive either adjuvant chemotherapy and hormonal therapy or adjuvant hormonal therapy alone (Sprarano et al., 2006). On the other hand, based on a 70-gene molecular signature, the MINDACT (Microarray in Node-negative Disease may Avoid ChemoTherapy) trial randomizes patients with a low risk molecular prognosis and a high-risk clinical prognosis to the use of clinicopathologic criteria or gene signature in treatment decisions for the possible avoidance of chemotherapy (MINDACT, 2006). These two trials have an important implication for the future individualized treatments for thousands of breast cancer patients (Swain, 2006). The Oncotype DX® used in the TAILORx trial is a reverse-transcriptase-polymerase-chain-reaction (RT-PCR) assay based on 21 genes, while the MINDACT trial employs a 70-gene molecular signature derived from the microarray (Paik et al., 2004, 2006; van't Veer et al., 2002; Van de Vijver et al., 2002).

Despite of difference between the two platforms employed in the diagnostic devices for molecular targets used in the two trials, both assays belong to a group of the IVDMIA (FDA, 2006) based on the selected differentially expressed genes for detection of the patients with the molecular targets. In addition, to reduce the variation in measuring the expression levels as well as the between-patient variability of expression levels, the IVDMIAs do not usually use all genes during the development stage. Therefore, identification of the differentially expressed genes between different groups of patients is the key to the accuracy and reliability of the devices for molecular targets. Once the differentially expressed genes are identified, the next task is to search an optimal representation or algorithm which provides the best discrimination ability between the patients with molecular targets and those without the targets. The current validation procedure for diagnostic device is for the assay based on one analyte. However, the IVDMIAs are in fact the parallel assays based on the intensities of multiple analytes. As a result, the current approach to assay validation for one analyte may not be appropriate and is inadequate for validation of IVDMIAs.

An enrichment design (Chow and Liu, 2004) for the targeted clinical trials is the one for which the patients with positive diagnosis for the molecular targets are randomized to receive the test drug or the control. However, because no IVDMIA can provide the perfectly correct diagnosis, some patients with positive diagnosis may not actually have the molecular targets. Consequently, the treatment effect of the test drug for the patients with targets may be under-estimated. On the other hand, estimation of the treatment effect based on the data from the targeted clinical trials needs to take into consideration the variability associated with the estimates of accuracy of the IVDMIA such as positive predictive value and false positive rate obtained from the effectiveness of the clinical trials of the IVDMIA.

In the next section, commonly used approaches to identification of differentially expressed genes are reviewed. Also included in this section is the

discussion of the relative merit and disadvantages of current methods. A set of interval hypotheses, which takes into consideration of the minimal biological meaningful expression level, is proposed. Based on the interval hypotheses, a two one-sided tests procedure is proposed. A discussion of the optimal representation or an algorithm of the IVDMIA based on the expression levels of the selected differentially expressed genes for the best diagnosis of the molecular targets is provided in Section 3. Also included in this section is a recommendation for determination of the number of genes to be included in the IVDMIA. In Section 4, the deficiency of the current validation for one analyte used for the IVDMIA is discussed. In addition, the issues and challenges for validation of the IVDMIA are also addressed in this section. Bias in estimation of the treatment effect of the test drug in the targeted clinical trials is discussed in Section 5. Approaches to obtaining the unbiased estimator of the treatment effect for patients with the molecular target and its variance are also given in this section. The assessment of IVDMIA and targeted clinical trials and some concluding remarks are given in Section 6.

2. IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

For a given gene, the fold change is defined as the ratio of average expression level of the gene, which is measured by the intensity under one condition (say tested or patients with a certain disease), to that under another condition (say controlled or normal subjects without the disease). A gene is declared to be differentially expressed if the observed fold change either exceeds a pre-specified threshold or is below a pre-determined lower threshold. We refer to this procedure as the *fixed fold-change rule*. The fixed fold-change rule does not take into consideration the variation in estimation of the average intensity. In addition, it is not in the framework of hypothesis testing and therefore the probability associated with errors for decision-making can not be quantified and/or assessed. On the other hand, most current available statistical methods for identification of differentially expressed genes such as the t-test, permutation t-test, or significance analysis of microarray (SAM) are in fact based on the following traditional hypotheses testing for equality (see, e.g., Dudoit et al., 2002; Simon et al., 2003; Tusher et al., 2001; Wang and Ethier, 2004):

$$H_0 : \mu_{Di} - \mu_{Ni} = 0 \quad \text{vs.} \quad H_a : \mu_{Di} - \mu_{Ni} \neq 0, \tag{1}$$

where $i = 1, \ldots, G$ and μ_{Ti} and μ_{Ni} are the true average expression levels on the log-scale (base 2) of gene i of the patients with the molecular targets and the normal subjects without the molecular targets, respectively.

The traditional hypotheses testing for equality is only to detect whether the difference in the average expression levels is 0 between the tested and controlled conditions. It fails to take into account the magnitudes of the biologically meaningful fold changes. In addition, due to simultaneously testing thousands of genes at the same time, with a small number of replicated samples, the false positive rate for identifying differentially expressed gene is extremely high. Therefore, various methods are proposed to resolve this issue. Basically, they are applications of multiple comparison procedures to use some arbitrarily selected stringent cut-off of

p-values to control false discovery rate (Benjamini and Hochberg, 1995; Hochberg and Tamhane, 1987) or to apply a combination of less stringent p-value for traditional hypotheses testing and the fixed fold change rule (MAQC Consortium, 2006). However, all of these methods fail to take into account both magnitudes of biologically meaningful fold changes and statistical significance simultaneously.

Since the objective is to identify the differentially expressed genes, the hypothesis for identifying differentially expressed genes should be formulated as the alternative hypothesis. On the other hand, gene i is said to be differentially expressed if the difference in average expression levels between the tested and controlled samples is either greater than a minimal biologically meaningful limit C_i (over-expressed) or smaller than a maximal biological meaningful limit $-C_i'$ (under-expressed). As a result, the hypothesis for identifying differential expressed genes between the tested and controlled samples can be formulated as the following hypotheses:

$$\begin{aligned} H_0 &: -C_i' \le \mu_{iD} - \mu_{iN} \le C_i \quad \text{vs.} \\ H_1 &: \mu_{iD} - \mu_{iN} < -C_i' \quad \text{or} \quad \mu_{iD} - \mu_{iN} > C_i, \quad i = 1, \ldots, G \end{aligned} \tag{2}$$

The parameter space for H_0 is $[-C_i', C_i]$, which represents the interval of no differential expression. On the other hand, the parameter space of the alternative hypothesis is the union of the intervals of over-expression (C_i, ∞) and under-expression $(-\infty, -C_i')$. In general, each gene should have its own differential expression limits and the differential expression limits does not have to be symmetric about 0. However for the sake of illustration, without loss of generality, in what follows, we assume that the differential expression limits are the same and are symmetric about 0. The interval hypotheses for differentially expressed gene can be then formulated as

$$H_0 : |\mu_{iD} - \mu_{iN}| \le C \quad \text{vs.} \quad H_1 : |\mu_{iD} - \mu_{iN}| > C, \quad i = 1, \ldots, G, \tag{3}$$

where C is some biologically meaningful differential expression limit.

Furthermore, the interval hypotheses can be decomposed into two sets of one-sided hypotheses:

$$\begin{aligned} H_{0U} &: \mu_{iD} - \mu_{iN} \le C \quad \text{vs.} \quad H_{1U} : \mu_{iD} - \mu_{iN} > C \quad \text{or} \\ H_{0L} &: \mu_{iD} - \mu_{iN} \ge -C \quad \text{vs.} \quad H_{1L} : \mu_{iD} - \mu_{iN} < -C, \quad i = 1, \ldots, G \end{aligned} \tag{4}$$

The first set of hypotheses is to verify whether the difference in average expression level between the tested and controlled samples for gene i is higher than the pre-specified upper differential expression limit for over-expression. The second set of hypotheses is to evaluate whether the difference in average expression levels between the tested and controlled samples for gene i is lower than the pre-determined lower differential expression limit for under-expression.

Since the parameter space of the alternative hypothesis in (3) is the union of the parameter spaces of the two one-sided hypotheses given in (4), H_0 in (3) is rejected at the α level of significance if and only if, either H_{0U} or H_{0L} is rejected at the $\alpha/2$ level of significance. In other words, under normal assumption, a two

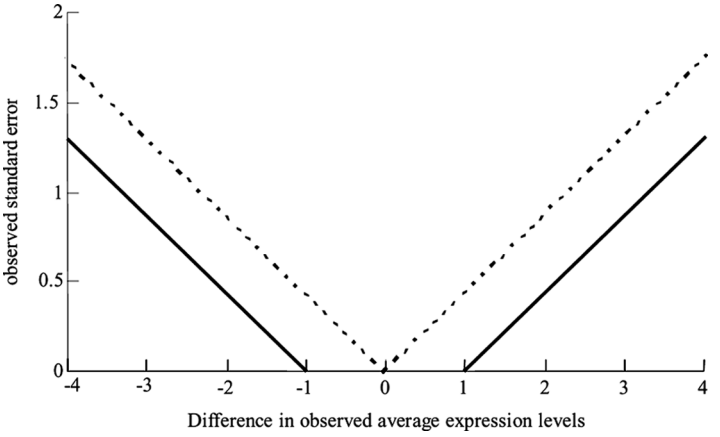

Figure 1 Rejection regions of the two one-sided tests procedure (the shade area bounded by solid line) and the unpaired two-sample t-test (dashed line) for $C = 1$, $n_{iT} = n_{iC} = 5$, and the $\alpha = 0.05$ nominal level.

one-sided tests procedure rejects the null hypothesis of (3) and we conclude that gene i is differentially expressed between the tested and controlled samples at the α level of significance if

$$t_{Ui} = \frac{\overline{Y}_{iD} - \overline{Y}_{iN} - C}{\sqrt{s_{pi}^2\left(\frac{1}{n_{iD}} + \frac{1}{n_{iN}}\right)}} > t_{(\alpha/2, n_{iD}+n_{iN}-2)} \quad \text{or} \quad t_{Li} = \frac{\overline{Y}_{iD} - \overline{Y}_{iN} + C}{\sqrt{s_{pi}^2\left(\frac{1}{n_{iD}} + \frac{1}{n_{iN}}\right)}} < -t_{(\alpha/2, n_{iD}+n_{iN}-2)}, \tag{5}$$

where $\overline{Y}_{ik}$, and n_{ik} are the sample mean expression and sample size of gene i under treatment k, respectively and s_{pi}^2 is the pooled sample variance for gene i, where $i = 1, \ldots, G$ and $k = T, C$. Figure 1 gives the rejection region of the two one-sided tests procedure (the solid line) at the α level of significance for $C = 1$, and $n_{iD} = n_{iN} = 5$ together with the rejection region of the conventional two-sample t-test for the hypothesis of equality (the dash line). From Fig. 1, an interval of no differential expression is formulated in the acceptance region for the interval hypothesis while the acceptance region for the two-sample t-test contains a single point of 0. In addition, the rejection region of the two one-sided tests procedure is a subset of that of the two-sample t-test. Consequently, the two-sided tests procedure will reduce the probability of falsely identifying un-expressed genes differentially expressed. It is straightforward to verify that under the normality assumption, the power function of the two one-sided tests is symmetric at the average of C_i and C_i' and it is an α-level test. For gene with very low variations in expression levels, s_{pi}^2 may be extremely small and t_{Ui} and t_{Li} in Eq. (5) may be artificially inflated. Following the SAM approach (Tusher et al., 2001), an empirical constant can be added to s_{pi}^2 in order to avoid such scenarios.

3. OPTIMAL REPRESENTATION OF IN VITRO DIAGNOSTIC MULTIVARIATE INDEX ASSAYS

For an IVDMIA that can be clinically meaningful and its validation can be practically feasible, it must be parsimonious with a clinically meaningful threshold that can provide the best diagnostic accuracy for the molecular targets under investigation. In addition, the IVDMIA is in fact some form of parallel assays with many analytes, and hence these analytes can be treated as multiple diagnostic markers with expression levels being the measurements in the same unit. As a result, a linear representation of expression levels of the selected differentially expressed genes presents a reasonable approach to the diagnosis of the molecular targets. It follows that the result of any IVDMIA with a linear representation is a continuous variable with a pre-determined cut-off for diagnosis of the molecular target. Therefore, first, we need to determine the coefficients in the linear combination of the multiple markers not only to have the best discrimination ability for classification of patients with the minimal classification error but also to provide the best diagnostic accuracy. There are many indices for evaluation of diagnostic accuracy such as sensitivity, specificity, false positive (FP) rate, positive predictive value (PPV) and negative predictive value (NPV). However, these indices change when a different threshold is used. On the other hand, the area under the receiver operating characteristic (ROC) curve is a quantitative criterion for evaluation of the overall performance of diagnostic accuracy. As a result, we recommend using the generalized area under ROC curve based on multiple diagnostic markers for evaluation of the diagnostic accuracy of the IVDMIA (Su and Liu, 1993). Then, based on the area under the generalized ROC curve of the IVDMIA, a threshold can be determined to balance between the sensitivity and specificity for clinical application. However, even though the generalized ROC curve of the IVDMIA has a good overall performance, for certain thresholds at which the clinicians choose to operate, the sensitivity and specificity might not be optimal.

Suppose that a total of g differentially expressed gene has been selected for the IVDMIA. Let $\mathbf{Y}_{Dk}$ ($\mathbf{Y}_{Nk}$) be a g-vector of the expression levels of patient k with (without) the molecular targets, $k = 1, \ldots, n_D(n_N)$. Assume that $\mathbf{Y}_{Dk} \sim N(\boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D)$ and $\mathbf{Y}_{Nk} \sim N(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$, a linear representation of the IVDMIA has the form of $\mathbf{a}'\mathbf{Y}_{Dk}(\mathbf{a}'\mathbf{Y}_{Nk})$ that has the best diagnostic accuracy if it can provide the maximal area under the ROC curve. In other words, one needs to determine the coefficients in $\mathbf{a}$ such that $P(\mathbf{a}'\mathbf{Y}_{Dk} > \mathbf{a}'\mathbf{Y}_{Nk})$ is maximized. Su and Liu (1993) showed that the Fisher linear discrimination function provides the coefficients of the best linear combination

$$\mathbf{a}_\circ = (\boldsymbol{\Sigma}_D + \boldsymbol{\Sigma}_N)^{-1}(\boldsymbol{\mu}_D - \boldsymbol{\mu}_N). \tag{6}$$

These coefficients can not only minimize the classification error but also provide the largest area under the generalized ROC curve, which is given by

$$A = \Phi\Big(\sqrt{(\boldsymbol{\mu}_D - \boldsymbol{\mu}_N)'(\boldsymbol{\Sigma}_D + \boldsymbol{\Sigma}_N)^{-1}(\boldsymbol{\mu}_D - \boldsymbol{\mu}_N)}\Big), \tag{7}$$

where $\Phi(\cdot)$ is the distribution function of the standard normal random variable. A consistent estimate of A can be obtained by replacing the parameters with their

unbiased estimators, i.e., sample mean vectors $\overline{\mathbf{Y}}_D$ and $\overline{\mathbf{Y}}_N$ and sample covariance matrices $\mathbf{S}_D$ and $\mathbf{S}_N$ (Su and Liu, 1993). Reiser and Faraggi (1997) provided confidence interval for A. However, in the case of the IVDMIA derived from microarray experiments, the number of genes usually exceeds tens of thousands and number of patients is rarely in hundreds. Consequently, unstable estimation of the covariance matrices because of small sample sizes results in very poor prediction for the patient's status of the molecular targets (see, e.g., Simon et al., 2003). As a result, from the result of their cross-validation experiments, Simon et al. (2003) recommended the use of diagonal linear discriminate function (DLDF) or the compound covariate predictor (CCP) for their superior performance of correct classification over other methods. For the DLDF, not only the covariances among genes are set to be zero but also the homogeneity is assumed for the variances between the patients and normal subjects.

From Eq. (6), it can be seen that the estimators of the coefficients in $\mathbf{a}_\circ$ are proportional to the traditional t-statistic, which are also the coefficients used in the compound covariate predictor. Therefore, the more differentially expressed are the genes, the more weights of the genes are for the DLDF. In this regard, one could include all genes in the DLDF or CCP for the IVDMIA. However, if a gene is not differentially expressed between the patients with and without the molecular targets, it will have a small t-statistic and hence does not contribute to the prediction ability of the resulting DLDF or CCP. Therefore, during the early development stage of the IVDMIA, all possible genes should be included for identification of differentially expressed genes. However, for construction of the linear representation of the IVDMIA, those genes with no differential expressions should be dropped. Unfortunately, how many and which genes should be included in the linear representation still remain a great challenge to the researchers. One rule of thumb is that the number of genes and the genes to be included in the classifier should reach a balance between the practicality and amount of information required for an accurate diagnosis for the molecular targets. If there is unequivocal evidence that a certain biological pathway is involved in pathogenesis of a disease, then from a viewpoint of biology, all genes affecting this pathway should be included in the classifier. Suppose that the sample sizes are equal for the patients with and without the molecular targets. One measure that can be used for possible determination of the number of genes included in the classifier is the partial between-group distance (PBGD) defined as

$$\text{PBGD} = \frac{\sum_{i=1}^{g} \left(\overline{Y}_{iD} - \overline{Y}_{iN}\right)^2 / s_{pi}^2}{\sum_{i=1}^{G} \left(\overline{Y}_{iD} - \overline{Y}_{iN}\right)^2 / s_{pi}^2}. \tag{8}$$

The range of PBGD is from 0 to 1. Because most of genes tested during the early development stage of the IVDMIA are not differentially expressed and if we put $\left(\overline{Y}_{iD} - \overline{Y}_{iN}\right)^2 / s_{pi}^2$ into the numerator of PBGD in Eq. (8) according to its magnitude sequentially, then PBGD is an increasing function of the number of genes. In order to be clinically practical and to be feasible to be validated, one desirable characteristic of any IVDMIA is to provide a high diagnostic accuracy with a set of small number of genes. Under this ideal situation, PBGD is very steep and reaches

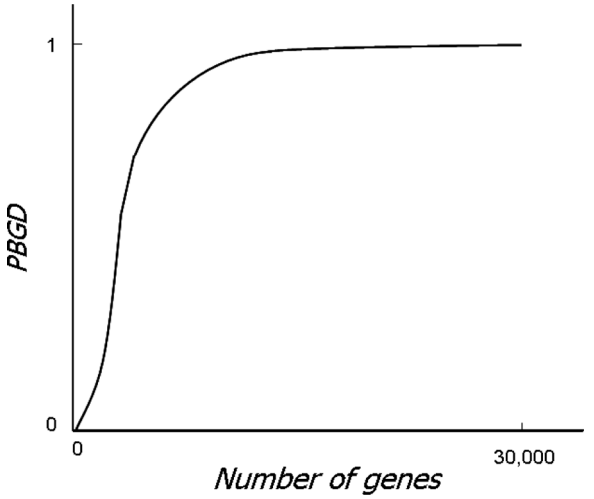

Figure 2 Number genes and partial between-group distance (PBGD).

the plateau of 1 very quickly as shown in Fig. 2. On the other hand, there might be several candidates for the classifier with similar diagnostic accuracy. Due to the principle of parsimony, treating the coefficients in the classifiers as fixed constants, based on the paired areas under the generalized ROC curves, one can apply the non-inferiority test to choose a classifier with the smallest number of genes but with an equivalent diagnostic accuracy (Li et al., 2006; Liu et al., 2006). However, the non-inferiority test based on the difference in paired areas of the generalized ROC curves derived from multiple markers requires further research.

4. VALIDATION OF IN VITRO DIAGNOSTIC MULTIVARIATE INDEX ASSAYS

As described above, Oncotype DX® used in the TAILORx trial is a RT-PCR assay based on 21 genes, while a 70-gene molecular signature derived from the microarray is used in the MINDACT trial. Therefore, IVDMIAs are parallel assays with multiple biomarkers and multiple medical decision points. It follows that validation of IVDMIA should address the performance and assay validation for each component as well as the overall quality performance of the whole IVDMIA (Frueh, 2006; Patterson et al., 2006). The FDA draft guidance suggests that for each target or expression pattern, the performance characteristics include assay sensitivity, reproducibility, validation of cut-off, reference range or medical decision point, assay range, and specificity (FDA, 2003). The FDA draft guidance also suggests consulting the guidelines on protocols for assay validation in clinical laboratory published by the Clinical Laboratory Standard Institutes (CLSI). However, these protocols are for a single analyte and are not suitable for complicated assay with multiple markers and a statistical algorithm for diagnosis. As a result, the assay validation of IVDMIA should employ different approaches although the principle of accuracy and precision remains the same (Canales et al., 2006; Ji and Davis, 2006). However, because the overall analytical performance

of the IVDMIA is determined by the performance of the individual component markers, at the minimum, the performance of each single gene should be evaluation by the approved guidelines on validation protocols issued by the CLSI.

Traditionally, one key issue for assay validation of the IVDMIA is the reference standards with known concentrations for establishment of the calibration curve, assessment of accuracy from recovery experiment, and evaluation of linearity and linear range of the IVDMIA. Recently, Shippy et al. (2006) investigated the relationship of the expression measurement of a transcript in a titration sample and the relationship between the signals of a given transcript in the two titration samples and that of each individual samples in the MAQC (Microarray Quality Control) study. They found that differences in normalization, platforms, and laboratory practices can lead to deviations from the mixing ratio expected in traditional assay validation and they proposed empirical measurements to estimate the true mRNA fraction in the titration samples. On the other hand, Tong et al. (2006) also examined the use of external RNA controls for assessment of the accuracy of the expression ratios between samples with known expression levels in the same MAQC study. They recommended a comprehensive study for modeling concentration response to determine the tolerance ranges for linear fit, slope and y-intercept for assay assessment, specificity in the context of false positives and false negatives. These studies by the investigators of the MAQC study indicate difficulty in obtaining the known concentration reference standards and assay validation for the IVDMIA based on the microarray platforms, and hence more research is needed for the challenges of validation of analytical aspects of the IVDMIA.

On the other hand, for a linear representation, the optimal algorithm to provide the best discrimination ability and diagnosis of molecular target for the IVDMIA is the diagonal linear discriminant function. Recall that the selected genes in DLDF are differentially expressed between the patients with and without the molecular target and weights are proportional to the t-statistics. Therefore the DLDF is an aggregate measure of expression levels with weights reflecting their relative contributions to the algorithm. But masking effects may occur while the relative unimportant genes with small weights become differentially expressed more than those with large weights. Once the weights are determined in the development stage, to avoid possible masking effect, the expression levels of each individual gene must exceed a pre-specific lower limit for the overall assay results to reach the threshold for a positive diagnosis of the molecular target. Theses pre-specific limits should be determined from the biological and clinical knowledge of relative roles of selected genes in the pathway of pathogenesis of the underlying disease.

Agreement and reproducibility measures are very important performance characteristics of IVDMIA and have recently drawn a lot of attention in the data generated from microarray experiments. For example, Dobbin et al. (2005), Irizarry et al. (2005), Larkin et al. (2005), and Members of the Toxicogenomic Research Consortium (2005) examined the agreement on measurements of gene expressions between laboratories and across different platforms. Testing the hypothesis of zero Pearson correlation coefficient (PCC) is the one of the most common statistical methods to assess comparability of gene expression levels between technical replicates within and across laboratories. However, to evaluate comparability on gene expressions within and between laboratories, it is to assess the agreement of the measurements of the technical replicates for the same genes of the same

samples. Hence objective for evaluation of comparability is to investigate the closeness or equivalence of gene expression levels between technical replicates of the same samples. Although Pearson correlation coefficient is an excellent statistic for evaluation of linear association, it is location and scale invariant. Hence it cannot detect changes in accuracy and precision and cannot be used for assessment of agreement of gene expression levels between technical replicates which requires evaluation of equivalence in both accuracy and precision. Therefore, hypothesis of zero linear correlation by Pearson correlation coefficient is not appropriate for evaluation of agreement of gene expression levels between technical replicates of the same samples.

On the other hand, the concordance correlation coefficient, proposed by Lin (1989, 1992) and Lin et al. (2002) is a product of Pearson correlation coefficient and a factor consisting of location and scale shifts. Therefore, it can be employed to evaluate the agreement of gene expression levels between the technical replicates of the same samples. In order to meet the minimal requirement of agreement, the hypothesis for assessment of agreement of gene expression levels between technical replicates should be formulated as the non-inferiority hypothesis which not only the linear association exceeds a pre-specified threshold but also the means and variability between technical replicates are equivalent within some pre-determined limits. Both asymptotic method and exact procedure based on Generalized Pivotal Quantities (GPQs) are available for interval estimation of concordance correlation coefficient for evaluation of agreement of gene expression levels between two technical replicates exceeds some minimal requirement of agreement (Lin, 1989; Liao et al., 2006).

5. EVALUATION OF EFFECTIVENESS OF TARGETED THERAPY

We first consider the situation where there is a validated IVDMIA for diagnosis of a particular molecular target involved with pathogenesis of the disease. In addition, we also consider a trial to compare the control treatment (c) with the molecularly targeted test treatment (t). For the purpose of illustration, we further assume that the primary endpoint is continuous. Table 1 gives population means by treatment and diagnosis of the molecular target. In Table 1, $\mu_{T+}, \mu_{C+}(\mu_{T-}, \mu_{C-})$ are the means of test and control groups for the patients with (without) the molecular target. For the enrichment design for the target clinical trials, patients with the diagnosis of the molecular target are randomized to receive the control or the test treatment. In addition, under the enrichment design, the parameter of interest for inference is $\mu_{T+} - \mu_{C+}$ that is the treatment effect of the test drug as compared to the control in the patients with the molecular target. Let $\overline{Y}_{T+}$ and $\overline{Y}_{C+}$ be the sample means of the test and control treatments for the patients with diagnosis of the molecular target respectively. Because no diagnostic test is perfect for the correct diagnosis of the molecular target without error, although under the enrichment design, patients with positive diagnosis still may not have the molecular target. It follows

$$E\left(\overline{Y}_{T+} - \overline{Y}_{C+}\right) = \gamma(\mu_{T+} - \mu_{C+}) + (1-\gamma)(\mu_{T-} - \mu_{C-}), \tag{9}$$

Table 1 Population means by treatment and diagnosis

	Molecular target			
Diagnosis	True status	Test	Control	Difference
$+(P_+)$	+ (PPV)	μ_{T+}	μ_{C+}	$\mu_{T+} - \mu_{C+}$
	− (FP)	μ_{T-}	μ_{C-}	$\mu_{T-} - \mu_{C-}$
$-(P_-)$	+ (FN)	μ_{T+}	μ_{C+}	$\mu_{T+} - \mu_{C+}$
	− (NPV)	μ_{T-}	μ_{C-}	$\mu_{T-} - \mu_{C-}$

$P+(P-)$ Positive (Negative) prevalence rate,
PPV: Positive predictive value
NPV: Negative predictive value
FP: False positive rate
FN: False negative rate

where γ is the positive predictive value (PPV) of the target. From Equation (9), the expected value of the difference in the sample means consists of two components. The first component is the treatment effect of the test drug in the patients with a positive diagnosis truly having the molecular target and the second component is the treatment effect of the patients with a positive diagnosis but in fact without out the molecular target. Under assumption that the test treatment is ineffective in the patients without the molecular target, it follows that $|\mu_{T+} - \mu_{C+}| > |\mu_{T-} - \mu_{C-}|$. It follows that, the difference in the sample means obtained under the enrichment design of the targeted clinical trials under-estimates the treatment effects of the test drug in the patients with the molecular target.

Although the estimate of the positive predictive value can be obtained independently from the clinical effectiveness trials of the IVDMIA (FDA, 2005), the true status for the individual patients with the molecular target in the targeted clinical trial is actually unknown. However, the information for the status of the molecular target is not completely missing because the estimate of positive predictive value is available. Let X_{Ti} be the latent variable that indicates the status of the molecular target of patient i in the test group; and $X_{Ti} = 1$ if the patients has the molecular target and $= 0$ otherwise, $i = 1, \ldots, n_T$. X_{ci} is similarly define for the patients in the control group, $i = 1, \ldots, n_C$. Then measurements of the primary endpoint, Y_{Ti}, for the patients in the test group are independently identically distributed as

$$Y_{Ti} \sim p_T\big(\cdot\,|\mu_{T+}, \sigma_T^2\big)^{XTi}\, q_T\big(\cdot\,|\mu_{T-}, \sigma_T^2\big)^{1-XTi}, \quad i = 1, \ldots, n_T,$$

where X_{Ti} is i.i.d Bernoulli random variable with probability of success γ and $p_T(\cdot)$ and $q_T(\cdot)$ are the normal probability density function for Y_{Ti}. Similarly,

$$Y_{Ci} \sim p_C\big(\cdot\,|\,\mu_{C+}, \sigma_C^2\big)^{XCi}\, q_C\big(\cdot\,|\,\mu_{C-}, \sigma_C^2\big)^{1-XCi}, \quad i = 1, \ldots, n_C,$$

where X_{Ci} is i.i.d Bernoulli random variable with probability of success γ and $p_C(\cdot)$ and $q_C(\cdot)$ are the normal probability density function for Y_{Ci}. It follows that the

likelihood function is proportional to

$$\left[p_T(\cdot \mid \mu_{T+}, \sigma_T^2)\right]^{\Sigma XTi}\left[q_T(\cdot \mid \mu_{T-}, \sigma_T^2)\right]^{nT-\Sigma XTi}\left[p_C(\cdot \mid \mu_{C+}, \sigma_T^2)\right]^{\Sigma XCi},$$
$$\left[q_T(\cdot \mid \mu_{C-}, \sigma_T^2)\right]^{nC-\Sigma Xci}, \tag{10}$$

EM algorithm (Dempster et al., 1977) or Bayesian inference using Gibbs sampling (Gelfand and Smith, 1990) can be used for both point and interval estimation for $\mu_{T+} - \mu_{C+}$. For EM algorithm, i.i.d. Bernoulli variables can be generated with the probability of success being the estimated positive predictive value. On the other hand, for Gibbs sampling, some informative prior for γ by incorporating the positive predictive value estimated from the clinical effectiveness trials for the IVDMIA can be used for the inference of the treatment effect of the targeted test drug.

Simon and Maitournam (2004) and Maitournam and Simon (2005) provide sample size determination for the targeted clinical trials for both continuous and binary endpoints. However, variability associated with estimates of positive predictive value, negative predictive value, false positive rate, and false negative rate is not considered in the sample size calculation and relative efficiency of the targeted clinical trials to the untargeted ones. On the other hand, for example, getfitnib is the specific inhibitor of the tyrosine kinase of epidermal growth factor receptor (EGFR) that is involved in the pathway of the pathogenesis of non-small cell lung cancer (NSCLC). However, the response rate of getfitnib in the patients with NSCLC is only about 10%. In addition, for another EGFR inhibitor erlotinib, the survival of the patients with NSCLC is correlated significantly with expression, polysomy, amplification, and mutation of EGFR. Therefore, multiple pathways with multiple targets may be involved for most of diseases. Consequently, in the foreseeable future, it is very likely that a cocktail of molecularly targeted agents will be employed to treat the diseases with multiple targets. Therefore, research on the innovative and novel designs and analyses for the targeted clinical trials in evaluation of multiple drugs for multiple molecular targets is urgently needed.

6. DISCUSSION

Currently, the inclusion and exclusion criteria for clinical trials are based on some clinical signs and symptoms or their corresponding measurements. However, as more molecular targets of the diseases are identified, the more frequently expression profiles of the molecular targets become inclusion and exclusion criteria, e.g., HercepTest® in diagnosis of overexpression of the *HER2 neu* gene for the treatment of Herceptin® in patients with invasive breast cancer. Microarray platform is the breakthrough technology that can simultaneously measure the genome-wide expression profiles of the pathways involved with the pathogenesis of the disease. But translation of microarray technology to the diagnostic devices for the molecular targets in the treatment of the disease by the molecularly targeted agents still faces many challenges. Because the goal of genomic composite biomarker classifiers or IVDMIA is to treat the patients with the molecular target using the molecularly targeted drugs and not to treat the patients without the target with ineffective and unnecessary treatments, clinical validation is as equally important as analytical validation of IVDMIA.

One of critical issues for clinical validation is the definition and availability of the gold standard for diagnosis of the molecular targets used for evaluation of sensitivity, specificity, positive predictive value, false positive rate, and ROC curve. Some investigators use the classifiers derived from other quantitative gene expression platforms e.g., RT-PCR, as the gold standard. However, in essence, these platforms are not the gold standard and classification error may also occur using these technology platforms for diagnosis of the same molecular target. As a result, almost all the parameters concerning the diagnostic accuracy can not be estimated without the gold standard. Under the situation without a gold standard, one can only assess the agreement or equivalence in diagnosis of the molecular target (Liu et al., 2002). However, equivalence in diagnosis between the test IVDMIA and the reference classifier based on other technological platform implies that both are accurate or both are inaccurate in diagnosis of the molecular target.

For the clinical effectiveness trial of the IVDMIA of the diagnostic accuracy of the molecular target, the inclusion and exclusion criteria for the patients should be exactly the same as those for the targeted clinical trials for evaluation of the efficacy and safety of the molecularly targeted agents. In addition, all procedures of the test IVDMIA evaluated in the clinical effectiveness trials and used for diagnosis in the target clinical (utility) trials should be pre-specified in the protocols and should be the same methods derived from the development stage of the classifier such as sample collection, RNA extraction, cDNA/cRNA synthesis, dye labeling, hybridization, and scanning, normalization procedures, and thresholds. In addition, reproducibility for the correct diagnosis such as within- and between-laboratory agreement should be also evaluated in the clinical effectiveness trial of the IVDMIA.

In development of any classifier, the prevalence rate must be taken into consideration. For example, since the misclassification rate of the DLDF is a function of the prevalence rate, determination of thresholds also depends upon the prevalence rate. On the other hand, because the molecularly targeted agents are specific inhibitors of their targets and may induce a large treatment effect in the patients with the molecular targets, hence targeted clinical trials are in general more efficient than the untargeted trials (Maitournam and Simon, 2005; Simon and Maitournam, 2004). However, if the prevalence rate of the target in the patient population is low, the recruitment period of the targeted clinical trials will be much longer than the untargeted ones. In addition the positive predictive value is proportional to the prevalence rate. Therefore if the prevalence rate of the targets is below 0.01, then the positive predictive value will be below 0.5. From Equation (9), the treatment effect of the molecularly target will be seriously under-estimated. However, when the prevalence rate is 0.1 and above, the false positive rate will decrease to below 10%. In this case, bias still exists but with a moderate magnitude. Furthermore, similar to gender or age, the genomic composite biomarker classifier is also another variable with the expression profiles to stratify patients into subgroups with and without the molecular targets. If the prevalence rate of a certain target is low, the number of patients in this subgroup will be very low. If follows that it might take a very long time to recruit the patients and the trial might not have sufficient power to prove the effectiveness of the molecularly targeted agent even the targeted clinical trial is more efficient. As a result, prevalence rate is a determining factor for development of molecularly targeted treatments. But how low is the prevalence rate? Is the personalized medicine for a subgroup of one patient with his or her

distinct signature attainable? Does a cocktail of molecularly targeted agents for multiple targets represent a feasible approach to targeted therapy? These are just a few challenges that one must ponder for the development of diagnostic multivariate indeed assays and molecularly targeted therapy.

ACKNOWLEDGMENTS

This research is partially supported by the Taiwan National Science Council Grant: NSC95 2118-M-002-007-MY2 granted to the first author.

REFERENCES

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society, Series B* 57:289-300.

Canales, R. D., Luo, Y., Willey, J. C., et al. (2006). Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature Biotechnology* 24:1115–1122.

Casciano, D. A., Woodcock, J. (2006). Empowering microarrays in the regulatory setting. *Nature Biotechnology* 24:1103.

Chow, S. C., Liu, J. P. (2004). *Design and Analysis of Clinical Trials*. New York: John Wiley and Sons.

Dalton, W. S., Friend, S. H. (2006). Cancer Biomarkers—an invitation to the table. *Science* 312:1165–1168.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1997). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39:1–38.

Dobbin, K. K., Beer, D. G., Meyerson, M., Yeatman, T. J., Gerald, W. L., Jacobson, J. W., et al. (2005). Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clinical Cancer Research* 11:565–573.

Dudoit, S., Yang, Y. H., Callow, M. J., Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12:111–139.

FDA (2003). Draft Guidance on *Multiplex Tests for Heritable DNA Markers, Mutations and Expression Patterns*. The U.S. Food and Drug Administration. Rockville, Maryland.

FDA (2005). Draft concept paper on Drug-Diagnostic Co-Development. The U.S. Food and Drug Administration. Rockville, Maryland.

FDA (2006). Draft Guidance on In Vitro Diagnostic Multivariate Index Assays. The U.S. Food and Drug Administration. Rockville, Maryland.

Frueh, F. W. (2006). Impact of microarray data quality on genomic data submissions to the FDA. *Nature Biotechnology* 24:1105–1107.

Gelfand, A. E., Smith, A. F. M. (1990). Sampling-based approaches to calculating densities. *Journal of the American Statistical Association* 85:398–409.

Hochberg, Y., Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: Wiley.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., et al. (2005). Multi-laboratory comparison of microarray platforms. *Nature Methods* 2:345–349.

Ji, H., Davis, R. W. (2006). Data quality in genomics and microarray. *Nature Biotechnology* 24:1112–1113.

Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R., Quackenbush, J. (2005). Independence and reproducibility across microarray platforms. *Nature Methods* 2:337–343.

Li, C. R., Liao, C. T., Liu, J. P. (2006). On the exact interval estimation for the difference in paired areas under the ROC curves. *Statistics in Medicine*, Published on-line on December 1, 2006.

Liao, C. T., Lin, C. Y., Liu, J. P. (2006). Non-inferiority tests based on concordance correlation coefficient for assessment of the agreement for gene expression data from microarray experiments. *Journal of Biopharmaceutical Statistics* 17:309–327.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255–268.

Lin, L. I. (1992). Assay validation using the concordance correlation coefficient. *Biometrics* 48:599–604.

Lin, L. I., Hedayat, A. S., Sinha, B., Yang, M. (2002). Statistical methods in assessing agreement: models, issues, and tools. *Journal of the American Statistical Association* 97:257–270.

Liu, J. P., Hsueh, H. M., Hsieh, E., Chen, J. J. (2002). Tests for equivalence or non-inferiority for paired binary data. *Statistics in Medicine* 21:231–245.

Liu, J. P., Ma, M. C., Wu, C. Y., Tai, J. Y. (2006). Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves. *Statistics in Medicine* 25:1219–1238.

Maitournam, A., Simon, R. (2005). On the efficiency of targeted clinical trials. *Statistics in Medicine* 24:329–339.

MAQC Consortium (2006). The MAQC project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* 24:1151–1161.

Members of the Toxicogenomic Research Consortium (2005). Standardization of global gene expression analysis between laboratories and across platforms. *Nature Methods* 2:351–356.

MINDACT Design and MINDACT trial overview. (2006). http://www.breastinternationalgroup.org/transbig.html. Accessed on June 5, 2006.

Paik, S., Shak, S., Tang, G., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine* 351:2817–2826.

Paik, S., Tang, G., Shak, S., et al. (2006). Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *Journal of Clinical Oncology* 24:1–12.

Patterson, T. A., Lobenhofer, E. K., Fulmer-Smentek, S. B., et al. (2006). Performance comparison of one-color and two-color platforms with the MAQC project. *Nature Biotechnology* 24:1140–1150.

Reiser, B., Faraggi, D. (1997). Confidence intervals for the general ROC criterion. *Biometrics* 53:644–652.

Shippy, R., Fulmer-Smentek, S., Jensen, R. V., et al. (2006). Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nature Biotechnology* 24:1123–1131.

Simon, R., Maitournam, A. (2004). Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* 10:6759–6763.

Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., Zhao, Y. (2003). *Design and Analysis of DNA Microarray Investigations*. New York: Springer.

Sprarano, J., Hayes, D., Dees, E., et al. (2006). Phase III randomized study of adjuvant combination chemotherapy and hormonal therapy vs. adjuvant hormonal therapy alone in women with previously resected axillary node-negative breast cancer with various levels of risk for recurrence (TAILORX Trial). http://www.cancer.gov/clinicaltrials/ECOG- PACCT-1. Accessed on June 5 2006.

Su, J. Q., Liu, J. S. (1993). Linear combination of multiple diagnostic markers. *Journal of the American Statistical Association* 88:1350–1355.

Swain, S. M. (2006). A step in the right direction. *Journal of Clinical Oncology* 24(23):1–2.

Tong, W., Lucas, A. B., Shippy, R., et al. (2006). Evaluation of external RNA controls for the assessment of microarray performance. *Nature Technology* 24:1132–1139.

Tusher, V. G., Tibshirani, R., Chu, G. (2001). Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of National Academy of Sciences* 98:5116–5121.

van't Veer, L. J., Dai, H., van de Vijver, M. J., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530–536.

Van de Vijver, M. J., He, Y. D., van't Veer, L. J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* 347:1999–2009.

Varmus, H. (2006). The new era in cancer research. *Science* 312:1162–1165.

Wang, S., Ethier, S. (2004). A generalized likelihood ratio test to identify differentially expressed genes from microarray data. *Bioinformatics* 20:100–104.