Determination of Sample Size for Validation Study in Pharmacogenomics

Youlan Rao, *Millennium, The Takeda Oncology Company*
Yoonkyung Lee, *The Ohio State University*
Jason C. Hsu, *The Ohio State University*

Technical Report No. 834

January, 2010

Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210-1247

Determination of Sample Size for Validation Study in Pharmacogenomics

Youlan Rao,* Yoonkyung Lee,† and Jason C. Hsu†

*Millennium, The Takeda Oncology Company, Cambridge, MA 02139, U.S.A.

email: youlan.rao@mpi.com

†Department of Statistics, The Ohio State University, Columbus, OH 43210, U.S.A.

email: yklee@stat.osu.edu and jch@stat.osu.edu

Abstract

Pharmacogenomics aims at co-development of a drug that targets a subgroup of patients for safety and efficacy and a device that identifies the responder group through their genetic variations. Development of such a prognostic device includes a training stage and a validation stage. The transition from the training stage to the validation stage typically involves change of platforms as a subset of potential genetic markers predictive of drug response are identified in the first stage and only those are used in the second stage. With the change in consideration, this paper concerns how to determine sample sizes for the validation stage to meet pre-specified sensitivity and specificity requirements in order to avoid futility of pharmacogenomic development. In particular, taking microarrays as a medical device, which measure gene expression levels, we show how to decide the numbers of subjects per group, replicated samples per subject, replicated probes per gene for the validation experiment. The change of platforms is taken into account in the sample sizes calculation by statistical modeling. Our formulation of sensitivity and specificity requirements calls for estimation of both measures. Lower bounds with carefully calibrated confidence levels can give appropriate sample size to meet the requirements. The procedure is illustrated in a proof-of-concept mice experiment.

Key Words: Change of platforms; Futility; Microarray; Pharmacogenomics; Sample size; Sensitivity; Specificity.

1 Introduction

Pharmacogenomics aims at the co-development of a drug that targets a subgroup of the patients, and a device that predicts whether a patient is in the subgroup. This subgroup can be responders to the drug, or patients free of serious adverse events (SAE). Typically, subgroups are discovered by comparing the genetic profiles of patients with different phenotypes, responders versus non-responders, or patients free from SAE versus patients who experience SAE (Wang et al., 2007). Microarrays could be used to develop such a diagnostic device for identification of subgroups.

The development in pharmacogenomics includes two major stages: a training stage and a validation stage. The purpose of the training stage is to identify a biomarker positive (G+) subgroup of patients and its complement, the biomarker negative (G-) subgroup. Biomarkers historically refer to substances in biological samples or measurements that indicate a person's disease state or response to a drug (Baker, 2005). They are crucial for efficient drug development, and there are various types of biomarkers such as disease biomarkers, surrogate endpoints, efficacy or outcome biomarkers. Disease biomarkers indicate the presence or likelihood of a particular disease in patients. For example, gene expression profiles linked with cancer can be taken as a disease biomarker in pharmacogenomics. The purpose of the validation stage is then to prove that the biomarker found in the training stage has sufficient sensitivity and specificity for clinical use, and to independently validate the efficacy and safety of the drug for the target G+ subgroup.

Transition from the training stage to the validation stage typically involves a change of platform as a subset of potential predictors of drug response are identified as a biomarker in the first stage and only those are used in the second stage. For example, MammaPrint (van 't Veer et al., 2002) is an FDA-approved microarray-based test to predict the likelihood of recurrence of breast cancer. In the training stage of MammaPrint, microarrays probing approximately 25,000 genes were used, while in the second stage of validation, microarrays probing 70 genes only were used. Generally, fewer genes are involved in a diagnostic device for use, and this change allows more replication of probes and more replicated samples from a subject in the validation stage than the training stage. To ensure that the diagnostic algorithm derived from the training stage applies to expressions measured in the validation stage, the platform change needs to be taken into account when planning a validation study and computing sample sizes required for the study. This issue is often overlooked in the process of developing a diagnostic device. As stated in FDA (2005b),

> When validating a gene or expression pattern, instead of a set of individual

biomarkers, a rigorous statistical approach should be used to determine the number of samples, and the methodology used for validation. It is recommended that the validation strategy be discussed in advance with FDA.

At the end of the training stage, the sensitivity and specificity of the diagnostic algorithm for a validation trial need to be estimated. If both the sensitivity and specificity are significantly high, pharmacogenomic development is recommended to proceed. Otherwise, further pharmacogenomic development is likely to be futile.

Generally speaking, sample size calculation is formulated as a problem of determining the number of subjects in a prospective experiment, where a random variable of interest is measured for each of the subjects, and its distribution is modeled with some unknown parameter. The sample size is then calculated so that inferences and decisions about the parameter can be correctly made. It is customary to calculate sample size based on power (Adcock, 1997). That is, some hypotheses of interest are specified in terms of the parameter prior to the experiment, and then the sample size is determined to achieve a desired power at a fixed type I error rate. Chow et al. (2003) elucidate this statistical approach to sample size calculation and provide its justification for different objectives in various clinical trials settings.

Consider the example of *abacavir* (brand name Ziagen), a potent antiretroviral for HIV-1. About 8% of the patients treated with *abacavir* develop a serious adverse event (SAE) of hypersensitive reaction. Mallal et al. (2002) reported on a retrospective study of *abacavir*-treated patients, looking for biomarkers in the HLA region that can screen out patients prone to this SAE. Three markers (HLA-B*5701, HLA-DR7, HLA-DQ3) were found to be highly associated with the occurrence of the hypersensitivity SAE. Martin et al. (2004) in a follow-up to the 2002 study also reported that five markers (HLA-B*5701, C4A6, HLA-DRB1*0701, HLA-DQ3, Hsp70-Hom M493T) are highly associated with the hypersensitive reaction. In accordance with the pharmacogenomic concept stated in FDA (2005a,b), a double-blind randomized study involving 1,956 patients from 19 countries was then conducted to validate the HLA-B*5701 biomarker (Hughes et al., 2008; Mallal et al., 2008). Their sample size for the validation study was calculated by specifying the power to detect a 4% drop in the SAE from the control group to the prospectively screened group. In addition to a significant reduction in the SAE rate, they reported 95% lower confidence bounds of 85.2% for sensitivity and 95.5% for specificity.

For multiple hypothesis testing with several parameters, appropriate definitions of power, type I error rate and the corresponding statistical test for the hypotheses of interest are necessary. Lee (2004) calculates sample size required for microarray experiments in which

finding differentially expressed genes between a treatment condition and a control condition is of interest. In her approach, the number of subjects is determined to achieve a desired individual power level for a given mean number of false positives of multiple hypotheses (per-familywise type I error rate), when the ratio of mean difference to standard deviation for individual hypotheses and the anticipated number of undifferentially expressed genes are specified in advance. Depending on the purpose of experiments, sample size can also be specified to obtain optimal confidence regions for multiple hypothesis testing. In particular, for multiple comparisons using Tukey's Multiple Pairwise Comparisons (MCA) method, the constrained Multiple Comparisons with the Best (MCB) method, and Dunnett's two-sided Multiple Comparisons with a Control (MCC) method, Hsu (1988) suggests to calculate sample size so that with a pre-specified probability the confidence intervals for mean differences cover the true parameter values and be sufficiently narrow.

For development of prognostic or diagnostic devices, it is sensible to determine sample size to achieve specified levels of sensitivity and specificity, as they are common measures of prediction accuracy for diagnostic rules. Pepe (2003) uses this strategy to calculate sample size in the context of developing a medical device to differentiate a diseased group from non-diseased group. In her approach, a device is considered having minimally acceptable performance when the true positive fraction (TPF) is at least some value, say TPF_0 and the false positive fraction (FPF) is at most, say FPF_0. Note that the true positive fraction is sensitivity and the false positive fraction is one minus specificity. By setting a hypothesis to statistically prove that the device is minimally acceptable, that is, setting $\text{H}_0 : \text{TPF} \leq \text{TPF}_0$ or $\text{FPF} \geq \text{FPF}_0$, the sample sizes for the two groups are chosen so that a positive conclusion would be drawn with a desired power at a specified type I error rate when the true TPF and FPF of the device are at some levels, TPF_1 and FPF_1. Here, $\text{TPF}_1(> \text{TPF}_0)$ and $\text{FPF}_1(< \text{FPF}_0)$ are the values in the alternative hypothesis specified in advance by researchers.

Determination of sample size for validation study in pharmacogenomics is more complicated than the traditional sample size problem since it involves multiple layers of sample sizes and device-specific parameters. To design a microarray experiment for a validation study, in addition to the number of subjects per group, the number of replicate samples for each subject, the number of replicates of each probe on the chip, the number of genes, and the number of probes for each gene need to be decided.

We propose in this paper that the sample sizes and device-specific parameters be determined so that the probability of sensitivity and specificity estimators being greater than some minimally acceptable values is sufficiently high. To outline our proposed procedure,

the number of genes as part of device-specific parameters is determined first by controlling a familywise type I error rate in multiple hypothesis testing for gene selection. Alternative gene selection procedures can be used if desired. A diagnostic rule is then built on the basis of the selected genes. Other device-specific parameters and the sample sizes are then calculated to meet pre-specified minimal sensitivity and specificity requirements for the diagnostic rule. Since the number of probes for each gene and the number of replicates of each probe are generally subject to spatial limitations of microarray platforms, typically there are upper bounds on theses parameters, depending on the number of selected genes. A possible range of the number of replicated samples for each subject may be limited as well because the volume of a biological sample drawn from each subject is finite.

The aforementioned change of platforms in the pharmacogenomic setting brings another complication that the distributions of training data and validation data are not the same. Only when the training and validation experiments are properly designed, as in the proof-of-concept mice experiment to be shown later, the change of the distributions between two stages can be handled appropriately through statistical modeling. Otherwise, it does not seem feasible to deal with the change in a statistically principled way.

Our formulation of sample size determination requires sensitivity and specificity estimates. We propose to use their confidence lower bounds based on the model for training and validation data. Compared to Pepe (2003)'s approach where TPF_1 and FPF_1 are prespecified, the proposed approach of using a confidence lower bound is more data-adaptive, and it also controls the probability of meeting the minimal sensitivity or specificity requirement directly. One may also use a model-based plug-in estimator of sensitivity and specificity. However, it will be shown that sample sizes calculated on the basis of a plug-in estimator may be too optimistic in some cases. The bias can be remedied by a confidence lower bound.

Section 2 describes the mathematical formulation and general steps for sample sizes calculation in pharmacogenomics. For simplicity, how to determine sample sizes for sensitivity requirement is discussed only. Specificity requirement can be similarly dealt with. Section 3 illustrates an application of the procedure in the proof-of-concept mice experiment. Section 4 presents a simulation study for numerical validation of the proposed sample sizes determination procedure followed by conclusions in Section 5.

2 Formulation of Sample Size Calculation

2.1 Basic Setting for Training and Validation Data

Suppose that there are K potential predictors of drug response in a training stage of a pharmacogenomic study. Let $X \in R^K$ be a random vector of the measurements of the predictors from a subject and $Y \in \{0, 1\}$ be an indicator of whether the subject is a responder $(Y = 1)$ or a non-responder $(Y = 0)$. Let $\mathcal{D}_n = \{(x_i, y_i)|i = 1, 2, \ldots, n\}$ denote the data with n subjects in the training stage, where x_i and y_i are the observed K covariates and subgroup label of the ith subject. The n observations (x_i, y_i) are assumed to be independent and identically distributed with some unknown distribution $P_{(X,Y)}$.

For validation data, let (X^*, Y^*) denote a new case in the second stage, where $X^* \in R^k$ is a k-dimensional random vector, and $Y^* \in \{0, 1\}$ is a subgroup label. Since only a subset of potential predictors of drug response are identified as a biomarker in the first stage and used in the second stage, k is typically much smaller than K. Let $Q_{(X^*,Y^*)}$ be the distribution of (X^*, Y^*), which is different from $P_{(X,Y)}$ because of the change of the platform. If microarrays are used to develop a diagnostic device, more replication of probes and more replicated samples from a subject are allowed in the validation stage. This change yields smaller variance parameters in $Q_{(X^*,Y^*)}$ than $P_{(X,Y)}$. A concrete example of such a change from $P_{(X,Y)}$ to $Q_{(X^*,Y^*)}$ is given in Section 3.

To take into account the change of platforms in a statistically principled way, we consider a model for the training data and derive a theoretically optimal diagnostic rule that predicts subgroup labels from the model for validation data (yet to be observed), reflecting the corresponding change.

Let $\phi_{\mathbf{V}}(x^*; \mathcal{D}_n)$ denote such a theoretically optimal rule for the validation data, which depends on some unknown model parameters. The subscript v is used to emphasize that the rule is for validation data. Use $\hat{\phi}_{\mathbf{V}}(x^*; \mathcal{D}_n)$ to denote the plug-in diagnostic rule with the model parameters replaced with estimates from the training data $\mathcal{D}_n$. In the validation stage, these parameter estimates are held fixed, and thus $\hat{\phi}_{\mathbf{V}}(x^*; \mathcal{D}_n)$ is considered fixed. The sensitivity of $\hat{\phi}_{\mathbf{V}}(x^*; \mathcal{D}_n)$ is then defined as the probability of correctly calling a subject a responder given the subject is a responder:

$$Sen(\hat{\phi}_{\mathbf{V}}) := P(\hat{\phi}_{\mathbf{V}}(X^*; \mathcal{D}_n) = 1|Y^* = 1).$$

For brevity, $Sen_{\mathbf{V}}$ is used to refer to $Sen(\hat{\phi}_{\mathbf{V}})$ in this paper.

Suppose that given a validation sample of m i.i.d pairs of $(X_j^*, Y_j^*), j = 1, 2, \ldots, m$ with $Y_j^* = 1$, the true sensitivity of $\hat{\phi}_{\mathbf{V}}$, $Sen_{\mathbf{V}}$, is estimated by a simple unbiased estimator

$\widehat{Sen}_{\mathbf{V}}$, the sample proportion of correctly predicting true positives by the diagnostic rule $\hat{\phi}_{\mathbf{V}}$ over the validation data. That is,

$$\widehat{Sen}_{\mathbf{V}} = \frac{1}{m}\sum_{j=1}^{m} I(\hat{\phi}_{\mathbf{V}}(X_j^*; \mathcal{D}_n) = 1 | Y_j^* = 1). \tag{1}$$

Note that $m \cdot \widehat{Sen}_{\mathbf{V}}$ follows a binomial distribution $B(m, Sen_{\mathbf{V}})$.

2.2 Minimal Sensitivity Requirement

Now consider the problem of finding the number of subjects m in the responder group for a validation study such that the probability of the estimated sensitivity $\widehat{Sen}_{\mathbf{V}}$ exceeding a pre-specified minimum level of sensitivity γ is at least $1-\beta$, that is,

$$P(\widehat{Sen}_{\mathbf{V}} \geq \gamma) \geq 1 - \beta. \tag{2}$$

If $\widehat{Sen}_{\mathbf{V}} \geq \gamma$, we say that the validation study is successful.

Figure 1 shows that given β and the true sensitivity $Sen_{\mathbf{V}}(> \gamma)$, the probability of successful validation $P(\widehat{Sen}_{\mathbf{V}} \geq \gamma)$ oscillates as a function of m and eventually surpasses $1-\beta$ as m increases. As shown in the figure, there can be more than one crossing points of $P(\widehat{Sen}_{\mathbf{V}} \geq \gamma)$ and $1-\beta$, but for large enough m, the inequality (2) holds. Hence, it is sensible to define the desired sample size m^* as the smallest number of subjects in the subgroup such that for any $m \geq m^*$, the sensitivity requirement (2) is satisfied, given the minimal sensitivity γ and the minimal probability of successful validation $1-\beta$. The following proposition shows that as long as the true sensitivity is greater than the minimal level γ, m^* is well-defined.

Proposition 1 *If $Sen_{\mathbf{V}} > \gamma$, then the sample size m^* is finite.*

Proof Since $m \cdot \widehat{Sen}_{\mathbf{V}}$ follows $B(m, Sen_{\mathbf{V}})$, by Hoeffding's inequality (Hoeffding, 1963),

$$P(m \cdot \widehat{Sen}_{\mathbf{V}} \leq m\gamma) \leq \exp(-2\frac{(m \cdot Sen_{\mathbf{V}} - m\gamma)^2}{m}) \text{ for } m\gamma < m \cdot Sen_{\mathbf{V}}.$$

So, if the probability upper bound is at most β, then (2) is satisfied, and the condition

$$\exp(-2\frac{(m \cdot Sen_{\mathbf{V}} - m\gamma)^2}{m}) \leq \beta$$

implies

$$m \geq -\frac{\log \beta}{2(Sen_{\mathbf{V}} - \gamma)^2}. \tag{3}$$

Thus m^* must be finite. □

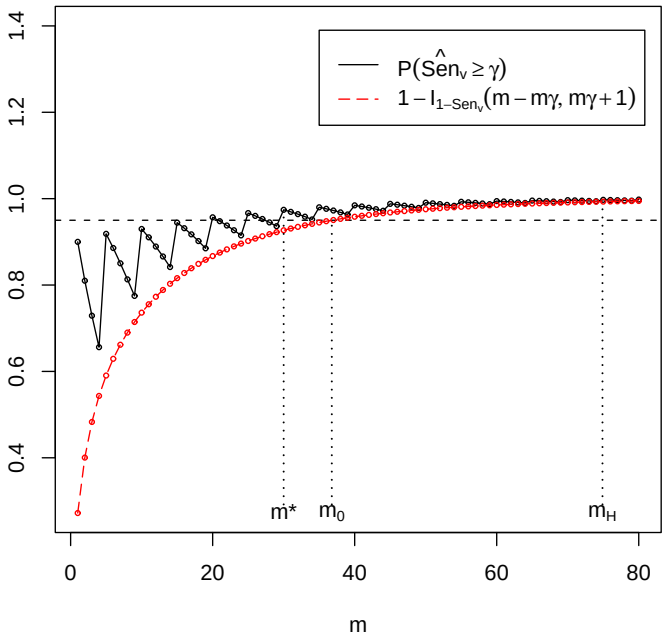

Figure 1: The probability $P(\widehat{Sen}_{\mathbf{V}} \geq \gamma)$ as a function of m and its lower bound $1 - I_{1-Sen_{\mathbf{V}}}(m-m\gamma, m\gamma+1)$ when the true sensitivity $Sen_{\mathbf{V}}$ is 0.90 and the minimal sensitivity level γ is 0.80. The horizontal dashed line indicates the minimal probability of successful validation $1-\beta$, with β fixed at 0.05. m_H is the upper bound of sample size given by Hoeffding's inequality in (3), m_0 is the root of $I_{1-Sen_{\mathbf{V}}}(m - m\gamma, m\gamma + 1) = \beta$, and m^* is the desired sample size. In this example, $m_H = 74.9$, $m_0 = 36.8$ and $m^* = 30$.

Let m_H denote the upper bound given by Hoeffding's inequality in (3). We can find m^* easily by backward search; starting from $\lceil m_H \rceil$, where $\lceil x \rceil$ is the ceiling of x, i.e. the smallest integer greater than or equal to x, and decreasing the integer by one each time until we reach the first integer for which the inequality (2) is not satisfied. However, the upper bound m_H tends to be very large, and hence this algorithm may not be efficient.

To sharpen the upper bound of m^*, we consider a continuous lower envelope of $P(\widehat{Sen}_{\mathbf{V}} \geq \gamma)$ by using the relationship between the cdf of binomial distribution and the regularized incomplete beta function. First observe that

$$\begin{aligned} P(m \cdot \widehat{Sen}_{\mathbf{V}} \geq m\gamma) &\geq P(m \cdot \widehat{Sen}_{\mathbf{V}} > \lfloor m\gamma \rfloor) \\ &= 1 - P(m \cdot \widehat{Sen}_{\mathbf{V}} \leq \lfloor m\gamma \rfloor - 1) \\ &= 1 - I_{1-Sen_{\mathbf{V}}}(m - \lfloor m\gamma \rfloor, \lfloor m\gamma \rfloor + 1), \end{aligned}$$

where $\lfloor x \rfloor$ is the floor of x. The last equality comes from the fact that for a binomial random variable X with $B(n,p)$, $P(X \leq k) = I_{1-p}(n-k, k+1)$ by integration by parts, where $I_x(a,b)$ is the cdf of $Beta(a,b)$ at x, also known as the regularized incomplete beta function. Since a and b in $I_x(a,b)$ can be real values, and

$$1 - I_{1-Sen_{\mathbf{V}}}(m - \lfloor m\gamma \rfloor, \lfloor m\gamma \rfloor + 1) \geq 1 - I_{1-Sen_{\mathbf{V}}}(m - m\gamma, m\gamma + 1), \tag{4}$$

by allowing m to be a real value, we obtain the right hand side of (4) as a continuous lower envelope of $P(\widehat{Sen}_{\mathbf{V}} \geq \gamma)$. The dashed line in Figure 1 depicts such a lower envelope.

For $Sen_{\mathbf{V}} \geq \gamma$, the function $1 - I_{1-Sen_{\mathbf{V}}}(m - m\gamma, m\gamma + 1)$ is shown to be strictly increasing in m. By equating the function to $1-\beta$ and solving for m, we get a unique solution m_0, which serves as an upper bound of m^*. That is, for any integer $m \geq \lceil m_0 \rceil$, the minimal sensitivity requirement (2) is satisfied. m_0 is usually much smaller than m_H as illustrated in Figure 1. With this smaller initial point, $\lceil m_0 \rceil$, the aforementioned backward search algorithm can be made more efficient. In summary, we start from $\lceil m_0 \rceil$ and decrease m by one until the inequality (2) does not hold. Then the m value for the second last step is m^*, the number of subjects for the responder group needed to meet the minimal sensitivity requirement. This algorithm is implemented in R and available at the second author's webpage.

Figure 2 illustrates how the sample size determined by the algorithm varies depending on the underlying true sensitivity when the probability of successful validation is 95%. Expectedly, the necessary sample size m^* decreases as the true sensitivity increases, and a larger sample size is required as the minimal level γ increases.

2.3 Estimation of True Sensitivity

The true sensitivity of the diagnostic rule $\hat{\phi}_{\mathbf{V}}$, i.e. $Sen_{\mathbf{V}}$ for the inequality (2), is usually unknown in practice, and it has to be estimated. As the diagnostic rule is derived from a model adjusting for the change of platforms, its true sensitivity can be estimated by plugging in estimated model parameters. However, it is well known that such a plug-in estimator exhibits an upward bias in estimating the true sensitivity, and hence the sample size calculated based on the estimator could be smaller than necessary. Taking a conservative approach, we propose to determine sample sizes by replacing $Sen_{\mathbf{V}}$ by its confidence lower bound.

For fixed m, the binomial distribution $B(m, Sen_{\mathbf{V}})$ is stochastically increasing in $Sen_{\mathbf{V}}$. So, if $P(X \geq m\gamma) \geq 1-\beta$ for a binomial variable X of $B(m,p)$ with $p < Sen_{\mathbf{V}}$, then (2) also holds because $m \cdot \widehat{Sen}_{\mathbf{V}}$ is stochastically greater than X. This is the rationale for

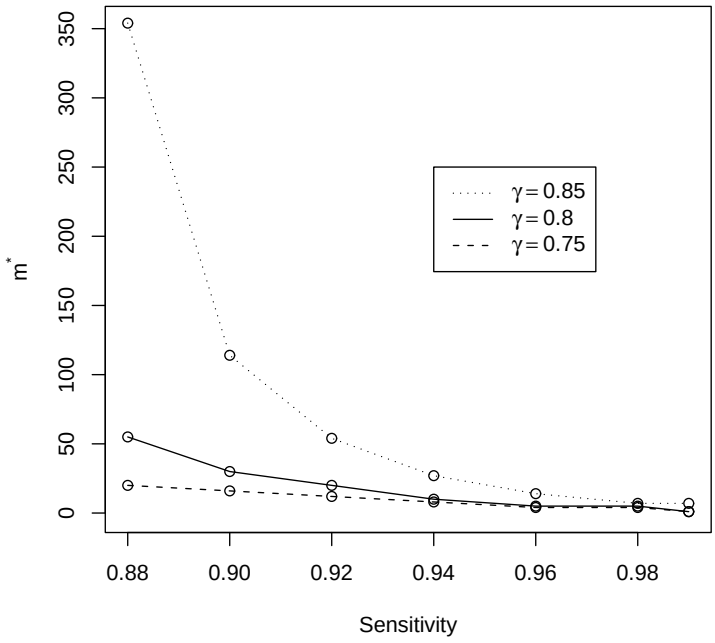

Figure 2: Sample sizes m^* necessary to meet the minimal sensitivity requirement of $\gamma = 75\%$, 80% and 85%, respectively, with 95% of confidence as the true sensitivity varies.

replacing the true sensitivity with its confidence lower bound. Note that as in Proposition 1, for a proper sample size m^*, the confidence lower bound has to be greater than γ. This condition provides a statistical criterion for futility in the setting of pharmacogenomic development.

3 Icelandic Mice Data Analysis

Hsu et al. (2009) describes an Icelandic mice experiment to discriminate five mouse strains. Four strains (conveniently labeled groups A, B, C, and D) have different mutations in Mitf (Microphthalmia transcription factor) gene and one strain (labeled group W) is a wild type. In the experiment, the five strains are differentiated on the basis of the expression levels of 99 genes regulated by Mitf gene in spleen tissues. To mimic the situation in a pharmacogenomic study, consider classifying mice from two strains, say group A and group W. Regard group A as the responder group $(Y = 1)$ and group W as the non-responder group

$(Y = 0)$. The 99 genes are taken as the potential genetic markers, from which a subset of genes predictive of responder or non-responder will be selected. This subset of genes are considered a biomarker. A diagnostic algorithm $\hat{\phi}_{\mathbf{V}}$ is then built based on these selected genes. It classifies mice into two groups: the biomarker positive subgroup ($\hat{\phi}_{\mathbf{V}} = 1$) and the biomarker negative subgroup ($\hat{\phi}_{\mathbf{V}} = 0$). The sensitivity of the algorithm is then the probability of predicting a mouse biomarker positive given the mouse is in group A. The specificity is the probability of predicting a mouse biomarker negative given the mouse is in group W.

Custom-designed NimbleGen microarrays were used in the training stage of the experiment. Each array had 12 mini-arrays, and the 99 genes were probed in each mini-array. For each gene, there were 32 probes, and each probe had two replicates. Design of the experiment followed randomization, replication and blocking principles, with four mice (subjects) per group, and four replicated biological samples per mouse. See Hsu et al. (2009) for details of the experimental design. Taking the training data from this experiment, we demonstrate the proposed procedure for determining sample sizes to achieve desired precision of sensitivity in a validation trial.

3.1 Construction of Discriminant Rule

For statistical modeling of the data, let x_{igmspr} denote the background-corrected, log transformed and normalized probe intensity of a mutated or wild type mouse for the ith gene $(i = 1, 2, \ldots, 99)$, the sth sample ($s = 1, 2, 3, 4$) for the mth mouse $(m = 1, 2, 3, 4)$ in group g $(g = A, B, C, D, W)$, the pth probe $(p = 1, 2, \ldots, 32)$, and the rth replicate $(r = 1, 2)$. The probe intensities can be modeled separately for each gene. Assume x_{igmspr} to follow a linear mixed effect model:

$$x_{igmspr} = \mu_i + \tau_{ig} + M_{im(g)} + S_{is(m(g))} + \pi_{p(i)} + \epsilon_{igmspr}, \tag{5}$$

where μ_i is the mean gene expression for the ith gene, τ_{ig} is the gth group effect on the ith gene, $M_{im(g)}$ is the mth subject effect in the gth group on the ith gene, $S_{is(m(g))}$ is the sth sample effect from the mth subject in the gth group on the ith gene, and $\pi_{p(i)}$ is the pth probe effect in the ith gene. We assume that ϵ_{igmspr} are independent and identically distributed with $N(0, \sigma^2_{i\epsilon})$ within each gene. For the subject effects, we assume that $M_{im(g)}$ are independent and identically distributed with $N(0, \sigma^2_{iM})$ regardless of the group. Similarly for the sample effects, $S_{is(m(g))}$ are assumed to be independent and identically distributed with $N(0, \sigma^2_{iS})$ regardless of the subject and group. The ϵ_{igmspr}, $M_{im(g)}$ and $S_{is(m(g))}$ are also assumed to be independent.

Treating the problem in a general setting, suppose that there are n_i genes, n_m subjects in each group, n_p probes for each gene, n_s replicated samples for each subject, and n_r replicates of each probe on the chip. The sample sizes n_m and n_s, and the device-specific parameters n_i, n_p and n_r can be different from the training stage to a validation stage. The linear mixed model in (5) now with these unspecified sample sizes and parameters n_m, n_s, n_p and n_r can characterize clearly the transition from $P_{(X,Y)}$ to $Q_{(X^*,Y^*)}$ due to the change of platforms. Hereafter, the superscripts t and v are used to indicate the change; t for training and v for validation. In the mice experiment, $n_m^t = 4$, $n_s^t = 4$, $n_i^t = 99$, $n_p^t = 32$, and $n_r^t = 2$.

Consider the normalized gene expression data from the mutant group A and the wild type group W. Genes that seem reasonably good in separating A from W were found by the average mean differences, and multiplicity adjustment was done by the resampling-based partitioning test procedure described in Hsu et al. (2009). The selected genes are RB1, USF1, Pu.1, Oa1, TPA1 and Bim, and their indices are 42, 45, 31, 29, 67, and 22, respectively. These six genes are used to build a diagnostic algorithm for discriminating the group A from the wild type W.

For simplicity, taking each mouse as a sampling unit, we consider prediction rules that use the average of probe intensities across the biological samples from the same mouse, and replicates as a summary measure of expression for each gene. The linear mixed effect model for the training experiment implies that the distribution of the average $\bar{X}^t_{igm\ldots}$ is

$$N(\mu_i + \tau_{ig}, \sigma^2_{iM} + \sigma^2_{iS}/n^t_s + \sigma^2_{i\epsilon}/(n^t_s n^t_p n^t_r)). \tag{6}$$

Suppose that a validation experiment has n^v_m subjects (mice) in each group, n^v_p probes for each gene, n^v_s replicated samples for each subject, and n^v_r replicates of each probe. If validation data are obtained under the same probabilistic mechanism as the training data other than the sample sizes and device-specific parameters, then the distribution of the (unobserved) validation data $\bar{X}^v_{igm\ldots}$ is given as

$$N(\mu_i + \tau_{ig}, \sigma^2_{iM} + \sigma^2_{iS}/n^v_s + \sigma^2_{i\epsilon}/(n^v_s n^v_p n^v_r)). \tag{7}$$

For convenience, the selected genes are relabeled so that their indices are from 1 to n^v_i, the number of genes used in the array for validation. In the mice experiment, $n^v_i = 6$. As in Section 2.1, let (X^*, Y^*) be a random vector for validation data. Then in this setting X^* consists of n^v_i components, and it is defined such that given $Y^* = g$, the distribution of $X^* := (X^*_1, X^*_2, \ldots, X^*_{n^v_i})^\top$ is the same as the distribution of $(\bar{X}^v_{1gm\ldots}, \bar{X}^v_{2gm\ldots}, \ldots, \bar{X}^v_{n^v_i gm\ldots})^\top$.

If we further assume independence across the selected genes, the optimal classification rule under the normal distribution setting is given by

$$\phi_{\mathbf{V}}(x^*) = \arg\min_g \sum_{i=1}^{n_i^v} \frac{(x_i^* - (\mu_i + \tau_{ig}))^2}{\sigma_{iM}^2 + \sigma_{iS}^2/n_s^v + \sigma_{i\epsilon}^2/(n_p^v n_s^v n_r^v)},$$

which is known as the diagonal linear discriminant analysis (DLDA). The decision rule $\phi_{\mathbf{V}}$ can be estimated by a plug-in rule,

$$\hat{\phi}_{\mathbf{V}}(x^*) = \arg\min_g \sum_{i=1}^{n_i^v} \frac{(x_i^* - (\hat{\mu}_i + \hat{\tau}_{ig}))^2}{\hat{\sigma}_{iM}^2 + \hat{\sigma}_{iS}^2/n_s^v + \hat{\sigma}_{i\epsilon}^2/(n_p^v n_s^v n_r^v)},$$

where $\hat{\mu}_i$, $\hat{\tau}_{ig}$, $\hat{\sigma}_{iM}$, $\hat{\sigma}_{iS}$, and $\hat{\sigma}_{i\epsilon}$ are estimates of μ_i, τ_{ig}, σ_{iM}, σ_{iS}, and $\sigma_{i\epsilon}$ from the linear mixed effect model (5) for the training data. We note that the estimates of the variance components are based on the observations from all five strains although the rule of our main interest is concerned with discriminating group A ($Y = 1$) from group W ($Y = 0$).

3.2 Sample Sizes Determination

In designing a validation experiment with microarrays, possible values of the device-specific parameters, the number of probes for each gene n_p^v, the number of replicates of each probe n_r^v and the size of replicated samples for each subject n_s^v are restricted due to the spatial and biological limitations mentioned in Section 1. In this mice experiment, n_p^v is supposed to be the same as $n_p^t = 32$, as we believe that the 32 probes chosen by a biologist are sufficiently sensitive in measuring the expression levels of each gene. Since fewer genes are probed in each microarray for the validation stage than the training stage (from $n_i^t = 99$ to $n_i^v = 6$), this change allows more space for replication of each probe. With 200 spots in each mini-array, n_r^v can be increased from $n_r^t = 2$ up to 33. For each mouse, its spleen tissue sample can only be divided into 4 to 6 pieces of replicated samples due to the fact that enough amount of a biological sample is required for efficient hybridization. This gives a range of 4 to 6 for n_s^v.

Given n_p^v, n_r^v, and n_s^v, consider calculation of the number of subjects n_m^v needed for validation of $\hat{\phi}_{\mathbf{V}}$ as a binary decision rule such that the sensitivity requirement (2) is met. In other words, we determine the sample size n_m^v so that $P(\widehat{Sen}_{\mathbf{V}} \geq \gamma) \geq 1 - \beta$, where $n_m^v \widehat{Sen}_{\mathbf{V}}$ follows $B(n_m^v, Sen(\hat{\phi}_{\mathbf{V}}))$ with $\hat{\phi}_{\mathbf{V}}$ depending on n_s^v, n_p^v, n_r^v, and the training data.

Under the assumption for the validation data in (7), direct calculation shows that the true sensitivity of $\hat{\phi}_{\mathbf{V}}$ is given by

$$Sen(\hat{\phi}_{\mathbf{V}}) = \Phi\left(\frac{(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)^\top \hat{\Sigma}_v^{-1}\left[\boldsymbol{\theta}_1 - \frac{1}{2}(\hat{\boldsymbol{\theta}}_0 + \hat{\boldsymbol{\theta}}_1)\right]}{\left[(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)^\top \hat{\Sigma}_v^{-1}\Sigma_v \hat{\Sigma}_v^{-1}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)\right]^{1/2}}\right), \tag{8}$$

where $\boldsymbol{\theta}_g = (\mu_1 + \tau_{1g}, \ldots, \mu_{n_i^v} + \tau_{n_i^v g})^\top$ for $g = 0, 1$, $\Sigma_v = \text{diag}(\sigma_{iM}^2 + \sigma_{iS}^2/n_s^v + \sigma_{i\epsilon}^2/(n_p^v n_s^v n_r^v),\ i = 1, 2, \ldots, i_n^v)$, and $\hat{\boldsymbol{\theta}}_g$ and $\hat{\Sigma}_v$ are plug-in estimates of $\boldsymbol{\theta}_g$ and Σ_v. To determine the sample size n_m^v, the true sensitivity of $\hat{\phi}_{\mathbf{V}}$ needs to be estimated. A simple plug-in estimator is

$$\widehat{Sen}_{\mathbf{V}}(\hat{\phi}_{\mathbf{V}}) = \Phi\left(\frac{(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)^\top \hat{\Sigma}_v^{-1}\left[\hat{\boldsymbol{\theta}}_1 - \frac{1}{2}(\hat{\boldsymbol{\theta}}_0 + \hat{\boldsymbol{\theta}}_1)\right]}{\left[(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)^\top \hat{\Sigma}_v^{-1}\hat{\Sigma}_v \hat{\Sigma}_v^{-1}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)\right]^{1/2}}\right) = \Phi\left(\frac{\hat{\delta}}{2}\right), \tag{9}$$

where $\hat{\delta}$ is a sample version of the Mahalanobis distance δ between the two normal distributions given as

$$\{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^\top \Sigma_v^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)\}^{1/2} = \left(\sum_{i=1}^{n_i^v} \frac{(\tau_{i1} - \tau_{i0})^2}{\sigma_{iM}^2 + \sigma_{iS}^2/n_s^v + \sigma_{i\epsilon}^2/(n_p^v n_s^v n_r^v)}\right)^{1/2}. \tag{10}$$

3.3 Confidence Lower Bound for Sensitivity

The simple plug-in estimator in (9) is generally observed to be biased upward, often yielding an optimistic estimate (McLachlan, 1992; Efron, 1983). Alternatively, we can use a confidence lower bound of $Sen(\hat{\phi}_{\mathbf{V}})$ with appropriately chosen level $1 - \alpha$. The effect of α on the sample sizes is investigated numerically in Section 4. Based on the numerical results, α will be calibrated to attain the desired level $1 - \beta$ for the probability of meeting the sensitivity requirement in (2).

Letting $\boldsymbol{b} = \hat{\Sigma}_v^{-1}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)$ and $\boldsymbol{a} = \frac{1}{2}(\hat{\boldsymbol{\theta}}_0 + \hat{\boldsymbol{\theta}}_1)$, which are held fixed in the validation stage, we have the true sensitivity of $\hat{\phi}_{\mathbf{V}}$ in (8) expressed as

$$Sen(\hat{\phi}_{\mathbf{V}}) = \Phi\left(\frac{\boldsymbol{b}^\top(\boldsymbol{\theta}_1 - \boldsymbol{a})}{\sqrt{\boldsymbol{b}^\top \Sigma_v \boldsymbol{b}}}\right). \tag{11}$$

To construct a confidence lower bound for $Sen(\hat{\phi}_{\mathbf{V}})$, it is sufficient to construct a confidence lower bound of

$$\eta_{a,b} := \frac{\boldsymbol{b}^\top(\boldsymbol{\theta}_1 - \boldsymbol{a})}{\sqrt{\boldsymbol{b}^\top \Sigma_v \boldsymbol{b}}}$$

since $\Phi(\cdot)$ is an increasing function.

To derive a confidence lower bound of $\eta_{a,b}$ given $\boldsymbol{a}$ and $\boldsymbol{b}$, we first consider an estimator of the form

$$\hat{\eta}_{a,b} = \frac{\boldsymbol{b}^\top(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{a})}{\sqrt{\boldsymbol{b}^\top \hat{\Sigma}_v \boldsymbol{b}}}.$$

For the numerator, $\hat{\boldsymbol{\theta}}_1$ follows $N(\boldsymbol{\theta}_1, \Sigma_t/n_m^t)$, where Σ_t is a diagonal matrix with entries $\sigma_{iM}^2 + \sigma_{iS}^2/n_s^t + \sigma_{i\epsilon}^2/(n_p^t n_s^t n_r^t)$, $i = 1, 2, \ldots, n_i^v$. For the denominator,

$$\boldsymbol{b}^\top \hat{\Sigma}_v \boldsymbol{b} = \sum_{i=1}^{n_i^v} b_i^2 (\hat{\sigma}_{iM}^2 + \hat{\sigma}_{iS}^2/n_s^v + \hat{\sigma}_{i\epsilon}^2/(n_p^v n_s^v n_r^v)).$$

From the AVOVA table for the linear mixed effect model (5), we have the following expected mean squares:

$$\begin{aligned} E(MSE_i) &= \sigma_{i\epsilon}^2, \\ E(MSS(GM)_i) &= \sigma_{iS}^2 n_g^t n_p^t n_r^t + \sigma_{i\epsilon}^2, \\ E(MSM(G)_i) &= \sigma_{iM}^2 n_g^t n_s^t n_p^t n_r^t + \sigma_{iS}^2 n_g^t n_p^t n_r^t + \sigma_{i\epsilon}^2, \end{aligned}$$

where MSE_i, $MSS(GM)_i$, and $MSM(G)_i$ are the mean squares for ϵ_i, $S_{i(m(g))}$, and $M_{i(g)}$, respectively. The variance components σ_{iM}^2, σ_{iS}^2, and $\sigma_{i\epsilon}^2$ are then estimated by the method of moment (Ravishanker and Dey, 2002). As a result, we have

$$\begin{aligned} \hat{\sigma}_{i\epsilon}^2 &= MSE_i, \\ \hat{\sigma}_{iS}^2 &= \frac{MSS(GM)_i - MSE_i}{n_g^t n_p^t n_r^t}, \\ \hat{\sigma}_{iM}^2 &= \frac{MSM(G)_i - MSS(GM)_i}{n_g^t n_s^t n_p^t n_r^t}. \end{aligned} \tag{12}$$

Hence

$$\boldsymbol{b}^\top \hat{\Sigma}_v \boldsymbol{b} = \sum_{i=1}^{n_i^v} b_i^2 (c_M MSM(G)_i + c_S MSS(GM)_i + c_E MSE_i),$$

where $c_M = \frac{1}{n_g^t n_s^t n_p^t n_r^t}$, $c_S = \frac{1}{n_g^t n_s^v n_p^t n_r^t} - \frac{1}{n_g^t n_s^t n_p^t n_r^t}$ and $c_E = \frac{1}{n_s^v n_p^v n_r^v} - \frac{1}{n_g^t n_s^v n_p^t n_r^t}$.

Since

$$\begin{aligned} MSE_i &\sim \frac{\sigma_{i\epsilon}^2}{df_E} \chi_{df_E}^2 \text{ with } df_E = n_g^t n_m^t n_s^t n_p^t (n_r^t - 1), \\ MSS(GM)_i &\sim \frac{\sigma_{iS}^2 n_g^t n_p^t n_r^t + \sigma_{i\epsilon}^2}{df_S} \chi_{df_S}^2 \text{ with } df_S = n_g^t n_m^t (n_s^t - 1), \\ MSM(G)_i &\sim \frac{\sigma_{iM}^2 n_g^t n_s^t n_p^t n_r^t + \sigma_{iS}^2 n_g^t n_p^t n_r^t + \sigma_{i\epsilon}^2}{df_M} \chi_{df_M}^2 \text{ with } df_M = n_g^t (n_m^t - 1), \end{aligned} \tag{13}$$

and they are mutually independent, by Satterthwaite approximation (Satterthwaite, 1946), $\boldsymbol{b}^\top \hat{\Sigma}_v \boldsymbol{b}$ follows approximately $(\boldsymbol{b}^\top \Sigma_v \boldsymbol{b}/df)\chi^2_{df}$ with

$$df = \frac{\left(\sum_{i=1}^{n_i^v} b_i^2 (c_M MSM(G)_i + c_S MSS(GM)_i + c_E MSE_i)\right)^2}{\sum_{i=1}^{n_i^v} \left(\frac{\left(b_i^2 c_M MSM(G)_i\right)^2}{df_M} + \frac{\left(b_i^2 c_S MSS(GM)_i\right)^2}{df_S} + \frac{\left(b_i^2 c_E MSE_i\right)^2}{df_E} \right)}.$$

By independence of $\hat{\boldsymbol{\theta}}_1$ and $\boldsymbol{b}^\top \hat{\Sigma}_v \boldsymbol{b}$, $\omega_b \hat{\eta}_{a,b}$ follows approximately t-distribution with degrees of freedom df and non-centrality parameter $\omega_b \eta_{a,b}$, where $\omega_b = \sqrt{\boldsymbol{b}^\top \Sigma_v \boldsymbol{b}/(\boldsymbol{b}^\top \frac{\Sigma_t}{n_m^t} \boldsymbol{b})}$. With ω_b estimated by $\hat{\omega}_b = \sqrt{\boldsymbol{b}^\top \hat{\Sigma}_v \boldsymbol{b}/(\boldsymbol{b}^\top \frac{\hat{\Sigma}_t}{n_m^t} \boldsymbol{b})}$, the distribution of $\hat{\eta}_{a,b}$ can be approximated by $t_{df}(\hat{\omega}_b \eta_{a,b})/\hat{\omega}_b$.

A $100(1-\alpha)\%$ confidence lower bound of $\eta_{a,b}$ is then obtained by testing $H_0 : \eta_{a,b} = \eta_0$ versus $H_a : \eta_{a,b} > \eta_0$ with $\hat{\eta}_{a,b}$ as a test statistic and inverting the acceptance region for $\eta_{a,b}$. To find the expression of the confidence lower bound, let η_{obs} be the observed value of $\hat{\eta}_{a,b}$. Denote the p-value for the one-sided test by $p(\eta_0) = P_{H_0}(\hat{\eta}_{a,b} \geq \eta_{obs})$. If $p(\eta_0) < \alpha$, H_0 is rejected. Under H_0, $\hat{\eta}_{a,b}$ follows $t_{df}(\hat{\omega}_b \eta_0)/\hat{\omega}_b$ approximately, and for fixed $\hat{\omega}_b$, $\hat{\eta}_{a,b}$ is stochastically increasing in η_0. Thus $p(\eta_0)$ is increasing in η_0, and the $100(1-\alpha)\%$ confidence lower bound for $\eta_{a,b}$ is given by the smallest possible value η_{LB} such that $p(\eta_{LB}) \geq \alpha$. When sample sizes and device-specific parameters for training data are large, df_E, df_S, and df_M tend to be large, which results in large degrees of freedom df. In this case, $\chi^2_{df}/df \approx 1$, and the random variable $\hat{\eta}_{a,b}$ can be further approximated by $N(\hat{\omega}_b \eta_{a,b}, 1)/\hat{\omega}_b$. Figure 3 shows the p-values of the hypothesis test, $p(\eta_0)$ for $\eta_{obs} = 1.94$, which is the observed value of $\hat{\eta}_{a,b}$ from the mice data. Note that the value of η_{obs} depends on sample size n_s^v and device-specific parameters n_p^v and n_r^v in the validation stage, and they are set to $n_s^v = 2$, $n_p^v = 32$, and $n_r^v = 2$ in this example. As shown in the figure, the 95% confidence lower bound of $\eta_{a,b}$ found by non-central t (denoted by η_{LB}^t) is less than that by normal approximation (denoted by η_{LB}^N). Generally, the normal approximation gives a less conservative confidence lower bound.

Given the lower bound of $\eta_{a,b}$, say, η_{LB}, the confidence lower bound of sensitivity $\Phi(\eta_{a,b})$ is given by $\Phi(\eta_{LB})$. Figure 4 shows sensitivity estimates of the discriminant rule for the mice data (A vs W) by using the simple plug-in estimator and $100(1-\alpha)\%$ confidence lower bounds given by the proposed method. Clearly, the plug-in sensitivity estimates are bigger than the confidence lower bounds. As the number of replicated samples per subject n_s^v increases, so does the true sensitivity of $\hat{\phi}_{\mathrm{V}}$ in (11). Accordingly, the sensitivity estimates increase with n_s^v as shown in the figure.

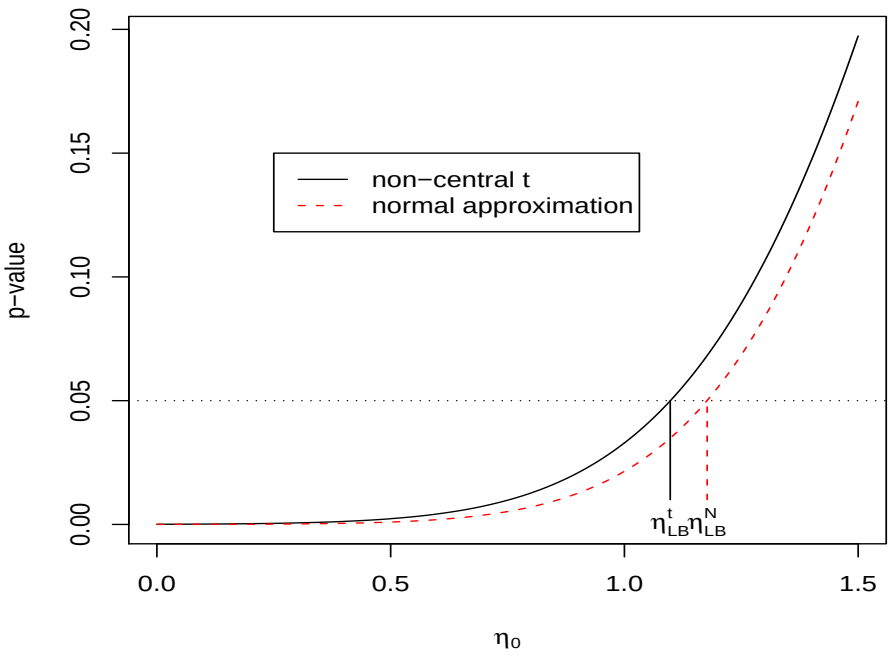

Figure 3: The p-values as a function of η_0 for testing $H_0 : \eta_{a,b} = \eta_0$ versus $H_a : \eta_{a,b} > \eta_0$ when $n_s^v = 2$, $n_p^v = 32$, $n_r^v = 2$, and thus $df = 46.11$, $\hat{\omega}_b = 2.15$ and $\eta_{obs} = 1.94$ from the mice experiment. The solid line is for non-central t-distribution approach and the dashed line is for normal approximation. The 95% confidence lower bounds of $\eta_{a,b}$ from non-central t and normal approximation are $\eta_{LB}^t = 1.098$ and $\eta_{LB}^N = 1.177$, respectively.

Given a sensitivity estimate either by a simple plug-in estimator or $100(1-\alpha)\%$ confidence lower bound as in Figure 4, now consider determination of sample size n_m^v. Figure 5 gives the combination of n_s^v and n_m^v necessary to meet the minimal sensitivity level of γ with at least 95% of probability when we vary γ from 0.75 to 0.95 in the same setting as Figure 4. The left panel is for the simple plug-in estimator while the right panel is for 95% confidence lower bound. For example, to attain a minimal level of sensitivity of 90% with $n_s^v = 2$, n_m^v has to be at least 20 from the left panel. Expectedly, as the minimal level of sensitivity increases, necessary sample sizes increase as well, and relatively large sample sizes are needed when confidence lower bounds are used to estimate sensitivity.

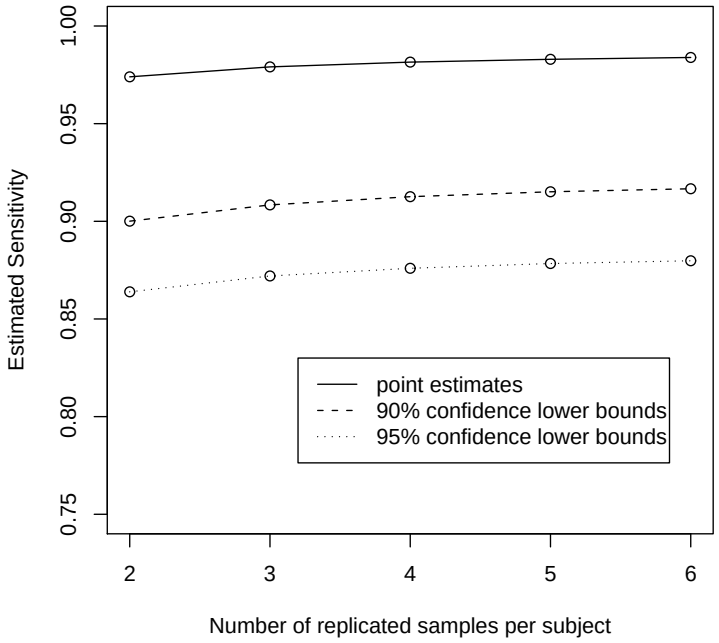

Figure 4: Sensitivity estimates of the discriminant rule for the mice data (A vs W) by the simple plug-in estimator and confidence lower bounds (90% and 95%) for varied number of replicated samples per subject when $n_r^v = 2$ and $n_p^v = 32$.

4 Numerical Study

We investigate the validity of the proposed sample sizes determination procedure by simulation. The procedure is designed for a microarray experiment in the second stage of pharmacogenomics. Under the linear mixed effect model assumption in (5) for probe level data, training and validation data are generated from normal distributions as in (6) and (7), respectively.

4.1 Simulation Module

For a numerical validation study, we take the following steps.

Step 1: Specify the number of genes n_i^v and true parameter values for means $\boldsymbol{\theta}_g = (\mu_1 + \tau_{1g}, \mu_2 + \tau_{2g}, \ldots, \mu_{n_i^v} + \tau_{n_i^v g})^\top$, $g = 0, 1$, and variance components σ_{iM}^2, σ_{iS}^2, and $\sigma_{i\epsilon}^2$, $i = 1, 2, \ldots, n_i^v$ in (6) and (7).

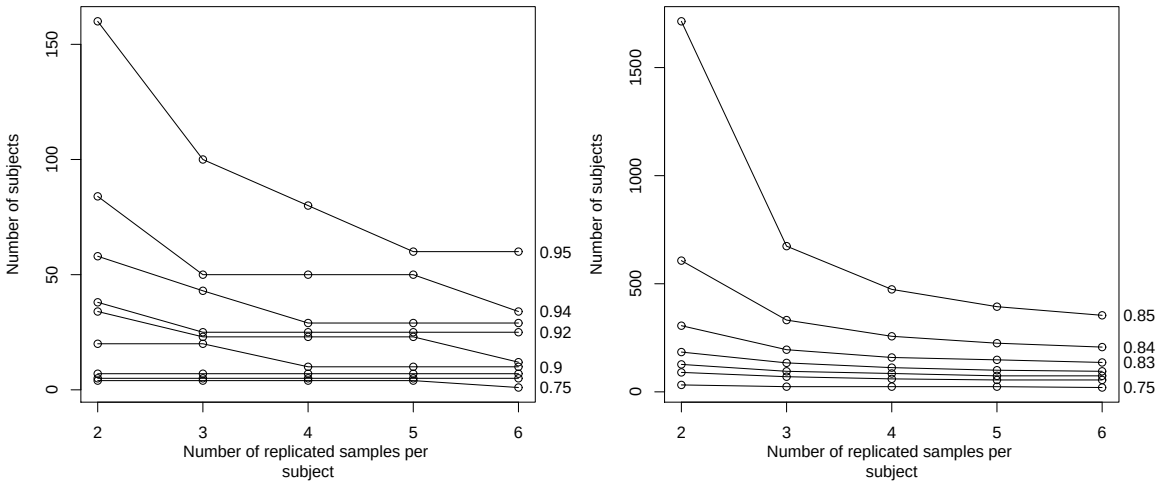

Figure 5: The required number of subjects and number of replicated samples per subject for various minimal sensitivity levels γ with 95% probability of successful validation when sensitivity is estimated either by the simple plug-in estimator (left panel) or by 95% confidence lower bounds (right panel). The γ values are in the right margin of each plot.

Step 2: After specifying sample sizes n_g^t, n_m^t, and n_s^t, and device-specific parameters n_p^t and n_r^t in the training stage, generate directly such summary statistics of training data as $\hat{\boldsymbol{\theta}}_g$ from $N(\boldsymbol{\theta}_g, \Sigma_t/n_m^t)$ and the values of mean squares MSE_i, $MSS(GM)_i$ and $MSM(G)_i$ from the scaled χ^2 distributions in (13). Given the mean squares, obtain the estimates of variance components $\hat{\sigma}^2_{i\epsilon}$, $\hat{\sigma}^2_{iS}$, and $\hat{\sigma}^2_{iM}$ from Equation (12).

Step 3: Regard $\hat{\boldsymbol{\theta}}_g$, $\hat{\sigma}^2_{i\epsilon}$, $\hat{\sigma}^2_{iS}$, and $\hat{\sigma}^2_{iM}$ from the previous step as if they were calculated from raw probe-level training data. After specifying the sample size n_s^v and the device-specific parameters n_p^v and n_r^v, calculate sample size n_m^v for a validation study to meet the minimal sensitivity requirement (2), as we did in the mice experiment in Section 3.

Step 4: Given n_m^v, compute the successful validation rate $P(\widehat{Sen}_{\mathbf{V}} \geq \gamma)$ using the fact that $n_m^v \widehat{Sen}_{\mathbf{V}} \sim B(n_m^v, Sen(\hat{\phi}_{\mathbf{V}}))$, where the true sensitivity of the plug-in classification rule $Sen(\hat{\phi}_{\mathbf{V}})$ from the training data in Step 2 is explicitly given by (8).

Step 5: Repeat the previous steps for multiple times to obtain an *unconditional* estimate of the successful validation rate $P(\widehat{Sen}_{\mathbf{V}} \geq \gamma)$ by averaging over replicates of training

data.

If the unconditional estimate at Step 5 is greater than $1-\beta$, the procedure of sample size determination is deemed valid. Otherwise, it indicates that the sample sizes specified by the proposed approach are not large enough to meet the minimal sensitivity requirement.

4.2 Results

To mimic the mice experiment, we set the true parameter values for means $\boldsymbol{\theta}_g$ and variance components σ^2_{iM}, σ^2_{iS}, and $\sigma^2_{i\epsilon}$, $i = 1, 2, \ldots, n_i^v$, to the actual estimates from the real data for two groups (A vs W) first and fixed $n_m^t = 20$, $n_s^t = 4$, $n_p^t = 32$, and $n_r^t = 2$. We varied n_s^v from 2 up to 6 and the confidence level $1-\alpha$ for a lower bound of the true sensitivity $Sen_{\mathbf{V}}$ from 90% down to 50% in the simulation study. The simple plug-in estimator of $Sen_{\mathbf{V}}$ in (9) was also considered for comparison. The desired level of success $1-\beta$ was set to 0.95 and the minimal level of sensitivity γ was 0.85 for the study. By taking the steps laid out, values of successful validation rate $P(\widehat{Sen}_{\mathbf{V}} \geq \gamma)$ were obtained for 100 replicates of training data simulated from the specified distribution.

Due to the skewness of the distribution of successful validation rates from the replicates, for each combination of n_s^v and $(1-\alpha)$, the distribution of $P(\widehat{Sen}_{\mathbf{V}} \geq \gamma)$ is summarized by the median as shown in Table 1. The table also shows the ideal sensitivity of $\Phi(\delta/2)$ with δ in (10), which is the maximally achievable sensitivity (independent of data) in the normal setting, and the median of true sensitivity in (8) for the 100 replicates as a reference. The interquartile range (IQR) of the true sensitivity across 100 replicates is from 0.012 to 0.017, and the IQR of successful validation rates is from 0.007 to 0.018. With such a high true sensitivity level (around 0.97) in this setting, we observe that there is only minor difference between the true sensitivity and its estimates (not shown in the table), and therefore estimating it by either a simple plug-in estimator or a confidence lower bound does not make much difference in sample sizes. Hence, the median values of successful validation rates even with the plug-in estimator are all greater than the desired level of success 0.95. With the simple plug-in estimator of the true sensitivity of the normal discriminant rule, the median number of subjects n_m^v required was 7 for different values of n_s^v.

In the second simulation study, the true mean parameters in Step 1 were set to be 0.14 closer than the actual estimates from the mice data so that the true sensitivity is lower than the previous setting. Except the mean parameters, all other factors were kept the same. Table 2 summarizes the median and IQR values of successful validation rates for the new setting. As seen in the table, when the true sensitivity is lower (around 92% to 94%),

The level $100(1-\alpha)\%$	The number of replicated samples per subject (n_s^v)				
for confidence lower bound	2	3	4	5	6
90	0.994	0.996	0.997	0.997	0.996
85	0.993	0.995	0.994	0.995	0.996
80	0.992	0.993	0.995	0.995	0.995
75	0.991	0.993	0.994	0.995	0.995
70	0.990	0.992	0.993	0.994	0.995
65	0.990	0.991	0.993	0.994	0.993
60	0.988	0.991	0.993	0.992	0.993
55	0.987	0.991	0.991	0.992	0.992
50	0.987	0.991	0.991	0.991	0.992
Plug-in estimate in (9)	0.987	0.990	0.991	0.991	0.992
Ideal sensitivity $\Phi(\frac{\delta}{2})$	0.974	0.980	0.982	0.983	0.984
True sensitivity in (8)	0.972	0.977	0.980	0.982	0.983

Table 1: Median values of successful validation rates $P(\widehat{Sen}_{\mathbf{V}} \geq 0.85)$ by the sample size calculation procedure for different values of n_s^v and confidence level $(1-\alpha)$ from 100 replicates of simulated training data when the true sensitivity is around 97% to 98%. Ideal sensitivity and median values of true sensitivity in (8) are in the last two rows.

The level $100(1-\alpha)\%$	The number of replicated samples per subject (n_s^v)				
for confidence lower bound	2	3	4	5	6
90	0.999 (0.020)	0.998 (0.022)	0.997 (0.019)	0.997 (0.028)	0.998 (0.028)
80	0.988 (0.065)	0.987 (0.061)	0.987 (0.067)	0.986 (0.062)	0.987 (0.054)
75	0.979 (0.068)	0.982 (0.087)	0.979 (0.065)	0.982 (0.058)	0.983 (0.054)
70	0.970 (0.081)	0.972 (0.080)	0.973 (0.064)	0.970 (0.074)	0.969 (0.068)
65	0.964 (0.113)	0.966 (0.087)	0.960 (0.080)	0.963 (0.073)	0.965 (0.065)
60	**0.951** (0.119)	**0.951** (0.096)	0.956 (0.078)	0.962 (0.069)	0.967 (0.069)
56	0.938 (0.123)	0.947 (0.092)	0.956 (0.076)	0.962 (0.075)	0.961 (0.078)
55	0.937 (0.116)	0.945 (0.089)	0.956 (0.076)	0.956 (0.079)	0.959 (0.078)
50	0.928 (0.127)	0.947 (0.086)	**0.950** (0.082)	**0.953** (0.075)	**0.955** (0.069)
Plug-in estimate	0.928 (0.125)	0.947 (0.090)	0.948 (0.082)	0.953 (0.070)	0.955 (0.069)
Ideal sensitivity	0.929	0.937	0.942	0.945	0.947
True sensitivity	0.924 (0.035)	0.933 (0.033)	0.938 (0.032)	0.941 (0.032)	0.943 (0.031)

Table 2: Median values (and IQR in parentheses) of successful validation rates $P(\widehat{Sen}_{\mathbf{V}} \geq 0.85)$ by the sample size calculation procedure for different values of n_s^v and confidence level $(1-\alpha)$ when the true sensitivity is around 92% to 94%. Ideal sensitivity and median values (and IQR) of true sensitivity in (8) are in the last two rows.

successful validation rates vary more widely depending on the sensitivity estimator used. The IQR of true sensitivity in (8) across 100 replicates is around 0.036, and that of successful validation rates is in the range of 0.02 to 0.13. The highlighted values in the table are the smallest (unconditional) successful validation rate that exceeds the desired level of 95% for each n_s^v, where a success is defined as $\widehat{Sen}_{\mathbf{V}}$ being at least 85%. 50% confidence lower bound and the simple plug-in estimator generally produce similar successful validation rates. When $n_s^v = 2$ or 3, the confidence level needs to be at least 60% to guarantee a successful validation experiment 95% of the time. Just for comparison with the higher sensitivity setting, when the true sensitivity is estimated by the simple plug-in estimator, the median numbers of subjects n_m^v necessary for a validation experiment increased to 27, 20, 20, 14, and 14 for $n_s^v = 2$, 3, 4, 5, and 6, respectively.

Setting the true mean parameters in Step 1 to be 0.2 closer than the actual estimates from the real experiment, we examined further the impact of the underlying true sensitivity on necessary sample sizes for a validation experiment. Table 3 shows the results when

The level $100(1-\alpha)\%$	The number of replicated samples per subject (n_s^v)				
for confidence lower bound	2	3	4	5	6
90	0.998 (0.104)	0.993 (0.093)	0.993 (0.099)	0.990 (0.068)	0.989 (0.074)
85	0.983 (0.136)	0.972 (0.116)	0.976 (0.114)	0.970 (0.110)	0.970 (0.102)
84	0.976 (0.142)	0.971 (0.109)	0.974 (0.098)	0.962 (0.113)	0.970 (0.110)
83	0.969 (0.145)	0.969 (0.114)	0.966 (0.118)	0.965 (0.115)	0.968 (0.112)
82	0.964 (0.165)	0.964 (0.129)	0.956 (0.119)	0.963 (0.123)	0.965 (0.114)
81	0.959 (0.193)	0.960 (0.114)	0.954 (0.118)	0.960 (0.126)	0.959 (0.112)
80	**0.953** (0.176)	**0.954** (0.138)	**0.954** (0.125)	**0.959** (0.124)	**0.956** (0.116)
75	0.931 (0.169)	0.928 (0.127)	0.939 (0.143)	0.930 (0.150)	0.938 (0.118)
50	0.842 (0.205)	0.870 (0.166)	0.876 (0.167)	0.874 (0.149)	0.898 (0.136)
Plug-in estimate	0.842 (0.203)	0.870 (0.179)	0.876 (0.163)	0.885 (0.148)	0.898 (0.133)
Ideal sensitivity	0.896	0.907	0.912	0.915	0.917
True sensitivity	0.896 (0.028)	0.906 (0.027)	0.911 (0.027)	0.915 (0.027)	0.917 (0.027)

Table 3: Median values (and IQR) of successful validation rates $P(\widehat{Sen}_{\mathbf{V}} \geq 0.85)$ by the sample size calculation procedure for different values of n_s^v and confidence level $(1-\alpha)$ when the true sensitivity is around 90% to 91%. Ideal sensitivity and median values (and IQR) of true sensitivity in (8) are in the last two rows.

the true sensitivity is around 90% to 91%. The highlighted values are again the smallest successful validation rate exceeding the desired level 95% for each n_s^v. The IQR of the true sensitivity level for 100 replicates is around 0.03, and that of successful validation rates is in the range of 0.06 to 0.24, clearly showing larger variability than the settings with higher sensitivity values. The table suggests that 80% confidence level is sufficient to obtain a proper sample size for n_m^v while using the simple plug-in estimates for sensitivity (nearly equivalent to 50% of confidence) results in success rates much smaller than 95%. The low success rate of the plug-in estimator is attributed to the fact that it gives a smaller sample size than necessary.

For reference, Table 4 provides median estimates of subjects n_m^v required in a validation experiment under the same setting as Table 3 from the 100 replicates.

Lowering the true sensitivity further down to 87% to 89% by setting the true mean parameters 0.23 closer than the actual estimates required about 95% confidence level for sensitivity estimation to ensure that the rate of successful validation is more than 95%. The

The level $100(1-\alpha)\%$	The number of replicated samples per subject (n_s^v)				
for confidence lower bound	2	3	4	5	6
90	317	164	127	107	94
85	174	104	80	74	67
84	160	94	74	67	60
83	147	94	74	60	57
82	134	87	67	60	54
81	127	80	67	54	54
80	**120**	**74**	**60**	**54**	**54**
75	94	60	54	47	40
50	40	34	27	20	20
Plug-in estimate	40	34	27	20	20

Table 4: Median number of subjects n_m^v required from 100 replicates of simulated training data for different values of n_s^v and confidence level $(1-\alpha)$ of lower bounds of the true sensitivity.

results are not shown here, but successful validation rates varied significantly more than the other settings with IQR values in the range of 0.22 to 0.62. As an extreme setting, when the mean parameters were set so that the true sensitivity is slightly above the pre-specified minimal level $\gamma = 0.85$, a large fraction of simulated data turned out to be statistically futile with confidence lower bounds for sensitivity smaller than 0.85.

The results from the simulation study can be used to calibrate the confidence level $1-\alpha$ for estimation of the true sensitivity to attain the desired success level $1-\beta$. Generally, the lower the true sensitivity level is, the higher confidence level is necessary for estimation of sensitivity. In particular, the simulation results suggest that under the specified setting, a simple plug-in estimator (approximately 50% of confidence) would be sufficient when the true sensitivity is around 97% or above, and 60% would be sufficient for 93% of sensitivity while 80% and 90% of confidence would be necessary for 90% and 88% of sensitivity, respectively.

5 Conclusion

The main objective of this paper is to calculate sample size for a validation study to meet pre-specified sensitivity and specificity requirement, in order to avoid futility of pharma-

cogenomic development. Change of platforms is taken into account in the sample sizes calculation by statistical modeling. The proposed formulation for meeting minimal sensitivity and specificity requirements calls for estimation of both measures. Their confidence lower bounds can substitute the unknown true values in the sample size calculation procedure. However, the confidence level has to be calibrated for appropriate sample sizes to ensure that the probability of a successful validation experiment exceeds a desired level. Our simulation study shows the relationship between the underlying sensitivity and the required confidence level in a normal distribution setting. The results can be used as a practical guideline to set the level of confidence adaptively.

As a future direction, robustness of the proposed procedure can be further investigated to see how sensitive the sample size calculation procedure is to the model assumptions. In principle, our approach can be extended to a general scenario where more complex diagnostic rules than DLDA are considered to account for potential correlations among genes.

Acknowledgements

Hsu's research was supported in part by the NSF Grant No. DMS-0505519 and a grant from the Icelandic Science and Technology Council.

References

Adcock, C. J. (1997). Sample size determination: a review. *The Statistician 46(2)*, 261–283.

Baker, M. (2005). In biomarkers we trust? *Nature Biotechnology 23(3)*, 297–304.

Chow, S. C., J. Shao, and H. S. Wang (2003). *Sample Size Calculations in Clinical Research*. Chapman & Hall.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association 78*, 316–331.

FDA (2005a). *Drug-Diagnostic Co-development Concept Paper*. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH), Office of Contracting and Procurement (OCP), U. S. Food and Drug Administration.

FDA (2005b). *Pharmacogenomic Data Submission: Guidance for Industry*. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH), U. S. Food and Drug Administration.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58(301)*, 13–30.

Hsu, J. (1988). Sample-size computation for designing multiple comparison experiments. *Computational Statistics and Data Analysis 7(1)*, 79–91.

Hsu, J. C., Y. Rao, Y. Lee, J. Chang, K. Bergsteinsdottir, M. K. Magnusson, T. Wang, and E. Steingrimsson (2009). Design and analysis of microarray experiments for pharmacogenomics. In A. Dmitrienko, A. C. Tamhane, and F. Bretz (Eds.), *Multiple Testing Problems in Pharmaceutical Statistics*, CRC Biostatistics Series. Chapman & Hall.

Hughes, S., A. Hughes, C. Brothers, W. Spreen, and D. Thorborn (2008). PREDICT-1 (CNA106030): the first powered, prospective trial of pharmacogenetic screening to reduce drug adverse events. *Pharmaceutical Statistics 7*, 121–129.

Lee, M. (2004). *Analysis of microarray gene expression data*. Kluwer Academic Publishers.

Mallal, S., D. Nolan, C. Witt, G. Masel, A. M. Martin, C. Moore, D. Sayer, A. Castley, C. Mamotte, D. Maxwell, I. James, and F. T. Christiansen (2002). Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *The Lancet 359*, 727–732.

Mallal, S., E. Phillips, G. Carosi, J.-M. Molina, C. Workman, J. Tomažič, E. Jägel-Guedes, S. Rugina, O. Kozyrev, J. F. Cid, P. Hay, D. Nolan, S. Hughes, A. Hughes, S. Ryan, N. Fitch, D. Thorborn, and A. Benbow (2008). HLA-B*5701 screening for hypersensitivity to abacavir. *New England Journal of Medicine 358*, 568–579.

Martin, A. M., D. Nolan, S. Gaudieri, C. A. Almeida, R. Nolan, I. James, F. Carvalho, E. Phillips, F. T. Christiansen, A. W. Purcell, J. McCluskey, and S. Mallal (2004). Predisposition to abacavir hypersensitivity conferred by HLA-B*5701 and a haplotypic Hsp70-Hom variant. *Proceedings of the National Academy of Sciences 101*, 4180–4185.

McLachlan, G. (1992). *Discriminant analysis and statistical pattern recognition*. Wiley-Interscience.

Pepe, M. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Ravishanker, N. and D. Dey (2002). *A first course in linear model theory*. Chapman & Hall/CRC.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin 6*, 110–114.

van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature 415*, 530–536.

Wang, S., R. O'Neill, and H. Hung (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics 6*, 227–244.