Ann. Occup. Hyg., Vol. 51, No. 7, pp. 611–632, 2007

doi:10.1093/annhyg/mem045

A Comparison of Several Methods for Analyzing Censored Data

PAUL HEWETT[1]* and GARY H. GANSER[2]

[1]Exposure Assessment Solutions, Inc., Morgantown, West Virginia; [2]Department of Mathematics, West Virginia University, Morgantown, West Virginia

Received 5 March 2007; in final form 17 August 2007

The purpose of this study was to compare the performance of several methods for statistically analyzing censored datasets [i.e. datasets that contain measurements that are less than the field limit-of-detection (LOD)] when estimating the 95th percentile and the mean of right-skewed occupational exposure data. The methods examined were several variations on the maximum likelihood estimation (MLE) and log-probit regression (LPR) methods, the common substitution methods, several non-parametric (NP) quantile methods for the 95th percentile and the NP Kaplan–Meier (KM) method. Each method was challenged with computer-generated censored datasets for a variety of plausible scenarios where the following factors were allowed to vary randomly within fairly wide ranges: the true geometric standard deviation, the censoring point or LOD and the sample size. This was repeated for both a single-laboratory scenario (i.e. single LOD) and a multiple-laboratory scenario (i.e. three LODs) as well as a single lognormal distribution scenario and a contaminated lognormal distribution scenario. Each method was used to estimate the 95th percentile and mean for the censored datasets (the NP quantile methods estimated only the 95th percentile). For each scenario, the method bias and overall imprecision (as indicated by the root mean square error or rMSE) were calculated for the 95th percentile and mean. No single method was unequivocally superior across all scenarios, although nearly all of the methods excelled in one or more scenarios. Overall, only the MLE- and LPR-based methods performed well across all scenarios, with the robust versions generally showing less bias than the standard versions when challenged with a contaminated lognormal distribution and multiple LODs. All of the MLE- and LPR-based methods were remarkably robust to departures from the lognormal assumption, nearly always having lower rMSE values than the NP methods for the exposure scenarios postulated. In general, the MLE methods tended to have smaller rMSE values than the LPR methods, particularly for the small sample size scenarios. The substitution methods tended to be strongly biased, but in some scenarios had the smaller rMSE values, especially for sample sizes <20. Surprisingly, the various NP methods were not as robust as expected, performing poorly in the contaminated distribution scenarios for both the 95th percentile and the mean. In conclusion, when using the rMSE rather than bias as the preferred comparison metric, the standard MLE method consistently outperformed the so-called robust variations of the MLE-based and LPR-based methods, as well as the various NP methods, for both the 95th percentile and the mean. When estimating the mean, the standard LPR method tended to outperform the robust LPR-based methods. Whenever bias is the main consideration, the robust MLE-based methods should be considered. The KM method, currently hailed by some as the preferred method for estimating the mean when the lognormal distribution assumption is questioned, did not perform well for either the 95th percentile or mean and is not recommended.

Keywords: censored data analysis; limit-of-detection

INTRODUCTION

As exposure limits decrease and exposure controls improve an increasingly frequent occurrence is the left-censored dataset; that is, a dataset where one or more measurements are less than the field limit-of-detection (LOD; i.e. the laboratory LOD divided by the sample volume) for a particular combination of sampling method, flow rate and sample time. Left-censored datasets tend to occur whenever there is a high LOD relative to exposures;

*Author to whom correspondence should be addressed.
Tel: 001 304 685 7050; e-mail: phewett_2006_07@oesh.com

the exposures span several orders of magnitude; or the sample time is short or the flow rate is low, resulting is a small sample volume. Published censored data analysis (CDA) methods fall into four general categories: substitution methods, log-probit regression (LPR) methods, maximum likelihood estimation (MLE) methods and non-parametric (NP) methods. Within each category or family, there are several variations, usually developed to reduce transformation bias (discussed later) or for the situation where it is suspected that the underlying distribution departs significantly from the assumed lognormal distribution in the hope of reducing the bias or improving overall accuracy (defined as bias plus precision) when estimating the 'mean concentration' (Helsel, 2005). This has resulted in numerous peer-reviewed articles that offer sometimes contradictory guidance regarding which is the preferable method. Information on the accuracy of upper percentile compliance statistics, such as the 95th percentile, when calculated from a censored dataset, is difficult to find. Helsel (2005), author of a recently published text devoted entirely to CDA, strongly recommended using the NP Kaplan–Meier (KM) method to estimate the mean whenever the dataset is $<50\%$ censored and 'robust' parametric methods in other instances. Last, the simple substitution methods, which continue to be commonly used, are often condemned in the literature in preference to these other methods.

We distinguish between 'simple-censored' and 'complex-censored' datasets. A simple-censored dataset contains one or more measurements censored at a single LOD, or two or more LODs, but all at the low end. In contrast, a complex-censored dataset contains measurements censored at two or more LODs with uncensored measurements scattered in between (Bullock and Ignacio, 2006). In our experience, most censored datasets appear to be of the simple-censored variety: a single laboratory is used and the exposure profile for the exposure group is reasonably stable. Complex datasets occur whenever investigators combine data from several studies where different laboratories were used or combine data from exposure groups that were collected across some broad span of time during which several laboratories were utilized. Furthermore, with large datasets comes the possibility that the data reflect different exposure distributions as exposure profiles are unlikely to be stable for periods of more than a year (Symanski *et al.*, 1996).

In this paper we address the following questions:

- What is the method bias and overall accuracy when estimating the 95th percentile or mean of a lognormal or contaminated lognormal exposure profile?
- Assuming that the underlying exposure profile is reasonably lognormal, which CDA methods should be considered?
- Which method should be used for complex-censored datasets and/or when we suspect that the underlying exposure distribution departs significantly from single lognormal assumption?
- Is there an 'omnibus' method that should be the first choice, regardless of the sample size, (observed) percent censored, complexity of censoring or variability in the data?

To address these questions, we estimated the bias and root mean square error (rMSE) for each of the CDA methods when estimating a commonly used compliance statistic (i.e. the 95th percentile) and the exposure profile mean (often used in environmental evaluations and epidemiological studies). The bias is a function not only of the CDA method employed but is also a function of the true geometric standard deviation (GSD), true percent censored and the sample size. The rMSE for each method is an estimate of the overall accuracy (i.e. overall imprecision), which is a function of both bias and precision.

While earlier studies of CDA methods (discussed later) are helpful for sorting out the above issues, they often had limitations. Obviously, the very early studies will be dated with respect to newer methods. Most studies, even the relatively recent studies, focused on only a limited number of methods. Many of the studies averaged their results across all of the posited scenarios and underlying distributions making it difficult to impossible to sort out the results for a specific method and type of underlying distribution. Most of the studies considered only one or a few samples sizes and/or only a few levels of variability. Many of the studies generated only 500–1000 simulated datasets per method, which can result in variable and misleading bias and rMSE values, which we in this study attempted to avoid by using 100 000 as the simulation sample size. Several studies examined the robustness of the CDA methods by challenging them with data drawn from identical 'contaminated lognormal' distributions, that is, distributions created by combining two lognormal distributions. However, these contaminated distributions were similar in shape to a standard lognormal and thus not a severe challenge for the methods. Last, all of the studies focused exclusively or primarily on estimating the mean of a lognormal, right-skewed distribution. Our main interests are the upper percentiles estimates, such as the 95th percentile exposure (Mulhausen and Damiano, 1998; Bullock and Ignacio, 2006), and while a few of the studies passingly addressed upper percentiles, none did so comprehensively enough for us to draw reliable inferences.

We recognize that many aspects of our computer simulations repeat work reported in earlier papers.

For example, we will address the analysis of both lognormal and contaminated lognormal exposure profiles, examine the effect of LODs in ranges of 1–50% and 50–80% of the underlying distribution and cover sample sizes ranging from 5–100. Furthermore, we will contrast and compare the most common examples from all four families of CDA methods, nearly all of which have been addressed by one or more previous papers. However, we feel these repeats are justified as our results provide interesting and sometimes surprising comparisons between the various methods.

We purposefully did not address confidence intervals in this study. Neither did the majority of the previous investigators. Some authors have examined and promote the use of the bootstrap or jackknife methods for devising confidence intervals (Shumway *et al*., 2002; Helsel, 2005). Helsel (2005) devotes a chapter to calculating parameter confidence intervals for censored datasets and discusses the various alternatives.

Last, we suspected at the outset that it was unlikely that a single method would always perform better than the others, regardless of the exposure scenario, which indeed proved true. Therefore, our goal is to inform the reader regarding the ability of each of the selected methods to extract unbiased and precise estimates of the usual parameters from datasets where the underlying distribution is either lognormal or contaminated lognormal, as well as from both simple- and complex-censored datasets. To accomplish this goal, we devised several exposure scenarios with the expectation that the reader will select the scenarios that best describe their experience and focus on the methods that perform best under these conditions.

CENSORED DATA ANALYSIS METHODS

Published or recommended methods for analyzing such datasets tend to fall into four categories or families:

- substitution methods,
- LPR methods,
- MLE methods, and
- NP methods.

Substitution methods

The three common substitution methods are LOD, LOD/2 and LOD/$\sqrt{2}$ substitution, although the choice of the substitution fraction is largely arbitrary. Substitution of each 'less than value' with the LOD continues to occur, despite the numerous recommendations against, with the invariable justification that such a practice is conservative (LOD substitution tends to result in a 'conservative' or positive bias for the mean and a negative bias for variability. The reduction in variability tends to result in a negative bias for the 95th percentile (which is calculated from the sample GM and sample GSD)). LOD/2 substitution appears to be the CDA method of choice in the epidemiological literature whenever large, complex-censored datasets are used to construct a job-exposure matrix (Hornung and Reed, 1990; Glass and Gray, 2001). When estimating the true geometric mean (GM) and GSD, Hornung and Reed (1990) recommended using LOD$\big/\sqrt{2}$ substitution whenever it was suspected that the underlying GSD is <3, and LOD/2 substitution otherwise. But when estimating the mean, they recommended the LOD/2 method provided the percent censoring was $<50\%$. Their paper is frequently referenced to justify using the LOD/2 substitution method.

All of the substitution methods are biased, and this bias will be a function of the true GSD, the true percent censored and the sample size. As the sample size increases, the bias asymptotically approaches a fixed value. El-Shaarawi and Esterby (1992) derived formulae for directly calculating the large sample bias for the mean when using substitution methods, given known values for the GM, GSD and the true percent censored. But since their formulae cannot be used to determine the bias for small sample sizes, bias when estimating upper percentiles, or the overall accuracy (bias plus precision) of the estimates and cannot be applied to complex-censored or contaminated lognormal distributions, we included the substitution methods in our simulations.

LPR methods

In occupational health, the LPR method has long been recommended (Hawkins *et al*., 1991; Mulhausen and Damiano, 1998) for analyzing censored data. The data, including the LODs, are sorted and plotted using log-probability plotting paper. This method, the LPR method, is based on following relationship, which is derived from the *Z*-value equation:

$$y_i = \hat{\mu}_y + \hat{\sigma}_y \cdot \Phi^{-1}(p_i),$$

where $y_i = \ln(x_i)$ and $\Phi^{-1}(p_i)$ refers to the inverse cumulative normal distribution for plotting position p_i. This is a linear equation which is solved for the non-LOD pairs of y_i and $\Phi^{-1}(p_i)$. The sample GM and sample GSD are estimated using the exponential of the intercept and slope, respectively. Blom's formula is typically used for calculating the *i*th plotting position: $p_i = (i - 3/8)/(n + 1/4)$.

A variation on this method is the robust LPR (LPR_r) method, which is felt to be less susceptible to departures from the lognormal assumption (Kroll and Stedinger, 1996) and avoids transformation bias [transformation bias refers to the bias in the estimate of the mean that results when the mean (on the concentration scale) is calculated from the sample GM and sample GSD (which are estimated using

the log-transformed data); the minimum variance unbiased estimator equation (Mulhausen and Damiano, 1998) is typically used with complete (i.e. uncensored) samples to reduce this bias] since the mean can be estimated using the simple arithmetic mean. With LPR_r, the missing values, represented by the non-detects, are predicted using the 'initial' values of the sample GM and sample GSD determined using LPR. The 'final' sample estimates are calculated by combining the predicted values and the detects and analyzing the dataset in the conventional fashion for the sample GM and GSD (and from these calculating the sample 95th percentile) and for the sample arithmetic mean (thus avoiding transformation bias).

In principle, neither of the LPR and LPR_r methods should be applied to complex-censored datasets. Helsel and Cohn (1988) devised an *ad hoc* method that can be applied to a complex-censored dataset by evenly spreading the LODs throughout the lower portion of the dataset. This method, which we will call the robust, multiple LOD LPR method (LPR_{rm}), was recommended by Helsel and Cohn (1988) whenever the lognormal distribution assumption is in doubt. The methodology is complex, so the reader is referred to the original articles. Helsel (2005) provided a worked example of the LPR_{rm} method, which he referred to as the 'robust regression on order statistics' method.

MLE methods

The MLE method is often considered the gold standard provided the data are well described by a lognormal distribution. The sample GM and GSD are those values that maximize the likelihood function:

$$\mathrm{LF} = \prod_{i=k+1}^{n} pdf(\ln x_i \mid \ln \mathrm{GM}, \ln \mathrm{GSD}) \cdot \prod_{j=1}^{k} cdf(\ln x_j \mid \ln \mathrm{GM}, \ln \mathrm{GSD}),$$

where n = sample size (including both censored and uncensored data), k = number of censored data, pdf refers to the probability density function and cdf refers to the cumulative distribution function. Since there is no close-form solution to this equation, Finkelstein and Verma (2001) recommended using the solver-function available in most spreadsheets to find the optimal solution. They provided an easy to follow example that can be extended to virtually any sample size and degree of censoring.

If the underlying distribution is felt to depart significantly from the lognormal distribution assumption, Kroll and Stedinger (1996) recommended using robust maximum likelihood estimation (MLE_r) where the MLE method is used to derive the initial estimates of the sample GM and GSD. As with LPR_r, the missing values (i.e. the censored exposures) are predicted using the initial sample GM and GSD and then combined with the detects. The final sample GM and GSD, as well as the simple arithmetic mean (thus avoiding transformation bias), are calculated in the conventional manner using the combined dataset.

While in principle either MLE or MLE_r can be applied to complex-censored datasets, we decided to create an additional variation (MLE_{rm}) which is identical to LPR_{rm}, except that the MLE method, rather than the LPR method, is used to generate the initial estimates of the sample GM and GSD. Otherwise, all other calculations are the same as those used for the LPR_{rm} method (see the citations for LPR_{rm}, such as Helsel, 2005).

The last variation on the MLE method is that devised by Succop *et al.* (2004). The MLE method is used to derive initial estimates of the sample GM and GSD. These estimates are used to estimate the cdf for each unique LOD in the dataset. Each non-detect is then replaced with what the Succop *et al.* called 'the most probable value' concentration, which corresponds to the predicted concentration at half of the cdf value for the LOD (for example, if an LOD is situated at the 10th percentile for the lognormal exposure profile predicted using the MLE method, each non-detect is simply replaced with the concentration predicted to occur at the 5th percentile). The mean of new dataset, which consists of the uncensored and the most probable values, is then estimated using the standard arithmetic mean formula. The authors did not estimate, via computer simulation, the bias or accuracy for this method, but after comparing these most probable values to laboratory values (where the laboratory was persuaded to provide measurements below the LOD), the authors concluded that this method is preferable to the simple substitution methods. We implemented the authors' method (referred to here as MLE_{mpv}) and after replacing all LOD values with the most probable values estimated the GM, GSD, 95th percentile and mean using the standard statistical formulae. We should note, however, that Succop *et al.* never claimed that their method could be applied to accurately estimating any parameter other than the mean.

NP methods

KM quantile and mean. Schmoyer *et al.* (1996) and She (1997) recommended applying the KM method, originally intended for application to the right-censored data that occur in prospective epidemiological studies and in clinical trials, to the left-censored concentration data encountered in environmental studies. The KM method is basically a NP method based on the empirical cdf [for uncensored datasets, the KM method produces quantiles that are identical to those produced by the empirical cdf method for all sample sizes except those where *np*

(where n = sample size and p = proportion: for example, $p = 0.95$ for the 95th percentile) equals an integer; in these cases, the empirical cdf method assigns the proportion to the ranked value for that integer, while the KM method assigns the proportion to the next higher ranked value]. Its main advantage is the ability to estimate the mean in the presence of non-detects, without relying upon a distributional assumption. The KM method is available in many statistics programs, but because it was originally intended for right-censored datasets, the exposure data must be 'flipped' before analysis (i.e. the left-censored dataset must be converted to a right-censored dataset). Helsel (2005) provided an example of the calculations and recommended it in preference to 'all other methods' whenever the (observed) percent censored is <50%. To implement this method, we wrote computer code that does not require flipping of the data. KM can also be used to estimate the median and other quantiles. For the 95th percentile, the minimum sample size should be 20.

Quantiles. NP methods for estimating quantiles (i.e. NP percentiles) do not require an assumption regarding the shape of the underlying distribution and are therefore considered to be robust to departures from the lognormal or any other distributional assumption. For the 95th quantile, that is, the NP sample 95th percentile, at least 19 or 20 measurements are required, depending upon the calculation method, and an observed percent censored that is no more than roughly 90% (depending upon the sample size). Hyndman and Fan (1996) present several methods, which they numbered Q1 through Q8, for calculating NP quantiles (see Appendix 1). The Systat (2007) statistical package default quantile method is 'Cleveland's method' (i.e. the Q5 method in Appendix 1). The SAS (2006) statistical package default method is the 'Empirical CDF with averaging' method (i.e. the Q2 method). The EPA recently recommended using the Q7 method. For these simulations, we selected the Q6 method as it was recommended by Gilbert (1987) and Helsel and Hirsch (2002):

- sort the data from low to high,
- calculate i = integer portion of $0.95(n + 1)$, and
- estimate the 95th percentile: $\hat{X}_{0.95} = x_i + (0.95(n+1) - i)(x_{i+1} - x_i)$.

Recommendations in the literature

Gilliom and Helsel (1986) applied several CDA methods—including LOD, LOD/2, LPR and MLE—to the analysis of randomly generated datasets of size 10, 25 and 50, drawn from lognormal, 'contaminated lognormal' (i.e. a mixture of two specific lognormal distributions), gamma and delta distributions. The datasets were censored at a single LOD set at 20%, 40%, 60% and 80% of the underlying distribution. Simulations were done at four levels of variability (i.e. four coefficients of variation) for each of the four types of distributions. For each of the 16 scenarios, 500 random datasets were generated. The rMSE summary statistics were calculated across all 16 simulations, making the results difficult to interpret for any one method and scenario, but the authors concluded that the LPR and MLE methods were, across all the distributions, the preferred methods for estimating the mean and 'median and interquartile range', respectively.

Helsel and Cohen (1988) extended the work of Gilliom and Helsel (1986) by considering multiple LODs and adding the LPR_{rm} method to the methods tested. They looked at only one sample size ($n = 25$) and three LODs, set at 20%, 40% and 80% of the underlying distribution. Roughly one-third of the measurements were assigned to each LOD. (If a randomly generated measurement was less than the assigned LOD, the measurement was truncated at the LOD.) Otherwise, their procedures were identical to those of Gilliom and Helsel (1986). The authors used the rMSE as the primary metric for comparison and concluded that the LPR_{rm} method is superior to all others when the underlying distribution departs from the lognormal distribution assumption.

Kroll and Stedinger (1996) compared the LPR, LPR_r, MLE and MLE_r methods when analyzing censored datasets drawn from the lognormal, contaminated lognormal [using the same contaminated lognormal distributions used by Gilliom and Helsel (1986) and Helsel and Cohen (1988)], gamma, delta and other distributions. They generated concentration datasets for samples sizes of 10, 25 and 50 and single LODs set at 10%, 20%, 40%, 60% and 80% of the underlying distribution. Because they summarized the rMSE results across all the distributions and their variations, as well as calculated ratios comparing the various methods, their results are difficult to interpret and to compare to the results of others. The authors used the rMSE as the metric for comparison and concluded that for censoring of 60% or less, the methods in general produced similar rMSE values, but that the MLE method was superior for the 80% censored scenarios. The robust methods—LPR_r and MLE_r—performed better when estimating the mean, with the MLE_r performing slightly better than the LPR_r method. However, for low to moderate censoring, the authors recommended the LPR and LPR_r methods over the MLE and MLE_r methods on the basis that they are easier to understand and implement.

Schmoyer *et al*. (1996) compared the KM method to the MLE method when doing hypothesis tests on the mean for datasets of size $n = 10$ drawn from lognormal, truncated normal and gamma distributions where the percent censored was roughly 25%, 50% and 75% of the underlying distribution. For each combination they generated 500 datasets. They presented their results in terms of power curves for a test

on the mean exposure. While they allowed that the interpretation of the power curves was subjective, they concluded that 'in general, the (KM) test seems better' than the MLE method. She (1997) compared the KM method to the LOD/2 substitution, LPR and MLE methods. She generated 1000 datasets of size $n = 21$ where each dataset had three censoring points randomly assigned from 10% to 80% of the underlying distribution (in increments of 10%). The bias and rMSE results varied, with the LOD/2 substitution occasionally performing better than the KM method. However, She rejected the LOD/2 as a valid method based on the perception that it 'has no statistical theoretical basis'. She concluded that the KM method 'performs as well as or better than' the MLE, LPR or LOD/2 substitution methods, making it an 'attractive alternative ... because it is non-parametric and quite robust when the distribution departs from normality (for the log-transformed data)'.

Shumway *et al*. (2002) compared the LPR_r and MLE methods when estimating the mean and variance from censored datasets drawn from the lognormal distribution. They looked at sample sizes of 20 and 50, with LODs set at 50% and 80% of the underlying distribution. They concluded that neither method was consistently better and that the choice depends upon the percent censored and departures from lognormality. They warned against combining different datasets as it would increase the probability that the data will come from a 'mixture of distributions', recommending instead the 'grouping of data into similar subsets'.

There are other articles, but the above appeared to be the most relevant. Textbooks on the statistical analysis of environmental data offer differing advice. Gibbons and Goleman (2001) reviewed the literature and concluded that the MLE method is the 'best overall estimator'. In what may be the only published textbook devoted to the topic of CDA, Helsel (2005) reviewed the literature and offered the following recommendations:

- for <50% censored use the KM method (for all sample sizes),
- for 50–80% censored use MLE_r or LPR_{rm} for sample sizes <50, and MLE for sample sizes >50, and
- for >80% censored report the NP exceedance fraction for the limit whenever the sample size is <50 and the NP upper percentiles (e.g. 90th or 95th) for sample sizes >50.

Helsel (2005) justified the universal application of the KM method for low to moderately censored datasets on the strength of the She (1997) and Shumway *et al*. (2002) papers, and the opinion that the KM method, since it does not require any distributional assumption, is robust to all situations where the true exposure profile departs significantly from the lognormal distribution assumption.

US agencies and organizations have published several monographs on the analysis of environmental data. The Environmental Protection Agency (EPA, 2006) offered the following general recommendations:

- if the percent censored is <15%, use substitution with zero, LOD/2, or the LOD, or use the MLE method,
- for 15–50% censored, use the MLE method, and
- for 50–90% censored, calculate the NP exceedance fraction for the limit.

The US Geological Survey agency (Helsel and Hirsch, 2002) published a guide to statistical methods in which the substitution methods were recognized to have good overall accuracy (i.e. low rMSE), but were not recommended because they tend to be biased and have no theoretical foundation. The MLE, MLE_r or LPR_r methods were recommended, the last two in particular when the lognormal distribution assumption is in doubt. The LPR_{rm} method was recommended whenever there are multiple LODs in the dataset. Frome and Wambach (2005) of Oak Ridge National Laboratory published an overview of the statistical methods that can be applied to censored datasets. They recommended the MLE method for general use and the KM method whenever the lognormal distribution assumption is in doubt.

In summary, it would appear that no one method has been recommended for all instances of sample size, degree of conformity to the lognormal distribution assumption and the degree of censoring. It also appears that the published computer simulations more often than not were limited in terms of the number of methods compared, the simulation sizes, the sample sizes selected and/or ranges for the percent censored. Our computer simulations are more comprehensive in regards to all of these issues and hopefully can be generalized to a wider range of actual scenarios.

COMPUTER SIMULATION METHODS

To estimate the bias and rMSE, we developed a computer program to generate and analyse censored datasets using the following CDA methods:

- Substitution: LOD, LOD/2 and LOD/$\sqrt{2}$
- Log-probit regression: LPR, LPR_r, LPR_{rm}.
- Maximum likelihood estimation: MLE, MLE_r, MLE_{rm} and MLE_{mpv}.
- Non-parametric methods: NP and KM.

For each artificial dataset, the estimates of the 95th percentile [most of the CDA methods lead to a sample GM and GSD, from which the sample 95th percentile can be estimated using the standard equation: $\hat{X}_{0.95} = \exp(\ln(\mathrm{GM}) + 1.645 \cdot \ln(\mathrm{GSD}))$] and mean

were compared to the true values. After the generation and analysis of 100 000 artificial-censored datasets, the average bias and rMSE (an estimate of the overall accuracy, discussed later) were calculated for each parameter.

To compare the methods, we devised the following three simulations (Table 1):

- Simulation 1: *n* ranged between 20 and 100, the true percent censored ranged between 1% and 50% and the true GSD ranged between 1.2 and 4,
- Simulation 2: *n* ranged between 20 and 100, the true percent censored ranged between 50% and 80% and the true GSD ranged between 1.2 and 4, and
- Simulation 3: *n* ranged between 5 and 19, the true percent censored ranged between 1% and 50% and the true GSD ranged between 1.2 and 4.

For each of these simulations, we devised the following four scenarios:

- Scenario I: a single lognormal distribution and a single LOD,
- Scenario II: a single lognormal distribution and three LODs,
- Scenario III: a contaminated lognormal distribution and a single LOD, and
- Scenario IV: a contaminated lognormal distribution and three LODs.

While there are other permutations that we could have devised (and did, as discussed later), we felt that the simulations and scenarios above represented a thorough testing of the various methods for analyzing censored data. Readers should be able to identify a familiar scenario and then determine the best method or methods.

For Simulation 1, we generated 100 000 artificial datasets from censored lognormal or censored contaminated lognormal distributions. The sample size for each dataset was randomly varied (using the uniform distribution) between 20 and 100 (inclusive). The percentage of the distribution that was censored was also randomly varied (using the uniform distribution), between 1% and 50% (inclusive). The laboratory LOD was then set at the concentration in the distribution corresponding to the percent censored.

Simulation 1 was repeated for each of four scenarios (see Table 1) and for each of the CDA methods. In Scenario I, a single lognormal distribution was assumed as well as a single laboratory. The GM was fixed at 1, while the GSD for the distribution was randomly varied between 1.2 and 4 (inclusive) using the uniform distribution. In Scenario II, a single lognormal distribution was also assumed, as in Scenario I, but three laboratories with different LODs were assumed to have been used, one each for approximately one-third of the samples. The LOD for each laboratory was randomly generated as described above. In Scenario III, a single laboratory was used; however, the underlying distribution was contaminated. A contaminated (i.e. non-lognormal) distribution was created by combining two lognormal distributions. The GM and GSD for each distribution were randomly generated from uniform distributions where the minimum and maximum values were 1 and 3, and 1.2 and 4, respectively. The fraction that the first distribution contributed to the overall distribution was also randomly varied by generating a fraction from a uniform distribution where the minimum and maximum were 0 and 1. (The fraction contributed by the second distribution was one minus this value.) In Scenario IV, a contaminated distribution was again

Table 1. Parameters used in Simulation 1

Simulation parameter	Scenario I	Scenario II	Scenario III	Scenario IV
	min–max	min–max	min–max	min–max
Sample size	20–100	20–100	20–100	20–100
Exposure profile distributions				
GM1[a]	1	1	1–3	1–3
GSD1	1.2–4	1.2–4	1.2–4	1.2–4
GM2	—	—	1–3	1–3
GSD2	—	—	1.2–4	1.2–4
$Distribution_1$ %[b]	100%	100%	0–100%	0–100%
Laboratory LOD as % of the exposure profile[c]				
$Laboratory_1$ LOD %	1–50%	1–50%	1–50%	1–50%
$Laboratory_2$ LOD %	—	1–50%	—	1–50%
$Laboratory_3$ LOD %	—	1–50%	—	1–50%

Simulations 2 and 3 were identical to Simulation 1 except that for Simulation 2 the LOD as a percent of the exposure profile ranged between 50% and 80% and for Simulation 3 the sample size ranged between 5 and 19.
[a]Since only a single distribution is used in Scenarios I and II, the GM was fixed at 1.
[b]$Distribution_2$ % will be 1—the percentage for $Distribution_1$.
[c]If three laboratories were used each was used ~1/3 of the time.

used, as just described, but with three laboratories and three LODs, as was described for Scenario II. Simulation 2 was identical to Simulation 1 above, except that the percentage censored varied between 50% and 80% (inclusive). Simulation 3 was also identical to Simulation 1, except that the sample sizes were allowed to vary between 5 and 19 (inclusive), rather than 20 and 100.

For each randomly generated dataset, the computer program did the following:

- determined if the dataset was invalid,
- determined if the dataset was completely uncensored,
- applied standard statistical methods to each valid, uncensored dataset, and
- applied the selected CDA method to each valid, censored dataset.

A dataset was invalid if all n measurements were censored or there were too few uncensored data. For the LPR-based and MLE-based methods, a valid dataset was one with at least three measurements and at least two of those must be uncensored. For the substitution methods, a valid dataset must have at least two measurements and at least one must be uncensored. The fraction of invalid datasets for each method was tracked by the program [for Simulations 1 and 2, the fraction of invalid datasets was typically $<0.1\%$; for Simulation 3, the typical fraction was $<0.5\%$ (the smaller sample sizes increased the likelihood of an invalid or completely censored dataset)]. If all n measurements were uncensored, the dataset was statistically analysed by the program using the standard statistical methods for estimating the GM, GSD, 95th percentile and mean (see Mulhausen and Damiano, 1998). The selected CDA method was applied whenever the dataset was valid and there was at least one LOD value. After the sample GM and GSD were calculated, the sample 95th percentile was calculated using the standard equation [most of the CDA methods lead to a sample GM and GSD, from which the sample 95th percentile can be estimated using the standard equation: $\hat{X}_{0.95} = \exp(\ln(\mathrm{GM}) + 1.645 \cdot \ln(\mathrm{GSD}))$]. For the LPR and MLE methods, the mean was estimated using the minimum variance unbiased estimator (mvue) equation (Mulhausen and Damiano, 1998), using $n - k$ and n as the sample size, respectively. (The mvue equation was designed to minimize the transformation bias that occurs when moving from the log scale to the concentration scale. Preliminary simulations suggested that using the full sample size (n) in the mvue equation was appropriate for the MLE-based methods, while for the LPR-based methods, a reduced sample size ($n - k$) results in less bias when used in the mvue equation.) For the substitution methods, the robust MLE or LPR methods and the $\mathrm{MLE}_{\mathrm{mpv}}$ method, the mean was estimated using the standard simple arithmetic mean formula. For the KM method, the mean was estimated without variation from the procedure outlined in Helsel (2005).

However, we estimated the mean regardless of the actual fraction of censored data [Helsel (2005) recommended that the KM method should not be used to estimate the mean whenever the dataset is $>50\%$ censored]. The KM and NP methods were used to estimate the 95th percentile in those simulations where the sample size was 20 or greater.

Once the sample 95th percentile and mean were calculated, the differences between the sample estimates and the true values were determined. After all 100 000 datasets were generated, the program calculated the average bias for each of the parameters across all 100 000 datasets:

$$\mathrm{Bias} = (\bar{x} - \theta),$$

where $\bar{x}$ is the mean of the 100 000 parameter estimates and θ is the true value. The program also calculated the rMSE, which is a combination of bias and precision:

$$rMSE = \sqrt{(\bar{x} - \theta)^2 + \frac{\sum(x - \bar{x})^2}{N - 1}}$$

where $N = 100\ 000$. The rMSE value can be considered a measure of overall imprecision or overall accuracy. For example, for a particular parameter, a proposed method may be biased but have lower variability. If the resulting rMSE is comparable to a gold standard method, the proposed method could be considered suitably accurate.

The bias and rMSE can be evaluated on the log scale (e.g. bias relative to $\ln X_{0.95}$) or the concentration scale (e.g. bias relative to $X_{0.95}$). Consistent with the overwhelming majority of other investigators, we chose to examine the bias and rMSE on the concentration scale, or relative to the true 95$^{\mathrm{th}}$ percentile and mean.

RESULTS AND DISCUSSION

The results for the Simulation and Scenario combinations are listed in Tables 2 through 13. For the bias comparison metric, the methods are ranked in order of the absolute value of the bias. For the rMSE comparison metric, the methods are ranked from low to high. The bias and rMSE are given in terms of the percent of the true value. Given that we used $N = 100\ 000$ for all simulations, the estimates of bias and rMSE should be fairly stable. Repeating the same simulation will generally produce a result that is within ± 0.2 of the table values. In contrast, when we reduce the simulation size to the $N = 500$ or $N = 1000$, as was used in most of the published studies on CDA, we observed that the bias and rMSE values frequently vary from the table values by plus or

Table 2. Simulation 1, Scenario I—single lognormal distribution and a single laboratory where the laboratory LOD is 1–50% of the true distribution; $20 \leq n \leq 100$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE_{rm}	−0.4	MLE_{mpv}	22.4	MLE_{rm}	0.0	MLE	17.9
LPR_{r}	0.4	Sub LOD$/\sqrt{2}$	22.6	MLE_{r}	0.1	LPR	18.4
Sub LOD/2	0.5	MLE_{rm}	23.0	LPR_{r}	0.1	MLE_{r}	19.5
MLE_{r}	0.5	Sub LOD/2	23.1	MLE_{mpv}	−0.1	MLE_{rm}	19.6
MLE	0.6	MLE_{r}	23.2	LPR_{rm}	−0.2	LPR_{r}	19.6
LPR_{rm}	1.4	MLE	23.3	LPR	0.5	LPR_{rm}	19.8
LPR	2.5	LPR_{r}	23.8	MLE	−0.5	MLE_{mpv}	19.8
KM	2.8	LPR_{rm}	24.1	Sub LOD$/\sqrt{2}$	0.6	Sub LOD$/\sqrt{2}$	19.9
MLE_{mpv}	−6.0	Sub LOD	24.2	Sub LOD/2	−1.8	Sub LOD/2	20.4
Sub LOD$/\sqrt{2}$	−7.8	LPR	24.9	Sub LOD	4.2	KM	20.5
Sub LOD	−13.5	KM	35.8	KM	4.2	Sub LOD	20.5
NP	15.2	NP	50.6				

Table 3. Simulation 1, Scenario II—single lognormal distribution and three laboratories where the LOD for each laboratory fell in the range of 1–50% of the true distribution; $20 \leq n \leq 100$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE	0.4	Sub LOD$/\sqrt{2}$	21.9	LPR_{rm}	0.0	LPR	16.5
Sub LOD/2	0.7	LPR_{r}	22.0	LPR	0.0	MLE	17.9
LPR_{rm}	−1.4	MLE_{rm}	22.1	MLE_{rm}	0.2	LPR_{rm}	19.6
MLE_{rm}	−2.3	MLE_{r}	22.3	MLE_{mpv}	−0.3	MLE_{rm}	19.6
MLE_{r}	−2.5	LPR	22.4	Sub LOD$/\sqrt{2}$	0.6	LPR_{r}	19.8
KM	2.9	Sub LOD/2	22.4	MLE	−0.7	KM	19.8
MLE_{mpv}	−4.1	LPR_{rm}	22.5	KM	1.4	MLE_{r}	19.9
LPR	−6.5	Sub LOD	22.6	MLE_{r}	1.4	MLE_{mpv}	19.9
LPR_{r}	−7.1	MLE	23.1	Sub LOD/2	−1.8	Sub LOD$/\sqrt{2}$	20.0
Sub LOD$/\sqrt{2}$	−7.4	MLE_{mpv}	23.5	LPR_{r}	3.2	Sub LOD	20.1
Sub LOD	−12.1	KM	34.2	Sub LOD	4.2	Sub LOD/2	20.1
NP	15.0	NP	52.1				

Table 4. Simulation 1, Scenario III—a contaminated lognormal distribution and a single laboratory where the laboratory LOD is 1–50% of the true distribution; $20 \leq n \leq 100$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
LPR_{r}	0.2	MLE_{mpv}	24.1	MLE_{rm}	−0.1	MLE	18.2
MLE_{r}	−0.3	Sub LOD$/\sqrt{2}$	24.4	MLE_{mpv}	−0.1	LPR	18.7
MLE	−0.5	Sub LOD/2	24.4	LPR_{r}	−0.1	MLE_{rm}	21.4
MLE_{rm}	−1.2	MLE_{rm}	24.7	MLE_{r}	−0.2	MLE_{mpv}	21.5
LPR_{rm}	1.3	MLE	24.9	LPR_{rm}	−0.4	Sub LOD$/\sqrt{2}$	21.8
Sub LOD/2	−1.6	MLE_{r}	25.2	Sub LOD$/\sqrt{2}$	1.1	MLE_{r}	21.8
LPR	1.8	LPR_{r}	25.7	Sub LOD/2	−1.4	Sub LOD/2	21.9
KM	4.6	Sub LOD	26.0	LPR	−1.5	LPR_{rm}	21.9
MLE_{mpv}	−7.2	LPR_{rm}	26.5	MLE	−2.3	LPR_{r}	22.6
Sub LOD$/\sqrt{2}$	−9.0	LPR	26.9	Sub LOD	4.3	KM	22.7
Sub LOD	−14.4	KM	39.9	KM	4.4	Sub LOD	22.9
NP	19.7	NP	64.4				

Table 5. Simulation 1, Scenario IV—a contaminated lognormal distribution and three laboratories where the LOD for each laboratory fell in the range of 1–50% of the true distribution; $20 \leq n \leq 100$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE	−0.7	Sub LOD/$\sqrt{2}$	23.8	MLE_{rm}	0.0	LPR	16.8
Sub LOD/2	−1.3	MLE_{rm}	24.1	LPR_{rm}	−0.2	MLE	18.1
LPR_{rm}	−2.2	Sub LOD/2	24.2	MLE_{mpv}	−0.4	LPR_{rm}	21.4
MLE_{r}	−3.2	LPR	24.3	Sub LOD/$\sqrt{2}$	1.0	LPR_{r}	21.5
MLE_{rm}	−3.3	LPR_{r}	24.3	Sub LOD/2	−1.3	MLE_{r}	21.6
KM	4.4	MLE_{r}	24.3	MLE_{r}	1.4	MLE_{mpv}	21.8
MLE_{mpv}	−5.4	LPR_{rm}	24.6	KM	1.6	KM	21.9
LPR_{r}	−7.9	MLE	24.7	LPR	−1.9	Sub LOD/$\sqrt{2}$	21.9
LPR	−7.9	Sub LOD	24.8	MLE	−2.4	MLE_{rm}	21.9
Sub LOD/$\sqrt{2}$	−8.7	MLE_{mpv}	25.3	LPR_{r}	3.1	Sub LOD/2	22.3
Sub LOD	−13.3	KM	39.7	Sub LOD	4.3	Sub LOD	22.4
NP	19.5	NP	62.7				

Table 6. Simulation 2, Scenario I—single lognormal distribution and a single laboratory where the laboratory LOD is 50–80% of the true distribution; $20 \leq n \leq 100$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE	0.0	MLE_{rm}	24.2	MLE	0.0	MLE	18.9
MLE_{r}	0.3	MLE	24.8	LPR	0.0	LPR	19.5
LPR_{r}	0.9	MLE_{r}	25.2	MLE_{rm}	0.3	MLE_{r}	19.8
LPR_{rm}	1.7	LPR_{r}	25.8	MLE_{r}	0.6	LPR_{r}	20.0
MLE_{rm}	−1.8	LPR_{rm}	26.4	Sub LOD/2	−0.7	MLE_{mpv}	20.0
KM	2.5	Sub LOD	26.6	MLE_{mpv}	−1.4	MLE_{rm}	20.1
LPR	3.5	Sub LOD/$\sqrt{2}$	27.7	LPR_{rm}	−1.5	LPR_{rm}	20.4
NP	14.9	LPR	28.0	LPR_{r}	1.6	Sub LOD/2	21.7
Sub LOD/2	−19.8	Sub LOD/2	28.3	Sub LOD/$\sqrt{2}$	12.0	Sub LOD/$\sqrt{2}$	24.5
Sub LOD	−20.7	MLE_{mpv}	29.6	KM	30.2	KM	38.2
Sub LOD/$\sqrt{2}$	−21.3	KM	33.5	Sub LOD	30.3	Sub LOD	38.3
MLE_{mpv}	−21.6	NP	49.8				

Table 7. Simulation 2, Scenario II—single lognormal distribution and three laboratories where the LOD for each laboratory fell in the range of 50–80% of the true distribution; $20 \leq n \leq 100$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE	0.2	MLE_{rm}	22.9	MLE	0.0	MLE	19.0
KM	2.9	LPR_{r}	22.9	MLE_{rm}	0.1	MLE_{rm}	19.6
LPR_{rm}	−3.1	LPR_{rm}	24.1	LPR_{rm}	−0.6	LPR_{rm}	19.8
LPR	−4.2	Sub LOD	24.3	Sub LOD/2	−0.8	MLE_{r}	19.9
LPR_{r}	−4.6	MLE	24.9	MLE_{mpv}	−1.7	MLE_{mpv}	20.6
MLE_{r}	4.8	Sub LOD/$\sqrt{2}$	26.3	MLE_{r}	3.4	Sub LOD/2	21.4
MLE_{rm}	−5.4	LPR	26.6	Sub LOD/$\sqrt{2}$	12.1	Sub LOD/$\sqrt{2}$	23.9
NP	15.0	MLE_{r}	26.7	LPR	17.9	LPR	25.8
Sub LOD	−17.8	Sub LOD/2	27.2	KM	19.8	KM	28.1
Sub LOD/2	−19.0	MLE_{mpv}	29.5	LPR_{r}	21.9	LPR_{r}	30.6
Sub LOD/$\sqrt{2}$	−19.8	KM	34.1	Sub LOD	30.5	Sub LOD	37.2
MLE_{mpv}	−21.8	NP	51.2				

Table 8. Simulation 2, Scenario III—a contaminated lognormal distribution and a single laboratory where the laboratory LOD is 50–80% of the true distribution; $20 \leq n \leq 100$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE	0.9	MLE_{rm}	26.7	Sub LOD/2	−0.2	MLE	20.1
MLE_{r}	1.0	MLE	27.7	LPR_{r}	−0.9	LPR	21.0
MLE_{rm}	−1.2	MLE_{r}	27.9	MLE_{r}	−1.0	MLE_{rm}	21.6
LPR_{r}	3.0	Sub LOD	28.2	MLE_{rm}	−1.1	MLE_{r}	21.6
LPR_{rm}	3.8	Sub LOD$/\sqrt{2}$	29.1	MLE	−2.3	LPR_{r}	21.8
KM	4.5	LPR_{r}	29.5	LPR	−3.2	MLE_{mpv}	21.8
LPR	5.4	Sub LOD/2	29.7	MLE_{mpv}	−3.2	LPR_{rm}	22.3
NP	19.7	LPR_{rm}	30.3	LPR_{rm}	−3.8	Sub LOD/2	22.7
Sub LOD	−21.1	MLE_{mpv}	31.3	Sub LOD$/\sqrt{2}$	12.4	Sub LOD$/\sqrt{2}$	25.9
Sub LOD/2	−21.2	LPR	31.8	Sub LOD	29.9	Sub LOD	39.0
Sub LOD$/\sqrt{2}$	−22.1	KM	40.0	KM	30.0	KM	39.1
MLE_{mpv}	−22.7	NP	62.3				

Table 9. Simulation 2, Scenario IV—a contaminated lognormal distribution and three laboratories where the LOD for each laboratory fell in the range of 50–80% of the true distribution; $20 \leq n \leq 100$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE	1.0	MLE_{rm}	25.4	Sub LOD/2	−0.1	MLE	19.9
LPR_{rm}	−2.1	LPR_{r}	25.7	MLE_{rm}	−1.3	LPR_{rm}	21.7
LPR	−3.5	LPR	26.2	MLE_{r}	2.0	MLE_{rm}	21.8
LPR_{r}	−3.6	Sub LOD	26.2	MLE	−2.3	MLE_{r}	21.8
KM	4.6	LPR_{rm}	26.9	LPR_{rm}	−2.8	MLE_{mpv}	22.4
MLE_{rm}	−5.1	MLE	27.7	MLE_{mpv}	−3.2	Sub LOD/2	22.9
MLE_{r}	5.7	Sub LOD$/\sqrt{2}$	28.1	Sub LOD$/\sqrt{2}$	12.3	LPR	24.6
Sub LOD	−18.5	Sub LOD/2	28.8	LPR	15.2	Sub LOD$/\sqrt{2}$	25.7
NP	19.7	MLE_{r}	29.5	KM	19.8	KM	29.4
Sub LOD/2	−20.3	MLE_{mpv}	31.2	LPR_{r}	20.4	LPR_{r}	31.1
Sub LOD$/\sqrt{2}$	−20.7	KM	40.0	Sub LOD	30.0	Sub LOD	37.6
MLE_{mpv}	−22.8	NP	61.6				

Table 10. Simulation 3, Scenario I—single lognormal distribution and a single laboratory where the laboratory LOD is 1–50% of the true distribution; $5 \leq n \leq 19$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE_{mpv}	1.3	Sub LOD$/\sqrt{2}$	54.3	LPR_{rm}	0.0	MLE	38.9
Sub LOD$/\sqrt{2}$	−1.8	Sub LOD/2	58.3	MLE_{rm}	0.0	LPR	40.2
MLE	3.7	Sub LOD	60.9	MLE_{r}	0.2	LPR_{r}	41.2
MLE_{rm}	4.1	MLE_{mpv}	61.3	MLE_{mpv}	0.2	MLE_{rm}	41.8
LPR_{r}	4.4	MLE	63.4	Sub LOD$/\sqrt{2}$	0.7	MLE_{r}	42.0
MLE_{r}	5.4	MLE_{rm}	63.7	LPR	0.8	KM	42.1
Sub LOD/2	6.8	MLE_{r}	66.7	LPR_{r}	0.9	Sub LOD$/\sqrt{2}$	42.7
Sub LOD	−7.9	LPR_{r}	69.0	Sub LOD/2	−1.9	Sub LOD/2	42.9
LPR_{rm}	11.5	LPR_{rm}	94.8	MLE	−2.1	MLE_{mpv}	43.1
LPR	12.6	LPR	110.5	KM	4.0	LPR_{rm}	43.2
				Sub LOD	4.4	Sub LOD	45.6

Table 11. Simulation 3, Scenario II—single lognormal distribution and three laboratories where the LOD for each laboratory fell in the range of 1–50% of the true distribution; $5 \leq n \leq 19$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE_{rm}	0.9	Sub LOD	50.8	MLE_{rm}	0.1	MLE	38.0
MLE_{mpv}	0.8	Sub LOD$/\sqrt{2}$	56.0	MLE_{mpv}	−0.3	LPR	38.0
LPR_{r}	−1.0	MLE_{rm}	59.6	LPR_{rm}	0.4	MLE_{rm}	41.0
Sub LOD$/\sqrt{2}$	−1.2	MLE_{mpv}	60.2	Sub LOD$/\sqrt{2}$	0.8	MLE_{mpv}	41.7
MLE_{r}	2.9	Sub LOD/2	60.6	MLE_{r}	1.5	MLE_{r}	42.1
MLE	3.1	MLE	60.7	LPR	1.6	Sub LOD	42.5
LPR	3.4	LPR_{r}	62.3	KM	1.6	LPR_{rm}	42.7
LPR_{rm}	4.1	MLE_{r}	63.8	Sub LOD/2	−1.8	LPR_{r}	42.8
Sub LOD	−6.8	LPR_{rm}	72.3	MLE	−2.3	KM	44.0
Sub LOD/2	7.2	LPR	90.6	LPR_{r}	4.0	Sub LOD/2	44.9
				Sub LOD	4.2	Sub LOD$/\sqrt{2}$	45.6

Table 12. Simulation 3, Scenario III—a contaminated lognormal distribution and a single laboratory where the laboratory LOD is 1–50% of the true distribution; $5 \leq n \leq 19$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE_{mpv}	0.5	Sub LOD	56.7	MLE_{mpv}	0.0	MLE	39.8
Sub LOD$/\sqrt{2}$	−2.6	Sub LOD$/\sqrt{2}$	58.1	LPR_{rm}	−0.1	LPR	40.8
MLE	3.4	Sub LOD/2	62.8	MLE_{r}	−0.1	MLE_{rm}	45.4
MLE_{rm}	4.0	MLE_{mpv}	64.4	MLE_{rm}	−0.3	LPR_{r}	45.8
MLE_{r}	5.5	MLE	65.8	LPR_{r}	0.7	MLE_{r}	45.8
Sub LOD/2	5.7	MLE_{rm}	68.0	LPR	−0.8	MLE_{mpv}	46.5
LPR_{r}	6.1	MLE_{r}	73.1	Sub LOD$/\sqrt{2}$	1.0	Sub LOD/2	46.6
Sub LOD	−8.6	LPR_{r}	100.9	Sub LOD/2	−1.5	Sub LOD	47.5
LPR_{rm}	13.2	LPR_{rm}	103.6	MLE	−3.4	Sub LOD$/\sqrt{2}$	47.8
LPR	13.7	LPR	121.3	Sub LOD	4.3	KM	47.9
				KM	4.5	LPR_{rm}	48.0

Table 13. Simulation 3, Scenario IV—a contaminated lognormal distribution and three laboratories where the LOD for each laboratory fell in the range of 1–50% of the true distribution; $5 \leq n \leq 19$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE_{mpv}	0.3	Sub LOD	54.8	MLE_{rm}	0.2	LPR	38.5
MLE_{rm}	0.8	Sub LOD$/\sqrt{2}$	57.4	LPR	0.2	MLE	38.9
LPR_{r}	−0.8	MLE	63.7	MLE_{mpv}	−0.2	LPR_{r}	46.2
Sub LOD$/\sqrt{2}$	−2.3	Sub LOD/2	64.3	LPR_{rm}	0.7	MLE_{rm}	46.5
MLE	2.3	MLE_{mpv}	65.5	Sub LOD$/\sqrt{2}$	0.9	Sub LOD	46.5
MLE_{r}	2.9	MLE_{rm}	66.4	Sub LOD/2	−1.3	Sub LOD$/\sqrt{2}$	47.4
LPR	3.9	LPR_{r}	66.8	MLE_{r}	1.4	Sub LOD/2	47.8
LPR_{rm}	4.9	MLE_{r}	70.6	KM	1.7	MLE_{mpv}	47.9
Sub LOD/2	6.0	LPR_{rm}	84.0	LPR_{r}	3.8	MLE_{r}	47.9
Sub LOD	−7.5	LPR	89.1	MLE	−4.0	LPR_{rm}	48.3
				Sub LOD	4.2	KM	49.2

minus several percentage points. Consequently, we feel that the bias and rMSE estimates in the tables are reliable.

When comparing methods, our view is that there is little practical difference between methods having an absolute bias that differs by only 1% or so, or a rMSE that differs by only 2% or so. Furthermore, it is worth mentioning that these are composite results for a wide range of distributions and censoring points created using the simulation parameter ranges specified in Table 1. It is likely that the rankings would change if the methods were challenged with a specific distribution, a specific LOD and a specific sample size. Our view is that a composite analysis is more informative regarding the performance that one should expect in the long run from each method.

Which comparison metric should be used: the bias or rMSE? The majority of the studies used the rMSE as a basis for comparing methods, a few used only bias and several provided both. We provide both metrics for the reader to consider, but lean toward the rMSE as the more informative metric. In principle, a method with the lowest rMSE has the best combination of both bias and precision. However, past investigators have often rejected the LOD/2 substitution method, even though the method had similar or lower rMSE values than their preferred method, on the basis that it had no theoretical basis. Our view is more utilitarian. While some methods do indeed appeal to a distributional assumption, we feel that because the true underlying exposure profile will never be identical in shape to this assumed distribution that all methods are essentially *ad hoc*. Therefore, those methods that have low rMSE values and are reasonably robust to departures from the unimodal, lognormal model should be preferred, with relative ties going to the method having the lowest bias.

Looking at both the rMSE and bias results, none of the CDA methods stood out as the 'single best method'. If anything, what was clear is that nearly every method will occasionally excel at estimating the 95th percentile or mean given some particular Simulation–Scenario combination. Given that there was no obvious overall winner, questions that could be addressed by inspecting the results in Tables 2 through 13 are:

- Which family of methods or specific method is generally superior overall?
- Which family of methods or specific method is generally superior given the types of distributions typically encountered in your experience?

For example, in our experience, a typical scenario is one where a single laboratory is used each year, resulting in a single censoring point, or LOD, assuming that the measurements all have roughly the same sample volume. Therefore, we are more inclined to weight the Scenario I results over the results of the other Simulation–Scenario combinations. Other investigators may have a different experience. For example, perhaps only a single laboratory is used, but the sample sizes are typically <20, suggesting that the Simulation 3 results will be more informative. Or perhaps the datasets are always complex and the lognormal distribution assumption is always in doubt, resulting from the combination of data from disperse areas and/or time periods, in which case the Scenario IV simulations are of interest.

Substitution methods

While the substitution methods have often been condemned (She, 1997; Helsel, 2005), their use continues. Our results, as expected, show that there is good reason for not using the LOD substitution method, as it consistently ranked near the bottom of the rankings in terms of both bias and rMSE, consistently underestimating the 95th percentile and overestimating the mean. In Simulation 1 (Tables 2–5), where the maximum percent censored was 50%, the LOD/2 and $\mathrm{LOD}/\sqrt{2}$ substitution methods did surprisingly well when estimating the 95th percentile, being consistently in the top half of the rMSE rankings. Neither of the LOD/2 and $\mathrm{LOD}/\sqrt{2}$ methods did well when estimating the mean (with the $\mathrm{LOD}/\sqrt{2}$ substitution method doing slightly better than the LOD/2 method), being consistently in the bottom half of the rankings. In Simulation 2 (Tables 6–9), where the percent censored ranged between 50% and 80%, the substitution methods were consistently in the middle rankings for rMSE when estimating the 95th percentile, but were nearly always at the bottom of the rankings for bias. For the mean, the substitution methods were consistently in the bottom of the rankings for both bias and rMSE. On the other hand, in Simulation 3 (Tables 10–13), where the sample sizes ranged between 5 and 19 (inclusive), the bias was variable, leading to no general observation, except to say that all three methods performed consistently well, using the rMSE metric, when estimating the 95th percentile.

Overall, the substitution methods—LOD, LOD/2 and $\mathrm{LOD}/\sqrt{2}$ substitution—did poorly when their bias is compared to the bias of the MLE-based and LPR-based methods (discussed later) for both the 95th percentile and mean, particularly in Simulation 2 where the distributions were highly censored. Of the three, LOD substitution was, as expected, the most severely biased and should be categorically avoided. However, there were scenarios, particularly those in Simulation 3, where the rMSE of the LOD/2 or $\mathrm{LOD}/\sqrt{2}$ methods was similar to or less than that of the higher order methods (i.e. the LPR-based or MLE-based methods), suggesting that the bias inherent in the substitution methods is somewhat offset by the reduced variability (i.e. increased precision) in the estimates.

LPR-based methods

While there were exceptions, the LPR-based methods tended to be in the middle to top half of the bias and rMSE rankings for Simulations 1 and 2. The LPR-based methods appear to be fairly robust when confronted with multiple LODs and/or contaminated distributions. The LPR_{rm} method tended to have lower bias than the LPR and LPR_r methods in the multiple LOD scenarios (Scenarios II and IV) in Simulations 1 and 2. Overall, the LPR-based methods were slightly lower in the rankings than the MLE-based methods, although exceptions frequently occurred. The LPR_{rm} method, which was designed solely with the intention of estimating the mean from complex datasets and when the single lognormal distribution assumption is in doubt, consistently had low bias when estimating the mean in all three simulations. However, the simpler LPR method consistently had lower rMSE values, suggesting that it is the superior method even when confronted with multiple LODs and/or contaminated distributions.

All three LPR-based methods did poorly when estimating the 95th percentile from small datasets (Simulation 3; Tables 10–13), frequently having both large bias and rMSE values. Overall, the LPR-based methods tended to have larger rMSE values when compared to the MLE-based methods (due to the occasional very large sample GSD). When estimating the 95th percentile, there was no consistent winner among the LPR-based methods, in which case the simplest to implement—that is, the LPR method—should probably be preferred. When estimating the mean, if only the rMSE is considered, the standard LPR method, while often more biased relative to the two so-called robust LPR-based methods, almost always had the lowest rMSE, regardless of the simulation or scenario. If bias for the mean estimate is a concern, the LPR_{rm} method should be considered.

MLE-based methods

With the exception of the MLE_{mpv} method, the MLE-based methods performed well in the single-distribution scenarios and were generally fairly robust in the multiple LOD and contaminated distribution scenarios. Typically, when estimating the 95th percentile, the MLE or MLE_r methods often ranked high for the bias metric, while for the mean, the MLE_{rm} method was often ranked very high. Regarding the rMSE metric, the MLE, MLE_r and MLE_{rm} methods were usually in the top half of the rankings. In Scenarios 1 and 2, the MLE_{mpv} method consistently ranked below the other MLE methods and most of the LPR-based methods for both bias and rMSE, particularly when estimating the 95th percentile. However, in the small sample size scenario (Scenario 3, Tables 10–13), the MLE_{mpv} method performed well, consistently exhibiting low bias and rMSE for the 95th percentile and low bias for the mean.

Overall, our choice would be either of the MLE or MLE_r methods when estimating the 95th percentile and the MLE or MLE_{rm} methods when estimating the mean. Since both the MLE_r and MLE_{rm} methods require additional manipulations, we would, if confined to a single choice, select the standard MLE method. Regarding the mean, the standard MLE method almost always had the lowest rMSE, regardless of the simulation or scenario. For the 95th percentile, the MLE method consistently appeared in the top half of the rankings for both bias and rMSE, particularly for the severely censored scenarios (Simulation 2; Tables 6–9) and appeared to be surprisingly robust when confronted with contaminated distributions (Scenarios III and IV) and complex-censored datasets (Scenarios II and IV). Given that the MLE_{mpv} method generally ranked lower than the other MLE methods for bias in Simulations 1 and 2, particularly for the complex-censored scenarios, and rarely exhibited clearly superior rMSE values, we see no compelling reason for using this method, even when estimating the mean (for which is was originally intended).

NP methods

*KM method for estimating the 95*th *percentile and the mean.* With occasional exceptions, the KM method was consistently in the middle to bottom half of the bias and rMSE rankings when estimating the 95th percentile. Regarding the mean, the KM method was, along with the LOD method, the worst of all the methods whenever there was a single LOD (Scenarios I and III), consistently yielding a strong positive bias for the mean. Its performance improved whenever there were multiple LODs (Scenarios II and IV), but still it remained consistently in the bottom half of the rankings. Helsel (2005) recommended against using the KM method whenever the observed percent censored was 50% or greater on the basis that the median cannot be estimated under such circumstances (and presumably the estimation of the mean also becomes problematic). In our simulations, we chose to ignore this restriction and estimated the mean and 95th percentile for any observed percent censored of $<95\%$. Running the simulations with Helsel's restriction did not improve the situation, and in fact, increased the absolute bias.

The KM method was selected by Helsel (2005) as the method of choice for estimating the mean for all sample sizes whenever the (observed) percent censored is $<50\%$. This recommendation appeared to be based solely upon two studies that endorsed the KM method as a reasonable alternative to the standard CDA methods. In one study, the investigators (Schmoyer *et al.*, 1996) concluded that the KM

method 'seems better' than the MLE method, but their conclusion was not free of equivocation and was focused on hypothesis testing rather than parameter estimation (as is the focus of our study). In the other study, the investigator (She, 1997) found that the LOD/2 substitution method often outperformed the KM method, but recommended the KM method over the substitution method because the LOD/2 method 'has no statistical theoretical basis'.

Interestingly, as we programmed the KM method, we immediately recognized that when estimating the mean exposure the KM method is mathematically identical to the worst of the substitution methods (LOD substitution), which was demonstrated in the results (e.g. see the tables for Scenarios I and III), resulting in a strong positive bias for the mean. If the non-detects are internal, that is, bounded on both sides by detects, the KM method, in principle, has a negative bias. When a dataset has both a terminal non-detect and one or more internal non-detects, the biases introduced tend to cancel, but with no guarantee of a near zero overall bias, as we see in the tables for Scenarios II and IV. Helsel (2005) felt that in most multi-LOD situations, the overall bias 'is not large', but did not determine the degree of the bias for any particular scenario. We found that with three LODs (i.e. Scenarios II and IV), the KM method performs better than the LOD method, but still does poorly when compared to the MLE-based and LPR-based methods, even for Scenario IV (i.e. the multiple LOD and contaminated distribution scenario) where one would expect it to outperform those methods that require a distributional assumption. Regarding the 95th percentile (whenever the sample size is 20 or more), the KM method performed better than the LOD method, but only infrequently outperformed the higher order methods.

Based on these results, we see no compelling reason to recommend the KM method for estimating either the mean or the 95th percentile, even when the dataset contains multiple LODs and is suspected to contain multiple distributions (i.e. is a contaminated lognormal). Furthermore, according to She (1997) for the KM to be successfully applied the censoring must be random: '. . .the probability that the measurement of an object is censored cannot depend on the value of censored variable'. Schmoyer *et al.* (1996) recognized this essential requirement when in their computer simulations the authors assumed that the censoring point or LOD was a random variable and generated a random LOD for each random measurement. However, with occupational exposure data (and we suspect environmental data as well), the probability that a measurement will be censored does indeed depend on the true 'value' or concentration: the censoring point is relatively fixed and any true concentration below that censoring point will result in a LOD measurement. These considerations make it difficult to envision an occupational scenario where the KM method could be applied, even if it consistently performed better in the computer simulations than the other methods (which it did not).

*Quantile methods for estimating the 95*th *percentile.* In terms of both bias and rMSE, the NP quantile method that we selected for estimating the 95th percentile (i.e. the Q6 method, list in Appendix 1) was either the worst or among the worst for both bias and rMSE, even for the contaminated distribution scenarios (Scenarios III and IV). Using the parameters for Simulation 1, Scenario IV, we tested the other quantile methods for estimating the 95th percentile (see Table 14 and Appendix 1). While all had less absolute bias than the Q6 method, the rMSE values were at the bottom of the rMSE rankings when compared to the parametric CDA methods (compare to the results in Table 5). This suggests that for the sample sizes considered in Simulation 1, Scenario IV (i.e. $20 \leq n \leq 100$), the LPR- and MLE-based methods are sufficiently robust as to outperform the NP quantile methods. This begs the question, at what point will the NP methods be superior?

We increased the severity of the contaminated distributions used in the Simulation 1–Scenario IV combination by increasing the range for the two GMs from 3-fold to 10-fold, but otherwise retained the other simulation parameters in Table 1. As expected,

Table 14. Simulation using Simulation 1, Scenario IV parameters, comparing the eight 95th percentile (quantile) estimation methods presented by Hyndman and Fan (1996)

95th Percentile			
Method	Bias (%)	Method	rMSE
Q1—empirical CDF	3.3	Q4—weighted average 1	30.0
Q2—empirical CDF w/averaging	3.9	Q7—weighted average 3	30.1
Q7—weighted average 3	−4.1	Q3—closest value	32.0
Q4—weighted average 1	−4.8	Q2—empirical CDF w/averaging	37.6
Q3—closest value	−4.9	Q1—empirical CDF	38.8
Q5—Cleveland	5.8	Q5—Cleveland	39.9
Q8—median quartile	10.1	Q8—median quartile	47.6
Q6—weighted average 2	19.4	Q6—weighted average 2	63.7

the bias and rMSE values for the NP quantile methods were virtually identical to those in Table 14. For the LPR- and MLE-based methods, the bias and rMSE values were again superior, little worse than the values in Table 5. However, when we increased the sample size range from 20–100 to 100–1000, holding all other Simulation 1, Scenario IV parameters the same, the bias and rMSE values for the NP quantile methods tended to approach those of the LPR- and MLE-based methods. This suggests that the NP methods could be applied to contaminated distributions, but only for very large sample sizes.

In summary, the NP methods, while they have the advantage of being applicable to any underlying exposure profile, whether or not that profile is close to some assumed distribution function, do not perform as well as the parametric methods for right-skewed exposure profiles for sample sizes of 100 or less, even for highly contaminated distributions. We feel that the contaminated distribution scenario that we developed was a rigorous test of the robustness of the parametric methods (i.e. the LPR- and MLE-based methods): the GM's for the two distributions were allowed to vary by a factor of three (or even 10, as described above), the GSD's were allowed to range between GSDs representing very low to extreme variability, and the percentage contribution of each distribution was allowed to range between 0% and 100%. Even so, the parametric methods appear to be sufficiently robust to these departures from the single lognormal distribution model that we routinely assume for our data.

Small (<20) and large (>100) sample sizes

Most of the above discussions apply to Simulations 1 and 2. At the smaller sample sizes used in Simulation 3, which unfortunately are all too common, the rankings could lead to different recommendations. For example, the MLE_{mpv} method, which did not fare well when confronted with larger datasets, consistently did very well for both the 95th percentile and mean, with both bias and rMSE at or near the top of the rankings for all four scenarios. However, the other MLE-based methods, which appear to be more universally applicable, also did well. The LPR-based methods tended to fall in the rankings, consistently ranking at or near the bottom for the rMSE. Furthermore, if the rMSE is taken as the premier comparison metric, it has to be noted that for this set of simulations the substitution methods did very well when estimating the 95th percentile.

The above discussions apply to either small or moderate sample sizes. As discussed earlier in the discussion regarding the quantile results, we repeated the simulations, but this time increasing the sample size range from 20–100 to 100–1000, holding all other Simulation 1, Scenario IV parameters the same. For these larger sample sizes, the bias and rMSE results for the LPR- and MLE-based methods tend to converge, suggesting that method-related differences should be a minor consideration when estimating either the 95th percentile or the mean. When estimating the 95th percentile, in the contaminated distribution scenarios the NP quantile methods tended to slightly outperform the MLE-based methods in terms of bias, but not rMSE, when challenged with a contaminated distribution, suggesting that the NP methods should not in such instances be considered an automatic alternative to the higher order methods when estimating the 95th percentile for large datasets. For sample sizes <100, the LPR- and MLE-based methods are sufficiently robust to be preferable to the NP quantile methods.

Opportunities for improvement

The purpose of this paper is to present and discuss the computer simulation results and not to identify any real or imagined defects in the various CDA methods. However, since all of the methods are essentially *ad hoc* and none can be considered the acknowledged universal choice, there must be room for improvement. For example, the effect of the plotting position formula on the LPR-based methods could be reexamined, although Helsel and Cohn (1988) stated that it had little effect. The plotting position method for the LPR_{rm} method is considerably different than that used in the LPR and LPR_r methods, being based upon the Weibull plotting positions rather than formulae that assume an underlying normal distribution (for the log-transformed values) (see Helsel, 2005). We found that the LPR_{rm} method causes identical plotting positions to be assigned to adjacent single, unique non-detects. Is this correct or is this a defect? Perhaps there are superior plotting position schemes for all of the LPR-based methods. Furthermore, all of the LPR-based methods rely upon standard linear regression with its independent–dependent variable assumption. Perhaps the utility of methods that do not assume that all measurement error resides with the dependent variable—such as major axis regression or reduced major axis regression—should be examined. For both the LPR and MLE methods, we adopted a simple scheme that allowed the use of the mvue equation (Mulhausen and Damiano, 1998) to reduce the transformation bias: for the LPR method we used the number of detects as the sample size in the mvue equation and for the MLE method we used the full sample size. This scheme was based upon preliminary computer simulations and seemed to work well, particularly as the sample size increased. However, a more sophisticated scheme for adjusting the sample size for the mvue equation might improve the bias and rMSE of these methods when estimating the mean.

One of the themes of this paper is that there will be datasets where the unimodal, lognormal distribution assumption is inappropriate, but to date, there is little guidance on how to make this determination. Certainly, subjective graphical techniques—log-probability plots and histograms—are helpful, but thus far all of the objective goodness-of-fit procedures assume or require a complete or uncensored dataset. A censored data goodness-of-fit method, perhaps consisting of both a subjective graphical test and an objective statistical test, where the outcome is a decision to use a standard versus robust version of a method (e.g. MLE versus MLE_r or MLE_{rm}) might prove to be useful.

None of the methods studied here account for the residual information in a complex dataset: the LOD and the laboratory in use for each measurement (whether a detect or a non-detect). It is conceivable that a superior CDA method could be devised for complex datasets that makes use of this discarded information.

Additional simulations

There are obviously numerous simulations and scenarios that we could have devised. We have already mentioned other simulations where we increased the sample size range or increased the severity of the contaminated distributions to examine to usefulness of the NP quantile methods when confronted with contaminated distributions. In addition, we repeated Simulation 1, but this time increasing the range for the percent censored in the four Scenarios from 1–50% to 1–80%. While there were some changes in the rankings, our conclusions above remain the same.

Hornung and Reed (1990) suggested that the LOD/2 substitution method could be used in epidemiological studies where the mean exposure is of interest and the percent censored is $<50\%$. We repeated Simulation 1, Scenario I, but this time restricting the range of GSDs to 1.2–2, 2–3 and 3–4. For the low GSD range, we found that the $\mathrm{LOD}/\sqrt{2}$ method, while slightly biased, had an rMSE value only slightly greater than those of the LPR- and MLE-based methods. For the medium GSD range, we found that both methods, while again slightly biased, had rMSE values slightly greater than those of the LPR- and MLE-based methods. For the high GSD range, we found that the LOD/2 method was virtually unbiased for the mean and had an rMSE value slightly greater than the LPR- and MLE-based methods. These findings are consistent with those of Hornung and Reed (1990) and suggest that the use of a judiciously selected substitution method when estimating the mean is not unreasonable.

Effect of exposure measurement error

To our knowledge, none of the published papers on CDA have considered the potentially confounding effect of measurement error. In our own computer simulations, we assumed that the randomly generated exposures were (i) measured precisely and (ii) that the measurements were not rounded or truncated. Therefore, our conclusions, and those of the referenced studies, are strictly applicable to ideal exposure measurement systems.

A sampling and analytical method produces an estimate of the true concentration. The accuracy (referring to the combination of bias and precision) of these estimates vary with the mass collected, the analyte and the analytical method. Due to variability in the sampling pump flow rate and manufacturing variation in the sampling device (e.g. the sampling device used to obtain a respirable dust sample), the mass collected will be an estimate of the true mass per unit volume at the location sampled. An analytical method is used to estimate the true mass of the analyte collected by the sampling device. The relative variability in the analytical method increases with decreasing collected mass. The overall effect of these factors on the method's total coefficient of variation (CV_t) is demonstrated in Appendix 2 for sampling respirable dust (i.e. respirable particulate mass or RPM).

The second issue involves the rounding or truncation of the measurements. The sample volume is typically rounded to two or three significant figures. More importantly, the analyte mass for samples near the laboratory LOD is reported using one significant digit. As the detected mass increases, the laboratory result will generally have two or three significant digits. Furthermore, instead of rounding, a laboratory may use truncation to obtain the necessary number of significant digits. All of this adds additional uncertainty to the reported mass detected per sample and the eventual calculated concentration.

We repeated Scenario I (see Table 1) for Simulations 1, 2 and 3 for the sampling of RPM and taking into account measurement error (but ignoring for now the potential effect of rounding or truncation). We assumed that each filter is blanked corrected using a separate blank per sample. Other assumptions regarding filter weighing precision and the variability associated with the flow rate and the sampling device are described in Appendix 2. In all of the previous Scenario I simulations (see Table 1), the GM was fixed at 1, and the field LOD varied according to the percentage of the distribution that was censored. Here the field LOD was fixed at 0.037 mg m^{-3} (see Appendix 2), which required that the GM vary according to the percentage of the distribution that was censored. Otherwise, all of the simulation conditions presented in the Computer Simulation Methods section for Scenario I apply. True concentrations were generated as before, but this time a 'measurement' was simulated

by adding measurement error to the true concentration using the following equation:

$$x' = x \cdot (1 + Z_r \cdot \mathrm{CV_t}(x)),$$

where x' = measured concentration, x = true concentration (a random value generated from a lognormal distribution), Z_r = random Z-value and $\mathrm{CV_t}(x)$ = the sampling and analytical method $\mathrm{CV_t}$ at x (see Appendix 2).

The results are presented in Tables 15–17. Interestingly, while the bias tended to increase for all methods, the rMSE often changed little or even decreased. (After considering the issue, we recognized that the since the $\mathrm{CV_t}$ increases with decreasing concentration, it is more likely that a true non-detect will be 'measured' as a detect and less likely that a detect will be measured as a non-detect. This results in a positive bias for the GM and a negative bias for the GSD, with an overall positive bias for both the 95th percentile and the mean with little change in the rMSE. This effect becomes more pronounced as the percent censored increases.) Although the relative ranking of an individual method often changed, the relative ranking of the LPR- and MLE-based methods remained much the same. Therefore, our general conclusions remain unchanged.

The results of the modified Scenario I simulations indicate that the general effect of measurement error is an increase in the bias for both the 95th percentile and the mean, but with little to no change to the rMSE, leaving unchanged our general conclusions regarding the relative rankings of the methods, at least when sampling respirable dust. However, we suspect that this will be true for substances other than respirable dust (i.e. RPM), given that the flow rate CV and analysis CV used in the RPM analysis are similar to those for most sampling and analytical methods. In conclusion, the general effect of measurement error on the comparisons of the various CDA methods should be minimal. The potentially confounding effect of rounding or truncation will be investigated at a later date.

RECOMMENDATIONS

Since in reality occupational exposure profiles cannot be truly lognormal, our view is that all of the CDA methods discussed here are *ad hoc*. While some have a practical basis (substitution, KM, NP) and the remainder appeal to the notion that the data are reasonably well described by a theoretical distribution function (the LPR- and MLE-based methods), all are data analysis tools that we use in the hope that the results will be reasonably close to the truth. In occupational health, we rely heavily on the lognormal distribution assumption for summarizing our right-skewed datasets and for approximating the shape of the true exposure profile. However, unlike the lognormal distribution function, the actual exposure profile for any particular workplace has an upper boundary and is an unknown function of the physical parameters of the workplace and work practices of the employees, and not a function of the GM and GSD. Even if the underlying exposure profile is reasonably lognormal, we will never know the true GSD or the true percentage of the underlying distribution that lies below the field LOD, so that overall it can be difficult to determine which of the simulation and scenario combinations devised here best fit our situation.

Our results show that for the simulations and scenarios postulated an 'omnibus' CDA method does not yet exist. In our view, a 'preferred' CDA method is one that has both low bias and rMSE for the exposure profile parameter of interest and is robust whenever the true underlying distribution departs from the lognormal distribution model. Our personal preference when estimating the 95th percentile is to use a MLE-based method (with the exception of the $\mathrm{MLE_{mpv}}$ method). The MLE-based methods appear to be fairly robust, especially when compared to the supposedly robust NP methods and preferable to the LPR-based methods which tend to have larger rMSE values, particularly when the sample size is <20. Within the MLE-based methods, the standard MLE method comes closest to being an omnibus method, and therefore receives our recommendation as the preferred method.

Another selection factor to consider is the ease of calculation (or accessibility). The substitution methods are by far the easiest to implement and when dealing with large datasets, such as when constructing a job-exposure-matrix for the mean exposure, are certainly expedient and may be reasonably accurate, as was suggested by Hornung and Reed (1990), depending upon the true (but unknown) underlying GSD and percent censored. The LPR-based methods are more complicated, particularly the $\mathrm{LPR_r}$ and $\mathrm{LPR_{rm}}$ methods as these involve *ad hoc* methods and as such are difficult to automate via a programming language. The MLE method is easily accessible using the solver-function of most computer spreadsheets (Finkelstein and Verma, 2001), but some manual manipulation of the data is required. The $\mathrm{MLE_r}$ and $\mathrm{MLE_{rm}}$ methods require additional manual manipulation and are therefore also difficult to program.

Procedures using r-code have been published for implementing the more complicated methods—for example, MLE and KM—but even here the user is required to become proficient in the statistical programming language of r-plus (Frome and Wambach, 2005). Many statistical packages include the KM method for right-censored data, which necessitates some manipulation of both the data and results in order to apply the method to left-censored data. Consequently,

Table 15. Simulation 1, Scenario I (with measurement error)—single lognormal distribution and a single laboratory where the laboratory LOD is 1–50% of the true distribution; $20 \leq n \leq 100$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE_{rm}	1.7	MLE_{mpv}	22.5	MLE	0.4	MLE	17.9
LPR_{r}	2.5	Sub LOD$/\sqrt{2}$	22.8	Sub LOD$/\sqrt{2}$	−0.8	LPR	18.4
MLE_{r}	2.8	MLE_{rm}	22.9	LPR_{rm}	1.0	MLE_{rm}	19.6
MLE	2.8	Sub LOD	23.7	MLE_{mpv}	1.1	LPR_{rm}	19.7
Sub LOD/2	3.5	MLE	23.7	MLE_{rm}	1.1	MLE_{r}	19.7
LPR_{rm}	3.5	MLE_{r}	23.7	MLE_{r}	1.2	MLE_{mpv}	19.7
MLE_{mpv}	−3.6	LPR_{r}	24.1	LPR_{r}	1.5	LPR_{r}	19.9
LPR	4.3	LPR_{rm}	24.7	LPR	1.5	Sub LOD$/\sqrt{2}$	20.0
KM	4.7	Sub LOD/2	24.9	Sub LOD$/\sqrt{2}$	1.7	Sub LOD/2	20.6
Sub LOD$/\sqrt{2}$	−5.3	LPR	25.2	Sub LOD	5.3	KM	20.8
Sub LOD	−11.4	KM	34.1	KM	5.3	Sub LOD	20.9
NP	16.9	NP	52.3				

Table 16. Simulation 2, Scenario I (with measurement error)—single lognormal distribution and a single laboratory where the laboratory LOD is 50–80% of the true distribution; $20 \leq n \leq 100$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
MLE_{rm}	2.4	MLE_{rm}	24.4	Sub LOD/2	3.6	MLE	19.6
MLE_{r}	4.3	Sub LOD	25.4	LPR_{rm}	4.1	LPR	20.2
LPR_{r}	5.0	MLE_{r}	25.5	MLE_{mpv}	5.1	MLE_{mpv}	20.5
MLE	5.4	MLE	25.8	LPR	5.5	LPR_{rm}	20.9
LPR_{rm}	6.0	Sub LOD$/\sqrt{2}$	26.1	MLE_{rm}	6.7	MLE_{r}	20.9
KM	6.6	LPR_{r}	26.6	LPR_{r}	6.8	MLE_{rm}	20.9
LPR	7.3	Sub LOD/2	26.7	MLE	6.8	Sub LOD/2	21.1
Sub LOD/2	−13.4	MLE_{mpv}	26.7	MLE_{r}	6.8	LPR_{r}	21.2
MLE_{mpv}	−15.6	LPR_{rm}	27.7	Sub LOD$/\sqrt{2}$	15.9	Sub LOD$/\sqrt{2}$	26.5
Sub LOD$/\sqrt{2}$	−16.3	LPR	28.4	Sub LOD	33.0	KM	40.4
Sub LOD	−17.1	KM	33.8	KM	33.0	Sub LOD	40.5
NP	19.0	NP	51.7				

Table 17. Simulation 3, Scenario I (with measurement error)—single lognormal distribution and a single laboratory where the laboratory LOD is 1–50% of the true distribution; $5 \leq n \leq 19$; $1.2 \leq \text{GSD} \leq 4$

95th Percentile				Mean			
Method	Bias (%)	Method	rMSE	Method	Bias (%)	Method	rMSE
Sub LOD$/\sqrt{2}$	0.8	Sub LOD	52.4	Sub LOD/2	−0.9	MLE	38.5
MLE_{mpv}	3.7	Sub LOD$/\sqrt{2}$	55.6	LPR_{rm}	0.9	LPR	40.5
MLE	6.1	Sub LOD/2	59.7	MLE	−1.0	MLE_{rm}	41.7
Sub LOD	−6.1	MLE_{mpv}	61.0	MLE_{r}	1.0	MLE_{r}	41.9
LPR_{r}	6.4	MLE	62.2	MLE_{rm}	1.1	LPR_{r}	41.9
MLE_{rm}	6.6	MLE_{r}	65.0	MLE_{mpv}	1.3	MLE_{mpv}	42.1
MLE_{r}	7.2	MLE_{rm}	65.8	Sub LOD$/\sqrt{2}$	1.8	LPR_{rm}	42.9
Sub LOD/2	9.8	LPR_{r}	70.9	LPR	2.0	Sub LOD	43.2
LPR_{rm}	13.1	LPR_{rm}	85.6	LPR_{r}	2.2	KM	43.2
LPR	13.9	LPR	96.2	KM	5.2	Sub LOD$/\sqrt{2}$	43.7
				Sub LOD	5.3	Sub LOD/2	44.0

Table 18. Recommended methods based on the rMSE results

Distribution assumption	Sample size	Parameter	
		$X_{0.95}$	Mean
1–50% Censored			
Reasonably lognormal	Small n, $5 \leq n \leq 19$	MLE	MLE (MLE_{rm}) LPR
	Large n, $20 \leq n \leq 100$	MLE_{rm} (MLE, MLE_r)	MLE MLE_{rm} LPR (LPR_{rm})
Contaminated lognormal	Small n, $5 \leq n \leq 19$	MLE	MLE LPR
	Large n, $20 \leq n \leq 100$	MLE (MLE_r MLE_{rm})	MLE (MLE_r MLE_{rm}) LPR
50–80% Censored			
Reasonably lognormal	Large n, $20 \leq n \leq 100$	MLE	MLE
Contaminated lognormal	Large n, $20 \leq n \leq 100$	MLE (MLE_{rm})	MLE (MLE_{rm})

Methods in parentheses are roughly equivalent in performance to the recommended method.

apart from any opinions regarding the preferred method, ease of use and accessibility are bound to be factors in the final selection of a CDA method.

One obvious solution to these dilemmas is to simply eliminate or reduce the need for a CDA method through the judicious selection of an analytical method that has a low LOD and/or the collection of larger volume samples. Furthermore, for those laboratories that routinely report LOQs rather than LODs, they should be requested to additionally report the traditional LOD and the mass detected between the LOD and the LOQ. While a measurement between the LOD and LOQ is admittedly less reliable than a measurement above the LOQ, the loss of information when using the LOQ makes it even more difficult to estimate the lognormal parameters for the underlying distribution (Eduard, 2002; Helsel, 2005).

We agree with Shumway *et al*. (2002) that the temptation to combine disparate data—that is, data collected from different plants or similar exposure groups or from periods when different laboratories were used—should be resisted. The resulting datasets are often complex, with multiple censoring points, and probably were drawn from more than one underlying distribution. It is unreasonable, in our view, to expect a CDA method to extract a highly accurate estimate of a parameter of interest under such circumstances, and a faith that a NP method will indeed do so is, as our simulations reveal, somewhat misplaced. However, if the analysis of a complex dataset is required and it is strongly suspected that the underlying distribution departs significantly from the unimodal lognormal model, our recommendation is to use the standard MLE method for estimating either the 95th percentile (or other upper percentiles) and mean. The results in this study suggest that little is gained from the increased complexity of the MLE_r and MLE_{rm} methods (or the LPR_r and LPR_{rm} methods).

Finally, Table 18 summarizes our recommendations for sample sizes of 100 or less. The listed methods are felt to be roughly equal in performance. Our preference is to select from a particular family of methods when at all possible. Since the MLE-based methods did consistently well in all of the scenarios, our table consists primarily of the MLE-based methods. However, a similar table could be constructed that would hold primarily LPR-based methods (except for perhaps when n is small). In our view, the standard MLE method comes closest to being an omnibus method. The so-called robust versions of the MLE- and LPR-based methods did not consistently result in superior performance when challenged with contaminated lognormal distributions. Due to the failure of the NP quantile methods and the KM method to perform better than the LPR- and MLE-based methods when confronted with contaminated distributions (for the scenarios postulated in this paper), we do not recommend their inclusion in any such table.

APPENDIX 1. QUANTILE CALCULATUION METHODS

The following table lists the NP quantiles presented by Hyndman and Fan (1996), the corresponding quantiles offered by two commercial statistics packages and the necessary calculations. Hyndman and Fan recommended the Q8 method. Q1 through Q3 use a step function approach. However, the Q2 method uses a averaging method whenever $0.95n$ equals an integer. The Q4 through Q8 methods use linear interpolation to estimate the 95th percentile whenever k is not an integer. The default quantile methods for the Systat Version 11 and the SAS Version 9 statistics programs are the Q5 and Q2 methods, respectively.

APPENDIX 2. CALCULATION OF MEASUREMENT ERROR AND THE MINIMUM DETECTABLE CONCENTRATION

Overall method accuracy is usually summarized using the propagation of errors formula (Kogut *et al*., 1997):

$$CV_t = \sqrt{CV^2_{pump} + CV^2_{sampler} + CV^2_{analysis}},$$

Hyndman and Fan (1996) quantile method	Systat Version 11	SAS Version 9	Intermediate calculations[a]	95th Percentile calculation
Q1	Empirical CDF	QNTLDEF 3	$k = 0.95n$, $i = \text{Floor}(k)$	If $(k - i) > 0$ then $X_{0.95} = x_{i+1}$, if $(k - i) = 0$ then $X_{0.95} = x_i$
Q2	Empirical CDF with averaging	QNTLDEF 5 (default)	$k = 0.95n$, $i = \text{Floor}(k)$	If $(k - i) > 0$ then $X_{0.95} = \frac{1}{2}(x_{i+1} - x_i)$, if $(k - i) = 0$ then $X_{0.95} = \frac{1}{2}(x_{i+1} - x_i)$
Q3	Closest value	QNTLDEF 2	$k = 0.95n$, $i = \text{Round}(k)$	$X_{0.95} = x_i$
Q4	Weighted average 1	QNTLDEF 1	$k = 0.95n$, $i = \text{Floor}(k)$	$X_{0.95} = x_i + (k - i)(x_{i+1} - x_i)$
Q5	Cleveland's method (default)		$k = 0.95n + \frac{1}{2}$, $i = \text{Floor}(k)$	$X_{0.95} = x_i + (k - i)(x_{i+1} - x_i)$
Q6	Weighted average 2	QNTLDEF 4	$k = 0.95(n + 1)$, $i = \text{Floor}(k)$	$X_{0.95} = x_i + (k - i)(x_{i+1} - x_i)$
Q7	Weighted average 3		$k = 0.95(n - 1) + 1$, $i = \text{Floor}(k)$	$X_{0.95} = x_i + (k - i)(x_{i+1} - x_i)$
Q8			$k = 0.95(n + \frac{1}{3}) + \frac{1}{3}$, $i = \text{Floor}(k)$	$X_{0.95} = x_i + (k - i)(x_{i+1} - x_i)$

[a]The 'Floor(k)' function returns the highest integer less than or equal to k. The 'Round(k)' function uses Banker's Rounding.

where CV_t = total coefficient of variation for a sampling and analytical method; CV_{pump} = fractional variation due to random variations in the pump flowrate; $CV_{sampler}$ = fractional variation due to the manufacturing variation in the sampling device and $CV_{analysis}$ = fractional variation due to the analytical determination of the analyte mass.

The CV_{pump} and $CV_{sampler}$ are generally considered to be independent of the true concentration. The $CV_{analysis}$ is not independent of the true concentration, resulting in the following general equation for CV_t:

$$CV_t(x) = \sqrt{CV_{pump}^2 + CV_{sampler}^2 + \left[\left(\frac{\sigma_{mass}}{Q \cdot T}\right) \Big/ x\right]^2},$$

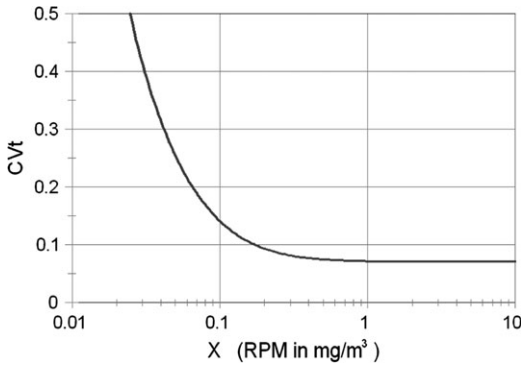

Fig. 1. Total coefficient of variation (CV_t) calculated as a function of true concentration when sampling respirable dust (i.e. RPM).

where σ_{mass} = the standard deviation of the analytical system; Q = flowrate; T = averaging time for the measurement and x = true concentration.

Using as an example, the sampling of respirable dust (respirable particulate mass, RPM) (and ignoring any uncorrectable particle size distribution effects), the total coefficient of variation at different concentrations of RPM can be estimated. Although recent studies (Kogut *et al*., 1997) have shown slightly lower values, the CV_{pump} has traditionally been given a value of 0.05. The $CV_{sampler}$ for the Dorr-Oliver 10-mm nylon cyclone has been estimated by Kogut *et al*. (1997) to be 0.023, but in this example we will use 0.05 as reported by Bartley *et al*. (1994). The σ_{mass} will vary with the analyte and analytical method. For a single weighing of a filter used in respirable dust sampling, a typical σ_{mass} is 0.005 mg. Let us assume that the mass collected on each filter is also blank corrected (i.e. each sample filter has a matching blank filter). Since both the sample filter and the matched blank are pre- and post-weighed, a total of four weighings are required to estimate the mass collected on the sample filter, resulting in an overall σ_{mass} of 0.010 mg. Finally, the low rate and averaging time will be set at the standard values of 1.7 Lpm (for RPM) and 480 min, resulting in the following equation:

Figure 1 shows the relationship between the CV_t and the true RPM concentration. At the higher

$$CV_t(x) = \sqrt{0.05^2 + 0.05^2 + \left[\frac{0.010\,\text{mg}}{0.0017\,\text{m}^3\,\text{min}^{-1} \cdot 480\,\text{min}} \cdot \frac{1}{x}\right]^2}.$$

concentrations, the CV_t is relatively constant. Sufficient mass is collected that the contribution of the $CV_{analysis}$ becomes insignificant compared to the fixed variability due to the sampling pump and sampler. At low concentrations, the $CV_{analysis}$ predominates and steadily increases with decreasing collected mass. (The curve does not remain flat forever as the concentration increases. Very high concentrations will tend to result in overloaded samplers, which will drive the CV_t upwards.)

A CV_t curve can be determined for any analyte and sampling method, and will have a shape similar to that in Fig. 1. According to the NIOSH (Abell and Kennedy, 1997), a reasonably accurate method should have a true CV_t that is <0.128 over the range of 10–200% of the exposure limit. However, at the method's field LOD—that is, the laboratory LOD divided by the sample volume, also called the minimally detectable concentration—the CV_t can be much greater. The field LOD for sampling RPM can be estimated using the σ_{mass}:

$$\mathrm{LOD} = \frac{3 \cdot \sigma_{\mathrm{mass}}}{Q \cdot T}.$$

$$\mathrm{LOD} = \frac{3 \cdot 0.010\,\mathrm{mg}}{0.0017\,\mathrm{m^3\,min^{-1}} \cdot 480\,\mathrm{min}} = 0.037\,\mathrm{mg\,m^{-3}}.$$

The factor of three forces the field LOD to be three standard deviations above the mean weight change for an unused filter (the true mean weight change is zero). Three standard deviations above the mean instrument response is the traditional method for determining the analytical LOD (Abell and Kennedy, 1997).

At the field LOD the CV_t is 0.34, indicating that there is a considerable amount of method variability, or what is commonly referred to as measurement error. This suggests that at and below the LOD, it is highly likely that a true non-detect—that is, a true concentration that is less than the LOD—could be reported as a detect, above the LOD, due to measurement error. The reverse it also true, a true detectable concentration could be reported as a non-detect.

REFERENCES

Abell MT, Kennedy ER. (1997) A computer program to promote understanding of the monitoring evaluation guidelines used at NIOSH. Am Ind Hyg Assoc J; 58: 236–41.

Bartley DL, Chen C, Song R *et al*. (1994) Respirable aerosol sampler performance testing. Am Ind Hyg Assoc J; 55: 1036–46.

Bullock WH, Ignacio JS, editors. (2006) A strategy for assessing and managing occupational exposures. 3rd edn. Fairfax, VA: American Industrial Hygiene Association.

Eduard W. (2002) Estimation of mean and standard deviation (letter to the editor). Am Ind Hyg Assoc J; 63: 4.

El-Shaarawi AH, Esterby SR. (1992) Replacement of censored observations by a constant: an evaluation. Water Res; 26: 835–44.

Environmental Protection Agency (EPA). (2006) Data quality assessment: statistical methods for practitioners, EPA QA/G-9S. Washington, DC: Environmental Protection Agency.

Finkelstein MM, Verma DK. (2001) Exposure estimation in the presence of nondetectable values: another look (see AIHAJ 63:4 2002 for letters to the editor). AIHAJ; 62: 195–8.

Frome EL, Wambach PF. (2005) Statistical methods and software for the analysis of occupational exposure data with non-detectable values. Oak Ridge, TN: Oak Ridge National Laboratory ORNL/TM-2005/52.

Gibbons RD, Goleman DE. (2001) Statistical methods for detection and quantification of environmental contamination. New York: John Wiley and Sons, Inc.

Gilbert RO. (1987) Statistical methods for environmental pollution monitoring. New York: Van Nostrand Reinhold.

Gilliom RJ, Helsel DR. (1986) Estimation of distributional parameters for censored trace level water quality data 1. Estimation techniques. Water Resour Res; 22: 135–46.

Glass DC, Gray CN. (2001) Estimating mean exposures from censored data: exposure to benzene in the Australian petroleum industry. Ann Occup Hyg; 45: 275–82.

Hawkins NC, Norwood SK and Rock JC, editors. (1991) A strategy for occupational exposure assessment. Fairview, VA: American Industrial Hygiene Association.

Helsel DR, Cohn TA. (1998) Estimation of descriptive statistics for multiply censored water quality data. Water Resour Res; 24: 1997–2004.

Helsel DR, Hirsch RM. (2002) Statistical methods in water resources. Department of the Interior, United States Geological Survey, Reston, Virginia. Available from: http://water.usgs.gov/pubs/twri/twri4a3/.

Helsel DR. (2005) Nondetects and data analysis. New York: John Wiley & Sons, Inc.

Hornung RW, Reed LD. (1990) Estimation of average concentration in the presence of nondetectable values. Appl Occup Environ Hyg; 5: 46–51.

Hyndman RJ, Fan Y. (1996) Sample quantiles in statistical packages. Am Stat; 50: 361–5.

Kogut J, Tomb TF, Parobeck PS *et al*. (1997) Measurement precision with the coal mine dust personal sampler. Appl Occup Environ Hyg; 12: 999–1006.

Kroll CN, Stedinger JR. (1996) Estimation of moments and quantiles using censored data. Water Resour Res; 32: 1005–12.

Mulhausen J, Damiano J, editors. (1998) A strategy for assessing and managing occupational exposures. 2nd edn. Fairfax, VA: American Industrial Hygiene Association.

SAS. (2006) SAS/STAT software, Version 9. SAS Institute, Inc. Available from: http://www.sas.com.

Schmoyer RL, Beaucamp JJ, Brandt CC *et al*. (1996) Difficulties with the lognormal model in mean estimation and testing. Environ Ecol Stat; 3: 81–97.

She N. (1997) Analyzing censored water quality data using a nonparametric approach. J Am Water Res Assoc; 33: 615–24.

Shumway RH, Azari RS, Kayhanian M. (2002) Statistical approaches to estimating mean water quality concentrations with detection limits. Environ Sci Tech; 36: 3345–53.

Succop PA, Clark S, Chen M *et al*. (2004) Imputation of data values that are less than a detection limit. J Occup Environ Health; 1: 436–41.

Symanski E, Kupper LL, Kromhout H, Rappaport SM. (1996) An investigation of systematic changes in occupational exposure. Am Ind Hyg Assoc J; 57: 724–35.

Systat. (2007) Systat, Version 11. Systat Software, Inc Available from: http://www.systat.com.