Published in final edited form as:

Stat Med. 2010 February 28; 29(5): 597–613. doi:10.1002/sim.3823.

Hybrid Pooled—Unpooled Design for Cost-Efficient Measurement of Biomarkers

Enrique F. Schisterman*, Albert Vexler, Sunni L. Mumford, and Neil J. Perkins

Division of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH/DHHS, 6100 Executive Boulevard, Rockville, Maryland 20852, U.S.A.

Summary

Evaluating biomarkers in epidemiological studies can be expensive and time consuming. Many investigators use techniques such as random sampling or pooling biospecimens in order to cut costs and save time on experiments. Commonly, analyses based on pooled data are strongly restricted by distributional assumptions that are challenging to validate because of the pooled biospecimens. Random sampling provides data that can be easily analyzed. However, random sampling methods are not optimal cost-efficient designs for estimating means. We propose and examine a cost-efficient hybrid design that involves taking a sample of both pooled and unpooled data in an optimal proportion in order to efficiently estimate the unknown parameters of the biomarker distribution. In addition, we find that this design can be utilized to estimate and account for different types of measurement and pooling error, without the need to collect validation data or repeated measurements. We show an example where application of the hybrid design leads to minimization of a given loss function based on variances of the estimators of the unknown parameters. Monte Carlo simulation and biomarker data from a study on coronary heart disease are used to demonstrate the proposed methodology.

Keywords

Maximum likelihood; Measurement error; Pooling; Random sampling; Receiver operating characteristics; Sampling design

1. Introduction

Epidemiological studies increasingly take advantage of new and sophisticated biomarkers to more effectively measure exposures and predictors of interest. However, assaying biomarkers can be expensive and labor intensive, and even the least expensive individual assays may be infeasible to analyze in large cohorts. As a result, it is advantageous to find cost-efficient strategies for sampling. Dorfman [1], Faraggi *et al.* [2], Mumford *et al.* [3] Schisterman and Vexler [4], Vexler *et al.* [5], Weinberg and Umbach [6], and Zhang and Gant [7], all discuss pooling and random sampling as two approaches commonly used by investigators to reduce overall cost. Simple random sampling involves choosing and testing a random subset of available samples. In this case, several individual biospecimens are ignored; however, estimation of the operating characteristics of these individual biospecimens is straightforward. Pooling, on the other hand, involves randomly grouping and physically mixing individual biological samples. Assays are then performed on the smaller number of pooled samples. Thus, pooling reduces the number of measurements

*schistee@mail.nih.gov.

without ignoring any individual biospecimens. Theoretically, the measurements of pooled observations are the average of the individual measurements, and have lent themselves well to normality and gamma assumptions in the literature [3;5;8].

Faraggi [2] describes that in certain situations the pooling design is more efficient than random sampling because it is applied to a greater number of individual biospecimens. However, Mumford *et al.* [3] and Vexler *et al.* [5] describe the tradeoff between random sampling and pooling and describe the conditions where random sampling is more efficient than pooling. For example, pooling is much less efficient in situations where the arithmetic mean is not an efficient estimator, e.g. in the case of log-normally distributed data. Moreover, when data are subject to a limit of detection, pooling cannot be recommended as the efficient design when more than 50% of the data are censored.

For a normally distributed biomarker, Faraggi *et al.* [2] show that when the mean is the only parameter of interest, the most cost-efficient strategy would be to pool all of the biospecimens and measure only one sample. That approach obviously does not provide a reasonable estimate of the variance. When estimating the variance of a biomarker, random sampling is simple and efficient and pooling is unnecessary per Liu and Schisterman [9]. Since we are most often concerned with estimation of both the mean and the variance in order to make inferences, it seems a natural progression to consider a hybrid design, one that utilizes both pooling and random sampling.

Previous authors have explored the effects of measurement error under both pooling and random sampling designs. In both designs, measurement error can be caused by problems related to instrumentation, unstable environmental conditions, operator subjectivity, etc. which Fuller [10], Coffin and Sukhatme [11], Schisterman *et al.* [12], and Perkins and Schisterman [13] discuss in more detail. Both designs are similarly affected by measurement error because the biospecimens, pooled and unpooled, are measured via the same process. Dunn [14], Freedman *et al.* [15], and Thurigen *et al.* [16] describe methods requiring repeated measurements or validation data to correct for measurement error. Faraggi [17], Liu *et al.* [18], Reiser [19], and Schisterman *et al.* [12] show the impact of measurement error on estimation. However, obtaining and assaying these replications can be difficult because of cost or other limitations [20]. In general, when no repeated measurements are available, additive measurement errors cannot be evaluated.

A form of error specific to the pooling design lies in the physical process of pooling biospecimens (e.g. effects of temperature, instrument, and technician variability) where the measurements of pooled observations deviate from the average of the individual measurements. We refer to this specific type of measurement error as pooling error. Obviously, measurements of individual specimens would be unaffected by pooling error as they are not pooled. As with other forms of error, pooling error may have a detrimental impact on effect estimation.

In this paper, we consider a hybrid sampling design, explore situations where pooling error and measurement error are present, and efficiently estimate the mean and variance of a normally distributed biomarker while accounting for these errors (Section 2). Additionally, this design allows for estimation of measurement error without requiring repeated sampling. Monte Carlo simulations are used to illustrate the developed techniques (Section 3). The presentation is limited to normally distributed data, but could be extended to other distribution functions as well. The optimal proportions of pooled and unpooled samples can be determined with respect to the given loss function. This methodology is exemplified in Section 4, by estimating the area under the receiver operating characteristics (ROC) curve, where the loss function of interest is the variance of the maximum likelihood estimator

(MLE) of the area under the curve (*AUC*). (See references by Faraggi and Reiser [21] and Shapiro [22] for details regarding ROC curves and the *AUC*.) The optimal proportion of pooled samples that minimizes the loss function is found. An application to cholesterol biomarker data from a study on coronary heart disease is presented in Section 5. In Section 6, we conclude with a discussion of our results.

2. Pooled—Unpooled Hybrid Design

In this section, MLEs are developed under the hybrid design, utilizing both pooled and unpooled samples, for several potential scenarios including no error, measurement error, pooling error, and measurement and pooling error. In each case, it is necessary to know the number of individual biological specimens available (N) and the number of assays that can be afforded (n) due to budget and/or time limitations.

2.1 Review of Pooled Design and Simple Random Sampling Design

Suppose that independent identically distributed (*iid*) biomarker levels $X_i \sim N(\mu_x, \sigma_x^2)$ can be observed. A random sample of size n can be drawn from the original population of N , and corresponding MLEs,

$$\widehat{\mu}_x = \frac{1}{n} \sum_{i=1}^n X_i, \quad \widehat{\sigma}_x^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\mu}_x)^2,$$

and their variances can be directly obtained, where specifically $\text{var}(\widehat{\mu}_x) = \frac{\sigma_x^2}{n}$.

Pooled samples are obtained by randomly grouping individual samples into groups of size p , physically combining specimens of the same group and testing each group as a single observation. Assuming that the pooled sample measurements reflect the individual samples' average we denote pooled samples of group size p by

$$Z_i = \frac{1}{p} \sum_{j=(i-1)p+1}^{ip} X_j, \quad i=1 \dots n = \frac{N}{p}.$$

By the additive property of the normal distribution, we obtain

$$Z_i \sim N\left(\mu_x, \frac{\sigma_x^2}{p}\right).$$

Obviously, the MLEs based on the pooled data are

$$\widehat{\mu}_x = \frac{1}{n} \sum_{i=1}^n Z_i, \quad \widehat{\sigma}_x^2 = p \cdot \left\{ \frac{1}{n} \sum_{i=1}^n (Z_i - \widehat{\mu}_x)^2 \right\},$$

where the variance of the estimator of μ is simply $\text{var}(\widehat{\mu}_x) = \frac{\sigma_x^2}{np}$.

One can show that the σ -estimators based on $\{Z_1, \dots, Z_n\}$ and $\{X_1, \dots, X_n\}$ are identically distributed as demonstrated by Liu and Schisterman [9]. However, in this simplest case involving pooling, the variance of the mean estimator is always smaller than that based on

$$\text{random sampling } \left(\frac{\sigma_x^2}{n} \leq \frac{\sigma_x^2}{p} \right).$$

2.2 Hybrid Design with Measurement Error

In this section, we consider the case where both pooled and unpooled samples are similarly affected by measurement error. Let us assume measurement error e_i^M is $e_i^M \sim N(0, \sigma_M^2)$. The actual measurements of the unpooled and pooled samples are *iid* random variables

$$X'_i = X_i + e_i^M \text{ and } Z'_i = Z_i + e_i^M = \frac{1}{p} \sum_{j=(i-1)p+1}^{ip} X_j + e_i^M \text{ respectively, so } X'_i \sim N(\mu_x, \sigma_x^2 + \sigma_M^2) \text{ and } Z'_i \sim N\left(\mu_x, \frac{\sigma_x^2}{p} + \sigma_M^2\right).$$

Next, consider a hybrid sample S , which is a combination of both pooled and unpooled samples, i.e. S is defined as $\{X_1, \dots, X_{[an]}, Z_1, \dots, Z_{[(1-\alpha)n]}\}$, where $\alpha \in [0, 1]$ represents the proportion that comes from a random sample. Note that $\alpha n + (1-\alpha)np \leq N$, but here we assume that $\alpha n + (1-\alpha)np = N$, if $\alpha < 1$, and that N and n are fixed based on budget and time constraints and the size of the cohort. In this paper we choose to optimize α , though since α and p are connected by a one-to-one mapping, it is equivalent to optimize by p .

The hybrid design with measurement error provides data $S' = \{X'_1, \dots, X'_{[an]}, Z'_1, \dots, Z'_{[(1-\alpha)n]}\}$. The MLEs of μ_x , σ_x^2 and σ_M^2 , based on S' can be derived from the following equations:

$$\begin{aligned} \hat{\mu}_x &= \frac{\alpha n \bar{X}' \left(\frac{\hat{\sigma}_x^2}{p} + \hat{\sigma}_M^2 \right) + (1-\alpha) n \bar{Z}' \left(\hat{\sigma}_x^2 + \hat{\sigma}_M^2 \right)}{\alpha n \left(\frac{\hat{\sigma}_x^2}{p} + \hat{\sigma}_M^2 \right) + (1-\alpha) n \left(\hat{\sigma}_x^2 + \hat{\sigma}_M^2 \right)}, \quad \hat{\sigma}_x^2 = \frac{p}{p-1} \left(\frac{\sum_{i \leq \alpha n} (X'_i - \hat{\mu}_x)^2}{\alpha n} - \frac{\sum_{i \leq (1-\alpha)n} (Z'_i - \hat{\mu}_x)^2}{(1-\alpha)n} \right), \\ \hat{\sigma}_M^2 &= \frac{1}{(1-\alpha)n} \sum_{i \leq (1-\alpha)n} (Z'_i - \hat{\mu}_x)^2 - \frac{\hat{\sigma}_x^2}{p}, \end{aligned}$$

where $\bar{X}'$ denotes the mean of $X'_1, \dots, X'_{[an]}$, and $\bar{Z}'$ denotes the mean of $Z'_1, \dots, Z'_{[(1-\alpha)n]}$. These estimates are asymptotically distributed as

$$\sqrt{n} \begin{pmatrix} \hat{\mu}_x - \mu_x \\ \hat{\sigma}_x^2 - \sigma_x^2 \\ \hat{\sigma}_M^2 - \sigma_M^2 \end{pmatrix} \sim N_3(0, \Sigma),$$

where Σ is the inverse of the Fisher Information Matrix, I , leading to the following variances:

$$\text{var}(\hat{\mu}_x) = \frac{\left(\frac{\sigma_x^2 + \sigma_M^2}{p}\right) \left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)}{\alpha n \left(\frac{\sigma_x^2}{p} + \sigma_M^2\right) + (1-\alpha)n(\sigma_x^2 + \sigma_M^2)},$$

$$\text{var}(\hat{\sigma}_x^2) = \frac{2p^2}{(p-1)^2} \frac{\alpha \left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2 + (1-\alpha)(\sigma_x^2 + \sigma_M^2)^2}{\alpha(1-\alpha)n}.$$

(For details, see Appendix A1.)

Proposition 2.2.1—The values $\alpha = 0$ (pooled samples only) and

$$\alpha = \frac{N^2 \sigma_x^4 + N(n+N) \sigma_x^2 \sigma_M^2 + nN \sigma_M^4}{N^2 \sigma_x^4 + 2N^2 \sigma_x^2 \sigma_M^2 + (2nN - n^2) \sigma_M^4}$$

minimize $\text{var}(\hat{\mu}_x)$ and $\text{var}(\hat{\sigma}_x^2)$, respectively. (Proof in Appendix A2).

Commonly, $\text{var}(\hat{\mu})$ and $\text{var}(\hat{\sigma})$ are components of a given loss-function that we would like to minimize with respect to α . In this case, it is generally not recommended for α to be exactly equal to 0 or 1. Formal notations and proof schemes mentioned in the Appendices can be used to obtain the optimal α under a stated loss-function. In Section 4, we show the application of this methodology for the case where the loss-function of interest is the variance of the *AUC* estimator, a function of both $\text{var}(\hat{\mu})$ and $\text{var}(\hat{\sigma})$.

2.3 Hybrid Design with Pooling Error

Pooling error is a specific type of measurement error that occurs as a result of the process of physically pooling the samples together and may introduce additional variability. We denote this error as e_i^p and assume $e_i^p \sim N(0, \sigma_p^2)$. Subsequently, we assume that the pooled

observations have the form $Z_i = \frac{1}{p} \sum_{j=(i-1)p+1}^{ip} X_j + e_i^p$, where $Z_i \sim N\left(\mu_x, \frac{\sigma_x^2}{p} + \sigma_p^2\right)$. The maximum likelihood estimators of μ_x , σ_x^2 , and σ_p^2 , and the corresponding Fisher Information Matrix, based on the hybrid design with pooling error only, are described in detail in Appendix A3 and Appendix A4. In contrast with section 2.1, $\hat{\mu}_x$ based on only pooled samples subject to pooling error, $\{Z_1, \dots, Z_n\}$, can be less efficient than that based on the random sample $\{X_1, \dots, X_n\}$. In this case, only the pooled samples and not the random samples are subject to error, making the equations slightly simpler.

The pooling-error-only case has vast applicability where research is based on well established biomarkers with thoroughly developed and precise measurement techniques. This is more often than not the case, as most studies assume no or a negligible amount of measurement error. However, the extra handling and processing that is required to pool samples and realize the efficiency benefits of the hybrid design could cause a non-negligible level of additional variability.

2.4 Measurement Error and Pooling Error

In order to allow for both pooling and measurement errors, the hybrid sample must consist of three different samples to be able to estimate all of the variance components (i.e.

σ_x^2 , σ_p^2 , and σ_M^2), otherwise the terms are not identifiable. Specifically, these three samples include one unpooled and two groups of pooled samples with different pooling group sizes, say p_1 and p_2 with $p_1 \neq p_2$. The observed sample S'' consists of:

$$\left\{ X'_i = X_i + e_i^M, 1 \leq i \leq \alpha_1 n \right\}, \left\{ Z'_i = \frac{1}{p_1} \sum_{j=(i-1)p_1+1}^{ip_1} X_j + e_i^p + e_i^M, 1 \leq i \leq \alpha_2 n \right\},$$

$$\left\{ Z''_i = \frac{1}{p_2} \sum_{j=(i-1)p_2+1}^{ip_2} X_j + e_i^p + e_i^M, 1 \leq i \leq (1 - \alpha_1 - \alpha_2)n \right\}$$

where $\alpha_1, \alpha_2 \in [0, 1]$, $\alpha_1 n + \alpha_2 n p_1 + (1 - \alpha_1 - \alpha_2) n p_2 = N$. Thus, the MLEs of $\mu_x, \sigma_x^2, \sigma_p^2$, and σ_M^2 , based on S^n can be obtained from the system of equations:

$$\widehat{\mu}_x = \frac{\alpha_1 n \bar{X}' A_1 A_2 + \alpha_2 n \bar{Z}' (\widehat{\sigma}_x^2 + \widehat{\sigma}_M^2) A_2 + (1 - \alpha_1 - \alpha_2) n \bar{Z}'' (\widehat{\sigma}_x^2 + \widehat{\sigma}_M^2) A_1}{\alpha_1 n A_1 A_2 + \alpha_2 n (\widehat{\sigma}_x^2 + \widehat{\sigma}_M^2) A_2 + (1 - \alpha_1 - \alpha_2) n (\widehat{\sigma}_x^2 + \widehat{\sigma}_M^2) A_1},$$

$$\widehat{\sigma}_x^2 = \frac{\sum_{i \leq \alpha_1 n} (X'_i - \widehat{\mu}_x)^2}{\alpha_1 n} - \widehat{\sigma}_M^2, \widehat{\sigma}_M^2 = \frac{p_1}{p_1 - 1} \left(\frac{\sum_{i \leq \alpha_2 n} (Z'_i - \widehat{\mu}_x)^2}{\alpha_2 n} - \frac{\sum_{i \leq \alpha_1 n} (X'_i - \widehat{\mu}_x)^2}{\alpha_1 n p_1} - \widehat{\sigma}_p^2 \right)$$

$$\text{and } \widehat{\sigma}_p^2 = \frac{\sum_{i \leq \alpha_2 n} (Z''_i - \widehat{\mu}_x)^2}{(1 - \alpha_1 - \alpha_2) n} - \frac{\sum_{i \leq \alpha_1 n} (X'_i - \widehat{\mu}_x)^2}{\alpha_1 n p_1} - \widehat{\sigma}_M^2 \frac{p_2 - 1}{p_2},$$

where $A_1 = \left(\frac{\widehat{\sigma}_x^2}{p_1} + \widehat{\sigma}_p^2 + \widehat{\sigma}_M^2 \right)$ and $A_2 = \left(\frac{\widehat{\sigma}_x^2}{p_2} + \widehat{\sigma}_p^2 + \widehat{\sigma}_M^2 \right)$.

These estimates asymptotically satisfy

$$\sqrt{n} \begin{pmatrix} \widehat{\mu}_x - \mu_x \\ \widehat{\sigma}_x^2 - \sigma_x^2 \\ \widehat{\sigma}_p^2 - \sigma_p^2 \\ \widehat{\sigma}_M^2 - \sigma_M^2 \end{pmatrix} \sim N_4(0, \Sigma),$$

where Σ is developed in Appendix A5. Thus, for example, we obtain that

$$\text{var}(\widehat{\mu}_x) = \frac{(\sigma_x^2 + \sigma_M^2) V_1 V_2}{n (\alpha_1 V_1 V_2 + \alpha_2 (\sigma_x^2 + \sigma_M^2) V_2 + (1 - \alpha_1 - \alpha_2) (\sigma_x^2 + \sigma_M^2) V_1)},$$

where $V_1 = \left(\frac{\sigma_x^2}{p_1} + \sigma_p^2 + \sigma_M^2 \right)$ and $V_2 = \left(\frac{\sigma_x^2}{p_2} + \sigma_p^2 + \sigma_M^2 \right)$.

For fixed N, n, p_1 and p_2 , α_1 and α_2 that minimize the variance of the mean are unique and can be calculated via conditional maximization. However, for other loss functions the solutions may not necessarily be unique.

3. Monte Carlo Simulations

In this section, Monte Carlo simulations are used to demonstrate the properties of the proposed methodology.

3.1 Measurement Error only

Normally distributed true biomarker levels, $N = 3000$ and $n = 1000$, were generated with $\mu=0$ and $\sigma_x^2=1$, which were subsequently combined with several scenarios of random measurement error, $\sigma_M^2=0.1, 0.4$, and 1.44 . Hybrid samples were formed from uncorrelated pooled and unpooled measurements in proportions of $\alpha = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$, and 0.9 .

Table 1 shows the results of these simulations and it seems clear that each estimator has Monte Carlo variance that is approximately equal to the derived theoretical variance. Table 1 confirms that the most efficient strategy for mean estimation utilizes the most pooled samples ($\alpha = 0.1$). In these simulations, the variance is minimized at small values of α but not always at the smallest value of α (0.2 for $\sigma_M^2=0.1$, 0.3 for $\sigma_M^2=0.4$, and 0.1 for $\sigma_M^2=1.44$). This discrepancy is due to the fact that the theoretical variance is calculated using a pooling size that does not have to be an integer value, but for the simulation, the value of p is rounded to the nearest integer value. One impact of the rounding is that it could potentially use more samples than are available. See Figure 1a.

For variance estimation, when measurement error exists, strategies that utilize more pooled samples are more efficient. Variance estimation is not as affected by rounding p as mean estimation. See Figure 1b. In addition, when measurement error is present, estimation is only possible when both pooled and unpooled samples are used. In the case where $\alpha = 0.1$, there is probably not a sufficient sample of unpooled data for proper estimation.

As with all maximum likelihood estimation, caution should be used when applying these methods due to questions of robustness, specifically regarding $\text{var}(\hat{\sigma}_x^2)$ used for optimization. Variance estimates are restricted to nonnegative values, however in the rare cases when unrestricted MLEs may result in $\hat{\sigma}^2 < 0$, we apply techniques used by Schisterman *et al.* and replace $\hat{\sigma}^2$ by a very small number [12]. Specifically, we replace $\hat{\sigma}_M^2$ with $1.4 \times \hat{\sigma}_x^2, 1.2 \times \hat{\sigma}_x^2, 1.1 \times \hat{\sigma}_x^2, 1.01 \times \hat{\sigma}_x^2, 1.005 \times \hat{\sigma}_x^2$, or $1.001 \times \hat{\sigma}_x^2$, and evaluate the appropriateness of the various alternatives. In addition, we compare these results to that

of replacing the variance with $\hat{\sigma}_M^2 = \frac{1}{N} \left| \frac{1}{(1-\alpha)n} \sum_{i \leq (1-\alpha)n} (Z_i - \hat{\mu}_x)^2 - \frac{\hat{\sigma}_x^2}{p} \right|$, as recommended by Vonesh and Carter [23]. In our case, the choice of replacement value did not have a large impact on the results.

3.2 Pooling Error Only

Again, normally distributed data was generated with $\mu=0$ and $\sigma_x^2=1$, but with varying pooling errors ($\sigma_p^2=0.3, 0.7$). Individual biospecimens available for testing ranged from ($N = 300, 200$), with varying numbers of assayed performed ($n = 150, 100, 50$). Hybrid samples were formed from uncorrelated pooled and unpooled measurements in proportions of ($\alpha = 0.25, 0.5, 0.75$). For values of α that do not imply integer values of p , the value of p in the simulation was rounded to the nearest integer value.

Table 2 displays the results of these simulations where it is clear that the Monte Carlo variance for the estimators is approximately equal to the derived theoretical variance. Table 2 also reinforces Proposition 2.3.1 when estimating the mean. Three cases are shown to demonstrate each of the possible optimal values of α . Case 1 corresponds to the situation where the pooling error is small in relation to the variability in the sample. Here the variance is minimized when more pooled samples are used (α is closest to zero). In Case 2, the

pooling error is greater than the sampling error and the variance is minimized when fewer pooled samples are used (α closest to 1). Case 3 is an example of a situation where the formula for the optimal α is used, and the simulation matches what we would expect. As expected for estimation of the variance, the simulation results show that for each different set of parameters, the sample which utilizes the greatest proportion of random samples (in this case where $\alpha = 0.75$) is always the most efficient.

In these cases, tests for normality (e.g. Kolmogorov Smirnov, Shapiro-Wilk) did not reject the hypothesis regarding the normal distribution of the maximum likelihood estimators repeated by Monte Carlo simulations.

4. An Application to AUC Estimation

A common measure of a biomarker's ability to discriminate between those with and without disease is the AUC as Faraggi and Reiser [21], Shapiro [22], and Wieand *et al.* [24] describe. Let X and Y denote diseased and healthy individual's test scores, respectively, and the $AUC = Pr(X > Y)$. Here, $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ lead to $AUC = \Phi(\delta)$, where

$$\delta = \frac{(\mu_x - \mu_y)}{\sqrt{(\sigma_x^2 + \sigma_y^2)}}.$$

Estimating AUC requires estimates of the mean and variance of both populations, so the optimal sampling strategy requires finding the proportion of pooled samples that minimizes the variance of the AUC estimator (the loss function of interest here). Without loss of generality and for demonstration purposes, we assume that only pooling error is in effect. The appropriate variances of our estimators from Section 2.3 (Appendix A3) are applied to the loss function of interest.

Specifically, one can show that the variance of the AUC estimator is:

$$Var(\widehat{AUC}) = Var(\Phi(\widehat{\delta})) \cong \phi(\widehat{\delta})^2 \cdot Var(\widehat{\delta}),$$

where

$$\begin{aligned} Var(\widehat{\delta}) = & \left(\frac{\partial \delta}{\partial \mu_x} \right)^2 \cdot Var(\widehat{\mu}_x) + \left(\frac{\partial \delta}{\partial \sigma_x} \right)^2 \cdot Var(\widehat{\sigma}_x) + \left(\frac{\partial \delta}{\partial \mu_y} \right)^2 \cdot Var(\widehat{\mu}_y) + \left(\frac{\partial \delta}{\partial \sigma_y} \right)^2 \cdot Var(\widehat{\sigma}_y) \\ & + \left(\frac{\partial \delta}{\partial \mu_x} \right) \left(\frac{\partial \delta}{\partial \sigma_x} \right) \cdot Cov(\widehat{\mu}_x, \widehat{\sigma}_x) + \left(\frac{\partial \delta}{\partial \mu_y} \right) \left(\frac{\partial \delta}{\partial \sigma_y} \right) \cdot Cov(\widehat{\mu}_y, \widehat{\sigma}_y) \end{aligned}$$

(See for details [3].)

Assume observations are of the form S'_x and S'_y described in Section 2.3. Applying the variances of the μ and σ estimators found in Section 2.3 (Appendix A3), we can directly find the values of α that minimizes the variance of the AUC estimator,

$$\begin{aligned} \alpha_x = & \frac{\sigma_x^4 (\mu_x - \mu_y)^2 (\sigma_x^2 - \sigma_{z_x}^2) \pm \sigma_x^3 \sigma_{z_x} (\mu_x - \mu_y) (\sigma_x^2 + \sigma_y^2) \sqrt{2(\sigma_x^2 - \sigma_{z_x}^2)}}{(\sigma_x^2 - \sigma_{z_x}^2) \left\{ \sigma_x^2 (\mu_x - \mu_y)^2 (\sigma_x^2 - \sigma_{z_x}^2) - 2\sigma_{z_x}^2 (\sigma_x^2 + \sigma_y^2)^2 \right\}}, \\ \alpha_y = & \frac{\sigma_y^4 (\mu_x - \mu_y)^2 (\sigma_y^2 - \sigma_{z_y}^2) \pm \sigma_y^3 \sigma_{z_y} (\mu_x - \mu_y) (\sigma_x^2 + \sigma_y^2) \sqrt{2(\sigma_y^2 - \sigma_{z_y}^2)}}{(\sigma_y^2 - \sigma_{z_y}^2) \left\{ \sigma_y^2 (\mu_x - \mu_y)^2 (\sigma_y^2 - \sigma_{z_y}^2) - 2\sigma_{z_y}^2 (\sigma_x^2 + \sigma_y^2)^2 \right\}}, \end{aligned}$$

where $\sigma_{z_x}^2 = \frac{\sigma_x^2}{p} + \sigma_{p_x}^2$, $\sigma_{z_y}^2 = \frac{\sigma_y^2}{p} + \sigma_{p_y}^2$, $\text{var}(e_x^p) = \sigma_{p_x}^2$, and $\text{var}(e_y^p) = \sigma_{p_y}^2$, where x_p and y_p respectively correspond to the pooled samples for cases and controls.

In general, for a loss function corresponding to the variance of our parameters, we can find the optimal value of α which minimizes the loss function and can be evaluated as a function of unknown parameters. Section 5 illustrates this theoretical result with a practical example.

5. Example

A study of coronary heart disease at the Cedars-Sinai Medical Center investigated the ability of a cholesterol biomarker to discriminate between individuals who recently survived a myocardial infarction (cases) and those with no previous history of myocardial infarction (controls). Biomarker measurements were collected on 40 cases and 40 controls and blood samples were pooled in groups of 2, keeping cases and controls separate, and then re-measured. We therefore have individual measurements for 40 cases and 40 controls, as well as measurements for 20 samples of two pooled cases and 20 samples of two pooled controls. This allowed us to organize hybrid samples by taking combinations of individual and pooled measurements. The mean ($\pm$ SD) was 205.5 ($\pm$ 42.3) for unpooled control samples and 207.9 ($\pm$ 35.2) for pooled control samples. The mean ($\pm$ SD) was 226.8 ($\pm$ 41.7) for unpooled case samples and 223.2 ($\pm$ 37.3) for pooled case samples. The standard deviation of the differences between pooled measurements and the average of corresponding individual measurements was 6.7 for cases and 7.4 for controls.

For the purposes of this example, we assume that only pooling error is in effect. Under these assumptions, we apply the results from Section 4 for determining the optimal proportions of pooled and unpooled samples to maximize efficiency when estimating the *AUC*. Using the empirical means and variances, we find that $\alpha = 0.4$ for the cases and $\alpha = 0.35$ for the controls to minimize the variance of the *AUC* estimator when only 20 assays are tested ($n=20$). In other words, it is recommended that hybrid samples include 8 individual and 12 pooled ($p=2$) samples for the cases and 7 individual and 13 pooled samples for the controls to minimize the variance of the *AUC* estimator. More control samples should be pooled since the estimated pooling error is slightly lower among the controls than the cases. Using the developed methodology, we are also able to estimate the amount of pooling error. In particular, we took a hybrid sample of both cases and controls using the optimal levels of α . The hybrid sample was distributed as follows (mean ($\pm$ SD)): unpooled control samples 201.0 ($\pm$ 31.5), pooled control samples 205.1 ($\pm$ 30.8), unpooled case samples 229.4 ($\pm$ 43.1), and pooled case samples 220.8 ($\pm$ 34.8). The MLEs were then estimated using the proposed methodology, and we found that

$\hat{\mu}_x = 223.5$, $\hat{\sigma}_x^2 = 1662.4$, $\hat{\sigma}_{p_x}^2 = 283.4$, $\hat{\mu}_y = 203.7$, $\hat{\sigma}_y^2 = 859.4$, and $\hat{\sigma}_{p_y}^2 = 449.3$. The *AUC* estimator was found to be 0.65 with a variance of 0.0050.

By implementation of the hybrid design in a small simulation study testing only 20 samples, we were able to reduce the variability of the *AUC* estimator by 30% as compared to using only a random sample (hybrid design: $A\hat{U}C = 0.621$, $\text{Var}(A\hat{U}C) = 0.0049$; random sample: $A\hat{U}C = 0.641$, $\text{Var}(A\hat{U}C) = 0.0071$). As compared to the full sample, the hybrid design used only half the number of samples with only a 37% loss of efficiency, whereas the random sample lost 96% (full sample: $A\hat{U}C = 0.640$, $\text{Var}(A\hat{U}C) = 0.0036$).

6. Discussion

We present methodology for a cost-efficient hybrid sampling design that combines both pooled and unpooled samples in optimal proportions to efficiently estimate the unknown

parameters of the biomarker distribution. Maximum likelihood estimators and their asymptotic distributions are given for various situations where errors are present. Application of this design also makes possible estimation of measurement error without repeated sampling. The hybrid design is similar to repeated measures but also retains, and in some cases improves, the efficiency of estimating μ_x and σ_x^2 . The information on repeated measures can only be used to estimate σ_M^2 , whereas use of the hybrid design can be used to estimate all parameters of interest. Any outside information available on the source and quantity of the errors can improve implementation of the hybrid design and subsequent estimation.

Misspecification of the measurement error model is of course a concern for estimation of α , and in turn for estimation of the main parameters of interest. Researchers often have substantial knowledge on the measurement of similar biomarkers, or biomarkers measured in a similar manner, and the laboratory conditions under which pooling will be performed (e.g. mechanical versus technician). This knowledge aids model specification. As mentioned, the pooling-error-only model is especially appropriate in cases where research is based on well-established and precise measurement techniques. In practice, it is best to allow for both types of error, especially if no prior knowledge of the biomarkers and the measurement process exists.

It should be noted that owing to the complexity of estimating the biomarker distribution from observed convoluted data, generalized likelihood or nonparametric inferences based only on pooled data may not be feasible, since the distribution of the observed averages involves convolution of p random variables of a biomarker distribution as discussed by Vexler *et al.* [8] In several cases, like those mentioned above, the hybrid design can be based on pre-specified proportions of pooled and unpooled samples (e.g. $\alpha = 0.5$). The definition and approximation of optimal values of the proportions with respect to a given loss function can also belong to statements of problems that forego the hybrid design execution. In this article, several rules are analyzed which aid in determining the optimal proportion of unpooled samples. These rules are dependent on parameters (e.g. σ_x^2, σ_p^2) that are unknown at the design stage. It is usually assumed that the experimenter has *a priori* knowledge regarding the distribution of measurements, and good initial values of the unknown parameters are chosen based on this knowledge. This information might be available from previous related experiments, preliminary studies, or pre-testing of some kind. This information could also be taken from a subsample of the N available samples, with some modifications to the estimators. Since the unknown parameters are evaluated in previous experimentation, one might argue correctly that uni-stage designs are in fact the second stage of a less formal multistage experiment. Thus, following the design literature (e.g. Liu and Chi [25], Bellhouse *et al.* [26], Sitter and Wu [27], and Racine *et al.* [28]) the problem of approximating the optimal proportions of pooled and unpooled samples can be considered by using Bayesian methods, learning samples with a fixed size, or two-stage design techniques, etc. Two-stage designs can, in particular, be based on sequential procedures that make relevant test decisions using the first stage, utilizing a non pre-specified number of observations. Note also that the relationship between the bias of estimation of optimal values of α and the corresponding bias of main parameter estimation can be a major factor in defining the appropriate size of the learning sample. Our limited simulation results show that reasonable estimates of the optimal values of α can be obtained from relatively small to moderate-sized learning samples. In fact, learning samples with sizes proportional to the square root of the number of measured assays (i.e. $\propto n^{1/2}$), (Liu and Chi [25], Bellhouse *et al.* [26], Sitter and Wu [27], and Racine *et al.* [28]) provide estimators of μ_x and σ_x^2 with variances comparable to those of the μ_x and σ_x^2 estimators considered in previous sections.

Consider, for example, the situations depicted in Table 1 of Section 3.1. To approximate the optimal evaluation of σ_x^2 , the values of α from Proposition 2.2.1, having form

$$\alpha = \frac{N^2\sigma_x^4 + N(n+N)\sigma_x^2\sigma_M^2 + nN\sigma_M^4}{N^2\sigma_x^4 + 2N^2\sigma_x^2\sigma_M^2 + (2nN - n^2)\sigma_M^4},$$

are estimated based on learning samples with different sizes and the 50/50 proportion of numbers of pooled and unpooled measurements. (The learning samples, in these cases, should also have pooled and unpooled data to estimate σ_M^2 .) Then, the hybrid samples are formed and the parameters estimated. Table 3 presents the corresponding Monte Carlo results.

Thus, comparing the Monte Carlo estimators of $\text{var}(\hat{\sigma}_x^2)$ in Table 1 and Table 3, we observe that the proposed estimators are quite robust to the choice of the learning sample size that is proportional to $(n^{1/2})$.

In this article, we demonstrate how to apply the hybrid design to estimate the *AUC*, but the methodology could also be applied to other functions such as hypothesis testing, regression, etc. In the present paper, we analyze normally distributed biomarkers, and the efficiency of the pooling design depends in large part on the distribution of interest. However, other common biomarker distributions could be evaluated in a similar manner. For nonparametric cases, the proposed estimators can be considered as the least square estimators that asymptotically have the variances presented in this paper. We consider the situation where only measurement error or only pooling error is present and the number of available biospecimens (N) is fixed. Hence, the pooling size is defined by the recommended optimal proportion (α) and the number of samples we can afford to test (n). In a similar manner, we could analyze the situation where the pooling size is fixed. However, when both measurement and pooling errors are present, p_1 and p_2 are considered fixed beforehand to allow for the optimization of α_1 . The proposed estimators could also be modified to allow for differential errors that are a function of the pooling group size or levels of the biomarker, or errors due to mixing in non-identical proportions.

In conclusion, the hybrid sampling design combines the benefits of both pooling and random sampling under strict cost limitations, while making possible the estimation of measurement error without repeated samples. Only a small to moderate-sized training sample is needed in order to successfully implement the hybrid design.

Acknowledgments

The authors thank the editor and reviewers for their valuable comments. This work was supported by the Intramural Research Program of the National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development.

Appendix

A1. Hybrid Design with Measurement Error

The log likelihood function based on S' has the form of:

$$\begin{aligned} \ell(\mu_x, \sigma_x, \sigma_M) = & -n \ln(\sqrt{2\pi}) - \alpha n \ln\left(\sqrt{\sigma_x^2 + \sigma_M^2}\right) - \frac{\sum_{i \leq \alpha n} (X'_i - \mu_x)^2}{2(\sigma_x^2 + \sigma_M^2)} \\ & - (1 - \alpha)n \ln\left(\sqrt{\frac{\sigma_x^2}{p} + \sigma_M^2}\right) - \frac{\sum_{i \leq (1-\alpha)n} (Z'_i - \mu_x)^2}{2\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)}. \end{aligned}$$

The derivatives of the log likelihood function are

$$\begin{aligned} \frac{\partial \ell}{\partial \mu_x} &= \frac{\sum_{i \leq \alpha n} (X'_i - \mu_x)}{\sigma_x^2 + \sigma_M^2} + \frac{\sum_{i \leq (1-\alpha)n} (Z'_i - \mu_x)}{\frac{\sigma_x^2}{p} + \sigma_M^2}, \\ \frac{\partial \ell}{\partial \sigma_x} &= \frac{-\alpha n \sigma_x}{\sigma_x^2 + \sigma_M^2} + \frac{\sum_{i \leq \alpha n} (X'_i - \mu_x) \sigma_x}{(\sigma_x^2 + \sigma_M^2)^2} - \frac{(1-\alpha)n \sigma_x}{p\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)} + \frac{\sum_{i \leq (1-\alpha)n} (Z'_i - \mu_x)^2 \sigma_x}{p\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2}, \\ \frac{\partial \ell}{\partial \sigma_M} &= \frac{-\alpha n \sigma_M}{\sigma_x^2 + \sigma_M^2} + \frac{\sum_{i \leq \alpha n} (X'_i - \mu_x)^2 \sigma_M}{(\sigma_x^2 + \sigma_M^2)^2} - \frac{(1-\alpha)n \sigma_M}{\frac{\sigma_x^2}{p} + \sigma_M^2} + \frac{\sum_{i \leq (1-\alpha)n} (Z'_i - \mu_x)^2 \sigma_M}{\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2}. \end{aligned}$$

Setting the first derivatives equal to zero and solving yields the following system of equations

$$\begin{aligned} \widehat{\mu}_x &= \frac{\alpha n \bar{X}' \left(\frac{\widehat{\sigma}_x^2}{p} + \widehat{\sigma}_M^2 \right) + (1-\alpha)n \bar{Z}' (\widehat{\sigma}_x^2 + \widehat{\sigma}_M^2)}{\alpha n \left(\frac{\widehat{\sigma}_x^2}{p} + \widehat{\sigma}_M^2 \right) + (1-\alpha)n (\widehat{\sigma}_x^2 + \widehat{\sigma}_M^2)}, \quad \widehat{\sigma}_x^2 = \frac{p}{p-1} \left(\frac{\sum_{i \leq \alpha n} (X'_i - \widehat{\mu}_x)^2}{\alpha n} - \frac{\sum_{i \leq (1-\alpha)n} (Z'_i - \widehat{\mu}_x)^2}{(1-\alpha)n} \right), \\ \widehat{\sigma}_M^2 &= \frac{\sum_{i \leq (1-\alpha)n} (Z'_i - \widehat{\mu}_x)^2}{(1-\alpha)n} - \frac{\widehat{\sigma}_x^2}{p}. \end{aligned}$$

The second derivatives of the log likelihood function are

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \mu_x^2} &= \frac{-\alpha n}{(\sigma_x^2 + \sigma_M^2)} - \frac{(1-\alpha)n}{\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)}, \\
\frac{\partial^2 \ell}{\partial \sigma_x^2} &= \frac{-\alpha n}{(\sigma_x^2 + \sigma_M^2)} + \frac{2\alpha n \sigma_x^2}{(\sigma_x^2 + \sigma_M^2)^2} + \frac{i \leq \alpha n}{(\sigma_x^2 + \sigma_M^2)^2} - \frac{4\sigma_x^2 \sum (X'_i - \mu_x)^2}{i \leq \alpha n} - \frac{4\sigma_x^2 \sum (X'_i - \mu_x)^2}{(\sigma_x^2 + \sigma_M^2)^3} \\
&\quad - \frac{(1-\alpha)n}{p\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)} + \frac{2(1-\alpha)n\sigma_x^2}{p^2\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2} + \frac{i \leq (1-\alpha)n}{p\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2} - \frac{i \leq (1-\alpha)n}{p^2\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^3} \\
&\quad - \frac{\sum (Z'_i - \mu_x)^2}{i \leq (1-\alpha)n} - \frac{4\sigma_x^2 \sum (Z'_i - \mu_x)^2}{i \leq (1-\alpha)n} - \frac{4\sigma_x^2 \sum (Z'_i - \mu_x)^2}{\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^3}, \\
\frac{\partial^2 \ell}{\partial \sigma_M^2} &= \frac{-\alpha n}{(\sigma_x^2 + \sigma_M^2)} + \frac{2\alpha n \sigma_M^2}{(\sigma_x^2 + \sigma_M^2)^2} + \frac{i \leq \alpha n}{(\sigma_x^2 + \sigma_M^2)^2} - \frac{i \leq \alpha n}{(\sigma_x^2 + \sigma_M^2)^3} - \frac{(1-\alpha)n}{\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)} + \frac{2(1-\alpha)n\sigma_M^2}{\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2} \\
&\quad + \frac{\sum (Z'_i - \mu_x)^2}{i \leq (1-\alpha)n} - \frac{4\sigma_M^2 \sum (Z'_i - \mu_x)^2}{i \leq (1-\alpha)n} - \frac{4\sigma_M^2 \sum (Z'_i - \mu_x)^2}{\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^3}, \\
\frac{\partial^2 \ell}{\partial \mu_x \partial \sigma_x} &= \frac{-2\sigma_x \sum (X'_i - \mu_x)}{i \leq \alpha n} - \frac{2\sigma_x \sum (Z'_i - \mu_x)}{i \leq (1-\alpha)n} - \frac{2\sigma_x \sum (Z'_i - \mu_x)}{p\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2}, \\
\frac{\partial^2 \ell}{\partial \mu_x \partial \sigma_M} &= \frac{-2\sigma_M \sum (X'_i - \mu_x)}{i \leq \alpha n} - \frac{2\sigma_M \sum (Z'_i - \mu_x)}{i \leq (1-\alpha)n} - \frac{2\sigma_M \sum (Z'_i - \mu_x)}{\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2}, \\
\frac{\partial^2 \ell}{\partial \sigma_x \partial \sigma_M} &= \frac{4\sigma_x \sigma_M \sum (X'_i - \mu_x)^2}{i \leq \alpha n} - \frac{4\sigma_x \sigma_M \sum (X'_i - \mu_x)^2}{(\sigma_x^2 + \sigma_M^2)^3} + \frac{2(1-\alpha)n\sigma_x \sigma_M}{p\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2} - \frac{4\sigma_x \sigma_M \sum (Z'_i - \mu_x)^2}{i \leq (1-\alpha)n} - \frac{4\sigma_x \sigma_M \sum (Z'_i - \mu_x)^2}{p\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^3}
\end{aligned}$$

Since these estimates are MLE's, we have where $\sqrt{n} \begin{pmatrix} \widehat{\mu}_x - \mu_x \\ \widehat{\sigma}_x^2 - \sigma_x^2 \\ \widehat{\sigma}_M^2 - \sigma_M^2 \end{pmatrix} \sim N(0, I^{-1})$, where

$$I = \begin{bmatrix} \frac{\alpha}{(\sigma_x^2 + \sigma_M^2)} + \frac{(1-\alpha)}{\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)} & 0 & 0 \\ 0 & \frac{\alpha}{2(\sigma_x^2 + \sigma_M^2)^2} + \frac{(1-\alpha)}{2p^2\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2} & \frac{\alpha}{2(\sigma_x^2 + \sigma_M^2)^3} + \frac{(1-\alpha)}{2p\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2} \\ 0 & \frac{\alpha}{2(\sigma_x^2 + \sigma_M^2)^2} + \frac{(1-\alpha)}{2p\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2} & \frac{\alpha}{2(\sigma_x^2 + \sigma_M^2)^2} + \frac{(1-\alpha)}{2\left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2} \end{bmatrix}.$$

Hence,

$$\begin{aligned}
\text{Var}(\widehat{\mu}_x) &= \frac{(\sigma_x^2 + \sigma_M^2) \left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)}{\alpha n \left(\frac{\sigma_x^2}{p} + \sigma_M^2\right) + (1-\alpha)n(\sigma_x^2 + \sigma_M^2)}, \\
\text{Var}(\widehat{\sigma}_x^2) &= \frac{2p^2}{(p-1)^2} \frac{\alpha \left(\frac{\sigma_x^2}{p} + \sigma_M^2\right)^2 + (1-\alpha)(\sigma_x^2 + \sigma_M^2)^2}{\alpha(1-\alpha)n}.
\end{aligned}$$

A2. The proof of Proposition 2.2.1

Since $P = \frac{N - \alpha n}{(1 - \alpha)n}$, we can write

$$\text{Var}(\hat{\mu}_x) = \frac{(\sigma_x^2 + \sigma_M^2)(\sigma_x^2(1 - \alpha)n + (N - \alpha n)\sigma_M^2)}{\alpha(1 - \alpha)n\sigma_x^2 + \alpha(N - \alpha n)\sigma_M^2 + (1 - \alpha)(N - \alpha n)(\sigma_x^2 + \sigma_M^2)},$$

and hence

$$\begin{aligned} \frac{\partial}{\partial \alpha} \text{Var}(\hat{\mu}_x) &= \frac{\left\{ \sigma_x^2(1 - \alpha)n + \sigma_M^2(N - \alpha n) \right\} \left\{ -n(\sigma_x^2 + \sigma_M^2) \right\} - \left\{ \sigma_x^2(1 - \alpha)n + \sigma_M^2(N - \alpha n) \right\} \left\{ -N\sigma_x^2 - n\sigma_M^2 \right\}}{\left[\sigma_x^2(1 - \alpha)n + \sigma_M^2(N - \alpha n) \right]^2} \\ &= \frac{\sigma_x^2 \sigma_M^2 (N - n)^2}{\left[\sigma_x^2(1 - \alpha)n + \sigma_M^2(N - \alpha n) \right]^2} > 0. \end{aligned}$$

Thus, $\text{Var}(\hat{\mu}_x)$ is a strictly increasing function. Therefore, the optimal alpha level is then found where $\alpha = 0$, which corresponds to taking a pooled sample.

While $P = \frac{N - \alpha n}{(1 - \alpha)n}$, we have

$$\begin{aligned} \text{Var}(\hat{\sigma}_x^2) &= \frac{2(N - \alpha n)^2}{(N - n)^2} \frac{\alpha \left(\frac{\sigma_x^2(1 - \alpha)n}{(N - \alpha n)} + \sigma_M^2 \right)^2 + (1 - \alpha)(\sigma_x^2 + \sigma_M^2)^2}{\alpha(1 - \alpha)n} \\ &= \frac{2\alpha \left\{ \sigma_x^2(1 - \alpha)n + \sigma_M^2(N - \alpha n) \right\}^2 + 2(1 - \alpha)(N - \alpha n)^2(\sigma_x^2 + \sigma_M^2)^2}{(N - n)^2 \alpha(1 - \alpha)n} \\ &= \frac{\sigma_x^4 \{ 2(1 - \alpha)(N^2 - 2\alpha nN + \alpha n^2) \} + 2\sigma_x^2 \sigma_M^2 \{ 2(1 - \alpha)(N - \alpha n)N \} + \sigma_M^4 \{ 2(N - \alpha n)^2 \}}{(N - n)^2 \alpha(1 - \alpha)n}, \end{aligned}$$

and hence

$$\frac{\partial}{\partial \alpha} \text{Var}(\hat{\sigma}_x^2) = \frac{-\sigma_x^4 N^2 (1 - \alpha)^2 - 2\sigma_x^2 \sigma_M^2 N^2 (1 - \alpha)^2 + \sigma_M^4 (N - \alpha n)(2\alpha N - N - \alpha n)}{(N - n)^2 \alpha^2 (1 - \alpha)^2 n}.$$

In this case, $\frac{\partial}{\partial \alpha} \text{Var}(\hat{\sigma}_x^2) = 0 \iff$

$$\begin{aligned} A\alpha^2 + B\alpha + C &= 0, \quad A = -N^2\sigma_x^4 - 2N^2\sigma_x^2\sigma_M^2 - (2nN - n^2)\sigma_M^4, \quad B = 2N^2(\sigma_x^2 + \sigma_M^2)^2, \\ C &= -N^2(\sigma_x^2 + \sigma_M^2)^2. \end{aligned}$$

There are two possible solutions:

$$\alpha_{0,1} = \frac{N(\sigma_x^2 + \sigma_M^2)^2(-N\sigma_x^2 - n\sigma_M^2)}{-N^2\sigma_x^4 - 2N^2\sigma_x^2\sigma_M^2 - (2nN - n^2)\sigma_M^4}, \quad \alpha_{0,2} = \frac{N(\sigma_x^2 + \sigma_M^2)^2(-N\sigma_x^2 - (2N - n)\sigma_M^2)}{-N^2\sigma_x^4 - 2N^2\sigma_x^2\sigma_M^2 - (2nN - n^2)\sigma_M^4}.$$

We can show that the first solution $\alpha_{0,1}$ is strictly less than 1. Write

$$\begin{aligned}\alpha_{0,1} &= \frac{N(\sigma_x^2 + \sigma_M^2)^2(-N\sigma_x^2 - n\sigma_M^2)}{-N^2\sigma_x^4 - 2N^2\sigma_x^2\sigma_M^2 - (2nN - n^2)\sigma_M^4} \\ &= \frac{-N^2\sigma_x^4 - N(n+N)\sigma_x^2\sigma_M^2 - nN\sigma_M^4}{-N^2\sigma_x^4 - 2N^2\sigma_x^2\sigma_M^2 - (2nN - n^2)\sigma_M^4} \\ &= \frac{N^2\sigma_x^4 + N(n+N)\sigma_x^2\sigma_M^2 + nN\sigma_M^4}{N^2\sigma_x^4 + 2N^2\sigma_x^2\sigma_M^2 + (2nN - n^2)\sigma_M^4}.\end{aligned}$$

Now, if we consider the coefficients for each term of $\alpha_{0,1}$, we can show that the numerator is less than the denominator and therefore the solution $\alpha_{0,1}$ is less than 1, i.e. since the coefficients for σ_x^4 in the numerator and denominator are equivalent, the coefficients for $\sigma_x^2\sigma_M^2$ are $N(n+N) < 2N^2$, the coefficients for σ_M^4 are $nN < 2nN - n^2$, and hence $\alpha_{0,1} \leq 1$.

Consider

$$\begin{aligned}\alpha_{0,2} &= \frac{N(\sigma_x^2 + \sigma_M^2)^2(-N\sigma_x^2 - (2N - n)\sigma_M^2)}{-N^2\sigma_x^4 - 2N^2\sigma_x^2\sigma_M^2 - (2nN - n^2)\sigma_M^4} \\ &= \frac{-N^2\sigma_x^4 - (3N^2 - nN)\sigma_x^2\sigma_M^2 - (2N^2 - nN)\sigma_M^4}{-N^2\sigma_x^4 - 2N^2\sigma_x^2\sigma_M^2 - (2nN - n^2)\sigma_M^4} \\ &= \frac{N^2\sigma_x^4 + (3N^2 - nN)\sigma_x^2\sigma_M^2 + (2N^2 - nN)\sigma_M^4}{N^2\sigma_x^4 + 2N^2\sigma_x^2\sigma_M^2 + (2nN - n^2)\sigma_M^4}.\end{aligned}$$

Here the coefficients for σ_x^4 in the numerator and denominator are equivalent, the coefficients for $\sigma_x^2\sigma_M^2$ are $3N^2 - nN > 2N^2$, the coefficients for σ_M^4 are $N(2N - n) > n(2N - n)$, and hence $\alpha_{0,2} > 1$.

Therefore, the optimal alpha is $\alpha_{0,1}$.

The proof of Proposition 2.2.1 is complete.

A3. Hybrid Design with Pooling Error

The log likelihood function based on sample S is

$$\ell(\mu_x, \sigma_x, \sigma_p) = -\alpha n \ln \sigma_x - \frac{\sum_{i \leq \alpha n} (X_i - \mu_x)^2}{2\sigma_x^2} - (1 - \alpha)n \ln \left(\sqrt{\left(\frac{\sigma_x^2}{p} + \sigma_p^2 \right)} \right) \sigma_z - \frac{\sum_{i \leq (1 - \alpha)n} (Z_i - \mu_x)^2}{2 \left(\frac{\sigma_x^2}{p} + \sigma_p^2 \right)}$$

To derive the maximum likelihood estimators of μ_x , σ_x , and σ_p , we calculate:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu_x} &= \frac{\sum_{i \leq \alpha n} (X_i - \mu_x)}{\sigma_x^2} + \frac{\sum_{i \leq (1-\alpha)n} (Z_i - \mu_x)}{\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)}, \\ \frac{\partial \ell}{\partial \sigma_x} &= \frac{-\alpha n}{\sigma_x} + \frac{\sum_{i \leq \alpha n} (X_i - \mu_x)^2}{\sigma_x^3} - \frac{(1-\alpha)n\sigma_x}{p\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)} + \frac{\sum_{i \leq (1-\alpha)n} (Z_i - \mu_x)^2 \sigma_x}{p\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^2}, \\ \text{and } \frac{\partial \ell}{\partial \sigma_p} &= \frac{-\sigma_p(1-\alpha)n}{\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)} + \frac{\sigma_p \sum_{i \leq (1-\alpha)n} (Z_i - \mu_x)^2}{\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^2}.\end{aligned}$$

Setting these first derivatives equal to zero and solving yields the following system of equations

$$\begin{aligned}\widehat{\mu}_x &= \frac{\alpha n \bar{X} \left(\frac{\widehat{\sigma}_x^2}{p} + \widehat{\sigma}_p^2\right) + (1-\alpha)n \bar{Z} \widehat{\sigma}_x^2}{\alpha n \left(\frac{\widehat{\sigma}_x^2}{p} + \widehat{\sigma}_p^2\right) + (1-\alpha)n \widehat{\sigma}_x^2}, \\ \widehat{\sigma}_x^2 &= \frac{1}{\alpha n} \sum_{i \leq \alpha n} (X_i - \widehat{\mu}_x)^2, \\ \text{and } \widehat{\sigma}_p^2 &= \frac{1}{(1-\alpha)n} \sum_{i \leq (1-\alpha)n} (Z_i - \widehat{\mu}_x)^2 - \frac{\widehat{\sigma}_x^2}{p},\end{aligned}$$

where $\bar{X}$ denotes the mean of $X_1, \dots, X_{[\alpha n]}$ and $\bar{Z}$ denotes the mean of $Z_1, \dots, Z_{[(1-\alpha)n]}$. Consider the second derivatives

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \mu_x^2} &= \frac{-\alpha n}{\sigma_x^2} - \frac{(1-\alpha)n}{\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)}, \\ \frac{\partial^2 \ell}{\partial \sigma_x^2} &= \frac{\alpha n}{\sigma_x^2} - \frac{\sum_{i \leq \alpha n} (X_i - \mu_x)^2}{\sigma_x^4} - \frac{(1-\alpha)n}{p\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)} \\ &\quad + \frac{2(1-\alpha)n\sigma_x^2}{p^2\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^2} + \frac{\sum_{i \leq (1-\alpha)n} (Z_i - \mu_x)^2}{p\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^2} - \frac{4\sigma_x^2 \sum_{i \leq (1-\alpha)n} (Z_i - \mu_x)^2}{p^2\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^3}, \\ \frac{\partial^2 \ell}{\partial \sigma_p^2} &= \frac{-(1-\alpha)n}{\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)} + \frac{2(1-\alpha)n\sigma_p^2}{\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^2} + \frac{\sum_{i \leq (1-\alpha)n} (Z_i - \mu_x)^2}{\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^2} - \frac{4\sigma_p^2 \sum_{i \leq (1-\alpha)n} (Z_i - \mu_x)^2}{\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^3}, \\ \frac{\partial^2 \ell}{\partial \mu_x \partial \sigma_x} &= \frac{-2 \sum_{i \leq \alpha n} (X_i - \mu_x)}{\sigma_x^3} - \frac{2\sigma_x \sum_{i \leq (1-\alpha)n} (Z_i - \mu_x)}{p\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^2}, \\ \frac{\partial^2 \ell}{\partial \mu_x \partial \sigma_p} &= \frac{-2\sigma_p \sum_{i \leq (1-\alpha)n} (Z_i - \mu_x)}{\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^2} \text{ and } \frac{\partial^2 \ell}{\partial \sigma_x \partial \sigma_p} = \frac{2(1-\alpha)n\sigma_x\sigma_p}{p\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^2} - \frac{4\sigma_x\sigma_p \sum_{i \leq (1-\alpha)n} (Z_i - \mu_x)^2}{p\left(\frac{\sigma_x^2}{p} + \sigma_p^2\right)^3}.\end{aligned}$$

Since these estimates are MLEs, we use the classical asymptotic properties which lead to the result $\sqrt{n}[\hat{\mu}_x - \mu_x \quad \hat{\sigma}_x^2 - \sigma_x^2 \quad \hat{\sigma}_p^2 - \sigma_p^2]^T \xrightarrow{n \rightarrow \infty} N(0, I^{-1})$, where the Fisher Information Matrix, I , is the negative expected value of the 2nd derivatives divided by n as $n \rightarrow \infty$, i.e.

$$I = - \lim_{n \rightarrow \infty} \frac{1}{n} E \begin{bmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & 0 & 0 \\ 0 & \frac{\partial^2 \ell}{(\partial \sigma_x^2)^2} = \frac{\partial^2 \ell}{4\sigma_x^2(\partial \sigma_x)^2} & \frac{\partial^2 \ell}{\partial \sigma_x^2 \partial \sigma_p^2} = \frac{\partial^2 \ell}{4\sigma_x \sigma_p \partial \sigma_x \partial \sigma_p} \\ 0 & \frac{\partial^2 \ell}{\partial \sigma_x^2 \partial \sigma_p^2} & \frac{\partial^2 \ell}{(\partial \sigma_p^2)^2} = \frac{\partial^2 \ell}{4\sigma_p^2 \partial(\sigma_p^2)} \end{bmatrix}.$$

It is clear that asymptotically these estimators are distributed as

$$\sqrt{n} \begin{pmatrix} \hat{\mu}_x - \mu_x \\ \hat{\sigma}_x^2 - \sigma_x^2 \\ \hat{\sigma}_p^2 - \sigma_p^2 \end{pmatrix} \sim N_3(0, \Sigma),$$

where Σ is the inverse of the Fisher Information Matrix.

A4. The Proof of Proposition 2.3.1

The optimal proportion of pooled and unpooled samples can be chosen in accord with a given loss function. We consider this loss function to be the variance of the mean estimator, the variance of the variance estimator, or a linear combination of both. When we are only interested in estimation of the mean, the proportion of pooled samples that minimizes the variance of the mean estimator depends on the size of the pooling error in relation to the variability of the sample.

Proposition 2.3.1

In the context of the variance $\text{var}(\hat{\mu}_x)$ of the mean estimator the optimal proportion α of unpooled samples satisfies the rule:

$$\text{Optimal } \alpha = \begin{cases} 0 & \text{if } \frac{\sigma_p^2}{\sigma_x^2} \leq \left(\frac{N-2n}{N}\right), \\ 1 & \text{if } \frac{\sigma_p^2}{\sigma_x^2} \geq 1, \text{ else} \\ \frac{\sigma_p^2 N + \sigma_x^2 (2n - N)}{n(\sigma_x^2 + \sigma_p^2)} & \end{cases};$$

whereas, in the context of the variance $\text{var}(\hat{\sigma}_x^2)$ of the variance estimator the optimal proportion is $\alpha = 1$.

As the pooling error gets larger in relation to the variability in the sample, it is recommended to perform measurements on a larger portion than random sampling. The existence of pooling error limits the area where the pooling design is efficient. In accordance with the distribution function of the maximum likelihood estimators, one can show that if we are only concerned with estimation of the variance, then the random sampling strategy is always recommended. The variance of the $\hat{\sigma}_x^2$ estimator is minimized when only a random sample is taken ($\alpha = 1$).

Proof—We want to find the minimum of the variance of the mean estimator with respect to α . In accordance with the asymptotic distribution of $\hat{\mu}$ (see, Appendix A1), we have

$$n \cdot \text{Var}(\hat{\mu}) = \frac{n}{\frac{\alpha}{\sigma_x^2} + \frac{(1-\alpha)}{\sigma_x^2/p + \sigma_p^2}}$$

Since $P = \frac{N - \alpha n}{(1 - \alpha)n}$, we can write the variance as a function of α

$$n \cdot \text{Var}(\hat{\mu}) = \frac{\frac{\sigma_x^4(1-\alpha)n}{N-\alpha n} + \sigma_x^2\sigma_p^2}{\frac{\sigma_x^2\alpha(1-\alpha)n}{N-\alpha n} + \alpha\sigma_p^2 + (1-\alpha)\sigma_x^2},$$

and hence

$$\frac{\partial}{\partial \alpha} \text{Var}(\hat{\mu}) = \frac{\sigma_x^2\sigma_p^2(N - \alpha n) [\sigma_x^2(N - 2n + \alpha n) - \sigma_p^2(N - \alpha n)]}{[\sigma_x^2(1 - \alpha)N + \sigma_p^2\alpha(N - \alpha n)]^2}.$$

Since the denominator of $\frac{\partial}{\partial \alpha} \text{Var}(\hat{\mu})$ is positive, we learn that the behavior of $\text{Var}(\hat{\mu})$ depends on the part of the numerator, say $Y = \sigma_x^2(N - 2n + \alpha n) - \sigma_p^2(N - \alpha n)$. Specifically, if Y is positive then the derivative is also positive, and hence $\text{Var}(\hat{\mu})$ increases by α . Denote $Y = \alpha A + B$, $\alpha \in [0, 1]$, $A = n(\sigma_x^2 + \sigma_p^2)$, $B = \sigma_x^2(N - 2n) - \sigma_p^2N$.

$A > 0$, and hence Y increases. Since the line $Y = \alpha A + B$, $\alpha \in [0, 1]$ can be located only 1) over axis $\alpha \in [0, 1]$, 2) below the axis $\alpha \in [0, 1]$, or 3) across $\alpha \in [0, 1]$, we have 3 different cases:

1. If $B \geq 0$, i.e. $\sigma_p^2 \leq \sigma_x^2 \frac{N - 2n}{N}$, then the derivative $\frac{\partial}{\partial \alpha} \text{Var}(\hat{\mu})$ is strictly positive, $\text{Var}(\hat{\mu})$ increases as α increases so the optimal $\alpha = \arg \min \text{var}(\hat{\mu})$ is found at $\alpha = 0$;
2. If $A + B \leq 0$, i.e. $\sigma_p^2 \geq \sigma_x^2$. Then $\frac{\partial}{\partial \alpha} \text{Var}(\hat{\mu})$ is strictly negative and the optimal alpha is found at $\alpha = 1$, because $\text{Var}(\hat{\mu})$ decreases as α increases from 0 to 1;
3. If $\alpha A + B = 0$, then $\alpha = \frac{-B}{A} = \frac{\sigma_p^2 N + \sigma_x^2(2n - N)}{n(\sigma_x^2 + \sigma_p^2)}$.

This completes the proof of Proposition 2.2.1.

A5. Measurement Error and Pooling Error

In this case, the log likelihood function based on S'' has the form of

$$\begin{aligned} \ell(\mu_x, \sigma_x, \sigma_p, \sigma_M) = & -\alpha_1 n \ln(\sigma_x^2 + \sigma_M^2)^{\frac{1}{2}} - \frac{\sum_{i \leq \alpha_1 n} (X'_i - \mu_x)^2}{2(\sigma_x^2 + \sigma_M^2)} - \alpha_2 n \ln\left(\frac{\sigma_x^2}{p_1} + \sigma_p^2 + \sigma_M^2\right)^{\frac{1}{2}} \\ & - \frac{\sum_{i \leq \alpha_2 n} (Z'_i - \mu_x)^2}{2\left(\frac{\sigma_x^2}{p_1} + \sigma_p^2 + \sigma_M^2\right)} - (1 - \alpha_1 - \alpha_2)n \ln\left(\frac{\sigma_x^2}{p_2} + \sigma_p^2 + \sigma_M^2\right)^{\frac{1}{2}} \\ & - \frac{\sum_{i \leq (1-\alpha_1-\alpha_2)n} (Z''_i - \mu_x)^2}{2\left(\frac{\sigma_x^2}{p_1} + \sigma_p^2 + \sigma_M^2\right)}. \end{aligned}$$

Setting the first derivatives $\frac{\partial \ell}{\partial \mu_x}$, $\frac{\partial \ell}{\partial \sigma_x^2}$, $\frac{\partial \ell}{\partial \sigma_p^2}$ and $\frac{\partial \ell}{\partial \sigma_M^2}$ equal to zero and solving yields the system of equations regarding the maximum likelihood estimators.

The maximum likelihood approach provides that

$$\sqrt{n} \begin{bmatrix} \hat{\mu}_x - \mu_x & \hat{\sigma}_x^2 - \sigma_x^2 & \hat{\sigma}_p^2 - \sigma_p^2 & \hat{\sigma}_M^2 - \sigma_M^2 \end{bmatrix}^T \underset{\alpha_1 n, \alpha_2 n, (1-\alpha_1-\alpha_2)n \rightarrow \infty}{\sim} N(0, I^{-1}),$$

where

$$I = - \lim_{\alpha_1 n, \alpha_2 n, (1-\alpha_1-\alpha_2)n \rightarrow \infty} \frac{1}{n} E \begin{bmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & 0 & 0 & 0 \\ 0 & \frac{\partial^2 \ell}{(\partial \sigma_x^2)^2} & \frac{\partial^2 \ell}{\partial \sigma_x^2 \partial \sigma_p^2} & \frac{\partial^2 \ell}{\partial \sigma_x^2 \partial \sigma_M^2} \\ 0 & \frac{\partial^2 \ell}{\partial \sigma_x^2 \partial \sigma_p^2} & \frac{\partial^2 \ell}{(\partial \sigma_p^2)^2} & \frac{\partial^2 \ell}{\partial \sigma_p^2 \partial \sigma_M^2} \\ 0 & \frac{\partial^2 \ell}{\partial \sigma_x^2 \partial \sigma_M^2} & \frac{\partial^2 \ell}{\partial \sigma_p^2 \partial \sigma_M^2} & \frac{\partial^2 \ell}{(\partial \sigma_M^2)^2} \end{bmatrix}.$$

In order to avoid technical formal notations, we do not mention the cumbersome forms of the second derivatives that can be simply calculated.

References List

1. Dorfman R. The detection of defective members of large populations. *Annals of Mathematical Statistics*. 1943; 14:436–440.
2. Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. *Statistics in Medicine*. 2003; 22:2515–2527. [PubMed: 12872306]
3. Mumford SL, Schisterman EF, Vexler A, Liu A. Pooling biospecimens and limits of detection: effects on ROC curve analysis. *Biostatistics*. 2006; 7:585–598. [PubMed: 16531470]
4. Schisterman EF, Vexler A. To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection. *Paediatric and Perinatal Epidemiology*. 2008; 22:486–496. [PubMed: 18782255]
5. Vexler A, Liu A, Schisterman EF. Efficient design and analysis of biospecimens with measurements subject to detection limit. *Biometrical Journal*. 2006; 48:780–791. [PubMed: 17094343]
6. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics*. 1999; 55:718–726. [PubMed: 11314998]
7. Zhang SD, Gant TW. Effect of pooling samples on the efficiency of comparative studies using microarrays. *Bioinformatics*. 2005; 21:4378–4383. [PubMed: 16234321]

8. Vexler A, Schisterman EF, Liu A. Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures. *Statistics in Medicine*. 2008; 27:280–296. [PubMed: 17721905]
9. Liu A, Schisterman EF. Comparison of Diagnostic Accuracy of Biomarkers With Pooled Assessments. *Biometrical Journal*. 2003; 45:631–644.
10. Fuller, W. *Measurement Error Models*. New York: John Wiley and Sons; 1987.
11. Coffin M, Sukhatme S. Receiver operating characteristic studies and measurement errors. *Biometrics*. 1997; 53:823–837. [PubMed: 9333348]
12. Schisterman EF, Faraggi D, Reiser B, Trevisan M. Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error. *American Journal of Epidemiology*. 2001; 154:174–179. [PubMed: 11447052]
13. Perkins NJ, Schisterman EF. The Youden Index and the optimal cut-point corrected for measurement error. *Biometrical Journal*. 2005; 47:428–441. [PubMed: 16161802]
14. Dunn, G. *Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors*. New York: Oxford University Press; 1989.
15. Freedman LS, Fainberg V, Kipnis V, Midthune D, Carroll RJ. A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics*. 2004; 60:172–181. [PubMed: 15032787]
16. Thurigen D, Spiegelman D, Blettner M, Heuer C, Brenner H. Measurement error correction using validation data: a review of methods and their applicability in case-control studies. *Statistical Methods in Medical Research*. 2000; 9:447–474. [PubMed: 11191260]
17. Faraggi D. The effect of random measurement error on receiver operating characteristic (ROC) curves. *Statistics in Medicine*. 2000; 19:61–70. [PubMed: 10623913]
18. Liu K, Stamler J, Dyer A, McKeever J, McKeever P. Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol. *Journal of Chronic Diseases*. 1978; 31:399–418. [PubMed: 711832]
19. Reiser B. Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. *Statistics in Medicine*. 2000; 19:2115–2129. [PubMed: 10931515]
20. Dunson DB, Weinberg CR. Modeling human fertility in the presence of measurement error. *Biometrics*. 2000; 56:288–292. [PubMed: 10783808]
21. Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Statistics in Medicine*. 2002; 21:3093–3106. [PubMed: 12369084]
22. Shapiro DE. The interpretation of diagnostic tests. *Statistical Methods in Medical Research*. 1999; 8:113–134. [PubMed: 10501649]
23. Vonesh EF, Carter RL. Efficient inference for random-coefficient growth curve models with unbalanced data. *Biometrics*. 1987; 43:617–628. [PubMed: 3663819]
24. Wieand S, Gail MH, James BR, James KL. A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*. 1989; 76:585–592.
25. Liu Q, Chi GY. On sample size and inference for two-stage adaptive designs. *Biometrics*. 2001; 57:172–177. [PubMed: 11252594]
26. Bellhouse DR, Thompson ME, Godambe VP. Two-stage sampling with exchangeable prior distributions. *Biometrika*. 1977; 64:97–103.
27. Sitter RR, Wu CF. Two-stage design of quantal response studies. *Biometrics*. 1999; 55:396–402. [PubMed: 11318192]
28. Racine A, Grieve AP, Fluhler H, Smith AFM. *Bayesian Methods in Practice: Experiences in the Pharmaceutical Industry*. *Applied Statistics*. 1986; 35:93–150.

Figure 1.

a. Variance of the mean estimator for both rounded and unrounded values of p , for the case of $N=3000$, $n=1000$, $\sigma_x^2=1$, and $\sigma_m^2=0.1$.

b. Variance of the variance estimator for both rounded and unrounded values of p , for the case of $N=3000$, $n=1000$, $\sigma_x^2=1$, and $\sigma_m^2=0.1$.

Table 1

Monte Carlo simulation results for measurement error only case. N=3000, n=1000, $\sigma_x^2=1$, 2500 repetitions.

σ_M^2	α	p^*	μ		σ_x^2	
			Monte Carlo		Theoretical	
			Mean	Variance	Mean	Variance
0.1	0.1	3.2	-0.0003	0.00051	0.00044	0.057
	0.2	3.5	-0.0001	0.00047	0.00044	0.021
	0.3	3.9	0.0003	0.00049	0.00045	0.015
	0.4	4.3	0.0000	0.00051	0.00046	0.011
	0.5	5.0	0.0001	0.00050	0.00047	0.008
	0.6	6.0	0.0009	0.00052	0.00049	0.006
	0.7	7.7	0.0003	0.00057	0.00052	0.005
	0.8	11.0	0.0001	0.00058	0.00056	0.004
	0.9	21.0	-0.0001	0.00069	0.00067	0.003
	0.4	3.2	-0.0014	0.00080	0.00075	0.089
0.4	0.2	3.5	-0.0006	0.00081	0.00076	0.037
	0.3	3.9	-0.0004	0.00076	0.00078	0.025
	0.4	4.3	-0.0002	0.00087	0.00081	0.020
	0.5	5.0	0.0002	0.00083	0.00084	0.014
	0.6	6.0	0.0001	0.00087	0.00088	0.012
	0.7	7.7	-0.0001	0.00096	0.00094	0.009
	0.8	11.0	0.0000	0.00102	0.00102	0.009
	0.9	21.0	0.0009	0.00119	0.00115	0.010
	1.44	0.1	3.2	0.0005	0.00178	0.260
	0.2	3.5	0.0004	0.00194	0.00183	0.123
1.44	0.3	3.9	0.0006	0.00195	0.00187	0.085
	0.4	4.3	0.0004	0.00191	0.00191	0.071
	0.5	5.0	-0.0004	0.00191	0.00196	0.052
	0.6	6.0	-0.0001	0.00204	0.00202	0.048
	0.7	7.7	0.0006	0.00209	0.00209	0.046
	0.8	11.0	0.0000	0.00217	0.00218	0.046
	0.9	21.0	0.0009	0.00119	0.00115	0.010
	1.44	0.1	3.2	0.0005	0.00178	0.260
	0.2	3.5	0.0004	0.00194	0.00183	0.123
	0.3	3.9	0.0006	0.00195	0.00187	0.085
	0.4	4.3	0.0004	0.00191	0.00191	0.071

		μ				σ_x^2			
σ_M^2	α	p^*	Monte Carlo		Theoretical		Monte Carlo		Theoretical
			Mean	Variance	Mean	Variance	Mean	Variance	Variance
0.9	0.9	21.0	-0.0006	0.00234	0.00229	0.00229	1.00	0.065	0.063

* p used in the simulation is rounded so that the datasets can be formed.

Table 2

Monte Carlo simulation results for pooling error only case. 5000 repetitions. $\mu = 0$.

	N	n	σ_p^2	α	μ			σ_x^2			σ_p^2		
					Monte Carlo Mean	Monte Carlo Variance	Asymptotic Variance	Monte Carlo Mean	Monte Carlo Variance	Asymptotic Variance	Monte Carlo Mean	Monte Carlo Variance	Asymptotic Variance
Case 1 * optimal $\alpha=0$	300	100	0.3	0.25	0.0014	0.0087	0.0084	0.9967	0.0818	0.0800	0.2920	0.0372	0.0371
				0.50	0.0003	0.0090	0.0089	0.9908	0.0393	0.0400	0.2968	0.0361	0.0356
		200	0.3	0.75	0.0009	0.0096	0.0094	0.9911	0.0259	0.0267	0.2963	0.0590	0.0579
				0.25	-0.0012	0.0175	0.0168	0.9408	0.1573	0.1600	0.3061	0.0748	0.0741
Case 2 optimal $\alpha=1$	300	100	0.7	0.50	-0.0020	0.0177	0.0178	0.9846	0.0809	0.0800	0.2935	0.0726	0.0712
				0.75	-0.0018	0.0195	0.0188	0.9687	0.0522	0.0533	0.2657	0.1146	0.1157
	200	50	0.7	0.25	0.0003	0.0120	0.0114	0.9803	0.0785	0.0800	0.6978	0.0590	0.0584
				0.50	-0.0012	0.0112	0.0109	0.9874	0.0397	0.0400	0.6956	0.0697	0.0676
Case 3 optimal $\alpha=0.8$	300	150	0.7	0.75	-0.0007	0.0105	0.0104	0.9877	0.0262	0.0267	0.6957	0.1273	0.0122
				0.25	0.0008	0.0247	0.0229	0.9271	0.1574	0.1600	0.6970	0.1152	0.1168
	200	50	0.7	0.50	0.0022	0.0233	0.0218	0.9725	0.0840	0.0800	0.6902	0.1425	0.1352
				0.75	0.0021	0.0230	0.0209	0.9650	0.0539	0.0533	0.6430	0.2409	0.2437
	300	150	0.7	0.25	0.0016	0.0083	0.0076	0.9777	0.0530	0.0533	0.7017	0.0404	0.0389
				0.50	0.0012	0.0076	0.0073	0.9926	0.0270	0.0267	0.7022	0.0464	0.0451
	200	100	0.7	0.75	0.0009	0.0073	0.0070	0.9897	0.0176	0.0178	0.6871	0.0815	0.0812
				0.25	0.0011	0.0112	0.0114	0.9867	0.0803	0.0800	0.6903	0.0598	0.0584
	200	100	0.7	0.50	0.0007	0.0108	0.0109	0.9839	0.0402	0.0400	0.6913	0.0682	0.0676
				0.75	-0.0002	0.0103	0.0104	0.9865	0.0266	0.0267	0.6952	0.1289	0.0122

* Case status refers to Proposition 2.3.1 (Appendix A4).

Table 3

Monte Carlo simulation results for Two-Stage Design. $N=3000$, $n=1000$, $\sigma_x^2=1$, 2500 repetitions. Learning samples were based on n' measurements from N' bioassays ($\alpha=0.5$, $p=2$).

Learning Samples			α		μ		σ_x^2		σ_M^2	
σ_M^2	N'	n'	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.4	45	30	0.8015	0.1464	-0.0011	0.0324	1.0026	0.1042	0.3981	0.0686
	60	40	0.8105	0.1398	0.0008	0.0326	1.0010	0.1051	0.4007	0.0693
	90	60	0.8142	0.1315	0.0012	0.0328	0.9977	0.1035	0.3999	0.0691
	150	100	0.8240	0.1173	0.0001	0.0328	1.0000	0.1017	0.3964	0.0668
	300	200	0.8293	0.0979	-0.0009	0.0329	0.9996	0.0995	0.3993	0.0642
	450	300	0.8354	0.0860	-0.0004	0.0316	0.9997	0.0977	0.3999	0.0652
1.44	600	400	0.8374	0.0766	0.0010	0.0330	0.9993	0.0977	0.3995	0.0633
	45	30	0.7521	0.1477	0.0008	0.0470	1.0048	0.2506	1.4359	0.2055
	60	40	0.7479	0.1411	-0.0003	0.0462	1.0003	0.2455	1.4373	0.1989
	90	60	0.7504	0.1324	-0.0005	0.0471	0.9988	0.2374	1.4370	0.1921
	150	100	0.7381	0.1192	-0.0008	0.0456	0.9969	0.2390	1.4383	0.1825
	300	200	0.7236	0.0968	0.0011	0.0454	1.0058	0.2200	1.4330	0.1679
	450	300	0.7244	0.0836	0.0003	0.0463	0.9995	0.2128	1.4358	0.1591
	600	400	0.7237	0.0748	-0.0011	0.0458	1.0034	0.2138	1.4344	0.1596