https://doi.org/10.1016/j.csda.2017.08.010

Sample size determination for high dimensional parameter estimation with application to biomarker identification

Binyan Jiang[a], Jialiang Li[b,*]

[a] *Department of Applied Mathematics, Hong Kong Polytechnic University Hum Hung, Kowloon, Hong Kong*
[b] *Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546*

Abstract

We consider sample size calculation to obtain sufficient estimation precision and control the length of confidence intervals under high dimensional assumptions. In particular, we intend to provide more general results for sample size determination when a large number of parameter values need to be computed for a fixed sample. We consider three design approaches: normal approximation, inequality method and regression method. These approaches are applied to sample size calculation in estimating the Net Reclassification Improvement (NRI) and the Integrated Discrimination Improvement (IDI) for a diagnostic or screening study. Two medical examples are also provided as illustration. Our results suggest the regression method in general can yield a much smaller sample size than other methods.

Keywords: Bernstein inequality, Bonferroni inequality, IDI, NRI, Sample size calculation, Training sample

1. Introduction

Diagnostic or screening tests are used to detect the patient disease status in medical practice. The accuracy of these tests may be assessed by all kinds of

*Corresponding author
Email address: stalj@nus.edu.sg (Jialiang Li)

traditional statistical methods such as sensitivity and specificity (Zhou, Obuchowski and McClish, 2011). In recent population studies it becomes more and more imperative to evaluate the accuracy gain when new information such as new biomarker or new model structure has been added into the existing diagnostic procedure. In practice in order to study the general diagnostic accuracy performance of statistical methods, we must obtain an appropriate data set with a reasonable sample size. A study with inadequate sample size may not have sufficient statistical efficiency to achieve a meaningful finding. On the other hand, it may be wasteful and unethical to conduct a study with too large a sample size. There are abundant sample size calculation approaches for various statistical problems; see for example Chow, Wang and Shao (2007). Within the diagnostic medicine literature, one can find a comprehensive review for sample size calculation in chapter 6 of Zhou, Obuchowski and McClish (2011). Additionally, Obuchowski and Zhou (2002) considered sample size calculation for diseased and non-diseased subjects required for attaining a prespecified conditional power to test hypotheses regarding diagnostic accuracy measures. Li and Fine (2004) extended earlier sample size formula for case-control studies to prospective cohort studies and provided a justification for the commonly used prevalence inflation method. Steinberg, Fine and Chappell (2009) investigated sample size methods for positive and negative predictive values which may depend on the disease prevalence.

In general, sample size calculation is performed to meet certain optimality criteria, controlling either the Type I/II errors in a hypothesis test problem or the length and confidence level of a confidence interval in an estimation problem (Zhou, Obuchowski and McClish, 2011). Earlier authors (Pencina, D'agostino and Demler, 2012; Leening et al., 2014) usually prefer the confidence interval approach to make inference in lieu of the hypothesis test approach. In this paper we aim at designing the sample size to attain sufficient estimation precision and controlling the length of confidence intervals, and we will focus on sample size methods for the interval estimation.

Accuracy measures are frequently reported for biomarker studies where a

large number of tests are evaluated simultaneously using the same data set. See Li and Fine (2008) and Li, Jiang and Fine (2013) for examples. Sample size calculation thus needs to acknowledge the high-dimension feature of the data set. A common approach is to use the asymptotic distribution of the estimator. However, if the study does not admit very large number of subjects, asymptotic approximation may be questionable. Without assuming the asymptotic distribution, van der Laan and Bryan (2001) proposed an inequality approach to calculate sample size for the mean estimation using Bernstein's inequality. This method provides a bound for the sample size required for a fixed significance level and only requires the existence of second order moments. We will adopt a similar method of using probability inequalities to calculate the sample size needed to obtain certain estimation accuracy. We further propose a regression approach when a training set is available. This approach may be less conservative than the normal approximation and the inequality approach. A regression calibration method has been used recently in Dobbin and Song (2013) for the estimation of regression coefficients in proportional hazards models. However, the authors considered a deterministic sample size computation under a very complicated calibration model. In this paper, we propose three sample size calculation approaches. The first approach is based on the normal approximation while the second approach is based on the probability inequalities. These methods may lead to very large sample size requirement. A third approach based on regression calibration is also proposed and may provide more realistic sample sizes in practice.

Pencina, D'agostino and Vasan (2008) proposed two quantitative criteria based on reclassification to directly evaluate the extent which a new predictor improves classification performance: the net reclassification improvement (NRI) and integrated discrimination improvement (IDI). These new statistics received wide acceptance in health science research. Uno et al. (2013) and Li, Jiang and Fine (2013) extended the formulation of NRI and IDI to failure time outcomes and multicategory outcomes, respectively. Because the NRI and the IDI yield lucid probability assess on diagnostic accuracy improvement, they have both been

widely reported and discussed in medical literature since their creation (Pencina, D'agostino and Vasan, 2008). Recently Steyerberg et al. (2010) assessed the performance of prediction models using a variety of methods and metrics and suggested that the NRI and the IDI can gain insight into the value of adding a novel predictor to an established model; Pencina et al. (2012) compared the NRI, the IDI and the ROC curve under nested models and recommended to report these three measures together to characterize the performance of the final model as these three measures offered complementary information. Some authors suggest combining these reclassification statistics with various calibration measures and decision analytic measures to avoid spurious claims of improved prediction and erroneous clinical inference (Pencina, D'agostino and Steyerberg, 2011; Pencina, D'agostino and Demler, 2012; Leening et al., 2014; Kerr et al., 2014). However, very little research work is available on the design of an epidemiological study for the estimation of the NRI and the IDI. As an application of our approaches we obtain explicit sample size calculation for studies aiming to evaluate the NRI and the IDI.

In the rest of this paper we will first introduce three approaches for sample size calculation, followed by extensive simulations results and two medical examples when these approaches are applied to evaluate the NRI and IDI estimation. Some remarks will also be provided in the end of this paper.

2. Methods

Suppose we are interested in estimating a parameter $\theta = (\theta_1, \cdots, \theta_p)^T \in R^p$ with a sample of size n where $n << p$. This is the so-called large-p-small-n setting. We usually construct an estimator $\hat{\theta} = (\hat{\theta}_{n1}, \cdots, \hat{\theta}_{np})^T$ from the sample which may have nice asymptotic properties. The research question of this article is to design a sample size n such that the estimation errors of all the covariates are bounded by ϵ with high probability $1 - \alpha$.

2.1. Method 1: Normal Approximation

In this section we assume that the distribution or asymptotic distributions of $\sqrt{n}\left(\hat{\theta}_{nj}-\theta_j\right)$ is $N\left(0,\sigma_j^2\right)$ for $k=1,\cdots,p$. This is achievable for many parameter estimation problems. We may use such asymptotic results to compute the sample size. For large n, we have

$$P\left(\frac{\sqrt{n}|\hat{\theta}_{nj}-\theta_j|}{\sigma_j}>z_{\alpha/2}\right)<\alpha, \tag{1}$$

where z_α is the upper α quantile of the standard normal distribution. Let $\epsilon=|\hat{\theta}_{nj}-\theta_j|$ be the anticipated error margin. We obtain the following sample size formula

$$n^*=\frac{z_{\alpha/2}^2\sigma_j^2}{\epsilon^2}. \tag{2}$$

Using this formula ensures that the estimation error for θ_k is bounded by ϵ with probability $1-\alpha$. However, this formula is appropriate if we only study one parameter ($p=1$).

Now to extend the above formula for multiple parameters, we may apply the Bonferroni inequality

$$P\left(\max_j|\hat{\theta}_{nj}-\theta_j|>\epsilon\right)\leq\sum_{j=1}^{p}P\left(|\hat{\theta}_{nj}-\theta_j|>\epsilon\right),$$

and bound the error probability for each marker equally to be α/p. The j^{th} estimator satisfies

$$P\left(\frac{\sqrt{n}|\hat{\theta}_{nj}-\theta_j|}{\sigma_j}>z_{\alpha/(2p)}\right)\leq\alpha/p, \tag{3}$$

to achieve the overall error probability α. This leads to the generalization of the sample size formula (2) to be

$$n^*=\frac{z_{\alpha/(2p)}^2 v}{\epsilon^2}, \tag{4}$$

where $v=\max_j\sigma_j^2$.

2.2. Method 2: Inequality

Sometimes it is not desirable to apply normal approximation using the central limit theorem. The symmetric unimodal distribution may not be an appropriate shape to describe the estimated parameters in finite samples, especially when the true parameters are naturally bounded. To relax the distributional conditions, we consider probability inequalities that provide a uniform bound for the estimation error. These inequalities only need moment conditions and are more flexible in practical studies.

Suppose that each estimate may be written as $\hat{\theta}_{nj} = n^{-1}\sum_{i=1}^{n}\psi_{ij}$ where ψ_{ij} is an evaluable quantity computed from the ith subject such that $|\psi_{ij}| \leq M$ for some constant $M > 0$ and $\lim_{n\to\infty} E\hat{\theta}_{n,j} = \theta_j$. We denote $v_j = E(\psi_{ij} - \theta_j)^2$ and $v = \max_{1\leq j\leq p} v_j$.

Bernstein. Using Bernstein's inequality (Bennett, 1962), we have for the j^{th} estimator

$$P\left(|\hat{\theta}_{nj} - \theta_j| > \epsilon\right) = P\left(|\sum_{i=1}^{n}[\psi_{ij} - \theta_j]| > n\epsilon\right) \leq 2\exp\left(\frac{-n\epsilon^2}{2v_j + 2\frac{M\epsilon}{3}}\right).$$

To achieve an error probability bound α we may bound each estimation error probability with α/p using the Bonferroni correction.

Consequently we have the following sample size formula

$$n^* = \frac{2v + \frac{2M\epsilon}{3}}{\epsilon^2}\left(\log p + \log\frac{2}{\alpha}\right). \tag{5}$$

Bennett. Using Bennett's inequality (Bennett, 1962), we have for the j^{th} estimator,

$$P\left(|\hat{\theta}_{nj} - \theta_j| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon}{M}[(1 + \frac{v_j}{M\epsilon})\log(1 + \frac{M\epsilon}{v_j}) - 1]\right).$$

Note that right hand side of the above inequality is a decreasing function of v_j. To achieve an error probability bound α we have the following sample size formula

$$n^* = \left\{\frac{\epsilon}{M}\left[\left(1 + \frac{v}{M\epsilon}\right)\log\left(1 + \frac{M\epsilon}{v}\right) - 1\right]\right\}^{-1} \cdot \left(\log p + \log\frac{2}{\alpha}\right). \tag{6}$$

2.3. Method 3: Regression

We next propose a regression approach to the sample size calculation problem. Suppose we have a training data set with size n_0. Practically, such a training data set might be obtained either from existing literature or by conducting a pilot study. We assume that for $\sqrt{n}E|\hat{\theta}_{nj} - \theta_j| \to C_j < \infty, nE|\hat{\theta}_{nj} - \theta_j|^2 \to \sigma_j^2 < \infty$ as $n \to \infty$ for $j = 1, \ldots, p$. For θ_j, the computation procedure is given below:

1. Generate K random resamples from the training data set. For iteration $k = 1, \cdots, K$, we generate a random number N_{jk} from a uniform distribution over $[n_0/2, n_0]$ and obtain a sample of size N_{jk}. We denote estimators of θ_j computed based on the kth sample as $\hat{\theta}_{jk}$. We then use $\bar{\theta}_j = \frac{\sum_{k=1}^{K} \sqrt{N_{jk}}\hat{\theta}_{jk}}{\sum_{k=1}^{K} \sqrt{N_{jk}}}$ as a reference value and evaluate the estimation error $\epsilon_{jk} = |\hat{\theta}_{jk} - \bar{\theta}_j|$ for each simulated sample.
2. Note that for a given sample size N_{jk}, we have $E\epsilon_j = E|\hat{\theta}_{jk} - \theta_j| \approx C_j/\sqrt{N_{jk}}$ and asymptotically $Var(\epsilon_j) \propto 1/N_{jk}$. Motivated by this, after we acquire K pairs of (N_{jk}, ϵ_{jk}) from Step 1, we treat N_{jk} as predictor and ϵ_{jk} as response and fit the following regression model:
$$\epsilon_{jk} = \frac{b_j}{\sqrt{N_{jk}}} + \frac{e_{jk}}{\sqrt{N_{jk}}}, \qquad k = 1, \cdots, K, \tag{7}$$
where b_j is a parameter to be estimated and e_{jk} is a random error with mean 0. We estimate b_j by a weighted least squares estimation and obtain
$$\hat{b}_j = \frac{1}{K} \sum_{k=1}^{K} \sqrt{N_{jk}}\epsilon_{jk}.$$
3. Based on the fitted model (7), we can then compute the sample size N_j for any desired ϵ value. Specifically, let z^j be the $1 - \alpha/p$ quantiles of $\{\sqrt{N_{jk}}\epsilon_{jk} - \hat{b}_j, k = 1, \cdots, K\}$. By solving
$$\epsilon = \frac{\hat{b}_j}{\sqrt{N_j}} + \frac{z^j}{\sqrt{N_j}},$$
we have:
$$N_j = \left(\frac{\hat{b}_j + z^j}{\epsilon} \right)^2.$$

We then set the overall sample size to be

$$n = \max\{N_j : 1 \leq j \leq p\}. \tag{8}$$

Remark 1. *As kindly pointed out by an anonymous reviewer, our regression method can be framed under the bootstrap approach (Shao and Tu, 2012) by resampling the absoluted deviation ϵ_{jk}. We use different sample size N_{jk} in each iteration since we intend to treat it as a random variable. In this setting we may be able to obtain estimates with varying degree of efficiency and therefore resemble closely to a hetergeneous population. If the population is homogeneous equal sample size is sufficient as is usually adopted in bootstrap.*

2.4. Theoretical justification when the variances are unknown

In practice, σ_i^2, the variances of $\sqrt{n}\hat{\theta}_{ni}, i = 1, \ldots, p$, are usually unknown. However, in many cases σ_i^2 can be well estimated. In this section, we establish some theoretical results for the normal approximation method when a proper estimator $\hat{\sigma}_i^2$ is used in the above sample size calculation methods. Similar results can be obtained for the inequality methods and is omitted for space consideration.

Assumption 1 There exist m training samples (for example, samples from historical studies) and using these training samples, σ_i^2 can be well estimated by $\hat{\sigma}_i^2, 1 \leq i \leq p$ such that:

$$P(|\hat{\sigma}_i^2 - \sigma_i^2| \geq t) \leq C_1 \exp\{-C_2 m t^2\}.$$

for some positive constants C_1, C_2.

Assumption 2 There exists a positive constant B s.t. $B^{-1} \leq \min_{1 \leq i \leq p} \sigma_i^2 \leq \max_{1 \leq i \leq p} \sigma_i^2 \leq B$.

Assumption 1 states that the tail probability of the estimation error of $\hat{\sigma}_i^2$ is decreasing in a exponential rate for $1 \leq i \leq p$. This assumption is satisfied in many scenarios especially in cases where $\hat{\sigma}_i^2$ can be written as a U-statistic, such as the sample variance and the estimates based on regression residual sum of squares; see for example Merlevede, Peligrad and Rio (2009); Ravikumar et

al. (2011) and the references therein. Assumption 2 is to ensure that the σ_i^2's are estimable.

Given $\hat{\sigma}_i^2, 1 \le i \le p$, we then obtain a sample version of (4):

$$\hat{n}^* = \frac{z^2_{\alpha/(2p)}\hat{v}}{\epsilon^2}, \tag{9}$$

where $\hat{v} = \max_j \hat{\sigma}_j^2$.

Theorem 1. *Let n^* and $\hat{n}^*$ be defined as in (4) and (9) respectively. Under Assumptions 1 and 2, for $\alpha < 2\Phi(-1)$, we have*

$$\hat{n}^* - n^* = \mathrm{O}\left(\epsilon^{-2}\sqrt{\frac{\log^3 p}{m}}\right),$$

with probability greater than $1 - \mathrm{O}(p^{-M})$ for some constant $M > 1$.

Proof of Theorem 1 By Assumptions 1 and 2 , for any constant $M > 1$, by choosing $t = \sqrt{\frac{(M+1)\log p}{C_2 m}}$, we have with probability greater than $1 - O(p^{-M})$, $|\hat{\sigma}_i^2 - \sigma_i^2| \le \sqrt{\frac{(M+1)\log p}{C_2 m}}$ for $i = 1, \ldots, p$. Hence we have with probability greater than $1 - O(p^{-M})$,

$$\hat{n}^* - n^* = \frac{z^2_{\alpha/(2p)}(\hat{v} - v)}{\epsilon^2} = \frac{z^2_{\alpha/(2p)}}{\epsilon^2}\sqrt{\frac{(M+1)\log p}{C_2 m}}. \tag{10}$$

Since $\alpha < 2\Phi(-1)$, it can be easily shown that $z^2_{\alpha/(2p)} > 1$. By Lemma 11 in Liu, Lafferty and Wasserman (2009) we have that

$$\frac{\phi(z_{\alpha/(2p)})}{2z_{\alpha/(2p)}} \le \Phi(z_{\alpha/(2p)}) = \frac{\alpha}{2p}.$$

By taking a logarithm on both sides of the above inequality and after some simple calculation we obtain:

$$z^2_{\alpha/(2p)} \le 2\log\frac{2p}{\alpha}.$$

Theorem 1 is then proved by plugging the above inequality into (10).

Theorem 1 indicates that when σ_i^2 can be well estimated by $\hat{\sigma}_i^2$, the plug-in version (9) is asymptotically equivalent to (4) as long as the training sample size

m is of order $o(\epsilon^{-4}\log^3 p)$. With a high probability the sample size computed using an estimated variance is very close to the sample size based on known variance. The theorem only requires some mild conditions in practice. Using similar arguments in the proofs of (10) we can also show that:

Theorem 2. *Under the assumptions of Theorem 1 and assume that* $\log p \ll m \ll \epsilon^{-2}\log p$, *we have*

(i) $m \ll n^*$;

(ii)

$$\frac{\hat{n}^*}{n^*} = 1 + o_p\left(\sqrt{\frac{\log p}{m}}\right).$$

Theorem 2 characterizes the dependence of sample size estimation on the training data. Part (i) of this theorem implies that the historical or training sample for estimating the variance can be much smaller than the actual sample size needed for the study. Part (ii) provides an assessment on the approximation error rate.

3. Numerical examples: application to NRI and IDI estimation in biomarker identification

In this section we consider explicit sample size calculation for studies aiming to evaluate the net reclassification improvement (NRI) and integrated discrimination improvement (IDI). Both simulation study and real data analysis will be provided in the following.

3.1. NRI and IDI

We first introduce some notations. Suppose we intend to collect a sequence of samples $\{U_i, X_{i1}, \cdots, X_{ip}, Y_i : i = 1, \cdots, n\}$, where Y_i is the binary outcome taking value in $\{1, 2\}$, U_i is the baseline explanatory variable in the traditional risk prediction and X_{ij} is the j^{th} biomarker for the i^{th} subject. The baseline model $\mathcal{M}_1$ involves U only and the improved model $\mathcal{M}_{2j}$ involves X_j in addition to U. Usually a binary logistic regression model can be used for the model

construction and we may obtain model-based risk prediction for each subject as $\hat{\mathbf{p}}_i(\mathcal{M}) = (\hat{p}_{1i}(\mathcal{M}), \hat{p}_{2i}(\mathcal{M}))$ indicating the probabilities of belonging to the two categories (Li, Jiang and Fine, 2013). We denote the class sample size $n_m = \sum_{i=1}^{n} I(Y_i = m)$ for the m^{th} category, $m = 1, 2$.

Let S represent the NRI and R represent the IDI in the following presentation. Suppose we have p biomarkers and we intend to estimate NRI and IDI for these markers with the same sample of observations. For the j^{th} biomarker, we denote its NRI and IDI by S_j and R_j respectively, $j = 1, \cdots, p$. It can be shown that sample estimates $\hat{S}_j$ and $\hat{R}_j$ can be written as an independent sum

$$\begin{aligned}
&\hat{\mathsf{S}}_j = \\
&\sum_{i=1}^{n}\sum_{m=1}^{2} \frac{w_m}{n_m}\{I(\hat{p}_{mi}(\mathcal{M}_{2j}) = \max \hat{\mathbf{p}}_i(\mathcal{M}_{2j}), Y_i = m) \\
&-I(\hat{p}_{mi}(\mathcal{M}_1) = \max \hat{\mathbf{p}}_i(\mathcal{M}_1), Y_i = m)\} \\
&= \frac{1}{n}\sum_{i=1}^{n} S_{ij}, \qquad (11)
\end{aligned}$$

and

$$\begin{aligned}
&\hat{\mathsf{R}}_j = \\
&\sum_{i=1}^{n}\sum_{m=1}^{2} \frac{w_m}{n_m(1-n_m/n)}\{[\hat{p}_{mi}(\mathcal{M}_{2j}) - \overline{\hat{p}_m(\mathcal{M}_{2j})}]^2 \\
&-[\hat{p}_{mi}(\mathcal{M}_1) - \overline{\hat{p}_m(\mathcal{M}_1)}]^2\} \\
&= \frac{1}{n}\sum_{i=1}^{n} R_{ij}, \qquad (12)
\end{aligned}$$

where w_m is the class weight, $\overline{\hat{p}_m(\mathcal{M}_j)} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{mi}(\mathcal{M}_j)$ and $\hat{\mathbf{p}}_i(\mathcal{M}_j) = (\hat{p}_{1i}(\mathcal{M}_j), \hat{p}_{2i}(\mathcal{M}_j))$ is the probability assessment vector, usually computed from a fitted logistic regression model (Li, Jiang and Fine, 2013). We note that there may be multiple types of estimators for IDI. They are similar in practice and all lead to consistent estimation for the true parameters. What we display above is one that can be easily extended to multiple categories.

We intend to obtain a sample size using the three approaches introduced in this paper such that we may apply the above formulas to estimate the NRI and

the IDI values with certain accuracy. Next we provide a simulation study and two real data examples to evaluate the three methods.

3.2. Simulation

We consider the following true model:

$$\log \frac{P}{1-P} = \beta_0 + \beta_u U + \beta_1 X_1 + \cdots \beta_p X_p, \tag{13}$$

where P is the response probability given the covariates, U is a baseline variable and X_j's are high-dimensional covariates. We assume that $\beta_1, \cdots, \beta_{p_0}$ are non-zero coefficients and $\beta_j = 0$ when $j > p_0$. Such a sparsity assumption is usually adopted in high-dimensional data analysis. U is generated randomly from the standard normal distribution and $X = (X_1, \ldots, X_p)^T$ is generated randomly from a $p-$dimensional multivariate normal distribution.

We intend to evaluate the NRI and the IDI of each X_j for its contribution on the response probability in addition to the baseline variable U. In order to estimate such accuracy improvement parameters for all p markers with sufficient precision, we need to determine the sample size with appropriate statistical methods.

We carry out the sample size calculation using the following five approaches: (i) Normal approximation as in (4); (ii) Bernstein inequality as in (5); (iii) Bennett inequality as in (6); (iv) The regression approach as in (8). In all simulations, we set $\alpha = .05$ and $\epsilon = 0.05, 0.1$. Knowledge on the variance parameter is usually unavailable and we take a conservative choice by setting $v = 1$ in all formula since both NRI and IDI are difference of two bounded probabilities. For the regression approach, we consider training samples of size $n_0 = 200$.

Once the required sample size n^* is obtained from a particular method, we evaluate the estimation performance based on the computed sample size. We repeat the following procedure for $M_R = 200$ times: for the k^{th} ($1 \leq k \leq M_R$) replication, we randomly generate n^* observations from model (13). Using the generated data, we then evaluate the sample NRI for each marker and compute

the observed errors $\hat{\epsilon}_i^k = |\hat{S}_i^k - S_i^0|, i = 1, \ldots, p$. Here $\hat{S}_i^k$ is the NRI estimator for the i^{th} covariate based on the n^* random samples generated in the k^{th} replication, and S_i^0 is the true NRI value for the i^{th} covariate computed using a large number of Monte Carlo samples. The observed errors $\hat{\epsilon}_i^k, i = 1, \ldots, p$ for IDI are computed similarly. We then record the following quantities:

- mean $\hat{\epsilon} := \frac{1}{pM_R} \sum_{i=1}^{p} \sum_{k=1}^{M_R} \hat{\epsilon}_i^k$;
- $\hat{\epsilon}_{\max}$: mean of the maximum of $\hat{\epsilon}_i$ over M_R replications, i.e., $\hat{\epsilon}_{\max} = \frac{1}{M_R} \sum_{k=1}^{M_R} \max\{\hat{\epsilon}_i^k, 1 \le i \le p\}$;
- cover: the proportion of cases that all the p true NRI/IDI values are captured inside the interval formed by the estimate $\pm$ the error margin among M_R replications.
- avr.out: the average of the proportion of markers having estimation errors larger than ϵ among the cases that not all the p NRI/IDI values are covered inside the error margin. Clearly, we have avr.out:= $\frac{\sum_{k=1}^{M_R} \sum_{i=1}^{p} I\{\hat{\epsilon}_i^k > \epsilon\}}{p \sum_{k=1}^{M_R} I\{(\max_{1 \le i \le p} \hat{\epsilon}_i^k) > \epsilon\}}$. If all p estimation errors are smaller than ϵ for all M_R replications, we set avr.out= 0.

We consider the following cases of parameter specification:

Case 1. $p = 200$, $\beta_0 = 0.5, \beta_u = 0.5$, $(\beta_1, \beta_2, \beta_3) = (1.5, 1.5, 2)$ and $\beta_j = 0$ for $j = 4, \ldots, 200$. $U, X_1, \ldots, X_p$ are generated independently from $N(0, 1)$.

Case 2. $p = 600$, $\beta_0 = 0.5, \beta_u = 0.5$, $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6) = (0.5, 0.5, 0.5, 0.25, 0.25, 0.25)$ and $\beta_j = 0$ for $j = 7, \ldots, 600$. $U \sim N(0, 1)$ and $X = (X_1, \ldots, X_p)^T \sim N(\mathbf{0}, \mathbf{\Sigma})$ where $\Sigma = (\sigma_{ij})_{p \times p}$ and $\sigma_{i,j} = 0.5^{|i-j|}$ for $1 \le i, j \le p$.

Case 3. $p = 1000$, $\beta_0 = 0.5, \beta_u = 0.5$, $(\beta_1, \beta_2, \beta_3) = (3, 4, 5)$ and $\beta_j = 0$ for $j = 4, \ldots, 1000$. $U \sim N(0, 1)$ and $X = (X_1, \ldots, X_p)^T \sim N(\mathbf{0}, \mathbf{\Sigma})$ where $\Sigma = (\sigma_{ij})_{p \times p}$. We set $\sigma_{ii} = 1, 1 \le i \le p$, $\sigma_{i,i-1} = \sigma_{i-1,i} = 0.5$ for $2 \le i \le p$ and $\sigma_{ij} = 0$ otherwise.

Case 4. $\beta_0 = 0.5, \beta_u = 0.5$, $(\beta_1, \beta_2, \beta_3) = (1.5, 1.5, 2)$ and $\beta_j = 0$ for $j = 4, \ldots, p$. $U, X_1, \ldots, X_p$ are generated independently from $N(0, 1)$. We consider $p = 6, 20, 63, 200, 632, 2000, 6324, 20000$. These eight values for p are chosen

such that the logarithm of them are equally distributed. The designed error magin ϵ is set to be 0.05.

Cases 1 to 3 cover different covariance structures, different dimensions and magnitude of the nonzero β_j's. Case 1 has the simplest covariance structure (independence) and the magnitude of the nonzero β_j's are moderate. In Case 2 we consider a dense covariance matrix and the nonzero β_j's are set to be small. In Case 3 we use a partially sparse covariance matrix and the nonzero β_j's are set to be relatively large. Note that the covariance matrix in Case 2 can also be seen as a sparse matrix in the looser sense that most of its elements are very close to zero. We remark that sparse assumptions on the covariance matrix are commonly used in the high dimensional literature; see for example Bickel and Levina (2008). For the Breast Cancer study given in the real data analysis section, we obtain a sparsity of 0.912 (Jiang, 2015), indicating that the covariance matrix is very sparse in that over 90% of the off-diagonal elements are zero. Case 4 is designed to check the robustness of our methods. In addition, the setting that $p = 6$ and $p = 20000$ are close to settings of the two real data examples in the next section. Simulation results are given in Tables 1-3 and Figures 1 and 2. For better presentation, a logarithm transformation has been applied to the x-axis of Figures 1 and 2. We summarize the following key observations:

- Methods (i), (ii), (iii) and (iv) work well in that the $\hat{\epsilon}_{\max}$ values are all smaller than the desired error margins and the coverage probabilities for all cases are all satisfactory, larger than $1 - \alpha = 0.95$. The mean $\hat{\epsilon}$ values are very small, indicating the sample sizes are large enough to control the overall estimation errors. The "avr.out" values are also small, indicating for those cases where not all the p markers are estimated within the desired error margin, only a very small proportion of the p markers have estimation errors larger than ϵ.

- Compared to the inequality methods (ii) and (iii), the normal approximation method (i) is more desirable with relatively small sample size

and equally high coverage probability. When large sample assumption is satisfied as in our simulation settings, normal approximation method can provide more reasonable estimation of the sample size than the crude inequality adjustment.

- Sample sizes computed using the regression methods (iv) are remarkably smaller than those using other methods. For example, under Case 1, to estimate NRI closely enough to the true value, method (iv) requires 1863 observations, less than 10% of the sample sizes required by methods (ii) and (iii). The regression approach presents an appealing saving on the planned sample sizes and may be favored in practice. Figures 1 and 2 also indicate that all the methods are robust with respect to the change of p under case 4. In addition, regression method can provide a much more parsimonious upper bound for the required sample size when p increases.

*3.3. STAR*D data*

We consider a real example taken from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) project conducted in the United States (Rush et al., 2004; Kuk, Li and Rush, 2010). STAR*D is a multisite, prospective, randomized, multistep clinical trial of outpatients with nonpsychotic major depressive disorder. The study compares various treatment options for those who do not obtain a satisfactory response with citalopram, a selective serotonin reuptake inhibitor antidepressant. Details of the study have been described previously in Rush et al. (2004). The primary research question is whether an individual patient could respond to the treatment. The outcome variable is defined as the 16-item Quick Inventory of Depression Symptomatology (QIDS16) is reduced by 50%. Investigators were trying to study the relationship between the probability of response and a set of baseline measures.

The baseline variable U we consider in this paper is the initial QIDS16 score (`base`). We then intend to evaluate the added predictive accuracy for demographic variables such as `age`, `sex`, duration of chronic disease (`chr`), general

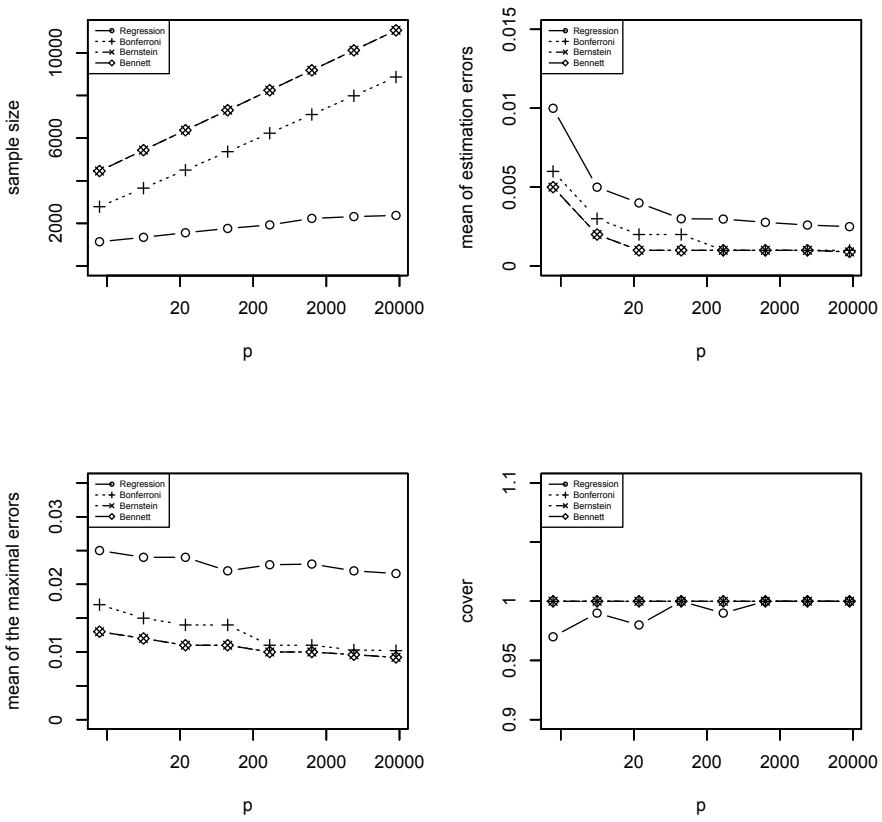

Figure 1: Simulation results for NRI under case 4. Top left: computed sample size using the five methods versus dimension p; Top right: mean $\hat{\epsilon}$ versus p; Bottom left: $\hat{\epsilon}_{\max}$ versus p; Bottom right: coverage rate versus p.

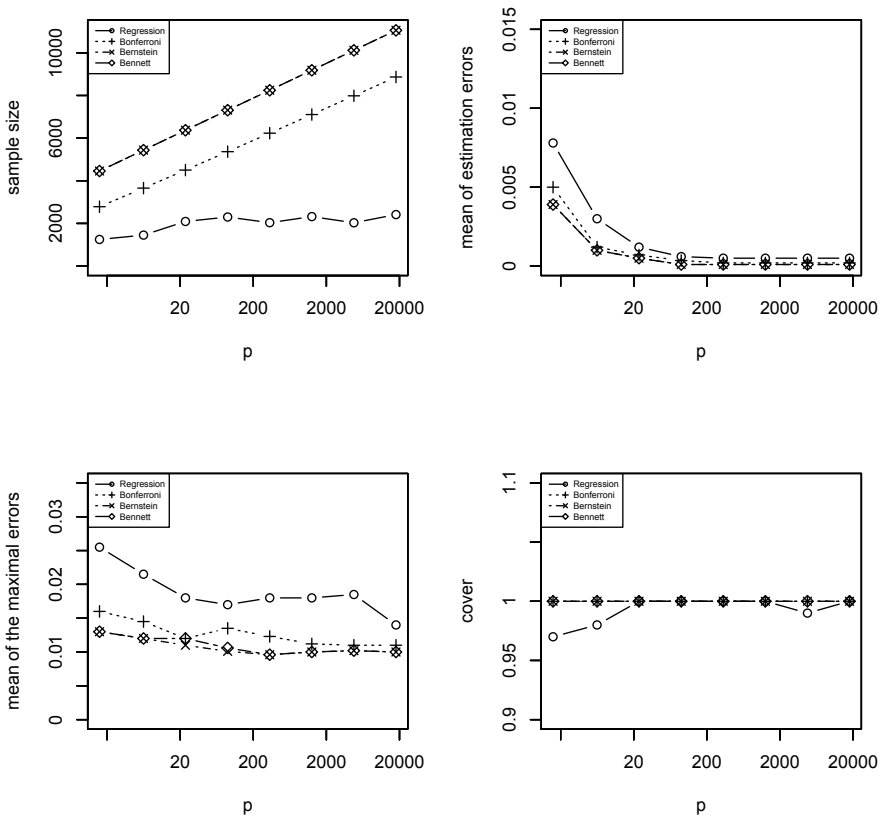

Figure 2: Simulation results for IDI under case 4. Top left: computed sample size using the five methods versus dimension p; Top right: mean $\hat{\epsilon}$ versus p; Bottom left: $\hat{\epsilon}_{\max}$ versus p; Bottom right: coverage rate versus p.

medical condition (`gmc`) and the anxiety/somatization factor (`anx`). The anxiety/somatization factor, derived from Cleary and Guy's factor analysis, includes six items from the original 17-item Hamilton Depression Rating Scale: the items for psychic anxiety, somatic anxiety, gastrointestinal somatic symptoms, general somatic symptoms, hypochondriasis, and insight. Furthermore, the change of QIDS score from baseline to week 2 (`change`) is also commonly used as a meaningful marker for the response (Kuk, Li and Rush, 2010). We consider the problem of estimating the NRI and IDI for $p = 6$ covariates. In the following we compute sample sizes adequate for estimating these accuracy parameters using the methods proposed in this paper.

Let S_0 and R_0 be estimates of NRI and IDI based on all the 2280 observations. We treat S_0 and R_0 as the true values since they are based on abundant samples and then randomly sampled $n_0 = 200$ observations to construct a training sample. We consider $\alpha = 0.05$ and $\epsilon = 0.05, 0.1$. Once a required sample size n^* is obtained, we then randomly sample n^* observations from the remaining 2280 observations with replacement and compute estimates of NRI and IDI. The above procedure is repeated for $M = 200$ times and we compute (a) mean $\hat{\epsilon}$; (b)$\hat{\epsilon}_{\max}$; (c) cover and (d) avr.out defined as in the simulation section. The results are summarized in Table 4.

Eyeballing Table 4, we notice that sample sizes computed using the regression methods perform very well and are most desirable. For example, based on method (iv), to estimate NRI as accurately as using the full sample, we only need 218 subjects, about 30% of those required by normal approximation and 20% of those required by inequality methods. Such findings agree with our previous simulation results.

3.4. Breast cancer study

Breast cancer is the second leading cause of deaths from cancer among women in the United States. Despite major progresses in breast cancer treatment, the ability to predict the metastatic behavior of tumor remains limited. The breast cancer study was first reported in van't Veer et al. (2002). 97 lymph

node-negative breast cancer patients 55 years old or younger participated in this study. Among them, 46 developed distant metastases within 5 years (metastatic outcome coded as 1) and 51 remained metastases free for at least 5 years (metastatic outcome coded as 0). Clinical risk factors (confounders) collected include age, tumor size, histological grade, angioinvasion, lymphocytic infiltration, estrogen receptor (ER), and progesterone receptor (PR) status. All these low-dimensional variables are considered as the baseline variable U in this study.

Expression levels for 24,481 gene probes were collected in the study. After the removal of genes with severe missingness there are still 24,188 genes. To quantify the accuracy improvement for predicting the disease using each gene on top of the baseline variables is a main scientific goal.

For this high-dimensional gene expression setting, we consider applying our methods to evaluate how large a sample is needed in a future study to estimate the NRI and IDI for all $p = 24188$ genes with adequate estimation accuracy. We set $\alpha = 0.05$, $\epsilon = 0.05, 0.1$. The sample sizes computed using the five different approaches are given in Table 5.

We notice that the results in this example agree with previous numerical results. The regression approach yield much smaller sample size requirement for the study. For example, we need roughly 452 subjects to estimate all 24,188 NRI values with a uniform error bound 0.1 at a 95% confidence level. To achieve the same accuracy, the normal approximation and inequality approaches would require tens of thousands of subjects.

4. Remarks

There could be further extensions for the proposed sample size calculation methods. Besides the Bonferroni correction, we may also adopt other adjustment to control the overall error probability. These correction methods may have better performance under some special assumptions. Also, the models considered in the regression approach of this paper could be fine-tuned with more complicated structure to achieve more reasonable results. From our sim-

ulations the simple nonlinear model seem to work well and could serve as a convenient starting point for more sophisticated design consideration.

The methods proposed in this paper can also be suitable for estimating parameters other than NRI and IDI under high-dimensional settings. When we need to deal with extremely large number of parameters, many familiar sample size calculation methods for standard parameter estimation must be modified in the same way as we have done this paper. Our proposed methods are expected to be useful for various large scale study designs. Among the three methods, regression method tends to produce relative smaller sample size requirement. The reduction is quite general and not limited to NRI and IDI. The regression method explicitly models the analytic relationship between sample size and estimation error and therefore the model prediction may be more informative than approximation and inequality methods. One may need to seek external information to obtain a sharper variance bound v used in these two approaches in order to improve them.

Acknowledgment

The authors would like to thank the referees for their valuable comments and suggestions. The authors would like to thank Professor Stephen Fienberg for his support and valuable comments. Jiang's research is partially supported by the Hong Kong RGC grant (PolyU 253023/16P). Li's work is partially supported by AcRF R-155-000-174-114.

References

Bennett, G., 1962. Probability inequalities for the sum of independent random variables. J. Am. Stat. Assoc. 57, 33–45.

Bickel, P., Levina, E., 2008. Covariance regularization by thresholding. Ann. Stat. 36, 2577–2604.

Chow, S., Wang, H., Shao, J., 2007. Sample Size Calculations in Clinical Research. CRC press.

Dobbin, K., Song, X., 2013. Sample size requirements for training high-dimensional risk predictors. Biostatistics 14, 639–652.

Jiang, B., 2015. An empirical estimator for the sparsity of a large covariance matrix under multivariate normal assumptions. Ann. Inst. Stat. Math. 67, 211–227.

Kerr, K., Wang, Z., Janes, H., McClelland, R., Psaty, B., Pepe, M., 2014. Net reclassification indices for evaluating risk prediction instruments: a critical review. Epidemiology 25, 114–121.

Kuk, A., Li, J., Rush, A., 2010. Recursive subsetting to identify patients in the star*d: a method to enhance the accuracy of early prediction of treatment outcome and to inform personalized care. J. Clin. Psychiatry 71, 1502–1508.

van der Laan, M., Bryan, J., 2001. Gene expression analysis with the parametric bootstrap. Biostatistics 2, 445–461.

Leening, M., Steyerberg, E., Calster, B., D'Agostino, R., Pencina, M., 2014. Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective. Stat. Med. 33, 3415–3418.

Li, J., Fine, J., 2004. On sample size for sensitivity and specificity in prospective diagnostic accuracy studies. Stat. Med. 23, 2537–2550.

Li, J., Fine, J., 2008. Roc analysis with multiple classes and multiple tests: methodology and its application in microarray studies. Biostatistics 9, 566–576.

Li, J., Jiang, B., Fine, J., 2013. Multicategory reclassification statistics for assessing improvements in diagnostic accuracy. Biostatistics 14, 382–394.

Liu, H., Lafferty, J., Wasserman, L., 2009. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. J. Mach. Learn. Res. 10, 2295–2328.

Merlevede, F., Peligrad, M., Rio, E., 2009. Bernstein inequality and moderate deviations under strong mixing conditions. High dimensional probability V: the Luminy volume, Institute of Mathematical Statistics. pp. 273–292.

Obuchowski, N., Zhou, X., 2002. Prospective studies of diagnostic test accuracy when disease prevalence is low. Biostatistics 3, 477–492.

Pencina, M., D'Agostino, R., Demler, O., 2012. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. Stat. Med. 31, 101–113.

Pencina, M., D'Agostino, R., Pencina, K., Janssens, A., Greenland, P., 2012. Interpreting incremental value of markers added to risk prediction models. Am. J. Epidemiol. 176, 473–481.

Pencina, M., D'Agostino, R., Steyerberg, E., 2011. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat. Med. 30, 11–21.

Pencina, M., D'Agostino, R., Vasan, R., 2008. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat. Med. 27, 157–172.

Ravikumar, P., Wainwright, M., Raskutti, G., Yu, B., 2011. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. Electron. J. Stat. 5, 935–980.

Rush, A., Fava, M., Wisniewski, S., Lavori, P., Trivedi, M., Sackeim, H., Thase, M., Nierenberg, A., Quitkin, F., Kashner, T., et al. 2004. Sequenced treatment alternatives to relieve depression. Control. Clin. Trials 25, 119–142.

Shao, J., Tu, D., 2012. The jackknife and bootstrap. Springer Science & Business Media.

Steinberg, M., Fine, J., Chappell, R., 2009. Sample size for positive and negative predictive value in diagnostic research using case-control designs. Biostatistics 10, 94–105.

Steyerberg, E., Vickers, A., Cook, N., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M., Kattan, M., 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology 21, 128.

Uno, H., Tian, L., Cai, T., Kohane, I., Wei, L., 2013. A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. Stat. Med. 32, 2430–2442.

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

Zhou, X., Obuchowski, N., McClish, D., 2011. Statistical Methods in Diagnostic Medicine. John Wiley & Sons.

Table 1: Simulation study to compare the five different approaches under case 1: (i) Normal approximation as in (4); (ii) Bernstein inequality as in (5); (iii) Bennett inequality as in (6); (iv) The regression approach as in (8). ϵ is the designed error margin . n^* is the sample size computed from the individual methods. $\hat{\epsilon}$ is the observed error margin for the estimation for all p markers.

NRI				
$\epsilon = 0.05$	(i)	(ii)	(iii)	(iv)
n*	5365	7310	7309	1863
mean $\hat{\epsilon}$	0.001	0.001	0.001	0.003
$\hat{\epsilon}_{\max}$	0.012	0.010	0.010	0.022
cover	1.000	1.000	1.000	1.000
avr.out	0.000	0.000	0.000	0.000
$\epsilon = 0.1$	(i)	(ii)	(iii)	(iv)
n*	1341	1857	1856	558
mean $\hat{\epsilon}$	0.005	0.003	0.003	0.008
$\hat{\epsilon}_{\max}$	0.029	0.024	0.024	0.045
cover	1.000	1.000	1.000	1.000
avr.out	0.000	0.000	0.000	0.000
IDI				
$\epsilon = 0.05$	(i)	(ii)	(iii)	(iv)
n*	5365	7310	7309	2510
mean $\hat{\epsilon}$	$<$0.001	$<$0.001	$<$0.001	0.001
$\hat{\epsilon}_{\max}$	0.012	0.010	0.010	0.017
cover	1.000	1.000	1.000	1.000
avr.out	0.000	0.000	0.000	0.000
$\epsilon = 0.1$	(i)	(ii)	(iii)	(iv)
n*	1341	1857	1856	497
mean $\hat{\epsilon}$	0.001	0.001	0.001	0.002
$\hat{\epsilon}_{\max}$	0.023	0.017	0.017	0.039
cover	1.000	1.000	1.000	1.000
avr.out	0.000	0.000	0.000	0.000

Table 2: Simulation study to compare the five different approaches under case 2: (i) Normal approximation as in (4); (ii) Bernstein inequality as in (5); (iii) Bennett inequality as in (6); (iv) The regression approach as in (8). ϵ is the designed error margin . n^* is the sample size computed from the individual methods. $\hat{\epsilon}$ is the observed error margin for the estimation for all p markers.

NRI				
$\epsilon = 0.05$	(i)	(ii)	(iii)	(iv)
n*	6192	8203	8202	2101
mean $\hat{\epsilon}$	0.001	0.001	0.001	0.002
$\hat{\epsilon}_{\max}$	0.012	0.010	0.010	0.021
cover	1.000	1.000	1.000	1.000
avr.out	0.000	0.000	0.000	0.000
$\epsilon = 0.1$	(i)	(ii)	(iii)	(iv)
n*	1548	2084	2083	412
mean $\hat{\epsilon}$	0.002	0.002	0.002	0.006
$\hat{\epsilon}_{\max}$	0.025	0.022	0.022	0.042
cover	1.000	1.000	1.000	1.000
avr.out	0.000	0.000	0.000	0.000
IDI				
$\epsilon = 0.05$	(i)	(ii)	(iii)	(iv)
n*	6192	8203	8202	1755
mean $\hat{\epsilon}$	$<$0.001	$<$0.001	$<$0.001	0.001
$\hat{\epsilon}_{\max}$	0.012	0.009	0.009	0.020
cover	1.000	1.000	1.000	1.000
avr.out	0.000	0.000	0.000	0.000
$\epsilon = 0.1$	(i)	(ii)	(iii)	(iv)
n*	1548	2084	2083	586
mean $\hat{\epsilon}$	0.001	0.001	0.001	0.002
$\hat{\epsilon}_{\max}$	0.024	0.019	0.020	0.039
cover	1.000	1.000	1.000	1.000
avr.out	0.000	0.000	0.000	0.000

Table 3: Simulation study to compare the five different approaches under case 3: (i) Normal approximation as in (4); (ii) Bernstein inequality as in (5); (iii) Bennett inequality as in (6); (iv) The regression approach as in (8). ϵ is the designed error margin . n^* is the sample size computed from the individual methods. $\hat{\epsilon}$ is the observed error margin for the estimation for all p markers.

NRI				
$\epsilon = 0.05$	(i)	(ii)	(iii)	(iv)
n*	6579	8619	8617	2421
mean $\hat{\epsilon}$	0.002	0.002	0.002	0.004
$\hat{\epsilon}_{\max}$	0.015	0.012	0.012	0.027
cover	1.000	1.000	1.000	1.000
avr.out	0.000	0.000	0.000	0.000
$\epsilon = 0.1$	(i)	(ii)	(iii)	(iv)
n*	1645	2190	2189	579
mean $\hat{\epsilon}$	0.005	0.004	0.004	0.010
$\hat{\epsilon}_{\max}$	0.035	0.024	0.026	0.061
cover	1.000	1.000	1.000	1.000
avr.out	0.000	0.000	0.000	0.000
IDI				
$\epsilon = 0.05$	(i)	(ii)	(iii)	(iv)
n*	6579	8619	8617	3032
mean $\hat{\epsilon}$	<0.001	<0.001	<0.001	<0.001
$\hat{\epsilon}_{\max}$	0.013	0.012	0.012	0.019
cover	1.000	1.000	1.000	1.000
avr.out	0.000	0.000	0.000	0.000
$\epsilon = 0.1$	(i)	(ii)	(iii)	(iv)
n*	1645	2190	2189	692
mean $\hat{\epsilon}$	0.001	<0.001	<0.001	0.002
$\hat{\epsilon}_{\max}$	0.033	0.021	0.021	0.039
cover	1.000	1.000	1.000	0.990
avr.out	0.000	0.000	0.000	0.001

Table 4: Real data analysis to compare the five different approaches using the STAR*D data: (i) Normal approximation as in (4); (ii) Bernstein inequality as in (5); (iii) Bennett inequality as in (6); (iv) The regression approach as in (8). ϵ is the designed error margin . n^* is the sample size computed from the individual methods. $\hat{\epsilon}$ is the observed error margin for the estimation for all p markers.

NRI				
$\epsilon = 0.05$	(i)	(ii)	(iii)	(iv)
n*	2784	4458	4457	1360
mean $\hat{\epsilon}$	0.012	0.010	0.010	0.015
$\hat{\epsilon}_{\max}$	0.024	0.020	0.020	0.032
cover	0.975	0.990	0.995	0.945
avr.out	0.167	0.167	0.167	0.182
$\epsilon = 0.1$				
n*	696	1133	1132	218
mean $\hat{\epsilon}$	0.018	0.016	0.016	0.026
$\hat{\epsilon}_{\max}$	0.038	0.034	0.034	0.054
cover	1.000	1.000	1.000	0.980
avr.out	0.000	0.000	0.000	0.167
IDI				
$\epsilon = 0.05$	(i)	(ii)	(iii)	(iv)
n*	2784	4458	4457	1374
mean $\hat{\epsilon}$	0.003	0.002	0.003	0.004
$\hat{\epsilon}_{\max}$	0.009	0.009	0.009	0.015
cover	1.000	1.000	1.000	0.990
avr.out	0.000	0.000	0.000	0.167
$\epsilon = 0.1$				
n*	696	1133	1132	454
mean $\hat{\epsilon}$	0.006	0.005	0.005	0.008
$\hat{\epsilon}_{\max}$	0.022	0.016	0.016	0.030
cover	1.000	1.000	1.000	0.985
avr.out	0.000	0.000	0.000	0.167

Table 5: Sample size calculation for the high-dimensional Breast Cancer data using the five different approaches: (i) Normal approximation as in (4); (ii) Bernstein inequality as in (5); (iii) Bennett inequality as in (6); (iv) The regression approach as in (8).

	(i)	(ii)	(iii)	(iv)
NRI (ϵ, α)				
$(0.05, 0.05)$	9013	11210	11208	1747
$(0.1, 0.05)$	2253	2848	2847	452
IDI (ϵ, α)				
$(0.05, 0.05)$	9013	11210	11208	6192
$(0.1, 0.05)$	2253	2848	2847	1420