Published in final edited form as:

Risk Anal. 2013 August ; 33(8): 1500–1509. doi:10.1111/j.1539-6924.2012.01939.x.

On determining the BMD from multiple outcomes in developmental toxicity studies when one outcome is intentionally missing

Julie S Najita^{1,2,*} and Paul J Catalano^{1,2}

¹Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215

²Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

Abstract

Public health concerns over the occurrence of developmental abnormalities that can occur as a result of prenatal exposure to drugs, chemicals, and other environmental factors has led to a number of developmental toxicity studies and the use of the benchmark dose (BMD) for risk assessment. To characterize risk from multiple sources, more recent analytic methods involve a joint modeling approach, accounting for multiple dichotomous and continuous outcomes. For some continuous outcomes, evaluating all subjects may not be feasible, and only a subset may be evaluated due to limited resources. The subset can be selected according to a prespecified probability model and the unobserved data can be viewed as intentionally missing in the sense that subset selection results in missingness that is experimentally planned.

We describe a subset selection model that allows for sampling pups with malformations and healthy pups at different rates, and includes the well-known simple random sample (SRS) as a special case. We were interested in understanding how sampling rates, which are selected beforehand, influence the precision of the BMD. Using simulations we show how improvements over the SRS can be obtained by oversampling malformations, and how some sampling rates can yield precision that is substantially worse than the SRS. We also illustrate the potential for cost-saving with oversampling. Simulations are based on a joint mixed effects model, and to account for subset selection, use of case weights to obtain valid dose response estimates.

Keywords

Benchmark dose; subset selection; dose response modeling; fetal toxicity

1. INTRODUCTION

Concern over the occurrence of birth defects and developmental abnormalities that may occur as a result of prenatal exposure to drugs, chemicals and other environmental factors has led to agencies such as the EPA and FDA to place emphasis on setting exposure limits in order to protect the public from the adverse effects of these substances. Quantitative risk assessment (QRA), the benchmark dose (BMD), first introduced by Crump⁽¹⁾, and its statistical lower limit, the BMDL, have become essential in the effort to identify exposure

*Address correspondence to: Julie S. Najita, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, MS CLS11007, 450 Brookline Avenue, Boston, MA 02215; jnajita@jimmy.harvard.edu.

limits. Developmental toxicity studies play a critical role in this risk assessment process and these studies have been described previously by others ⁽²⁾.

Briefly, a typical developmental toxicity experiment involves exposing a pregnant dam to a toxin at the peak of fetal organogenesis, with 20–30 dams assigned to 2–3 dose groups or a control in order to evaluate dose response. Just prior to normal delivery, the dams are sacrificed and the uterine contents are examined for fetal outcomes. In these studies, risk is often based on multiple events, which can be a mixture of binary and continuous outcomes, such as fetal malformation and birth weight. To evaluate risk involving both kinds of outcomes, joint models can be used to estimate the dose response and these estimates can then be used to determine a BMD and BMDL. These approaches are widely known and have frequently been applied to the analysis of binary and continuous outcomes ^(3,4,5,6,7,8,9,10) and are extensions for single outcomes ^(11,12,13). As developmental toxicity studies typically involve many litters, this means collecting data for a large number of animals because having complete data is usually a requirement of these approaches.

Multiple outcomes can include those that may be cheap and quick to assess, such as malformation determined by gross examination, and others, such as the size of the ocular cup or alveolar lung volume, which are time-consuming and expensive to measure. While evaluating offspring for multiple outcomes is common, not all pups may be evaluated for all of the outcomes and only a subset may be evaluated for a particular outcome due to practical limitations. As not all pups are evaluated, the observable but unobserved data can be viewed more broadly as intentionally missing in the sense that subset selection results in missingness that is planned because selection is under the control of the experimenter. Missing data can also arise for reasons that are unplanned, however this type of missingness is not the focus of this paper.

Intentionally missing data, as a result of selecting only a subset for evaluation, can occur for a number of reasons in developmental toxicity studies. For rodents, EPA test guidelines suggest evaluating one-half of each litter for skeletal alterations (bone and cartilage) and the rest of the litter for soft tissue anomalies. For rabbits, evaluating for internal structures of the head, eyes, brain, nasal passages and tongue is suggested for at least half of the fetuses. Thus, the subset may be a simple random sample (SRS) with a one-half proportion (50% SRS) in accordance with federal testing guidelines ⁽¹⁴⁾. However, a SRS may not always be statistically ideal despite its appeal as a straightforward sampling approach.

If exposure-related changes in the continuous outcome are more likely to occur in affected pups, the experimenter might preferentially select affected offspring in which to observe the outcome. With limited resources for detailed assessments, it may be desirable to select a larger proportion of pups having a malformation and a small proportion of healthy pups instead of a SRS. This type of sampling has arisen before in a developmental toxicity study where eye size and malformation of the eye were evaluated ⁽¹⁵⁾. In that study, eye size was measured in the majority of pups having microphthalmia and a small proportion of pups without malformation.

Selecting a subset that is nested within the main study can be considered part of a larger set of two-stage approaches known as two-phase sampling and has been used in other contexts ^(16,17,18). The subset at the second stage is not necessarily a representative subsample, and it is widely known from the survey literature that if one is interested in the continuous outcome alone, the analysis must account for the unequal probability of selection using case weights. In prior work, we demonstrated how this weighting approach can also be used to model both outcomes. The technical details and the use of weights to estimate dose response are described elsewhere ⁽¹⁹⁾.

Our main interest was in the precision of the BMD estimate and how one might choose the second-stage sampling proportions (subsampling rates) when planning a study. Naturally, because only a subset is evaluated for one of the outcomes in subsampling, there is less precision compared to the case where all subjects are evaluated. It is less obvious how much precision is lost with a subset and, as there might be many ways to select a subset of a given size, the various approaches might lead to very different precision. Our goal for this work is to illustrate how the precision of BMD estimates that are derived from these joint models is influenced by choice in subsampling rates. Using simulations, we show how some choices in the subsampling rates still give precision that is similar to a SRS and where precision is better than a SRS. We also give examples where there is substantially less precision. This paper is organized as follows. In Section 2, we describe a probability model for subset selection to formalize the notion of subsampling. In Sections 3 and 4, we present simulations that illustrate the effects of various subsampling approaches using a joint mixed effect model⁽¹⁹⁾. To facilitate the discussion, we briefly review the joint model and how the BMD's standard error (SE) was determined. We conclude with some recommendations.

2. PROBABILITY MODEL FOR SUBSET SELECTION

Here we present a probability model that serves as a convenient way to describe data arising from subsampling, as well as simple random sampling, and complete data. While one could envision more complex alternatives, the model is well-suited for demonstrating the relationships between subsampling rates (chosen by the experimenter) and BMD precision.

The mechanism generating the observed data is viewed in two steps. In the first step, a live pup that can be evaluated for malformation and the continuous outcome, S , is sampled at random from the population. Outcomes observable at this step are those one would typically observe in a traditional developmental toxicity study where viable pups are evaluated at the time that uterine contents are examined. The pups can then fall in one of two strata, based on observed malformation status. At this point, pups are not yet evaluated for the continuous outcome so S is not yet observed. At the second step, pups are selected for the subset from the stratum of affected offspring with probability p_1 and from the stratum of unaffected offspring with probability p_2 . Only pups selected for the subset are evaluated for the continuous outcome. For pups not selected, only malformation status is observed. Assuming there are a total of N live pups in the main study and M pups will be selected for the subset, under the model, the average value of M depends on the subsampling rates and the marginal probability of malformation (P_m):

$$E(M) = N(p_1 P_m + p_2 (1 - P_m)). \quad (1)$$

Clearly, there is not a unique way to choose a subset of size M because the experimenter is free to choose p_1 and p_2 . For example, one can sample a relatively large proportion of the study if p_1 is large and malformation is relatively common, or if p_2 is large and malformation is uncommon, or it can be a mixture of the strata. The case $p_1 = p_2 = 1$ corresponds to the complete data setting where the selection step is absent. When the sampling rates are the same but less than 1, $p_1 = p_2 = p < 1$ corresponds to the SRS case where the population-based relative frequencies are retained in the subsample. When the sampling rates differ, $p_1 > p_2$ corresponds to oversampling abnormalities while $p_1 < p_2$ corresponds to undersampling.

With subsampling, the observed data are a collection of data pairs for pups in the subset and malformation status for pups in the main study but not in the subset. Letting R indicate if a pup is included in the subset, there are four possible types of observations that reflect malformation status and whether both outcomes are observed:

$$P(R=1, \text{malf}, S=s) = p_1 P(\text{malf}, S=s)$$

$$P(R=1, \text{unaffected}, S=s) = p_2 P(\text{unaffected}, S=s)$$

$$P(R=0, \text{malf}) = (1 - p_1) P(\text{malf})$$

$$P(R=0, \text{unaffected}) = (1 - p_2) P(\text{unaffected})$$

When p_1 and p_2 differ, the observed data do not correspond to a representative sample from the population and maximum likelihood estimates are difficult due to the complexity of the full likelihood that results from the sampling design. Valid estimates are obtained by maximizing a partial likelihood (PL) where observations in the PL score are weighted by the inverse probability of selection. Estimates obtained by solving the weighted score equations can be viewed as weighted estimating equations (WEE) estimates and are asymptotically normally distributed⁽¹⁹⁾. Throughout this paper, the BMD estimates and associated standard errors (SE) we report are based on these WEE estimates.

3. EVALUATING SUBSAMPLING DESIGNS BY SIMULATION

To study the impact of subsampling, we wanted to compare the precision of BMD estimates obtained under subsampling with those that would be obtained had only a SRS of pups been evaluated for the continuous outcome. Our motivation was to compare BMD estimates obtained when one-half of pups are evaluated for the continuous outcome following federal testing guidelines with those that could be obtained by over- or undersampling affected pups. While testing guidelines suggest a one-half proportion, we examined the SRS approach more broadly, comparing over- and undersampling to SRS over a range of proportions. We also examined precision with subsampling versus having complete data as a way to relate the subsampling approach to the most expensive study where all offspring are evaluated.

We wanted to know for a given proportion (M/N), whether oversampling would yield better precision than SRS and if so, to quantify the improvement. We thought this would be important for planning purposes, as there might be resources for only a specific number of pups and given this constraint, and there might be interest in approaches giving better precision than the SRS. Additionally, one might want to know about subsampling rates yielding less precision than the SRS, so as to avoid them.

Recognizing that there is less precision with smaller subsets, we also sought ranges of M/N where oversampling was better and where benefit with oversampling was unlikely. We thought this might help to identify a subsampling strategy that was better over a wide range of subset sizes and might be a good all-around strategy.

We also hypothesized that precision with subsampling would depend on the prevalence of malformation and that greater improvements over the SRS could be obtained when the binary outcome (here malformation) was less common. This was based on the idea that with SRS, the subsample would contain mostly unaffected pups having higher fetal weights and far fewer malformations having lower fetal weights, resulting in less precision for estimating

the dose response for fetal weight and less precision for the BMD. Therefore, we looked at two scenarios, where malformations were more common and where they were less common, to understand how improvements over the SRS change with malformation rate.

To characterize differences in precision as a function of subset size, we examined subsampling over a range of M/N and noted improvements against the SRS. To compare designs, we wanted to relate precision with subsampling to the most expensive study of evaluating all offspring. Because precision achievable with any subsampling strategy could not exceed the precision possible with complete data, we used a relative measure of precision, taking the ratio of the SE with complete data (SE_c) to the SE under subsampling (SE_{sub}). Hence, the relative precision for any subsample would be less than 1 ($RP_c = SE_c / SE_{sub}$). We used a similar measure to compare precision with a subsampling design to SRS, setting $RP_{SRS} = SE_{SRS} / SE_{sub}$. Thus, $RP_{SRS} > 1$ indicates the subsampling design has better precision than the SRS while $RP_{SRS} < 1$ indicates less precision. We examined the relative precision for values of M/N ranging 0.3 to 0.8. There were two reasons for restricting subsets to this range. First, it seemed unlikely that an experimenter would select a proportion much lower than one-half because there might be interest in evaluating the continuous outcome alone as a univariate endpoint and a small subsample might be viewed as having limited utility on its own. Second, we wanted to avoid some of the technical difficulties that might be seen when fitting a model to a relatively small data set which we thought would be likely with a proportion of 0.30 or less and the purpose of including $M/N = 0.3$ was to illustrate some of those challenges.

From (1), it is obvious that there is not a unique choice of subsampling rates to select a given proportion of pups. To evaluate a variety of designs, the subsampling rates were chosen by first specifying a value for p_1 and solving (1) for p_2 to maintain a fixed value of M/N with a given P_m and repeating the process over a range of values for p_1 . Illustrating the precision over a wide range of subsampling rates was thought to be a reasonable way to describe how design choices are related to the BMD precision, although the intent was not to catalog all possible (p_1, p_2) pairs.

3.1 Dose-response model and BMD determination

To account for multiple outcomes, the dose response model we used is based on mixed effects models described previously^(20,10) and is a model we have used previously to illustrate the subsampling⁽¹⁹⁾. To choose parameter values for the simulation and to facilitate the discussion, we focus on malformation and fetal weight as these are typical primary endpoints for live offspring although other endpoints could be chosen because the subsampling approach can be applied quite generally.

Let Y denote malformation status and S for the continuous variable. To model Y , we used a latent variable, Y^* , so that (without loss of generality) there is a malformation, $Y = 1$, if $Y^* > 0$ and there is no malformation, $Y = 0$, if $Y^* \leq 0$. With dam (litter) i exposed to dose d_i we have

$$Y_{ik}^* = \alpha_{11} + \alpha_{12}d_i + \tau_1\theta_i + \varepsilon_{1ik} \quad (2)$$

$$S_{ik} = \alpha_{21} + \alpha_{22}d_i + \tau_2\theta_i + \varepsilon_{2ik} \quad (3)$$

$$\text{var}(\varepsilon_{ik}) = \begin{pmatrix} \delta^2 & \delta\sigma\rho \\ \delta\sigma\rho & \sigma^2 \end{pmatrix} \quad (4)$$

where ε_{ik} has a bivariate normal distribution with mean vector 0, $k = 1, \dots, n_i$ indicates a fetus within a litter and $i = 1, \dots, N_L$ refers to litter. The random effect θ_j , which is independent of ε_{ik} , allows for outcomes to be less variable among pups from the same litter and more variable among pups from different litters, commonly referred to as a litter effect. τ_1 and τ_2 allow for outcomes that may be on different scales and ρ allows for correlation between outcomes within a single fetus. We note that because the latent variable for malformation is unobservable, the latent mean and variance are not uniquely identifiable from binary data so δ is set equal to 1 for estimability^(3,21,19).

We defined an adverse outcome to be the occurrence of a malformation or a low fetal weight (weight below a threshold s_c) as in other reports^(3,7,9). Then the probability of an adverse outcome at dose d is

$$P(d) = P(Y_k = 1 \text{ or } S_k < s_c | d) = 1 - P(Y_k^* > 0, S_k > s_c | d).$$

Using additional risk to quantify risk beyond the background rate(s), we defined the risk function as $r(d) = P(d) - P(0)$ which can be expressed in terms of tail areas of a bivariate normal distribution (see Appendix). We then applied established methods^(1,22,23,24,25,26) to estimate the BMD, substituting WEE estimates of the dose response parameters in the risk function. Briefly, with a benchmark response (BMR) of 0.10, the BMD was obtained by solving for dose in the equation $r(d) = 0.10$ using numerical techniques since there is no analytical solution. The variance for the BMD was determined using the Delta Method⁽²⁵⁾. We note that the LR-based approach⁽²²⁾ that is available with maximum likelihood estimates cannot be applied here because WEE estimates are based on the weighted PL rather than the full likelihood. Valid BMDL's can be easily calculated from WEE estimates using the LED- or Delta Method-based approaches although we did not determine the BMDL in this study because we were mainly interested in variance estimates.

3.2 Details of the simulations

We constructed simulations to be consistent with the range of data seen in practice. For the dose response model, simulated experiments included four dosed groups and a control, assuming the rate of malformation, transformed by the inverse probit function, increased linearly with dose while mean fetal weight decreased linearly with dose. We used the joint dose response model in (2–4) to generate simulated data and standard methods mentioned in Section 3.1 to determine the BMD and its variance.

To evaluate our hypothesis that subsampling could yield greater improvements relative to SRS when malformation was less common, we examined BMD precision in two scenarios. In scenario A, malformation rates ranged from 7% (background) to 69% at the highest dose, giving an overall malformation rate of 35%. In scenario B, malformation rates ranged from 6% (background) to 50% at the highest dose for an overall rate of 25%. Other aspects of the model were identical in both scenarios. We assumed a mean fetal weight of 500 mg among controls, and comparing controls to those exposed to the highest dose, we assumed mean fetal weight was reduced by 300 mg. Values for the remaining parameters were chosen to produce modest within-litter correlations of 0.10 for each outcome and moderate within-fetus correlations of 0.60. Correlations were assumed to be constant over dose.

The parameter values used were $\alpha_{11} = 1.5$ and $\alpha_{12} = -2$ in scenario A and $\alpha_{11} = 1.6$, and $\alpha_{22} = -1.6$ in scenario B. For other parameters, values were set to $\alpha_{21} = 5$, $\alpha_{22} = -3$, $\tau_1 = \tau_2 = 0.33$, $\sigma = 1.0$, and $\rho = 0.56$. In all cases dose levels were spaced equally between the control and the maximum dose (set to 0, 0.25, 0.50, 0.75, and 1.0). The threshold, s_c , was set to the 5th percentile of fetal weight, giving a true BMD of 0.220 in scenario A and 0.262 in

scenario B. The number of live offspring in each litter was fixed at 10 and the number of dams assigned to each dose level was fixed at 30, corresponding to study sizes of $N=1,500$ pups.

Each subsampling design was defined by overall malformation rate (P_m), subset size (M), and sampling rates (p_1, p_2). For each subsampling design, we simulated an experiment in order to obtain the complete data set, then sampled the simulated data according to that subsampling design. Dose response estimates were obtained from the simulated subsample using weights that accounted for the subsampling design and the BMD was determined by substituting dose response estimates from the weighted analysis. To evaluate the precision for the BMD, we calculated the empirical SE (SD of the BMD estimates over all simulations) as well as model-based SE's. Each experiment was simulated 500 times to have enough precision in empirical estimates of the BMD SE so that observed differences between SE estimates due solely to sampling variability would be unlikely. Relative precision (RP_C, RP_{SRS}) was determined taking the ratios of empirical SE's. This process was repeated over a range of sampling rates p_1 and p_2 and over a range of proportions (M/N) and for each scenario.

4. RESULTS

SE estimates and relative precisions are presented in Tables I–II. For $M/N=0.3$ and 0.8 , some designs were not possible with the given N, P_m, M , and p_1 as it was not always possible to find sampling rates satisfying the constraints in (1). We first confirmed that precision with a subset improved as subset size increased. With SRS, precision relative to complete data (RP_C) steadily improved with subset size and was highest with the largest subset, as expected. With oversampling and undersampling, there was a trend toward improved precision with improvement varying somewhat within a type of subsampling design depending on the sampling rates that were chosen. For example with an 80% subset, RP_C ranged 0.80–0.93 in scenario A and 0.52–0.88 in scenario B. RP_C was lower with a 50% subset, ranging 0.46–0.70 in scenario A and 0.45–0.70 in scenario B. Looking at the range of relative precision by type of subsampling, RP_C in scenario A was 0.93 with oversampling and ranged 0.80–0.88 with undersampling with a 80% subset. With a 50% subset, RP_C ranged 0.64–0.70 with oversampling and 0.46–0.62 with undersampling. Results were similar in scenario B. The gains in precision with subset size were most evident with SRS and oversampling. For example in scenario A, RP_C increased from 0.43 to 0.91 as M/N increased from 0.3 to 0.8 with SRS. With oversampling, a similar trend was also seen with RP_C increasing from 0.50 to 0.93. We found a similar trend in scenario B ($RP_C=0.42$ –0.85 with SRS, $RP_C=0.44$ –0.88 with oversampling).

Next, looking at subsampling relative to SRS (RP_{SRS}), we found that oversampling provided precision at least as good as or better than SRS for a given subset size in nearly every design that we simulated. With the exception of one case (row 36, Table I), precision relative to the SRS ranged $RP_{SRS}=0.99$ –1.15 in scenario A. In addition, with a 50% subsample, oversampling gave precision that was numerically higher but essentially equivalent to the SRS, suggesting that oversampling would be a reasonable alternative to the SRS (see shaded rows, $M/N=0.50$). The higher precision with oversampling was evident in plots of RP_{SRS} and subset size (Figure 1).

In contrast, undersampling tended to give less precision than either SRS or oversampling however, the amount of precision that was lost varied, with only small losses when p_1 and p_2 were similar, and the greatest loss when there were large differences in the sampling rates (e.g., $RP_{SRS}=0.96$ row 10 versus $RP_{SRS}=0.57$ row 6, Table I). This heterogeneity was seen over a range of subset sizes and was not limited to a particular subset size (Figure 1).

While precision with oversampling was more homogeneous over a range of subset sizes, precision with undersampling was more heterogeneous and SE estimates were more diffuse, suggesting that precision is more sensitive to choice of p_1 and p_2 with undersampling and less sensitive with oversampling. For example when $M/N=0.50$, RP_{SRS} ranged from 1.01–1.10 in scenario A and 1.07–1.09 in scenario B with oversampling. In contrast, RP_{SRS} ranged 0.73–0.97 and 0.70–0.89 in scenarios A and B respectively with undersampling. Therefore, if subset size was dictated by limited resources, an oversampling design taking nearly all malformations would not fare any worse than sampling roughly equal proportions from each stratum. With undersampling, however, there was potential for a substantial negative impact on precision depending on the subsampling rates that were chosen.

Next, comparing scenarios A and B, larger improvements in precision with oversampling were seen in scenario B with smaller subsets, reflected by numerically larger values of RP_{SRS} , suggesting greater improvements with oversampling when the malformation rate was lower. For example, with $M/N=0.4$ and similar subsampling rates, we found there was greater precision relative to SRS in scenario B ($RP_{SRS}=1.22$) compared with scenario A ($RP_{SRS}=0.99$).

If evaluating each pup for the continuous outcome incurs a fixed cost, designs that yield the same precision as SRS but involve a smaller subset can be viewed as cost saving. The potential for cost-saving with oversampling was evident in both scenarios, with greater saving when the malformation rate was lower. To see this, we compared oversampling and a 60% subset with SRS and a 70% subset. The oversampling design required an average of 900 pups while the SRS design required 1050, so oversampling required 150 fewer pups on average. In scenario A, the precision with a 60% subset gave a SE that was only 6% larger than that with a 70% subset with SRS (SE=0.0103 row 19 versus SE=0.0097 row 11, Table I). With the lower malformation rate in scenario B, the precision with a 60% subset and oversampling gave nearly the same SE that could be obtained with SRS and a 70% subset (SE=0.0123 row 23 versus SE=0.0120 row 13, Table II).

Finally, our simulations used Fisher scoring, a standard approach for model fitting. In most cases, model fitting was successful for at least 90% of simulated data sets. However, model fitting was noticeably impacted in scenario B, where the malformation rate was lower, when $M/N=0.30$. With this smaller subset size, there were sufficient data to fit the model for 80% of simulated data sets when $p_1=0.7$ and $p_2=0.167$ and this percentage decreased further to 62% when $p_1=0.8$ and $p_2=0.133$ (data not shown). While a 30% subset would not likely be chosen for practical reasons, by constructing our simulations to specifically study convergence at low event rates and with small subsets, we verified that breakdown in model fitting is evident with a 30% subset.

5. DISCUSSION

Using a joint model for dose response, we sought to evaluate the relative precision of two subsampling strategies, oversampling and undersampling, compared with SRS, a common sampling technique. We found simulation to be a convenient way to assess potential designs and construct a matrix of reasonable designs which can help to guide experimental planning. The simulations focus on live outcomes because this was thought to be the simplest approach to make the comparison as the subsampling is intended for live outcomes. The simulations focus mainly on dose effects with live outcomes and it is well known that outcomes such as fetal weight can be related to litter size or embryolethality^(27,28,29,30,31). Adjustment for litter size can be through its inclusion as a model covariate.

Our simulations show that undersampling was more often less efficient than either SRS or oversampling. With undersampling, precision was sensitive to both subset size (the number of pups that are sampled) as well as the subsampling rates. In contrast, oversampling gave precision at least as good as SRS not just with a 50% subset, but over a range of subset sizes we considered. Importantly, oversampling was usually better than SRS at smaller subset sizes with more noticeable improvements observed when malformations were less common. This would suggest that if subset size is fixed, several reasonable options are available with oversampling when choosing subsampling rates.

In our simulations, M is modeled as a random variable however it is likely to be a fixed quantity in practice. The BMD SE estimated via simulation reflects the variability of M so may be somewhat larger than it would be if M was fixed.

While we found that oversampling performed better than undersampling as a design strategy, we are not suggesting that this will always be the case. The improvement we observed occurs because malformation is positively correlated with lower fetal weight. Effectively, the oversampling approach capitalizes on this relationship between the endpoints by increasing the chance of sampling pups with low fetal weights. This results in a greater likelihood of observing data at the extremes of the fetal weight distribution, allowing for greater stability in estimating dose response and therefore, improved precision for the BMD.

In addition, while improvements were earned with smaller subsets and lower malformation rates, some caution is warranted in extrapolating results. Problems that commonly occur with relatively small data sets, such as poor model convergence, can arise here as well and success is dependent on the malformation rate as well as the subsampling rates. Finally, our discussion of precision is meaningful only for valid estimates of the BMD so data analysis would still need to use weighted estimators in order to avoid bias, which could be significant if the within-stratum sampling rates differ substantially.

Acknowledgments

The authors thank the reviewers for helpful comments that improved the manuscript. This work was supported in part by grant ES07142 from the National Institute of Environmental Health Sciences of the US National Institutes of Health.

APPENDIX

Additional risk under the joint model

In the joint model, the probability of an adverse outcome at dose d is given by

$$P(d) = P(Y_k(d) = 1 \text{ or } S_k(d) < s_c) = P(Y_k^*(d) \leq 0 \text{ or } S_k(d) < s_c) = 1 - P(Y_k^*(d) > 0, S_k(d) > s_c).$$

Additional risk can be written as a difference of upper tail areas of two bivariate normal distributions.

$$\begin{aligned} r(d) &= P(d) - P(0) \\ &= Pr(Y_k^*(0) > 0, S_k(0) > s_c) \\ &\quad - Pr(Y_k^*(d) > 0, S_k(d) > s_c) \\ &= \int_{c_S(0)}^{\infty} \int_{c_Y(0)}^{\infty} \varphi_2(z_1, z_2; r_{YS}) dz_1 dz_2 - \int_{c_S(d)}^{\infty} \int_{c_Y(d)}^{\infty} \varphi_2(z_1, z_2; r_{YS}) dz_1 dz_2. \end{aligned}$$

where

$$c_s(d) = (s_c - \alpha_{21} - \alpha_{22}d) / \sqrt{\tau_1^2 + 1}, c_y(d) = (-\alpha_{11} - \alpha_{12}d) / \sqrt{\tau_2^2 + \sigma^2}, r_{ys} = (\sigma\rho + \tau_1\tau_2) / \sqrt{(\tau_1^2 + 1)(\tau_2^2 + \sigma^2)},$$

and $\phi_2(\cdot; r)$ is the bivariate normal probability density function with correlation r .

REFERENCES

1. Crump KS. A new method for determining allowable daily intakes. *Fundamental and Applied Toxicology*. 1984; 4:854–871. [PubMed: 6510615]
2. Manson, JM. Testing of Pharmaceutical Agents for Reproductive Toxicity. In: Kimmel, C.; Buelke-Sam, J., editors. *Developmental Toxicology*. 2nd Ed. New York: Raven Press; 1994. p. 379–402.
3. Catalano PJ, Ryan LM. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*. 1992; 87:651–658.
4. Catalano PJ, Scharfstein DO, Ryan LM, Kimmel CA, Kimmel GL. Statistical model for fetal death, fetal weight, and malformation in developmental toxicity studies. *Teratology*. 1993; 47:281–290. [PubMed: 8322222]
5. Fitzmaurice GM, Laird NM. Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*. 1995; 90:845–852.
6. Sammel MD, Ryan LM, Legler JM. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society Series B*. 1997; 59:667–678.
7. Regan MM, Catalano PJ. Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology. *Biometrics*. 1999; 55:760–768. [PubMed: 11315004]
8. Dunson DB, Chen Z, Harry J. A bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics*. 2003; 59:521–530. [PubMed: 14601753]
9. Geys H, Molenberghs G, Williams P. Two latent variable risk assessment approaches for mixed continuous and discrete outcomes from developmental toxicity data. *Journal of Agricultural, Biological and Environmental Statistics*. 2001; 6:340–355.
10. Gueorguieva RV, Agresti A. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*. 2001; 96:1102–1112.
11. Kodell RL, Howe RB, Chen JJ, Gaylor DW. Mathematical modeling of reproductive and developmental toxic effects for quantitative risk assessment. *Risk Analysis*. 1991; 11:583–590. [PubMed: 1780500]
12. Allen BC, Kavlock RJ, Kimmel CA, Faustman EM. Dose-response assessment for developmental toxicity III: Statistical models. *Fundamental and Applied Toxicology*. 1994; 23:496–509. [PubMed: 7867901]
13. Kavlock RJ, Allen BC, Faustman EM, Kimmel CA. Dose-response assessment for developmental toxicity IV: Benchmark doses for fetal weight changes. *Fundamental and Applied Toxicology*. 1995; 26:211–222. [PubMed: 7589910]
14. Office of Prevention Pesticides and Toxic Substances. Health effects test guidelines OPPTS 870.3700 Prenatal developmental toxicity study. United States Environmental Protection Agency; 1998. EPA publication no. 712-C-98-207.
15. Weller E, Long N, Smith A, Williams P, Ravi S, Gill J, et al. Dose-rate effects of ethylene oxide exposure on developmental toxicity. *Toxicological Sciences*. 1999; 50:259–270. [PubMed: 10478863]
16. Scott, A.; Wild, C. Population-based case-control studies. In: Pfeiffermann, D.; Rao, CR., editors. *Handbook of Statistics, Vol. 29B, Sample surveys: Inference and Analysis*. Oxford: Elsevier; 2009. p. 431–453.
17. Deming WE. An essay on screening, or on two-phase sampling, applied to surveys of a community. *International Statistical Review*. 1977; 45:29–37.
18. Shrout PE, Newman SC. Design of two-phase prevalence surveys of rare disorders. *Biometrics*. 1989; 45:549–555. [PubMed: 2765638]
19. Najita JS, Li Y, Catalano PJ. A novel application of a bivariate regression model for binary and continuous outcomes to studies of fetal toxicity. *Applied Statistics*. 2009; 58:555–573. [PubMed: 20357904]

20. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics*. 1994; 50:933–944. [PubMed: 7787006]
21. Heagerty PJ, Lele SR. A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*. 1998; 93:1099–1111.
22. Crump, KS.; Howe, RB. A review of methods for calculating statistical confidence limits in low dose extrapolation. In: Clayson, DB.; Krewski, D.; Monro, I., editors. *Toxicological risk assessment: Vol 1. Biological and statistical criteria*. Boca Raton, FL: CRC Press; 1985. p. 187-203.
23. Kimmel CA, Gaylor DW. Issues in qualitative and quantitative risk analysis for developmental toxicology. *Risk Analysis*. 1988; 8:15–20. [PubMed: 3375503]
24. Chen JJ, Kodell RL. Quantitative risk assessment for teratologic effects. *Biometrics*. 1989; 47:966–971.
25. Ryan LM. Quantitative risk assessment for developmental toxicity. *Biometrics*. 1992; 48:163–174. [PubMed: 1581483]
26. Gaylor D, Ryan L, Krewski D, Zhu Y. Procedures for calculating benchmark doses for health risk assessment. *Regulatory Toxicology and Pharmacology*. 1998; 28:150–164. [PubMed: 9927564]
27. Romero A, Villamayor F, Grau MT, Sacristan A, Ortiz JA. Relationship between fetal weight and litter size in rats: application to reproductive toxicology studies. *Reproductive Toxicology*. 1992; 6:453–456. [PubMed: 1463926]
28. Chen JJ. A malformation incidence dose-response model incorporating fetal weight and/or litter size as covariates. *Risk Analysis*. 1993; 13:559–564. [PubMed: 8259446]
29. Chen JJ, Gaylor DW. The correlations of developmental endpoints observed from 2,4,5-Trichlorophenoxyacetic acid exposure. *Teratology*. 1992; 45:241–246. [PubMed: 1631778]
30. Rai K, Ryzin JV. A dose-response model for teratological experiments involving quantal responses. *Biometrics*. 1985; 41:1–9. [PubMed: 4005367]
31. Williams DA. Dose-response models for teratological experiments. *Biometrics*. 1987; 43:1013–1016. [PubMed: 3427158]

(a) Scenario A. $P_m=0.35$.(b) Scenario B. $P_m=0.25$.**Fig. 1.**

Relative precision for various subsampling rates. Relative precision is defined as the ratios of the BMD SE comparing SRS to subsampling ($RP_{SRS} = SE_{SRS}/SE_{sub}$). $RP_{SRS} > 1$ indicates designs where there is higher precision with subsampling. Plots show relative precision in scenarios A and B.

Table 1

Relative precision of the BMD with subsampling with malformation rate $P_m=0.35$ and $N=1500$ pups (Scenario A). Each row corresponds to a distinct subsampling design. Column 3 indicates the type of subsampling that was simulated (U for undersampling, S for SRS, and O for oversampling). Model-based and empirical SE estimates were 0.008 and 0.008 respectively for the complete data case. For each design, 500 experiments were simulated.

row	M/N	sampling	p_1	p_2	M (mean)	model SE	emp SE	RP_c	RP_{SRS}
1	0.8	U	0.5	0.962	1201.9	0.0107	0.0100	0.80	0.88
2		U	0.6	0.908	1201.6	0.0099	0.0096	0.83	0.92
3		U	0.7	0.854	1200.8	0.0094	0.0091	0.88	0.97
4		S	0.8	0.800	1200.2	0.0085	0.0088	0.91	1.00
5		O	0.9	0.746	1199.8	0.0089	0.0086	0.93	1.02
6	0.7	U	0.2	0.969	1050.9	0.0172	0.0169	0.47	0.57
7		U	0.3	0.915	1050.7	0.0139	0.0148	0.54	0.66
8		U	0.4	0.862	1050.8	0.0120	0.0121	0.66	0.80
9		U	0.5	0.808	1051.3	0.0110	0.0103	0.78	0.94
10		U	0.6	0.754	1051.0	0.0103	0.0101	0.79	0.96
11		S	0.7	0.700	1050.9	0.0089	0.0097	0.82	1.00
12		O	0.8	0.646	1050.4	0.0098	0.0097	0.82	1.00
13		O	0.9	0.592	1050.7	0.0097	0.0094	0.85	1.03
14	0.6	U	0.2	0.815	900.7	0.0172	0.0172	0.47	0.62
15		U	0.3	0.762	900.8	0.0141	0.0152	0.53	0.70
16		U	0.4	0.708	900.6	0.0124	0.0125	0.64	0.85
17		U	0.5	0.654	901.7	0.0115	0.0111	0.72	0.95
18		S	0.6	0.600	901.7	0.0094	0.0106	0.75	1.00
19		O	0.7	0.546	900.8	0.0108	0.0103	0.78	1.03
20		O	0.8	0.492	898.3	0.0108	0.0106	0.75	1.00
21		O	0.9	0.438	897.8	0.0110	0.0105	0.76	1.01
22	0.5	U	0.2	0.662	755.9	0.0173	0.0173	0.46	0.73
23		U	0.3	0.608	751.4	0.0145	0.0155	0.52	0.81
24		U	0.4	0.554	750.9	0.0131	0.0130	0.62	0.97
25		S	0.5	0.500	750.0	0.0100	0.0126	0.63	1.00
26		O	0.6	0.446	749.1	0.0121	0.0115	0.70	1.10

row	M/N	sampling	p_1	p_2	M (mean)	model SE	emp SE	RP_ϵ	RP_{SRS}
27		O	0.7	0.392	748.6	0.0122	0.0118	0.68	1.07
28		O	0.8	0.338	748.2	0.0127	0.0121	0.66	1.04
29		O	0.9	0.285	749.4	0.0135	0.0125	0.64	1.01
30	0.4	U	0.2	0.508	601.4	0.0178	0.0180	0.44	0.79
31		U	0.3	0.454	599.0	0.0154	0.0160	0.50	0.89
32		S	0.4	0.400	598.3	0.0111	0.0143	0.56	1.00
33		O	0.5	0.346	599.4	0.0140	0.0132	0.61	1.08
34		O	0.6	0.292	599.7	0.0143	0.0132	0.61	1.08
35		O	0.7	0.239	599.7	0.0152	0.0145	0.55	0.99
36		O	0.8	0.185	598.9	0.0166	0.0165	0.48	0.87
37	0.3	O	0.2	0.354	449.4	0.0186	0.0183	0.44	1.01
38		O	0.3	0.300	449.1	0.0130	0.0184	0.43	1.00
39		O	0.4	0.246	449.2	0.0166	0.0160	0.50	1.15

Table II

Relative precision of the BMD with subsampling with malformation rate $P_m=0.25$ and $N=1500$ pups (Scenario B). Each row corresponds to a distinct subsampling design. Column 3 indicates the type of subsampling that was simulated (U for undersampling, S for SRS, and O for oversampling). Model-based and empirical SE estimates were 0.0097 and 0.0093 respectively for the complete data case. For each design, 500 experiments were simulated.

row	M/N	sampling	p_1	p_2	M (mean)	model SE	emp SE	RP_c	RP_{SRS}
1	0.8	U	0.3	0.967	1197.0	0.0167	0.0178	0.52	0.61
2		U	0.4	0.933	1197.7	0.0143	0.0148	0.63	0.74
3		U	0.5	0.900	1199.2	0.0129	0.0123	0.76	0.89
4		U	0.6	0.867	1199.4	0.0119	0.0118	0.79	0.92
5		U	0.7	0.833	1199.5	0.0113	0.0113	0.82	0.96
6		S	0.8	0.800	1200.2	0.0103	0.0109	0.85	1.00
7		O	0.9	0.767	1200.8	0.0106	0.0106	0.88	1.03
8	0.7	U	0.2	0.867	1047.4	0.0201	0.0200	0.47	0.60
9		U	0.3	0.833	1047.7	0.0168	0.0180	0.52	0.67
10		U	0.4	0.800	1048.1	0.0145	0.0151	0.62	0.79
11		U	0.5	0.767	1049.2	0.0132	0.0128	0.73	0.94
12		U	0.6	0.733	1049.9	0.0123	0.0125	0.74	0.96
13		S	0.7	0.700	1050.9	0.0107	0.0120	0.78	1.00
14		O	0.8	0.667	1051.9	0.0115	0.0116	0.80	1.03
15		O	0.9	0.633	1052.4	0.0113	0.0113	0.82	1.06
16	0.6	U	0.2	0.733	897.9	0.0202	0.0203	0.46	0.65
17		U	0.3	0.700	899.1	0.0171	0.0183	0.51	0.72
18		U	0.4	0.667	899.8	0.0149	0.0155	0.60	0.85
19		U	0.5	0.633	900.5	0.0138	0.0135	0.69	0.97
20		S	0.6	0.600	901.7	0.0114	0.0131	0.71	1.00
21		O	0.7	0.567	902.1	0.0126	0.0125	0.74	1.05
22		O	0.8	0.533	901.8	0.0125	0.0125	0.74	1.05
23		O	0.9	0.500	903.0	0.0124	0.0123	0.76	1.07
24	0.5	U	0.2	0.600	749.6	0.0202	0.0205	0.45	0.70
25		U	0.3	0.567	750.3	0.0175	0.0188	0.49	0.77
26		U	0.4	0.533	749.8	0.0156	0.0161	0.58	0.89
27		S	0.5	0.500	751.4	0.0123	0.0144	0.65	1.00

row	M/N	sampling	p_1	p_2	M (mean)	model SE	emp SE	RP_ϵ	RP_{SRS}
28		O	0.6	0.467	750.2	0.0141	0.0133	0.70	1.08
29		O	0.7	0.433	749.7	0.0139	0.0133	0.70	1.08
30		O	0.8	0.400	750.4	0.0139	0.0132	0.70	1.09
31		O	0.9	0.367	751.3	0.0141	0.0134	0.69	1.07
32	0.4	U	0.2	0.467	598.2	0.0211	0.0223	0.42	0.82
33		U	0.3	0.433	597.9	0.0189	0.0202	0.46	0.90
34		S	0.4	0.400	598.3	0.0136	0.0182	0.51	1.00
35		O	0.5	0.367	599.8	0.0161	0.0149	0.62	1.22
36		O	0.6	0.333	600.5	0.0160	0.0146	0.64	1.25
37		O	0.7	0.300	600.9	0.0162	0.0149	0.62	1.22
38		O	0.8	0.267	602.1	0.0166	0.0149	0.62	1.22
39		O	0.9	0.233	603.0	0.0175	0.0160	0.58	1.14
40	0.3	U	0.2	0.333	448.4	0.0215	0.0223	0.42	1.00
41		S	0.3	0.300	449.1	0.0158	0.0224	0.42	1.00
42		O	0.4	0.267	450.1	0.0192	0.0193	0.48	1.16
43		O	0.5	0.233	451.4	0.0191	0.0171	0.54	1.31
44		O	0.6	0.200	451.9	0.0195	0.0184	0.51	1.22
45		O	0.7	0.167	452.3	0.0203	0.0198	0.47	1.13
46		O	0.8	0.133	452.6	0.0209	0.0213	0.44	1.05