Published in final edited form as:

Paediatr Perinat Epidemiol. 2008 September ; 22(5): 486–496. doi:10.1111/j.1365-3016.2008.00956.x.

To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection

Enrique F. Schisterman and Albert Vexler

Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD, USA

Summary

Pooling of biological specimens has been utilised as a cost-efficient sampling strategy, but cost is not the unique limiting factor in biomarker development and evaluation. We examine the effect of different sampling strategies of biospecimens for exposure assessment that cannot be detected below a detection threshold (DT). The paper compares use of pooled samples to a randomly selected sample from a cohort in order to evaluate the efficiency of parameter estimates.

The proposed approach shows that a pooling design is more efficient than a random sample strategy under certain circumstances. Moreover, because pooling minimises the amount of information lost below the DT, the use of pooled data is preferable (in a context of a parametric estimation) to using all available individual measurements, for certain values of the DT. We propose a combined design, which applies pooled and unpooled biospecimens, in order to capture the strengths of the different sampling strategies and overcome instrument limitations (i.e. DT). Several Monte Carlo simulations and an example based on actual biomarker data illustrate the results of the article.

Keywords

biological samples; pooling design; sampling strategy; detection threshold; Monte Carlo simulations

1. Introduction

The use of biomarkers for exposure assessment is common in epidemiology. The power gained by using a large sample of individuals must be weighed against the cost of performing many assays. After reproducibility and variability are established for the biomarker, financial constraints usually limit further evaluation to small sets of samples. For example, the cost of a single assay measuring polychlorinated biphenyls (PCBs) is up to \$1000 so only small studies have been able to examine, for example, whether PCBs are associated with cancer or endometriosis.^{1,2} However, the imprecision of the results limits the conclusions that can be drawn for the suggested association.

Currently, two different approaches have been suggested to evaluate expensive biomarkers. Suppose we have biological specimens from a patient population A of size N , $A = \{A_1, A_2, \dots, A_N\}$, with test results $X = \{X_1, X_2, \dots, X_N\}$. One approach selects a random sample of the patient population $A^{(r)} = \{A_{k_1}, A_{k_2}, \dots, A_{k_n}\} \in A$, where $n (\leq N)$ is determined by a power calculation and $\{k_i, i = 1, \dots, n\}$ is a subsequence of set $\{1, 2, 3, \dots, N\}$ where assays are performed on the subset of specimens with observed results $\{X_{k_1}, X_{k_2}, \dots, X_{k_n}\}$.

Alternatively, a pooling strategy may be employed where two or more specimens are physically combined into a single 'pooled' unit for analysis. Thus, a greater portion of the population is assayed for the same price compared with the random sampling approach. The amount of information per assay increases so the number of assays needed to achieve equivalent information decreases.^{3–6} Formally, the samples from patient population A are randomly combined into $n = N/p$ pooled specimens of size p . The n pooled assays are considered the average of the contributing individual results, i.e.

$$\begin{aligned} X^{(p)} &= \{X_1^{(p)}, X_2^{(p)}, \dots, X_n^{(p)}\} \\ &= \left\{ \frac{1}{p}(X_{k_{11}} + \dots + X_{k_{1p}}), \frac{1}{p}(X_{k_{21}} + \dots + X_{k_{2p}}), \right. \\ &\quad \left. \dots, \frac{1}{p}(X_{k_{n1}} + \dots + X_{k_{np}}) \right\}, \end{aligned} \quad (1.1)$$

where $\{k_{1i}, i = 1, \dots, p\}, \dots, \{k_{ni}, i = 1, \dots, p\}$ are some disjoint subsequences of set $\{1, 2, 3, \dots, N\}$. Note that, this formal definition of pooled data is commonly accepted for methodological analyses and practical applications of the pooling design.^{3–6}

The concept of pooling biospecimens can be utilised in population-based epidemiological studies to explore the relationship between biomarker levels and outcome. The method's primary goal is in establishing distributional parameters for a specific biomarker. Consequently, pooling can be seen as a primary tool for case-control and cohort studies exploring discrete outcomes. The technique has been explored extensively in the literature starting with publications related to cost-efficient syphilis testing of World War II recruits.⁷ Weinberg and Umbach introduced pooling to estimate odds ratios for case-control studies.⁸ Faraggi *et al.*³ and Liu and Schisterman^{4,5} examined the inference of the effect of pooling on the area under the Receiver Operating Characteristic curve.

Cost is not the only limiting factor in biomarker evaluation. Instrument sensitivity may also be problematic. Another common complexity arises when some participants have levels below the detection threshold (DT).⁹ Under these circumstances, biomarker values at or above the DT are measured and reported, but values below the DT are unobservable. Formally, instead of X , we observe $Z = \{Z_1, Z_2, \dots\}$, such that

$$Z_i = \begin{cases} X_i, & \text{if } X_i \geq d; \\ \text{Not Available(N/A)}, & \text{if } X_i < d, \end{cases} \quad (1.2)$$

where d is a value of the DT. Similarly, for the pooling design, we observed $Z^{(p)} = \{Z_1^{(p)}, \dots, Z_n^{(p)}\}$, where

$$Z_i^{(p)} = \begin{cases} X_i^{(p)}, & \text{if } X_i^{(p)} \geq d; \\ \text{N/A}, & \text{if } X_i^{(p)} < d. \end{cases} \quad (1.3)$$

A variety of approaches have been used to analyse data with a lower DT. Substitution of $d/2$ or $d/\sqrt{2}$ for observations below the DT has been previously described.^{9–11} These values are based on the assumption of a normal ($d/2$) or lognormal ($d/\sqrt{2}$) distribution.¹² Lubin and colleagues proposed multiple imputation based on bootstrapping when the exposure distribution function is known.¹³ Recent work shows that substitution of $E(X | X < d)$ for data

below the DT allows for unbiased estimation of linear and, under certain conditions, logistic regression parameters.¹² Schisterman and colleagues have shown that unbiased estimates may also be obtained non-parametrically if data below the DT are replaced by zero for no intercept models and by an estimator of $E(X | X \geq d)$ for intercept models.¹²

The main objective of this paper is to examine parameter estimation and efficiency of the pooling approach compared with the random sampling approach for assays with a DT. In Section 2, we compare the numerical (quantifiable) information available and efficiency of each sampling scheme. In Section 3, we propose a mixed (unpooling–pooling) design, which takes advantage of the strengths of each approach. In Section 4, we present maximum likelihood techniques to be utilised with the different designs. In Section 5 we illustrate methods to account for pooling and random measurement error and in Section 6 we present our conclusions.

2. Evaluation of designs for biospecimens measurements with a DT: which design yields more numerical information?

The efficiency of pooling and random sampling are compared to determine which design yields more numerical information. Efficiency, here, weighs the available information against the inherent limitations of each design. For clarity, we assume that X has a normal distribution. However, the conclusions from this section are true for most distributions commonly used, including gamma.

Situation 1. The DT is below the mean ($X \sim N(\mu=0, \sigma_x^2=1)$), $\mu > d$)

Figure 1a plots the density function of the normally distributed biomarker X with a DT at $d = -1$. The shaded area corresponds to values of X below the DT where missing values would be reported. The unshaded area corresponds to reportable numerical values of X . In this case, as $\Pr\{X_1 < -1\} \approx 0.16$, the expected proportion of observations below the DT is approximately 16%. Pooling the specimens reduces the effective variance of biomarker X , i.e. by definition (1.1) the variance of the pooled samples is $\text{var}(X^{(p)}) = \sigma_x^2/p$ and the mean is 0.³ For the pooled samples, assuming $p = 2$, $\Pr\{X_1^{(p)} < -1\} \approx 0.08$ so approximately 8% of the pooled observations are below the DT as shown in Fig. 1c. Thus, the expected number of unobserved test results from the random sample design is about twice $((n \times \Pr\{X_1 < d\}) / (n \times \Pr\{X_1^{(p)} < d\}) \approx 0.16/0.08 = 2)$ the expected number of N/As under the pooling strategy.

The rationale for pooling in this case is to take advantage of the statistical properties of averages through physical implementation, i.e. the value of pooled specimens is the mean of the individual biomarker values.

The pooling strategy provides more numerical observations than the random sampling approach with equivalent initial sample size, because the pooled distribution $X^{(p)}$ with $\text{var}(X^{(p)}) = \sigma_x^2/p$ is more concentrated around the expectation $\mu = 0$. Hence, the pooling strategy is more efficient than the random sample in estimating the mean and variance. Note that if $d = -\infty$ the maximum likelihood estimators of μ based on full data Z and pooled data $Z^{(p)}$ have equal efficiency.³ In Situation 1, the ratio of the expected number of numerically observed test results of set Z to that of set $Z^{(p)}$ is $(N \times \Pr\{X_1 > d\}) / (n \times \Pr\{X_1^{(p)} > d\}) \geq N/n$, i.e. the number of numerically observed results from $Z^{(p)}$ increases relative to the observed numerical elements of Z . Moreover, although $(N \times \Pr\{X_1 > d\}) / (n \times \Pr\{X_1^{(p)} > d\}) > 1$, we cannot conclude that the observed pooled data $Z^{(p)}$ has less numerical information than the full data set Z .

Consider an example with four unpooled individual specimens: $X_1 = 3.1$, $X_2 = 3.5$, $X_3 = 4.0$ and $X_4 = 2.0$. If $DT = 3.0$, $Z = \{3.1, 3.5, 4.0, N/A\}$, where N/A signifies a value below the DT. If $p = 2$, the pooled samples include only two numerical observations:

$$X_1^{(2)} = \frac{1}{2}(X_1 + X_2) = 3.3, \quad X_2^{(2)} = \frac{1}{2}(X_3 + X_4) = 3.0,$$

yielding $Z^{(2)} = \{3.3, 3.0\}$. In this example, the value of X_4 is not ignored by the pooled data, which are less affected by the DT than the full sample.

Situation 2. The DT is above the mean ($X \sim N(\mu=0, \sigma_x^2=1)$, $\mu < d$)

Figures 1b and d depict when the location of the DT is above the mean of X and the situation previously described is reversed. If the DT is located at 1 for example, more pooled samples will have values less than the DT than unpooled samples. As shown in Fig. 1b, the amount of unobserved data (shaded area) is smaller in the unpooled data than in the pooled data. Hence, pooling is beneficial when the DT is below the mean and detrimental when it is above the mean.

Nevertheless, in Situation 2, the pooling strategy might still be more efficient than random sampling. Intuitively, the pooled observations might be more informative than the unpooled observations because each pooled observation is based on more than one test result.

Situation 3. The DT is far above the mean ($d \gg \mu = EX$)

When the DT is much greater than the mean biomarker value, the pooling strategy is completely inefficient because the pooled data are based upon substantially less numerical information than a random sample of unpooled data.

In order to demonstrate the conclusions from Situations 1–3 with respect to sample size, we generated random sample $X_1, \dots, X_{N=1000} \sim N(0,1)$ 5000 times for $p = 2$ and $p = 4$ with $d \in [-2,2]$. Figure 2 shows the average number of uncensored observations from the three types of samples: $\{Z_1, \dots, Z_N\}$ (whole sample); $\{Z_1, \dots, Z_{N/p}\}$ (random sample) and

$\{Z_1^{(p)}, \dots, Z_{N/p}^{(p)}\}$ (pooled sample) plotted against the values of the DT. These findings also apply to other than a normal distribution. For example, the gamma case shown in Fig. 3 ($\chi^2_{(20)}$ for unpooled data and $\chi^2_{(40)}$ for pooled data) leads to similar conclusions.

3. Pooled–unpooled design

Consider the situation where an assay is relatively inexpensive and could be measured for every participant. As previously stated, numerical values are not assigned when X is below the DT. However, knowledge of the data below the DT is important for inference. Richardson and Ciampi suggested imputing $E(X | X < d)$ for values below the DT.¹⁴ Because cost is not an impediment in this example, we propose to assay the individual specimens and then pool the specimens and assay the pooled samples as well. As described in Section 2, when an individual specimen with a value less than the DT is pooled, the pooled sample may have a numerical result. Therefore, the individual's X may be back-calculated (reconstructed) using the pooled results and the individual results from the other samples in the pool. The combination of pooling results with traditional unpooled measurements can produce numerical results for the maximum number of study participants, including some below the DT.

In this discussion we use $p = 2$ without loss of generality. Consider an individual k with an unpooled value $X_k < d$, yielding a test result $Z_k = N/A$. As shown previously, $X^{(p)} = \frac{X_k + X_l}{2}$ may be above the DT, given some individual l . Thus, X_k can be estimated by $2X^{(p)} - X_l$ every time that $X^{(p)} > d$ and $X_k < d$. If $X^{(p)} > d$ then X_l can be numerically measured in the unpooled data because $X_l > d$. Note that the samples to be pooled must be selected randomly to avoid dependency in the data.

By the combined application unpooled–pooled strategies, the value of some observations less than the DT can be calculated allowing $E(X | X < d)$ to be non-parametrically estimated using the method proposed by Richardson and Ciampi.¹⁴ We call the proposed technique the pooled–unpooled resampling design.

4. Maximum Likelihood Estimation (MLE)

4.1 MLE for pooled or random sampling design

Having introduced the pooled–unpooled hybrid design, we can utilise maximum likelihood to estimate unknown parameters of a biomarker's distribution. We can consider a biomarker that follows $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, N$, where estimation of the mean of X_i (the variance is assumed to be unknown) based on Z or $Z^{(p)}$ is of interest. Note that Z_i is not normally distributed because there will be a mass of data at the DT. Therefore, the direct classical maximum likelihood method based on normal data is unsuitable, and naively applying the least squares method to the available data leads to an estimator of $E(X_i, X_i \geq d) \neq \mu$. Instead, we utilise maximum likelihood estimators proposed by Gupta that are appropriate for this case because they yield asymptotically unbiased estimators of the mean and variance (see Appendix for estimation details).¹⁵ The target likelihood function [Appendix formula (A.1)] has two parts: a component related to the number of unobserved test results and a component for the numerically observed data (this part is similar to the classic MLE). The statistical properties and rate of convergence to unbiasedness of the maximum likelihood estimators are functions of the location of the DT. These estimators are based on likelihood theory; they maximise the Fisher information and are the most efficient. However, their accuracy and efficiency are direct results of the study design. We propose the obtained efficiency of $(\hat{\mu}, \hat{\sigma})$ in formula (A.2) of the Appendix. Suppose X_i has the standard normal distribution $N(0,1)$ Figure 4 displays the efficiency of the maximum likelihood estimators.

The estimations of μ and σ based on the pooled data are more efficient than those based upon the random sample up to $d < \mu$. However, if $d \gg \mu$, then the pooling strategy is not recommended.

We can estimate parameters based on data following a gamma distribution in a similar manner.¹⁶ The gamma-shape parameter of the pooled data is $p \times$ the shape parameter of unpooled data and the scale parameter of the pooled data is $1/p \times$ the scale parameter of unpooled data.³ The conclusions for the gamma case are similar to the normal.

4.2 MLE for pooled–unpooled design

The likelihood function in Section 4.1 is composed of two parts: one related to N/A -observed data (where $X < d$) and a second for numerically observed data (where $X \geq d$); estimation for pooled–unpooled resampled data has three kinds of data. The first sample (S_1) has only N/A elements. Test results in this sample were initially below the DT and have not been reconstructed by applying the pooling resampling. Thus, for all $k = 1, \dots, N$, we have

$$\begin{aligned}
\Pr\{X_k \in S_1\} &= \Pr\{X_l \notin S_1, X_k < d, \\
&\quad \frac{X_k + X_l}{2} < d\} + \Pr\{X_l \in S_1, X_k < d, \\
&\quad \frac{X_k + X_l}{2} < d\} = \Pr\{X_k < d, \frac{X_k + X_l}{2} < d\} \\
&= \Pr\{X_1 < d, \frac{X_1 + X_2}{2} < d\}
\end{aligned} \tag{3.1}$$

The second sample (S_2) has reconstructed elements. Test results in this sample were initially below the DT and have been reconstructed by applying the pooling resampling. Therefore, elements of set S_2 have distribution function

$$\begin{aligned}
F_R(u) &= \Pr\{X_k < u, X_k \in S_2\} = \Pr\{X_k < u, X_k < d, \frac{X_k + X_l}{2} > d\} \\
&= \Pr\{X_1 < u, X_1 < d, \frac{X_1 + X_2}{2} > d\}.
\end{aligned} \tag{3.2}$$

The last sample (S_3), as in Section 4.1, includes the numerically observed data. The likelihood function is a product of the densities that correspond to (3.1), (3.2) and the case, where numerical results were initially observed. We describe the likelihood in detail in Appendix formula (A.3).

To illustrate the proposed method, we generated a random sample $\{X_1, \dots, X_{N=300}\}$ 10 000 times from $N(\mu=1, \sigma_x^2=4)$. We applied the one-step resampling to each simulated sample for $d = 0.5$. The average number of observed numerical values in the simulation ($\#\{X_i > d\}$) was about 179.58 (which agrees with the estimate

$n \times \left(1 - \int_{-\infty}^d \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{1}{2}\left(\frac{u-\mu}{\sigma_x}\right)^2\right) du\right) = 179.6119$. After applying the resampling strategy, the average number of numerical elements in the Monte Carlo (MC) data was about 221.63 (which agrees with the estimate $n \times (1 - \Pr\{X_k \in S_1\}) = 221.569$). Thus, on average, the numerical information increased by approximately 25%. As would be expected, the MC variances of the μ and σ -estimators improved due to the resampling by about 24%. Similar results are attained for different values of d , where $d < \mu$.

4.3 Example

Cholesterol measurements were collected for 10 normal volunteers at a medical centre. The mean and standard deviation for total cholesterol were estimated to be ($\hat{\mu} = 200.73$) and ($\hat{\sigma} = 51.72$) respectively. The specimens were then randomly paired and the pooled specimens were assayed. For the purpose of demonstration, we artificially created a threshold (DT = 150) such that some numerical values could not be observed. In Table 1, we show the individual and pooled cholesterol values with and without the DT.

In this example, 20% of the individual measurements are below the threshold, whereas no pooled observations are below the DT. Applying the maximum likelihood method, the asymptotically unbiased mean and standard deviation were estimated to be ($\hat{\mu} = 196.13$) and ($\hat{\sigma} = 56.38$), respectively, from unpooled data with the DT. Although more costly, by assaying both pooled and unpooled specimens, we can reasonably estimate values below the DT (Table 1). Moreover, using both the reconstructed data and the unpooled data above the DT, the mean and the variance are estimated to be $\hat{\mu} = 198.99$ and $\hat{\sigma} = 53.44$

5. Pooling and random error in design

5.1 Pooling errors

Although definition (1.1) shows the theoretical notation for pooled data, practically, pooling biological specimens can lead to additive pooling errors. In this section we use the maximum likelihood method from Section 4 and revise definition (1.1) to

$$\begin{aligned} X^{(p)} &= \{X_1^{(p)}, X_2^{(p)}, \dots, X_n^{(p)}\} \\ &= \left\{ \frac{1}{p}(X_{k_{11}} + \dots + X_{k_{1p}}) + \varepsilon_1, \right. \\ &\quad \frac{1}{p}(X_{k_{21}} + \dots + X_{k_{2p}}) + \varepsilon_2, \\ &\quad \left. \dots \frac{1}{p}(X_{k_{n1}} + \dots + X_{k_{np}}) + \varepsilon_n \right\}, \end{aligned} \quad (5.1)$$

where pooling errors $\varepsilon_1, \dots, \varepsilon_n$ are independent $N(0, \sigma_\varepsilon^2)$ distributed random variables and $X \sim N(\mu, \sigma_X^2)$. Definition (5.1) accounts for the pooling errors which were ignored by definition (1.1). In order to investigate the robustness of our approach for addressing pooling errors, we executed MC simulations. Formally, we assumed that only $n = N/p$ measurements can be performed and compare:

1. Random sampling: We randomly choose $X_1, \dots, X_{N/p}$ from the full sample and observed $Z_1, \dots, Z_{N/p}$ because of the DT. The mean of X_i was estimated using the likelihood approach on the truncated data $\{Z_1, \dots, Z_{N/p}\}$.
2. Pooling: We randomly choose biospecimen sets of size p with

$$X_i^{(p)} = \frac{1}{p} \sum_{j=p(i-1)+1}^{pi} X_j + \varepsilon_i, \quad i=1, \dots, \frac{N}{p} \text{ and observed } Z_i^{(p)}, \quad i=1, \dots, \frac{N}{p} \text{ by (1.3). Again, the mean of } X_i \text{ was estimated using the MLE based on } Z_1^{(p)}, \dots, Z_{N/p}^{(p)}.$$

The accuracy of estimators $(\hat{\mu}, \hat{\sigma})$ of (μ, σ_X) is indicated by their MC variances. We assumed a biomarker distribution $X_i \sim N(\mu=1, \sigma_X^2=4)$, $i=1, \dots, N$, and generated a random sample $X_1, \dots, X_{N=300}$ 10 000 times for pool sizes $p=2, 4$ and for various DTs. Figure 5 presents the MC estimators of $E(\hat{\mu} - EX)^2$ and $E(\hat{\sigma} - \sigma_X)^2$, where $\hat{\mu}$, the estimator of EX , is based on full data $\{Z_1, \dots, Z_N\}$, the random sample $\{Z_1, \dots, Z_{N/p}\}$ and the pooled data $\{Z_1^{(p)}, \dots, Z_{N/p}^{(p)}\}$ with $\sigma_\varepsilon = \sigma_X/10$, $\sigma_\varepsilon = \sigma_X/5$, $\sigma_\varepsilon = \sigma_X/3$, $\sigma_\varepsilon = \sigma_X/2$ and $\sigma_\varepsilon = \sigma_X$.

The figure suggests that the conclusions in Sections 2 and 4 are correct for μ -estimation up to $\sigma_\varepsilon \leq \sigma_X/2$ [Figs 5(a.1)–(a.4)]. However, above $\sigma_\varepsilon \geq \sigma_X/5$, the σ_X -estimator using the pooled data has the largest variance. The likelihood function (A.1) should be changed because of pooling errors as shown in (5.1). The statistical properties of $\varepsilon_1, \dots, \varepsilon_n$ can be evaluated in a similar manner to a study by Schisterman *et al.*¹⁷

5.2 Random measurement error

In addition to pooling error, studies are also subjected to random measurement error as a function of instrument calibration. Random measurement error occurs as a result of random instrument variability. One can account for random measurement error in the pooled or random sample designs through the use of standard techniques previously developed in the literature.^{18,19} These techniques include utilising error models, regression calibration models, validation studies or replication data to estimate and adjust for random measurement error. In addition, while not explicitly described here, standard information reported by a laboratory such as the

coefficient of variation for the biomarker and reliability of the assay can be included in these models.

6. Discussion

In this paper, we examine pooling and random sampling as strategies to evaluate biospecimens with a DT. These types of data are common in epidemiological research and include two types of values: numerical and non-numerical (i.e. N/A). Because numerical values yield more information than missing data, it is a goal of any researcher to minimise the number of N/A observations. Accordingly, we have explored theoretical methods as well as simulations where a pooling design is more efficient than a random sample. In addition, we show that the efficiency of the pooling design is dictated by the location of the DT but is independent of the distributional assumptions (e.g. gamma, t-distribution, Lognormal). For all distributions, there is a range of DTs where the pooling strategy is more efficient than a random sample because the inference-based pooling design provides more numerical information. In fact, in some cases pooling is more efficient than using the full sample. We showed that whenever $EX > d$ (i.e. $>50\%$) pooling is always the most efficient sampling strategy, but other factors, such as the underlying distribution, must be considered when $EX < d$.

Certainly, a preliminary analysis of biospecimens with incomplete measurements, such as a test to see if $EX > d$, is appropriate. Towards this end, the unpooled–pooled strategy proposed in Section 3 is not only helpful for the evaluation of pooling errors but can also be applied to a first-stage data study. In addition, the efficiency of MLEs under each design can be evaluated.

Cost has been the main motivation for pooling biological specimens or to randomly select a subset of individuals to be assayed. However, we have shown that in some cases, even when the full data are available, estimations based on pooled data increase efficiency over the use of individual measurements when the assay has a DT. This is because of the greater number of observations above the DT under pooling, which can then be used in the estimation procedure. However, using unpooled data allows, for example, distributional assumptions to be tested, the location of the DT to be estimated and the expected number of observations below the DT. In addition, one is able to stratify the pooled samples by confounders in order to retain confounding and covariate information in the pooled samples. To take advantage of the strengths of each of these approaches, we proposed a pooled–unpooled resampling design. According to this design, in the first stage all the patient population (or a random sample of them) is measured individually, and in the second stage, the patient population is pooled in groups of size p and these pooled samples are assayed. By employing this approach, we are able to reconstruct data that were unobserved in the first stage due to the DT.

This simple approach that we propose captures the strengths of the statistical properties of the distribution function of the averages by physically grouping biological specimens in order to overcome the instrument limitations.

References

1. Laden F, Neas LM, Spiegelman D, Hankinson SE, Willett WC, Ireland K, et al. Predictors of plasma concentrations of DDE and pcbs in a group of U.S. women. *Environmental Health Perspectives* 1999;107:75–81. [PubMed: 9872720]
2. Louis GM, Weiner JM, Whitcomb BW, Sperrazza R, Schisterman EF, Lobdell DT, et al. Environmental PCB exposure and risk of endometriosis. *Human Reproduction* 2005;20:279–285. [PubMed: 15513976]
3. Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. *Statistics in Medicine* 2003;22:2515–2527. [PubMed: 12872306]

4. Liu A, Schisterman EF. Sample size and power calculation in comparing diagnostic accuracy of biomarkers with pooled assessments. *Journal of Applied Statistics* 2004;31:41–51.
5. Liu A, Schisterman E. Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biometrical Journal* 2003;45:631–644.
6. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology* 2005;16:73–81. [PubMed: 15613948]
7. Keeler E, Berwick D. Effects of pooled samples. *Health Laboratory Science* 1976;13:121–128. [PubMed: 1270261]
8. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics* 1999;55:718–726. [PubMed: 11314998]
9. Helsel, D. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. Hoboken, NJ: John Wiley & Sons, Inc; 2005.
10. Finkelstein M, Verma D. Exposure estimation in the presence of nondetectable values: another look. *AIHAJ* 2001;62:195–198. [PubMed: 11331991]
11. Hornung R, Reed L. Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene* 1990;5:46–51.
12. Schisterman EF, Vexler A, Whitcomb BW, Liu A. The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology* 2006;163:374–383. [PubMed: 16394206]
13. Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, et al. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental Health Perspectives* 2004;112:1691–1696. [PubMed: 15579415]
14. Richardson DB, Ciampi A. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology* 2003;157:355–363. [PubMed: 12578806]
15. Gupta AK. Estimation of the mean and standard deviation of a normal population from a censored sample. *Biometrika* 1952;39:260–273.
16. Chapman DG. Estimating the parameters of a truncated gamma distribution. *Annals of Mathematical Statistics* 1956;27:498–506.
17. 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]
18. Carroll, RJ.; Ruppert, D.; Stefanski, LA. *Measurement Error in Nonlinear Models*. Boca Raton, FL: Chapman & Hall; 1995.
19. Gustafson, P. *Measurement Error and Misclassification in Statistics and Epidemiology*. Boca Raton, FL: Chapman & Hall; 2004.

Appendix

Appendix

1. Estimation of mean and standard deviation using the Gupta method

Following Gupta's method,¹⁵ we obtain the MLE based on a sample with observations subject to a DT:

$$L(\Omega) = \frac{N_{\Omega}!}{(k_{\Omega}-1)!(N_{\Omega}-k_{\Omega})!} \cdot (\Pr\{X^{\Omega} < d\})^{k_{\Omega}} \cdot \prod_{i: X_i^{\Omega} \geq d} f_{\Omega}(X_i^{\Omega}), \quad (\text{A.1})$$

where f_{Ω} is a density function of X_1^{Ω} ; N_{Ω} is a size of set Ω ; k_{Ω} is a number of N/A -elements of set Ω ; ($\Omega = Z$, $X^{\Omega} = X_1$) or ($\Omega = \{Z_i, i = 1, \dots, n\}$, $X^{\Omega} = X_1$), or ($\Omega = Z^{(p)}$, $X^{\Omega} = X_1^{(p)}$), if the full

data or the random sample, or the pooled data are available, respectively. Especially, since $X_i \sim N(\mu, \sigma^2)$, we have

$$\Pr\{X_1 < d\} = \int_{-\infty}^d f_z(u) du, \quad \Pr\{X_1^{(p)} < d\} = \int_{-\infty}^d f_{z^{(p)}}(u) du,$$

$$f_z(u) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2} \quad \text{and} \quad f_{z^{(p)}}(u) = \frac{\sqrt{p}}{\sqrt{2\pi}\sigma} e^{-\frac{p}{2}\left(\frac{u-\mu}{\sigma}\right)^2}.$$

Thus, $L(\Omega)$ is a function of unknown parameters μ and σ , say $L(\Omega) = L(\mu, \sigma; \Omega)$.

The target estimators $\hat{\mu}, \hat{\sigma}$ of μ, σ (where $(\mu, \sigma) = \arg \max_{(\mu, \sigma)} L(\mu, \sigma; \Omega)$) are numerical solutions of system

$$\left\{ \frac{\partial \log(L(\hat{\mu}, \hat{\sigma}; \Omega))}{\partial \hat{\mu}} = 0, \quad \frac{\partial \log(L(\hat{\mu}, \hat{\sigma}; \Omega))}{\partial \hat{\sigma}} = 0. \right.$$

2. Asymptotic efficiency of the MLE

The variances of considered estimators $(\hat{\mu}, \hat{\sigma})$ of (μ, σ) can be found by inverting the Fisher information matrix. Using Gupta,¹⁵ we obtain, depending on pooled/unpooled database,

$$\lim_{N \rightarrow \infty} N \begin{bmatrix} \text{var}(\hat{\mu}) & \text{cov}(\hat{\mu}, \hat{\sigma}) \\ \text{cov}(\hat{\mu}, \hat{\sigma}) & \text{var}(\hat{\sigma}) \end{bmatrix} = \sigma^2 w_j \begin{bmatrix} p_j + \varphi(\eta_j) G(\eta_j) & \frac{1}{g_j^{1/2}} (\varphi(\eta_j) + \eta_j \varphi(\eta_j) G(\eta_j)) \\ \frac{1}{g_j^{1/2}} (\varphi(\eta_j) + \eta_j \varphi(\eta_j) G(\eta_j)) & \frac{1}{g_j} (2p_j + \eta_j \varphi(\eta_j) + \eta_j^2 \varphi(\eta_j) G(\eta_j)) \end{bmatrix}^{-1}, \quad (\text{A.2})$$

$$\eta_j = \frac{g_j^{1/2}(d - \mu)}{\sigma}; p_j = \int_{\eta_j}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du; G(\eta) = \frac{e^{-\frac{1}{2}\eta^2}}{\int_{-\infty}^{\eta} e^{-\frac{1}{2}u^2} du} + \eta$$

where $\varphi(u)$ is the standard normal

density function $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ and if $(\hat{\mu}, \hat{\sigma})$ are based on

$$\begin{cases} \text{full data } Z \text{ then } j=1, g_1=1, w_1=1; \\ \text{pooled data } Z^{(p)} \text{ then } j=2, g_2=p, w_2=1; \\ \text{random sample } \{Z_i, i=1, \dots, n\} \text{ then } \\ j=3, g_3=1, w_3=p. \end{cases}$$

3. Maximum likelihood corresponds to the resampling strategy

In accord with Section 3.2, we have

$$\begin{aligned}
L(\mu, \sigma_X; S_1, S_2, S_3) &\propto (\Pr\{X_1 \in S_1\})^{m_1} \\
&\left(\prod_{i: X_i \in S_2} \frac{d}{du} F_R(u) \Big|_{u=X_i} \right) \times \frac{1}{\sigma_X^{N-m_1-m_2}} \\
&\exp\left(-\frac{1}{2\sigma_X^2} \sum_{i: X_i \in S_3} (Z_i - \mu)^2\right) =: L_c(\mu, \sigma_X),
\end{aligned} \tag{A.4}$$

where m_1 and m_2 are number of elements of data S_1 and S_2 respectively; by applying (3.1), (3.2) and convolution transforms, we obtain

$$\begin{aligned}
\Pr\{X_1 \in S_1\} &= \Pr\left\{X_1 < d, \frac{X_1 + X_2}{2} < d\right\} \\
&= \frac{1}{2\pi\sigma_X^2} \int_{-\infty}^d \int_{-\infty}^{2d-u} \exp\left(-\frac{(t-\mu)^2}{2\sigma_X^2}\right) dt \\
&\exp\left(-\frac{(u-\mu)^2}{2\sigma_X^2}\right) du, \\
\frac{d}{du} F_R(u) \Big|_{u=X_i} &= \frac{d}{du} \Pr\{X_1 < u, X_1 < d, \frac{X_1 + X_2}{2} > d\} \Big|_{u=X_i} \\
&= \frac{1}{2\pi\sigma_X^2} \frac{d}{du} \left(\int_{-\infty}^{\min(u,d)} \int_{2d-t}^{\infty} \exp\left(-\frac{(y-\mu)^2}{2\sigma_X^2}\right) dy \right. \\
&\left. \exp\left(-\frac{(t-\mu)^2}{2\sigma_X^2}\right) dt \right) \Big|_{u=X_i} \\
&= \frac{1}{2\pi\sigma_X^2} \int_{2d-X_i}^{\infty} \exp\left(-\frac{(y-\mu)^2}{2\sigma_X^2}\right) dy \\
&\exp\left(-\frac{(X_i-\mu)^2}{2\sigma_X^2}\right).
\end{aligned}$$

Now, the maximum likelihood estimators are $(\widehat{\mu}, \widehat{\sigma}_X) = \arg \max_{(\mu, \sigma_X)} \log(L_c(\mu, \sigma_X))$.

4. Maximum likelihood that allows for the pooling errors

The general maximum likelihood function is

$$\begin{aligned}
L(\mu, \sigma_X, \sigma_\varepsilon; \{Z_1, \dots, Z_{n_1}\} \cap \{Z_1^{(p)}, \dots, Z_{n_2}^{(p)}\}) &\propto \\
&\left(\frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^d e^{-\frac{1}{2\sigma_X^2}(u-\mu)^2} du \right)^{m_1} \times \left(\frac{1}{\sqrt{2\pi(\sigma_X^2/p + \sigma_\varepsilon^2)}} \right. \\
&\left. \int_{-\infty}^d e^{-\frac{1}{2(\sigma_X^2/p + \sigma_\varepsilon^2)}(u-\mu)^2} du \right)^{m_2} \frac{1}{\sigma_X^{n_1-m_1}(\sigma_X^2/p + \sigma_\varepsilon^2)^{n_2/2-m_2}} \times \\
&\exp\left(-\frac{1}{2\sigma_X^2} \sum_{i: Z_i \neq N/A} (Z_i - \mu)^2 - \frac{1}{2(\sigma_X^2/p + \sigma_\varepsilon^2)} \sum_{i: Z_i^{(p)} \neq N/A} (Z_i^{(p)} - \mu)^2\right) =: L_c(\mu, \sigma_X, \sigma_\varepsilon),
\end{aligned} \tag{A.3}$$

where m_1 and m_2 are number of N/As of sets $\{Z_i, i = 1, \dots, n_1\}$ and $\{Z_i^{(p)}, i = 1, \dots, n_2\}$ respectively. Therefore, the maximum likelihood estimators are

$$(\widehat{\mu}, \widehat{\sigma}_X, \widehat{\sigma}_\varepsilon) = \arg \max_{(\mu, \sigma_X, \sigma_\varepsilon)} \log(L_c(\mu, \sigma_X, \sigma_\varepsilon)).$$

Figure 1.
Normally distributed data constrained by a detection threshold (shaded area represents unobserved data). DT, detection threshold.

Figure 2.
The Monte Carlo averages of the numerical sample sizes. DT, detection threshold.

Figure 3. Chi-square distributed data constrained by a detection threshold (shaded area represents unobserved data). DT, detection threshold.

Figure 4.

Efficiency of the maximum likelihood estimators: $\lim_{N \rightarrow \infty} \log(N \times \text{var}(\hat{\mu}))$ and $\lim_{N \rightarrow \infty} \log(N \times \text{var}(\hat{\sigma}))$ are plotted by graphs (a) and (b) respectively. Curves (---), (—) and (·····) correspond to databases Z , $Z^{(p)}$ and $\{Z_i, i = 1, \dots, n\}$ respectively.

Figure 5. Logarithm of the Monte Carlo estimators of $E(\hat{\mu} - \mu)^2$ and $E(\hat{\sigma} - \sigma_X)^2$ [graphs (a) and (b) respectively], where the pooling error is in effect. Curves (-----), (—) and (.....) correspond to databases Z , $Z^{(p)}$ (with pooling errors) and $\{Z_i, i = 1, \dots, n\}$ respectively.

Table 1
Numerical reconstruction of the values below the detection threshold (DT)

A	Original unpoolled cholesterol results	Pooled cholesterol samples in set of $p = 2$	Cholesterol values above the putative DT = 150	Pooled cholesterol values above the putative DT = 150	Reconstructed unpoolled values below the DT
1	138.2		N/A		$2 \times X_1^{(p)} - X_2 = 122.6$
2	263.2	$X_1^{(p)} = 192.9$	263.2	$X_1^{(p)} = 192.9$	
3	153.2		153.2		
4	186.4	$X_2^{(p)} = 164.1$	186.4	$X_2^{(p)} = 164.1$	
5	182.3		182.3		
6	214.0	$X_3^{(p)} = 211.6$	214.0	$X_3^{(p)} = 211.6$	
7	144.3		N/A		$2 \times X_4^{(p)} - X_8 = 140.5$
8	192.5	$X_4^{(p)} = 166.5$	192.5	$X_4^{(p)} = 166.5$	
9	292.6		292.6		
10	240.6	$X_5^{(p)} = 286.3$	240.6	$X_5^{(p)} = 286.3$	

$X_j^{(p)}$ are the cholesterol test results by pooling $(Ap(j-1)+1 \cap Apj), j \geq 1$.