Ordinal dose-response modeling approach for the phthalate syndrome

Todd D. Blessinger\textsuperscript{a,d,*}, Susan Y. Euling\textsuperscript{b,d}, Lily Wang\textsuperscript{e}, Karen A. Hogan\textsuperscript{a,d}, Christine Cai\textsuperscript{a,d}, Gary Klinefelter\textsuperscript{c,d}, Anne-Marie Saillenfait\textsuperscript{f}

\textsuperscript{a}Center for Public Health and Environmental Assessment (CPHEA), Mail code 8623R, USA
\textsuperscript{b}Office of Children's Health Protection (OCHP), Mail code 1107T, USA
\textsuperscript{c}National Health and Environmental Effects Research Lab (NHEERL), USA
\textsuperscript{d}United States (US) Environmental Protection Agency, Washington, DC 20460, USA
\textsuperscript{e}CPHEA, US Environmental Protection Agency, Mail code B243-01, Research Triangle Park, NC 27711, USA
\textsuperscript{f}Institut National de Recherche et de Sécurité, Vandoeuvre, France

Abstract

**Background:** The phthalate syndrome (PS) is a collection of related male reproductive developmental effects, ranging in severity, that have been observed in rats after gestational exposure to developmentally-toxic phthalates. For statistical purposes, the PS is defined as a single endpoint and one dose-response analysis is conducted, rather than conducting multiple analyses on each individual endpoint.

**Objective:** To improve dose-response modeling approaches for the PS and other syndromes of effects by accounting for differing severity levels among the endpoints.

**Methods:** Ordinal dose-response modeling was performed on PS data from a published study of diisobutyl phthalate (DIBP) gestational exposure to male Sprague-Dawley rats. To incorporate PS endpoint severity, the endpoints were categorized into ordinal levels based on the expected impact of male developmental endpoint’s on fertility. Then, a benchmark dose was estimated for each ordinal level. A bootstrap procedure was used to account for the nested nature of the data, and a sensitivity analysis was performed to assess the bootstrap results. A comparison of the estimates between the ordinal and the dichotomous model was performed.

**Results:** The ordinal version of the log-logistic model applied to the data categorized by PS endpoint severity level provided benchmark dose estimates that were closer to each other in value and had lower variability than the traditional dichotomous application. The sensitivity analysis confirmed the validity of the bootstrap results.

\*Corresponding author at: United States (US) Environmental Protection Agency, Washington, DC 20460, USA.
blessinger.todd@epa.gov (T.D. Blessinger).

Declaration of Competing Interest
None.
Conclusion: The ordinal dose-response modeling method accounts for severity differences among dichotomous PS endpoints, can be expanded in the future to include more severity levels, and can be used in both single and cumulative phthalate risk assessments.

Keywords
Phthalate syndrome; Ordinal data analysis; Dose-response modeling; Male reproductive development; Nested model; Diisobutyl phthalate

1. Introduction
Phthalates esters (referred to as “phthalates”) are used as plasticizers and additives for PVC products, nitrocellulose, and concrete (US EPA, 2015). As a result of these uses, phthalates are found in numerous consumer products such as nail polish, makeup, and carpets (US EPA, 2015). Exposure to developmentally toxic phthalates, phthalates with three to seven (or eight) carbon atoms in the backbone of the alkyl side chain, can lead to developmental effects (CPSC, 2014). These phthalates include di-n-pentyl (DPENP) (diamyl phthalate), butylbenzyl (BBP), dibutyl (DBP), diisobutyl (DIBP), dihexyl (DHEXP), di(2-ethyl-hexyl) (DEHP), dicyclohexyl (DCHP), and diisononyl (DINP) (US CPSC, 2014). Further, the human population, including pregnant women, are exposed to multiple phthalates (Koch et al., 2017; Zota et al., 2014; Wittassek et al., 2009; Ye et al., 2008). Therefore, developmentally toxic phthalates are chemicals of concern for children because of the combination of increased cumulative phthalate exposure levels and increased susceptibility during a critical window of development is expected to lead to a greater risk.

Gestational exposure to one or more developmentally toxic phthalates can lead to a set of male reproductive developmental health effects collectively called the phthalate syndrome (PS) in rats (reviewed in Foster, 2006; Fisher et al., 2003; Foster et al., 2001). The phthalate syndrome is characterized by malformations of male reproductive organs (e.g., epididymis, vas deferens, seminal vesicles, prostate, and external genitalia) such as hypospadias and cryptorchidism and developmental effects such as retention of nipples/areolae (sexually dimorphic structures in rodents) and lack of masculinization of the perineum during development, leading to a reduced anogenital distance (AGD) (US CPSC, 2014). The more severe PS endpoints of the male reproductive tract (i.e., malformations) are observed at higher phthalate doses whereas changes in AGD and nipple/areolae retention can be observed at lower phthalate doses as well. Critical windows of exposure for the PS have been defined in rat studies. Gestation is the most sensitive, neonatal and peripubertal exposure is less sensitive, and adulthood is the least sensitive window of exposure (US CPSC, 2014). The modes of action for the PS effects observed in rats after gestational phthalate exposure are due to a decrease in fetal testicular androgens, decreased INSL3 signaling, and fetal germ cell developmental effects (e.g., multinucleated gonocytes) that have not been defined at the molecular level (Makris et al., 2013; van den Driesche et al., 2015; Boekelheide et al., 2009; Howdeshell et al., 2017).

In response to evidence that humans are exposed to multiple phthalates, the National Academy of Sciences convened experts to provide recommendations regarding cumulative
risk, resulting in the *Phthalates and Cumulative Risk Assessment* report (NRC, 2008). The committee recommended that a cumulative risk assessment be performed using an approach that evaluates common adverse effects from exposure to phthalates as well as to other chemicals that affect androgen activity. In 2008, Congress banned children’s toys and child care articles containing more than 0.1% of DEHP, DBP, and BBP. The U.S. Consumer Product Safety Commission (CPSC)’s Chronic Hazard Advisory Panel (CHAP) on Phthalates and Phthalate Alternatives developed a risk assessment for eight developmentally toxic phthalates (US CPSC, 2014; Review by Lioy et al., 2015), leading to a ban on children’s toys and child care articles containing more than 0.1% of five additional phthalates (DINP, DPENP, DHEXP, DCHP, and DIBP). In 2014, EPA added seven phthalates that are still in commerce to their Toxic Substances Control Act (TSCA) Work Plan List for Chemical Assessments list (US EPA, 2014) to determine whether any additional assessment is needed with regard to TSCA-specific uses or exposure scenarios.

There have been efforts to perform dose-response modeling by defining the PS as a single endpoint, as opposed to modeling each individual endpoint separately. In this approach, PS animals are defined as those exhibiting none vs. one or more PS endpoints. In one such evaluation of the PS data after perinatal (part of gestation and part of lactation) di(2-ethylhexyl)phthalate (DEHP) exposure to rats, Gray et al. (2009) designated an animal as having the PS if it had at least one of the PS effects, and found a statistically significant increase in the PS at doses ranging from 11 to 300 mg/kg-day. However, modeling data as dichotomous does not account for differences in severity that exist among PS effects. For example, the consequence of some malformations (e.g., hypospadias, undescended testis) can lead to abnormal reproductive function, impacting fertility. By contrast, other PS endpoints lead to abnormal developmental effects (e.g., retained areolas or nipples) that do not have anticipated impacts on fertility. Therefore, an analysis that incorporates these differences is preferred.

Dose-response modeling of ordinal outcomes has been applied in some recent cases, for example, using U.S. EPA’s CatReg software (see, for example, Krewski et al., 2010). In these analyses, CatReg was applied as a meta-analysis in which results from multiple endpoints, often observed across multiple studies, were assigned severity scores based on expert judgement of biological severity and the scores were combined into a single analysis. Chen and Chen (2014) proposed a method to assign ordinal levels based on a latent continuous variable. In addition, the NRC *Phthalates and Cumulative Risk Assessment* report (NRC, 2008; App D) conducted an analysis in which five continuous PS endpoints that are shared between antiandrogen and phthalate chemicals were modeled as part of an antiandrogen chemical cumulative risk assessment case study. The observed responses of five selected endpoints were transformed based on the desirability of response, as determined using consensus expert opinion, and then combined into a unitless score on the interval from 0 to 1, which was analyzed as a continuous, rather than dichotomous, endpoint.

For the analysis presented here, a straightforward extension of the dichotomous log-logistic model to the ordinal case was applied to the individual animal data from a study in which multiple PS endpoints were assessed in male Sprague-Dawley rats exposed to DIBP during...
gestation (Saillenfait et al., 2008). The ordinal scale was defined by designating two levels of severity for PS endpoints based on knowledge about its relationship to male fertility.

In risk assessment, development of human health risk values, such as reference doses (RfDs) used by EPA, involves the fitting of dose-response models suited to the data type, as well as selection of a low level of response (benchmark response, or BMR) that supports identifying an exposure level (i.e., a benchmark dose, or BMD) that is health protective. Often, the lower confidence bound on the BMD, called the BMDL, is used as a point of departure (POD), which is divided by a series of uncertainty factors to determine a RfD (U.S. EPA, 2002, 2012). When conducting dose-response modeling to develop a risk value for a dichotomous endpoint, a BMD is estimated at only one BMR. However, for the phthalate syndrome, the range of biological significance across the spectrum of different severity PS endpoints supports the development of different BMRs based on the anticipated severity level. Therefore, the BMDs for the two ordinal levels were estimated at different BMRs. The results from the ordinal analysis were then compared with those from dichotomous characterizations of the same data.

2. Materials and methods

2.1. Individual animal PS dataset

The timing of DIBP exposure and endpoint assessment for the Saillenfait et al. (2008) study design are illustrated in Fig. 1. Additional details about the study design, animals, the DIBP and control dose groups, and endpoint assessment methods can be found in Saillenfait et al. (2008). In this study, males were exposed from gestation days (GD) 12–21 by gavage to DIBP and were then sacrificed and necropsied at one of two adult ages, either postnatal day (PND) 76–86 (~postnatal week 11–12, PNW, two males in each litter) or on PND 111–122 (~PNW 16–17, the remaining males in each litter), and were examined for the following endpoints –

- Nipple and areola retention
- Gross abnormalities of external and internal genitalia
- Testes descent
- Histopathological lesions in the testes and epididymis (on either PND 76–86 or on PND 111–122)
- Male reproductive organ weights

Individual animal data from Saillenfait et al. (2008) were provided by the study director, Dr. Saillenfait (co-author). These summary data were reported on a per-litter basis by treatment group in the published article.

2.2. Development of ordinal scale

PS endpoints that were measured as continuous variables, such as anogenital distance (AGD), age of preputial separation (PPS), and male reproductive organ weight decreases, but had no clinical or biological method for assigning ordinal level categories were not included in this analysis. The PS endpoints that were measured as dichotomous variables
were therefore included in this analysis. Then, among the dichotomous PS endpoints, they can be characterized as a collection of permanent male reproductive developmental endpoints that are associated with either zero to moderate expected impacts on fertility (level 1), or are associated with severe infertility (level 2). It is important to note that male fertility was not assessed in the Saillenfait et al. (2008) study; sperm quality, sperm quantity, and mating endpoints were not included. Instead, information from experts and the published literature about associations between male organ histopathological endpoints and infertility was utilized. Infertility is a severe effect because a lack of offspring is equivalent to mortality in the next generation. In order to designate an endpoint according to its expected impact on fertility, as delineated above, the expert judgment of toxicologists was used. The ordinal level binning decisions were based on discussions with experts in male development and reproductive toxicology, Gary Klinefelter (co-author), Anne-Marie Saillenfait (co-author), and NCEA’s Reproductive and Developmental Toxicity and Neurotoxicity Workgroup, and these decisions are consistent with recommendations of toxicological pathologists (Lanning et al., 2002). The expert evaluation excluded three male reproductive histopathology endpoints, amorphous crystalline material, distorted tubules, and dilated tubule lumen, based on knowledge that sometimes these endpoints may not necessarily be treatment-related (e.g., poor fixation). Therefore, the six animals that exhibited these three endpoints and no other PS endpoints were designated as not having the PS.

For the purposes of this analysis, the PS was defined as an ordinal outcome with one of three separate levels, as follows:

- **Level 0**: The animal has no PS endpoints (i.e., no male reproductive developmental effects and no male reproductive organ histopathological effects of score 2–5).
- **Level 1**: The animal has one or more PS endpoints that are associated with zero to moderate expected impacts on fertility, but has no PS endpoints associated with severe impacts on fertility, based on expert judgement.
- **Level 2**: The animal has one or more PS endpoints that are associated with severe infertility, based on expert judgement.

Animals sacrificed at PNW 11–12 were included in the analysis because all had histopathology examination data (Table 5, Saillenfait et al., 2008) that had gone through peer review of that article. In contrast, animals sacrificed at PNW 16–17 animals were not included because a smaller sample size was assessed for histopathology than the PNW 11–12 group and because the PNW 16–17 histopathology data was not all reported in the Saillenfait et al. (2008) article and therefore did not undergo peer review. To confirm the influence of the histopathological endpoints on the PS levels, the proportion of animals sacrificed at PNW 11–12 whose PS ordinal level classification was increased by the inclusion of histopathology endpoints was calculated, for each dose group.

In order to compare the results of the ordinal method to a dichotomous PS modeling method, the PS was modeled as a single dichotomous endpoint with two different definitions for PS response:
• Any PS: an animal was designated as PS if it had one or more PS endpoints, regardless of severity (analogous to combining levels 1 and 2).

• Severe PS: an animal was designated as PS if it had at least one PS endpoint that is associated with severe infertility (analogous to using only level 2).

2.3. Benchmark response selection to estimate benchmark doses

Among other considerations, BMR selections should generally identify levels of minimal biological significance, with a different BMR corresponding to each ordinal level to address the difference in severity among the PS endpoints. For demonstration purposes, BMRs of 5% and 1% extra risk were selected for ordinal levels 1 (developmental effects associated with zero to moderate anticipated impacts on fertility) and 2 (associated with a severe infertility), respectively, based partly on EPA recommendations regarding biological considerations for determining reference values (US EPA, 2012).

3. Theory and calculation

A log-logistic model was used to model the cumulative probability that the PS level was at least as high as the one observed. That is, for observed ordinal PS level \( Y \), dose \( x \), and level \( j \) for \( j = 1, 2 \), the cumulative probability was defined as

\[
P[Y \geq j | x] = \gamma_j + \frac{1 - \gamma_j}{1 + \exp[-\log(v/(1 - v)) + \beta(\theta_v, j - \log x)]},
\]

where \( v = \text{BMR}, \ \theta_v, j = \log(\text{BMD}_j) \) at BMR \( v \), \( \beta = \text{slope} \), and \( \gamma_j = \text{background response rate} \).

This model was chosen because it exhibits greater flexibility in model fitting than some other models such as the logistic. The model form in Eq. (1) does not use the more common intercept-slope parameterization, which is used in EPA’s Benchmark Dose Software (BMDS), but instead incorporates the log-BMD as an explicit parameter for ease of estimating the BMD. The relationship between the two model forms is described in more detail in Appendix A. Also observe that the slope was assumed to be constant across levels to maintain monotonicity of the cumulative probability per dose across levels.

The BMDs of interest in this application are the level 1 BMD\(_{0.05,1}\), \( \theta_{0.05,1} \), and the level 2 BMD\(_{0.01,2}\), \( \theta_{0.01,2} \). Model parameters were estimated using \( v = 0.01 \), so the BMD parameters were estimated initially as \( \theta_{0.01,1} \) and \( \theta_{0.01,2} \). Then \( \theta_{0.05,1} \) was estimated using the relationship

\[
\theta_{0.05,1} = \theta_{0.01,1} + \frac{c}{\beta}.
\]

where \( c = \log\left(\frac{0.05/0.95}{0.01/0.99}\right) \). The derivation of this relationship is provided in Appendix A.

As discussed in Section 2.3, the model defined in (1) assumes that the observations are independent, but, in reality, intra-litter correlation exist. Therefore, litter should be included in the model as a random effect because the set of litters that comprises the sample is a random selection from the population of litters. Several mixed model methods exist for modeling ordinal outcomes in the presence of random effects, but they generally require a
sufficiently large number of animals per random observation (e.g., EPA’s CatReg software, which uses Wald-type estimates). Only two animals per litter (those sacrificed at PNW 11–12) were included in the analysis, so the litters were not large enough to support using these methods. Therefore, a bootstrap resampling method was used instead.

Bootstrap resampling is a method of estimating parameters in which a large number of samples are selected with replacement from the original sample, and a parameter estimate is derived for each resample (see, for example, Efron and Tibshirani, 1994). The PS dataset consists of a two-level nested design, with animal nested within litter and litter treated as a random effect, so it was necessary to use a method of selecting samples that incorporates this design. Van der Leeden et al. (2007) presents three such methods with varying degrees of assumptions. The first two methods require restrictive assumptions on the analysis, such as a linear relationship between the response and model effects, which do not apply to our case. Therefore, the third method, called the “cases bootstrap,” was considered most applicable. Under this method, a large number of samples were selected with replacement from the original PS dataset in the following manner. Assuming that there were $n_i$ litters exposed to each dose $x_i$, for each bootstrap iteration samples were selected by 1) selecting $n_i$ litters with replacement from the set of litters exposed to dose $x_i$ and 2) for each of the litters selected in the first step, selecting 2 animals with replacement from that litter. Then, the model in (1) was fit to each bootstrap sample and estimates of the parameters computed, including an estimate of $\log(BMD)$. These steps were repeated $B$ times (in this case, $B = 10,000$) to yield $B$ samples and $\log(BMD)$ estimates, from which summary statistics such as the mean, standard deviation, and relevant percentiles were computed. The cases bootstrap procedure is described in more detail in Appendix B.

This bootstrap procedure is analogous to fitting the model defined by Eq. (1), where for each level $j$ and dose group $i$ the parameter $\theta_j$ is replaced with $\theta_j + u_{k(i)}$, where $u_{k(i)}$ is the random effect of the $k$th litter nested within the $i$th dose group, $i = 1, ..., n_i$. The selection of the bootstrap samples was done in SAS version 9.4, and the analysis was conducted in SAS Proc NLMIXED. The SAS code and description are provided at Dataset (2019).

Upon fitting the dose-response model to the bootstrap samples, simultaneous confidence intervals for the level 1 BMD$_{0.05}$ and level 2 BMD$_{0.01}$ were computed using an adjusted percentile method developed by Mandel and Betensky (2008). Details of this method are provided in Appendix B. The confidence intervals derived from this method have simultaneous coverage of their respective BMDs. Thus, for example, the probability that the 90% confidence intervals for the level 1 BMD$_{0.05}$ and level 2 BMD$_{0.01}$ both contain their respective parameters is at least 90%. In addition to the Mandel & Betensky intervals, the Bonferroni-adjusted confidence intervals were computed for comparison because the Bonferroni method generally is the most conservative adjustment method, yielding the widest possible simultaneous confidence intervals; thus, it provides a good measure of the most extreme case. Also, the non-simultaneous (“unadjusted”) confidence intervals were computed to demonstrate the narrowest possible intervals.

Because the litter sizes were so small, some sensitivity bootstrap analyses were conducted to assess the adequacy of the nested model bootstrap method used and determine if the effect
of litter was discernible in the results. In one analysis, the bootstrap method was applied to the original PS dataset disregarding litter. More specifically, a large number of samples (B=10,000) were selected with replacement from the PS dataset without regard to litter assignment, and the model in (1) was fit to each of these bootstrap samples and parameter estimates computed. In a second analysis, three permutations of the original PS dataset were selected with animals randomly assigned to litters, and the bootstrap analysis with the litter effect included was applied to each permuted sample. The results of these two analyses were compared to the original analysis results. In addition, to demonstrate the application of the model, the model in (1) was applied to the original PS dataset without regard to litter and with no resampling, and the model was fit and parameters estimated using normal-based methods. The equality of slopes across levels was tested in this analysis using an F-test, and the adequacy of the model was tested using a chi-square test.

For comparison to the ordinal dose-response results, a dichotomous dose-response analysis was conducted using the dichotomous PS definitions described in Section 2.2. The modeling was conducted in BMDS version 2.7 using the nested logistic model to incorporate the nested litter design (U.S. EPA, 2016; Allen et al., 1994; Kupper et al., 1986). For each analysis, the model was fit both with and without the dam weight measured immediately prior to dosing included as a covariate, and litter correlations were estimated. The model fit that yielded the lowest AIC value was selected for BMD estimation. In addition, to compare the ordinal PS results to the results from more traditional dose-response methods, selected individual PS endpoints were modeled using this dichotomous analysis.

4. Results

4.1. Groupings of PS endpoints into ordinal levels

The PS endpoints reported by Saillenfait et al. (2008) and included in this analysis (Table 1) were classified according to their developmental effects and their expected associations with fertility, either characterized as zero to moderate anticipated impacts (level 1) or severe impacts (level 2), based on expert input (see Section 2.2). For example, areola/nipple retention is a marker for decreased androgen activity following gestational exposure but is not associated with decreased fertility in adults.

Application of the ordinal PS classifications in Table 1 to the developmental reproductive data for male offspring evaluated at PNW 11–12 is summarized in Table 2. Incidence of the PS (levels 1 and 2 combined) increased in a dose-related manner from 0% in the control to 100% in the highest dose group (625 mg/kg-day), and incidence of PS level 2 increased up to 85% in the highest dose group.

The ordinal level was changed by including histopathology endpoints for 14 out of 114 animals (12%) sacrificed at PNW 11–12. Eleven (11), 2, and 1 animals had levels that increased from 0 to 1, 0 to 2, and 1 to 2, respectively, due to the presence of histopathology endpoints, the most common of these being azoospermia, oligospermia, sloughed cells, and seminiferous tubule degeneration-atrophy/hypoplasia (a list of these animals is provided in Appendix C). The 250 and 500 mg/kg-day groups had more of these animals than the 125 and 625 mg/kg-day groups, probably because the middle dose groups had a more even
spread across the ordinal levels than the extreme dose groups. The percent of animals with increased PS levels was considered non-negligible, so it was determined that the histopathology endpoints had a significant influence on the PS level assignments. Thus, the PNW 16–17-sacrificed animals were excluded from the analysis.

4.2. Dose-response analysis of ordinal PS dataset

4.2.1. Litter effect excluded—From the analysis of the ordinal dataset with litter effect excluded, the model exhibited an adequate fit to the data (Table 3; p = 0.213), and the slope was significantly greater than zero (p = 0.009), indicating a significant dose-response relationship. A visual inspection of the fit overlaid on the scatterplot of the dose-response data (Fig. 2ab) also indicates a strong dose-response relationship; the residual plots (Fig. 2cd) also confirm the appropriateness of the log-logistic model for this dataset.

The estimates and SEs of the level 1 BMD\textsubscript{05} and level 2 BMD\textsubscript{01} from this analysis and their SEs were similar in magnitude (Table 4, denoted “normal-based”), with the level 2 BMD\textsubscript{01} being slightly higher than the level 1 BMD\textsubscript{05}. Although the level 2 BMD\textsubscript{01} had slightly wider unadjusted and Bonferroni-adjusted confidence intervals (Table 4, denoted “normal-based”), mostly at the high end, the intervals were not especially different in width.

4.2.2. Bootstrap analysis—The level 1 BMD\textsubscript{05} and level 2 BMD\textsubscript{01} estimates from the 10,000 bootstrap runs (Table 4, labeled “bootstrap”) were higher than their corresponding normal-based BMD estimates because of skewness in the distributions of the bootstrap BMD estimates. Also, the bootstrap standard errors of the BMDs (Table 4, labeled “bootstrap”) were much higher than their normal-based counterparts because intra-litter correlation resulted in increased variability.

The histograms of the bootstrap distributions of the level 1 BMD\textsubscript{05} and level 2 BMD\textsubscript{01} (Fig. 3ab) exhibit significant right skewness, which is common for simulated BMD distributions in our experience. In addition, the distributions are multimodal. Besides the large mode left of center, they each have two additional, smaller modes near the high end. For both distributions, approximately 16% of the BMDs were represented in the two smaller peaks. It is not clear why the distributions exhibit these two extra modes, but it is possible that the extra modes are a result of the limitations of the dataset itself, in the context of the bootstrap sampling, and are not a reflection of the true distribution of the BMDs. The histograms of the bootstrap distributions of the log-transformed level 1 BMD\textsubscript{05} and level 2 BMD\textsubscript{01} (Fig. 3cd) are slightly left skewed, but much less skewed than the untransformed BMD distributions. Because the logarithm is an increasing function, the log-transformed BMDs also exhibit two extra modes at the high end. The departures of the untransformed BMD distributions from normality were substantial, making non-normal-based confidence intervals such as the Mandel-Betensky intervals particularly useful.

The Mandel & Betensky bootstrap intervals (Table 4, labeled “M&B”) were considerably wider than their respective normal-based intervals, especially at the high end, which reflects the increased variability due to the random litter effect. The Bonferroni intervals were wider than the unadjusted and Mandel & Betensky intervals, but overall the three sets of intervals are similar for each BMD, for both the 90% and 95% cases. A forest plot of the confidence
intervals (Fig. 4) demonstrates this similarity and the difference with the normal-based intervals.

For determining RfDs, the lower bounds of these intervals are often used as the PODs; the lower bounds (BMDLs) are used to account for the statistical error in the BMD estimate. The three bootstrap BMDLs in Table 4 are reasonably similar for each BMD, with up to a 13% difference. Ultimately, the Mandel & Betensky intervals are only slightly more conservative than the unadjusted intervals and provide reliable lower bounds to use as PODs.

4.2.3. Sensitivity bootstrap analysis—The sensitivity bootstrap analysis with the litter effect disregarded yielded level 1 BMD$_{05}$ and level 2 BMD$_{01}$ estimates of 204 and 220, respectively, which were 5–6% lower than the corresponding estimates in the original bootstrap analysis, and the SEs were 75.6 and 77.0, respectively, which were 17–18% lower than in the original bootstrap analysis. Although these statistics did not differ greatly across the two analyses, they differed enough to demonstrate that the inclusion of the litter effect was warranted.

The bootstrap analyses on the three permuted samples yielded BMD estimates that were very close (0–1.5% lower) to those from the original analysis and SEs that were somewhat close but slightly more discrepant (3–7% lower). These results demonstrate that the litter assignments in the originally observed dataset have a slight but noticeable influence on the estimated BMD distribution, thus verifying the influence of the litter effect. Furthermore, the differences for the permuted samples were smaller than for the analysis disregarding litter, thus providing additional evidence that the inclusion of the litter effect was warranted. A table of the results of the sensitivity bootstrap analyses and related histograms are provided in Appendix D.

4.3. Comparison to dose-response modeling of dichotomous datasets

The incidence data for the two dichotomous analyses are presented in Table 2 (“levels 1 & 2” for “any PS” and “level 2” for “severe PS”). Unlike the ordinal analysis, for which the BMD estimates were similar between the two severity levels, the BMD estimates from the two dichotomous PS analyses differed considerably. The “severe PS” BMD$_{01}$ estimate was approximately 2.7 times higher than the “any PS” BMD$_{05}$ estimate, and the “severe PS” BMDL$_{01}$ was approximately 2.3 higher than the “any PS” BMDL$_{05}$ (Table 5). In comparison, the bootstrap estimates of the level 1 BMD$_{05}$ and level 2 BMD$_{01}$ (215 and 234 mg/kg-day, respectively) were in the middle of the range from the “any PS” BMD to the “severe PS” BMD. Similarly, the Mandel-Betensky-adjusted 95% level 1 and level 2 BMDLs (98 and 101 mg/kg-day, respectively) were in the middle of the range from the “any PS” BMDL to the “severe PS” BMDL, although slightly closer to the “any PS” BMDL.

In addition, a number of PS endpoints listed in Table 1 were noted by Yost et al. (2019) to exhibit a significant dose-related trend. Several of these endpoints were modeled individually as dichotomous endpoints (Table 6), using the procedure described in Section 3. The BMR used for each endpoint was consistent with the level assignments in Table 1, and the BMDL was calculated with one-sided 95% coverage.
The BMDs of these endpoints ranged from 112 to 480 mg/kg-day, and the BMDLs ranged from 60 to 266 mg/kg-day. These wide ranges reflect the high variability that often arises when modeling multiple endpoints. One of the more biologically relevant endpoints, azoospermia/oligospermia, had one of the lowest BMDs, 117 mg/kg-day, and the lowest BMDL, 60 mg/kg-day, results which are very similar to those from the “any PS” dichotomous approach. In comparison, the bootstrap estimates of the level 1 BMD_{05} and level 2 BMD_{01} (215 and 234 mg/kg-day) and the corresponding Mandel-Betensky-adjusted 95% BMDLs (98 and 101 mg/kg-day) were in the middle of the range outlined by the individual endpoint results.

5. Discussion

The ordinal dose-response analysis of PS presented here improves upon the dichotomous analyses because it incorporates severity of effect, and it is a relatively simple extension of the dichotomous analysis. The strengths of the ordinal analysis method are that it provides BMD estimates that are closer to each other in value and have lower variability than the separate dichotomous analyses of the individual PS endpoints. The dichotomous analyses exaggerate variability by not incorporating modeling artifacts due to sparsity of data, especially with regard to the low litter sizes. By contrast, the ordinal PS dataset yielded less variable results because it had higher incidence rates among the middle dose groups and also because the equal slopes in the ordinal analysis draws the level 1 and level 2 curves, and thus the BMD estimates and BMDLs, closer together. This ordinal approach is flexible, as more than two severity levels can be included in other analyses, and it could be applied to dose response assessment of other phthalates and antiandrogens with common effects that may vary in co-occurrence. Therefore, the ordinal approach can be expanded to be broadly applicable to dose-response modeling of any set or syndromes of effects. In addition, the confidence intervals for the BMDs provide simultaneous coverage across the levels, so they can be used in a “post hoc” manner. In other words, either confidence interval, for level 1 BMD_{05} or level 2 BMD_{01}, can be used in subsequent risk assessment even though the data have already been collected and dose-response analysis conducted, and the probabilistic coverage of the interval will not be compromised.

The modeling approach had some limitations, including the lack of consideration of all the continuously measured PS endpoints measured in the Saillenfait et al. (2008) study and the relationships among the PS endpoints. Preliminary investigation of the DIBP data considered here indicated that some endpoints tended to co-occur, for example, for azoospermia, oligospermia, granulomatous inflammation, and interstitial mononuclear cells (data not shown). Only one animal had granulomatous inflammation without also having azoospermia or oligospermia, and no animal had interstitial mononuclear cells without also having the other three. This is not surprising because these four endpoints are mechanistically related. Additionally, the analysis excluded several biologically relevant continuous endpoints such as anogenital distance and reproductive organ weights. In the future, methods for including continuous endpoints in the analysis can be utilized, either by dichotomizing the continuous endpoints using biologically relevant cutoffs, or by doing a multivariate analysis of all PS endpoints that includes the actual observed continuous values.
These types of analyses have been explored previously for certain cases (e.g., Catalano et al., 1994; Regan and Catalano, 1999).

The dataset that was analyzed in this study had some strengths and limitations. The Saillenfait et al. (2008) study strengths for modeling included the assessment of numerous male reproductive developmental endpoints, had ~20 dams/dose group, and had available individual animal data, including histopathological examination of the testis and epididymis. The fact that this prenatal exposure toxicity study collected data on a large number of PS endpoints, allowed the model to integrate information on a wide array of male reproductive endpoints. However, the study did not include a direct measure of fertility in the animals, so the categorization had to rely on established associations between male reproductive endpoints, including histopathological endpoints, and male fertility. Also, the small number of animals (two) per litter allowed a very limited estimation of the random variation within litters. The long-established practice of culling of litters to 8–10 offspring on Day 4 resulted in the loss of valuable information. Although this practice was first implemented to reduce the influence of variable litter sizes on related outcomes, statistical methods have advanced sufficiently that re-evaluation of this practice is encouraged. Data on entire litters are essential for investigating complex outcomes such as those addressed here and as shown in other studies (for example, Blystone et al., 2010).

The ordinal analysis of PS conducted here was possible because of the availability of the incidence data for PS endpoints for individual animals, thus allowing the assignment of animals to ordinal levels. In addition, the Saillenfait et al. (2008) study data and methods were well-documented and thoroughly described. Unfortunately, many publications in the peer-reviewed literature do not provide individual animal data or this level of methodological detail. As a result, obtaining individual animal data requires contacting authors and relies on the willingness of authors to collaborate with other scientists, including allowing for additional data analyses. Integrated analysis of multiple endpoints is a growing need in toxicology, and having the data available on an individual animal level is necessary to do many types of these analyses such as multivariate analysis of endpoints. It would be beneficial if journals and scientists publish and share their individual animal data.

Future refinements of the dose-response approach for the PS include (1) further delineation of severity levels; (2) consideration of patterns of responses among the PS endpoints; and (3) inclusion of continuous endpoints using a multivariate analysis approach. As an example of further refinement of severity levels using the Saillenfait et al. (2008) data, level 1 can be split into two levels according to whether a PS endpoint has either (1) no or (2) some more moderate expected association with impacts on fertility. As an example of patterns of responses, analysis of the PS endpoints that co-occur or occur at the lowest doses could be informative for dose response modeling. Understanding the temporal patterns of expression over the course of development and into adulthood could inform precursors to severe effects. In a multivariate approach, both dichotomous and continuous PS outcomes (e.g., anogenital distance, reduced fetal testicular testosterone levels) could be modeled to evaluate the interaction of effects among all of the PS endpoints, as well as the degree of contribution of the individual effects to the PS dose response curve.
The *Phthalates and Cumulative Risk Assessment* report (NRC, 2008) recommended performing a cumulative assessment that included phthalates as well as antiandrogens based on shared male reproductive endpoints. Using data from one phthalate (DIBP), the analysis presented here addresses the required first step in this process – the development of a dose-response approach for the PS endpoints that commonly occur after gestational phthalate or antiandrogen exposure. Once PS modeling is performed within individual phthalate noncancer risk assessments, the next step is to further refine the methodology and develop data to enable performing a cumulative phthalate noncancer risk assessment.

In conclusion, the ordinal dose-response modeling analysis offers an alternative to dichotomous analysis of the PS in which multiple levels of endpoint severity can be incorporated and modeled at different response levels, all in a single analysis. The analysis demonstrated in this investigation included two levels of endpoint severity, but the ordinal analysis is sufficiently flexible so that more levels can be included. Furthermore, the analysis of male reproductive effects as a syndrome leads to more stable BMD estimates.

**Acknowledgments**

We thank EPA’s National Center for Risk Assessment’s (NCEA) Reproductive, Developmental, and Neurological Toxicology Workgroup, co-led by Susan Makris and Andrew Kraft, for their expertise and fruitful discussion regarding the severity levels of the male

**Disclaimer**

The views expressed are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Appendix A**: Log-logistic model

For the dichotomous case, the most common form of the log-logistic model used in toxicology is

\[ P[Y = 1|x] = \gamma + \frac{1 - \gamma}{1 + \exp[-(\alpha + \beta \log x)]} \]

where \( \alpha \) is the intercept, \( \beta \) is the slope, and \( \gamma \) is the background response rate. For example, this is the model form used in U.S. EPA’s Benchmark Dose Software (BMDS). For the ordinal case, this equation would be written

\[ P[Y \geq j|x] = \gamma_j + \frac{1 - \gamma_j}{1 + \exp[-(\alpha_j + \beta \log x)]} \]  (A1)

where \( \alpha_j \) and \( \gamma_j \) are the intercept and background response rate for ordinal level \( j \) and \( \beta \) is the slope. Using this model form, estimating the BMD requires expressing the BMD as a function of \( \alpha_j \) and \( \beta \). However, the model can be reparameterized so that the BMD is an explicit parameter in the model using the definition of the BMR, denoted \( \nu \).
The BMD associated with this BMR, denoted $BMD_\nu$, is related to the slope and intercept by the equation

$$\log BMD_\nu = \frac{1}{\beta} [-\alpha_j + \log(\nu/(1 - \nu))] .$$

Solving this equation for $\alpha$ and substituting it into (A.1) yields Eq. (1), with $\theta_\nu$ substituted for $\log BMD_\nu$. The BMD is log-transformed in the equation to improve the normal approximation of its estimate.

The model defined in Eq. (1) was fit to the ordinal PS data using $\nu = 0.01$, and the fit statistics were used to estimate the BMDs of interest, the level 1 BMD$_{0.01, 1}$ and the level 2 BMD$_{0.01, 2}$. The second of these, $\theta_{0.01, 2}$, was a parameter in the model that was fit, so its estimate and standard error (SE) were automatically output from the model run. $\theta_{0.01, 1}$ was estimated using its relationship to $\theta_{0.01, 1}$, described as follows. For each $x$, the value in the right-hand side of Eq. (1) must be the same across values of $\nu$, which after simplification implies that the inside of the exponent term in (1) must be the same across $\nu$. Applying this principle to $\nu = 0.01$ and 0.05 yields

$$-\log(0.05/0.95) + \beta (\theta_{0.01, 1} - \log x) = -\log(0.01/0.99) + \beta (\theta_{0.01, 1} - \log x) .$$

Solving this equation for $\theta_{0.01, 1}$ yields Eq. (2). The estimate and SE of $\theta_{0.01, 1}$ are listed in the Additional Estimates table of SAS Output A2, under the label “theta_int.”

SAS did not automatically compute the goodness-of-fit chi-square test statistic in SAS Proc NLMIXED, so additional code was written to do this computation. The formula used for the test statistic was

$$\chi^2 = \sum_{i=1}^{5} \sum_{j=0}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} .$$

2where $O_{ij}$ and $E_{ij}$ are the observed and expected incidence frequencies of ordinal level $j$ in the $i$th dose group. Here, $E_{ij} = n_i \cdot Pr_i(j)$, where $n_i$ is the number of animals in the $i$th dose group and $Pr_i(j)$ is the probability that an animal in the $i$th dose group has ordinal level $j$. The degrees of freedom for this test were calculated using the formula $df = r \times (c - 1) - k$, where $r$ is the number of dose groups, $c = 3$ is the number of levels of the ordinal response variable, and $k = 5$ is the number of parameters that are estimated in the model. This procedure is described in Agresti (2010; see Chapter 3).
Appendix B.: Description of bootstrap method and Mandel-Betensky intervals

A large number of bootstrap datasets were randomly selected using the following procedure. For dose $x_i$ suppose there are $J_i$ litters exposed to that dose, and let $Y_{jk}$ denote the PS incidence of the $k$th animal in the $j$th litter for $j = 1, \ldots, J_i$; $k = 1, 2$. That is, for $k = 1, 2$, $Y_{jk} = 1$ if the $k$th animal in litter has PS, and $Y_{jk} = 0$ otherwise. Select bootstrap samples as follows:

1. Randomly select a sample of size $J_i$ with replacement from the set of litters exposed to dose $x_i$ and denote this litter $j_m^*$ for $m = 1, \ldots, J_i$. Let $Y^*_{jm} = (Y^*_{1, jm}, Y^*_{2, jm})$ denote the vector of PS incidence in that litter. Note that because the sampling is with replacement, some litters may be selected multiple times, while others may not be selected at all.

2. From each litter $j_m^*$ selected in step 1, randomly select 2 observations from $Y^*_{jm}$, denoted $V^*_{jm} = (V^*_{1, jm}, V^*_{2, jm})$. Note that because the sampling is with replacement, the same observation may be selected twice.

This procedure results in the sample of PS observations $\{V^*_{jm} : m = 1, \ldots, J_i\}$ for dose $x_i$. When collected across all doses, these samples collectively constitute a dataset of the same nature and size as the original sample of PS observations with litter effect ignored. The dose-response model described in Section 3 was applied to each bootstrap sample and estimates of the level 1 BMD$_{05}$ and level BMD$_{01}$ were obtained. The collection of these estimates across all bootstrap samples constituted an estimated distribution of the BMDs.

Simultaneous confidence intervals for the level 1 BMD$_{05}$ and level 2 BMD$_{01}$ using an adjusted percentile bootstrap interval developed by Mandel and Betensky (2008). Here, the lower and upper interval bounds for each BMD were the percentiles of its respective bootstrap-derived BMD distribution, where the rank of each bootstrap sample was redefined based on the coordinate whose rank was most discrepant from its respective median. More explicitly, the sample ranks were redefined as follows:

1. For the $b$th bootstrap sample and $j = 1, 2$, let $r(b, j)$ = the rank of the $j$th BMD coordinate taken across all samples, and define the relative rank $r^*(b, j) = |r(b, j) - (B + 1)/2|$ and sign of the relative rank $s^*(b, j) = \text{sign}[r(b, j) - (B + 1)/2]$, where $\text{sign}(x) = 1$ if $x \geq 0$ and $-1$ if $x < 0$.

2. The final rank of the $b$th bootstrap sample was based on the redefined ranks across the BMD coordinates associated with that sample’s estimate. That is, its rank was defined as $r(b) = (B + 1)/2 + s^*(b, j) \cdot \max_j r^*(b, j)$, where the maximum in this expression is taken across the BMD coordinates in the $b$th sample.
Appendix C: Analysis results of assessing influence of histopathology endpoints on ordinal PS levels

Fourteen (14) animals out of a total of 114 (12%) had their PS levels increase because of the presence of histopathology endpoints; 11, 2, and 1 animals had levels that increased from 0 to 1, 0 to 2, and 1 to 2, respectively (see Table C1).

Table C1

Animals sacrificed at PNW 11–12 whose PS level increased due to the presence of histopathology endpoints, including their scores with histopathology excluded and included.

| Dose | Litter | Animal | Level w/o histopath | Level w/ histopath |
|------|--------|--------|---------------------|-------------------|
| 125  | 26     | 3      | 0                   | 2                 |
| 125  | 27     | 1      | 0                   | 1                 |
| 250  | 41     | 1      | 0                   | 1                 |
| 250  | 44     | 1      | 0                   | 1                 |
| 250  | 48     | 4      | 0                   | 1                 |
| 250  | 51     | 3      | 0                   | 1                 |
| 250  | 52     | 1      | 0                   | 1                 |
| 250  | 52     | 3      | 0                   | 1                 |
| 500  | 64     | 3      | 0                   | 1                 |
| 500  | 73     | 1      | 0                   | 1                 |
| 500  | 74     | 1      | 0                   | 1                 |
| 500  | 74     | 3      | 0                   | 2                 |
| 625  | 82     | 7      | 1                   | 2                 |
| 625  | 93     | 1      | 0                   | 1                 |

Appendix D: Results of sensitivity bootstrap analysis

See Table D1 and Fig. D1.

Table D1

The estimates, SEs, and 90% and 95% Mandel & Betensky confidence intervals of the BMDs from the original and sensitivity bootstrap analyses.

| Parameter         | Bootstrap Analysis        | Estimate | SE   | 90% CI       | 95% CI       |
|-------------------|---------------------------|----------|------|--------------|--------------|
| Level 1 BMD_{0.05} | No litter effect          | 203.6    | 75.62| (117.5, 393.8)| (104.7, 411.0)|
|                   | With litter effect: original sample | 214.8   | 91.81| (111.3, 428.3)| (98.40, 478.2)|
|                   | With litter effect: permuted sample 1 | 213.8  | 89.47| (108.7, 417.4)| (96.29, 474.8)|
|                   | With litter effect: permuted sample 2 | 211.9  | 85.69| (111.9, 410.4)| (97.22, 471.3)|
|                   | With litter effect: permuted sample 3 | 214.3  | 89.30| (112.7, 430.5)| (97.57, 478.0)|
| Level 2 BMD_{0.01} | No litter effect          | 219.8    | 77.00| (121.0, 402.9)| (105.1, 425.4)|
|                   | With litter effect: original sample | 233.6  | 93.01| (116.1, 459.2)| (100.6, 479.2)|
|                   | With litter effect: permuted sample 1 | 232.8  | 92.8 | (111.7, 435.5)| (94.18, 476.1)|

Environ Int. Author manuscript; available in PMC 2021 January 01.
| Parameter                        | Bootstrap Analysis    | Estimate | SE   | 90% CI       | 95% CI       |
|---------------------------------|-----------------------|----------|------|--------------|--------------|
| With litter effect: permuted sample 2 | 230.4                 | 88.52    | (113.9, 427.6) | (95.88, 472.8) |
| With litter effect: permuted sample 3 | 234.0                 | 91.24    | (115.3, 459.4)  | (97.42, 477.7) |

*Environ Int.* Author manuscript; available in PMC 2021 January 01.
Fig. D1.
Histograms from bootstrap analysis of the PS with the litter effect excluded. Shown are the histograms of the untransformed BMD_{0.05} for level 1 (a) and the untransformed BMD_{0.01} for level 2 (b), and of the log-transformed BMD_{0.05} for level 1 (c) and the log-transformed BMD_{0.01} for level 2 (d).
Appendix E.: SAS output for ordinal analysis of PS dataset

SAS Output E1:

Pertinent sections of the output from the ordinal analysis of the PS dataset in SAS Proc NLMIXED with the slopes allowed to differ across ordinal levels. The results from testing the equality of the slopes are listed in the Contrasts table.

Parameter Estimates

| Parameter | Estimate | Standard Error | DF  | t Value | Pr > | Alpha | Lower  | Upper  |
|-----------|----------|----------------|-----|---------|-------|-------|--------|--------|
| gamma2    | 0.02636  | 0.01973        | 4   | 1.34    | 0.2525| 0.05  | -0.02843| 0.08115|
| g1        | 0        | .              | 4   | .       | .     | .     |        |        |
| theta2    | 5.6835   | 0.2525         | 4   | 22.51   | >0.0001| 0.05  | 4.9824 | 6.3845 |
| theta1    | 1.0586   | 0.3321         | 4   | 3.19    | 0.0333| 0.05  | 0.1365 | 1.9807 |
| beta2     | 3.8990   | 0.7384         | 4   | 5.28    | 0.0062| 0.05  | 1.8488 | 5.9492 |
| beta1     | 7.7482   | 3.0030         | 4   | 2.58    | 0.0613| 0.05  | -0.5895| 16.0859|

Contrasts

| Label     | Num DF | Den DF | F Value | Pr > F |
|-----------|--------|--------|---------|--------|
| equal slopes | 1      | 4      | 2.86    | 0.1662 |

Additional Estimates

| Label     | Estimate | Standard Error | DF  | t Value | Pr > | Alpha | Lower  | Upper  |
|-----------|----------|----------------|-----|---------|-------|-------|--------|--------|
| gamma1    | 0.02636  | 0.01973        | 4   | 1.34    | 0.2525| 0.05  | -0.02843| 0.08115|
| gamma2    | 0.02636  | 0.01973        | 4   | 1.34    | 0.2525| 0.05  | -0.02843| 0.08115|
| beta1     | 3.8990   | 0.7384         | 4   | 5.28    | 0.0062| 0.05  | 1.8488 | 5.9492 |
| beta2     | 7.7482   | 3.0030         | 4   | 2.58    | 0.0613| 0.05  | -0.5895| 16.0859|
| theta1    | 5.0482   | 0.1694         | 4   | 29.79   | < 0.0001| 0.05 | 4.5778 | 5.5187 |
| theta2    | 5.6835   | 0.2525         | 4   | 22.51   | < 0.0001| 0.05 | 4.9824 | 6.3845 |
| log(beta1)| 1.3607   | 0.1894         | 4   | 7.18    | 0.0020| 0.05  | 0.8349 | 1.8865 |
| log(beta2)| 2.0475   | 0.3876         | 4   | 5.28    | 0.0062| 0.05  | 0.9714 | 3.1235 |

SAS Output E2:

Pertinent sections of the output from the ordinal analysis of the PS dataset in SAS Proc NLMIXED with the slopes assumed to be equal across ordinal levels. The estimates and SEs of the BMDs and other parameters of interest are listed in the Additional Estimates table.

Parameter Estimates

| Parameter | Estimate | Standard Error | DF  | t Value | Pr > | Alpha | Lower  | Upper  |
|-----------|----------|----------------|-----|---------|-------|-------|--------|--------|
| gamma2    | 0.011675 | 0.01796        | 5   | 0.93    | 0.3937| 0.05  | -0.02941| 0.06290|
| g1        | 0.01526  | 0.02998        | 5   | 0.51    | 0.6324| 0.05  | -0.06181| 0.09233|
| theta2    | 5.2359   | 0.2495         | 5   | 20.99   | <0.0001| 0.05 | 4.5946 | 5.8773 |
### Parameter Estimates

| Parameter | Estimate | Standard Error | DF | t Value | Pr > |t| | Alpha | Lower | Upper |
|-----------|----------|----------------|----|---------|-------|---|-------|-------|-------|
| thd       | 0.4277   | 0.1016         | 5  | 4.21    | 0.0084 | 0.05| 0.1664 | 0.6889 |
| bt        | 4.5913   | 1.1067         | 5  | 4.15    | 0.0089 | 0.05| 1.7466 | 7.4361 |

### Additional Estimates

| Label      | Estimate | Standard Error | DF | t Value | Pr > |t| | Alpha | Lower | Upper |
|------------|----------|----------------|----|---------|-------|---|-------|-------|-------|
| gamma1     | 0.03201  | 0.03700        | 5  | 0.87    | 0.4265 | 0.05| −0.06309 | 0.1271 |
| gamma2     | 0.01675  | 0.01796        | 5  | 0.93    | 0.3937 | 0.05| −0.02941 | 0.06290 |
| beta       | 4.5913   | 1.1067         | 5  | 4.15    | 0.0089 | 0.05| 1.7466 | 7.4361 |
| theta1     | 5.1678   | 0.2174         | 5  | 3.77    | <0.0001| 0.05| 1.9090 | 7.7267 |
| theta2     | 5.2359   | 0.2495         | 5  | 20.99   | <0.0001| 0.05| 4.5946 | 5.8773 |
| theta_int  | 4.8083   | 0.2993         | 5  | 16.07   | <0.0001| 0.05| 4.0390 | 5.5776 |
| L1 BMD05   | 175.53   | 38.1603        | 5  | 4.60    | 0.0058 | 0.05| 77.4359 | 273.62 |
| L2 BMD01   | 187.91   | 46.8806        | 5  | 4.01    | 0.0102 | 0.05| 67.3950 | 308.14 |

### References

Agresti A, 2010 Analysis of Ordinal Categorical Data, second ed Wiley.

Allen BC, Kavlock RJ, Kimmel CA, Faustman EM, 1994 Dose-response assessment for developmental toxicity. III. Statistical models. Fundam. Appl. Toxicol 23 (4), 496–509. 10.1006/faat.1994.1134. [PubMed: 7867901]

Blystone CR, Kissling GE, Bishop JB, Chapin RE, Wolfe GW, Foster PM, 2010 Determination of the di-(2-ethylhexyl) phthalate NOAEL for reproductive development in the rat: importance of the retention of extra animals to adulthood. Toxicol. Sci 116 (2), 640–646. 10.1093/toxsci/kfq147. [PubMed: 20484383]

Boekelheide K, Kleymenova E, Liu K, Swanson C, Gaido KW, 2009 Dose-dependent effects on cell proliferation, seminiferous tubules, and male germ cells in the fetal rat testis following exposure to di(n-butyl) phthalate. Microsc. Res. Tech 72 (8), 629–638. 10.1002/jemt.20684. [PubMed: 19165735]

Catalano P, Ryan L, Scharfstein D, 1994 Modeling fetal death and malformation in developmental toxicity studies. Risk Anal. 14 (4), 629–637. 10.1111/j.1539-6924.1994.tb00276.x. [PubMed: 7972963]

Chen CC, Chen JJ, 2014 Benchmark dose calculation for ordered categorical responses. Risk Anal. 34 (8), 1435–1447. 10.1111/risa.12167. [PubMed: 24444309]

Dataset, 2019 Data and output for Ordinal Dose-Response Modeling Approach for the Phthalate Syndrome 508. 10.23719/1503702.

Efron B, Tibshirani RJ, 1994 An Introduction to the Bootstrap. Chapman & Hall.

Fisher JS, Macpherson S, Marchetti N, Sharpe RM, 2003 Human ‘testicular dysgenesis syndrome’: a possible model using in-utero exposure of the rat to dibutyl phthalate. Hum. Reprod 18 (7), 1383–1394. 10.1093/humrep/deg273. [PubMed: 12832361]

Foster PM, 2006 Disruption of reproductive development in male rat offspring following in utero exposure to phthalate esters. Int. J. Androl 29 (1), 140–147. 10.1111/j.1365-2605.2005.00563.x. [PubMed: 16102138]

Foster PMD, Mylchreest E, Gaido KW, Sar M, 2001 Effects of phthalate esters on the developing reproductive tract of male rats. Hum. Reprod. Update 7 (3), 231–235. 10.1111/j.1600-0463.2001.tb05776.x. [PubMed: 11392369]
Gray LE Jr, Barlow NJ, Howdeshell KL, Ostby JS, Furr JR, Gray CL, 2009 Transgenerational effects of Di (2-ethylhexyl) phthalate in the male CRL:CD(SD) rat: added value of assessing multiple offspring per litter. Toxicol. Sci 110 (2), 411–425. 10.1093/toxsci/kfp109. [PubMed: 19482887]

Howdeshell KL, Hotchkiss AK, Gray LE Jr, 2017 Cumulative effects of anti-androgenic chemical mixtures and their relevance to human health risk assessment. Int. J. Hyg. Environ. Health 220 (2 Pt A), 179–188. 10.1016/j.ijheh.2016.11.007. [PubMed: 27923611]

Koch HM, Rüther M, Schütze A, Conrad A, Pälinke C, Apel P, Brüning T, Kolossa-Gehring M, 2017 Phthalate metabolites in 24-h urine samples of the German Environmental Specimen Bank (ESB) from 1988 to 2015 and a comparison with US NHANES data from 1999 to 2012. Int. J. Hyg. Environ. Health 220 (2 Pt A), 130–141. 10.1016/j.ijheh.2016.11.003. [PubMed: 27863804]

Krewski D, Chambers A, Birkett N, 2010 The use of categorical regression in modeling copper exposure–response relationships. J. Toxicol. Environ. Health A 73, 187–207. 10.1080/15287390903340781. [PubMed: 20077290]

Kupper LL, Portier C, Hogan MD, Yamamoto E, 1986 The impact of litter effects on dose-response modeling in teratology. Biometrics 42 (1), 85–98. 10.2307/2531245. [PubMed: 3719065]

Lanning LL, Creasy DM, Chapin RE, Mann PC, Barlow NJ, Regan KS, Goodman DG, 2002 Recommended approaches for the evaluation of testicular and epididymal toxicity. Toxicol. Pathol 30 (4), 507–520. 10.1080/01926230290105695. [PubMed: 12187942]

Lioy PJ, Hauser R, Gennings C, Koch HM, Mirkes PE, Schwetz BA, Kortenkamp A, 2015 Assessment of phthalates/phthalate alternatives in children’s toys and childcare articles: Review of the report including conclusions and recommendation of the Chronic Hazard Advisory Panel of the Consumer Product Safety Commission. J. Expo. Sci. Environ. Epidemiol 25 (4), 343–353. [PubMed: 25944701]

Makris SL, Euling SY, Gray LE Jr, Benson R, Foster P, 2013 Use of genomic data in risk assessment case study: I. Evaluation of the dibutyl phthalate male reproductive development toxicity data set. Toxicol. Appl. Pharmacol 271 (3), 336–348. 10.1016/j.taap.2010.09.006. [PubMed: 20849870]

Mandel M, Betensky RA, 2008 Simultaneous confidence intervals based on the percentile bootstrap approach. Comput. Stat. Data An 52 (4), 2158–2165. 10.1016/j.csda.2007.07.005.

National Research Council (NRC), 2008 Phthalates and Cumulative Risk Assessment: The Tasks Ahead. Committee on the Health Risks of Phthalates, Board on Environmental Studies and Toxicology, Division on Earth and Life Sciences. The National Academies Press, Washington, DC 10.17226/12528.

Regan MM, Catalano PJ, 1999 Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology. Biometrics 55 (3), 760–768. 10.1111/1526-441X.1999.00760.x. [PubMed: 11315004]

Saillenfait AM, Sabaté JP, Gallissot F, 2008. Diisobutyl phthalate impairs the androgen-dependent reproductive development of the male rat. Reprod. Toxicol 26 (2), 107–115. 10.1016/j.reprotox.2008.07.006. [PubMed: 18706996]

U.S. Consumer Product Safety Commission (CPSC), 2014 Chronic Hazard Advisory Panel (CHAP) Report on Phthalates and Phthalate Alternatives 2014. Available online at: https://www.cpsc.gov/s3fs-public/CHAP-REPORT-FINAL%20(1).pdf. Directorate for Health Sciences, Bethesda, MD.

U.S. Environmental Protection Agency (EPA), 2002 A review of the reference dose and reference concentration processes (pp. 1–192). (EPA/630/P-02/002F). Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. Available online at: https://www.epa.gov/sites/production/files/2014-12/documents/rfd-final.pdf.

U.S. EPA, 2012 Benchmark Dose Technical Guidance (EPA/100/R-12/001) Washington, DC: U.S. Environmental Protection Agency, Risk Assessment Forum. Available online at: https://www.epa.gov/sites/production/files/2015-01/documents/benchmark_dose_guidance.pdf.

U.S. EPA, 2014 Toxic Substances Control Act (TSCA) Work Plan for Chemical Assessments: 2014 Update. Office of Pollution Prevention and Toxics, October 2014. Available at: https://www.epa.gov/assessing-and-managing-chemicals-under-tscaswork-plan-chemical-assessments-2014-update.

U.S. EPA, 2015 Phthalate Action Plan. Available online at: https://www.epa.gov/sites/production/files/2015-09/documents/phthalates_actionplan_revised_2012-03-14.pdf.
van den Driesche S, McKinnell C, Calarrão A, Kennedy L, Hutchison GR, Hrabalkova L, Jobling MS, Macpherson S, Anderson RA, Sharpe RM, Mitchell RT, 2015 Comparative effects of di(n-butyl) phthalate exposure on fetal germ cell development in the rat and in human fetal testis xenografts. Environ. Health Perspect 123 (3), 223–230. 10.1289/ehp.1408248. [PubMed: 25514601]

van der Leeden R, Meijer E, 2007 Busing FMTA. Resampling Multilevel Models Ch. 11 of Handbook of Multilevel Analysis. Springer 10.1007/978-0-387-73186-5_11.

Wittassek M, Angerer J, Kolossa-Gehring M, Schäfer SD, Klockenbusch W, Dobler L, Günsel AK, Müller A, Wiesmüller GA, 2009 Fetal exposure to phthalates—a pilot study. Int. J. Hyg. Environ. Health 212 (5), 492–498. 10.1016/j.ijheh.2009.04.001. [PubMed: 19423389]

Ye X, Pierik FH, Hauser R, Duty S, Angerer J, Park MM, Burdorf A, Hofman A, Jaddoe VW, Mackenbach JP, Steegers EA, Tiemeier H, Longnecker MP, 2008. Urinary metabolite concentrations of organophosphorous pesticides, bisphenol A, and phthalates among pregnant women in Rotterdam, the Netherlands: the Generation R study. Environ. Res 108 (2), 260–267. 10.1016/j.envres.2008.07.014. [PubMed: 18774129]

Yost EE, Euling SY, Weaver JA, Beverly BEJ, Keshava N, Mudipalli A, Arzuaga X, Blessinger T, Dishaw L, Hotchkiss A, Makris SL, 2019 Hazards of diisobutyl phthalate (DIBP) exposure: a systematic review of animal toxicology studies. Environ. Int 125, 579–594. 10.1016/j.envint.2018.09.038. [PubMed: 30591249]

Zota AR, Calafat AM, Woodruff TJ, 2014 Temporal trends in phthalate exposures: findings from the National Health and Nutrition Examination Survey, 2001–2010. Environ. Health Perspect 122 (3), 235–241. 10.1289/ehp.1306681. [PubMed: 24425099]
Fig. 1.
Illustration of the temporal aspects of the dosing and phthalate syndrome endpoint evaluation in the Saillenfait et al. (2008) study. The figure is limited to the phthalate syndrome endpoints included in this univariate analysis. Sacrificed adult males were shaved and the ventral surface of the thorax was examined for the presence of areolas and/or nipples. The external and internal genitalia were examined for gross abnormalities and for the position of testes. Histopathology was conducted on testes and epididymides. In the Saillenfait et al. (2008) study, Sprague-Dawley female rats were mated to adult male rats of the same strain and from the same supplier. The day that vaginal smears were sperm positive was designated as gestation day 0 (GD 0). Pregnant females were randomly assigned to treatment groups based on body weight at GD 0. Pregnant females were treated once a day with either vehicle alone (olive oil), 500 mg/kg/day DBP, 125 mg/kg/day DIBP, 250 mg/kg/day DIBP, 500 mg/kg/day DIBP, or 625 mg/kg/day DIBP by gavage on GD 12–21. The number of dams per treatment group was 10–14. Postnatal day (PND) 0 was defined as the day of parturition. On PND 4, litters were culled to 10 pups, keeping as many males as possible. Litters with less than 8 pups were not maintained. See Saillenfait et al. (2008) for the numbers of animals and specific endpoints assessed at each life stage. Arrows with vertical bars indicate transitions between three lifestages: gestation, postnatal development, and adulthood. Note that the scale of time is not consistent across the three lifestages.
Fig. 2.
Plots from ordinal PS analysis with litter effect excluded. The scatter plot of cumulative incidence rate with line plot of fitted cumulative probability curve shown for level 1 (a) and level 2 (b), with the corresponding BMD estimate indicated by a solid square, and residual plots with zero line shown for level 1 (c) and level 2 (d).
Fig. 3.
Histograms from bootstrap analysis of the PS. Shown are the histograms of the untransformed BMD\textsubscript{05} for level 1 (a) and the untransformed BMD\textsubscript{01} for level 2 (b), and of the log-transformed BMD\textsubscript{05} for level 1 (c) and the log-transformed BMD\textsubscript{01} for level 2 (d).
Fig. 4.
Forest plots of BMD estimates and confidence intervals for level 1 for the normal-based and bootstrap analyses. Shown are the level 1 BMD\(_{0.05}\) 90\% (a) and 95\% (b) confidence intervals and the level 2 BMD\(_{0.01}\) with 90\% (c) and 95\% (d) confidence intervals, with estimates denoted by blue diamonds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Table 1

Proposed severity classifications for PS endpoints reported in Saillenfait et al. (2008), based on the expected impact on fertility (see Materials and Methods for details).

| Male developmental reproductive effects | Level 1 | Level 2 |
|----------------------------------------|---------|---------|
| • Areola/nipple Retention | | Cleft prepuce |
| • Cleft prepuce | | Hypospadias (mild, moderate, or severe) |
| | | Exposed os penis |
| | | Undescended testes (left, right, or both) |
| | | Small penis |
| | | Vaginal pouch |
| | | Prostate absent |
| | | Seminal vesicles abnormal |

**Epididymis histopathology**

- Interstitial mononuclear cells: grade 2–5
- Oligospermia: grade 2–4
- Grade 5 oligospermia, unilateral (i.e., in only one epid), w/o azoospermia in other epididymis; or azoospermia, unilateral, w/o grade 5 oligospermia in other epididymis
- Sloughed cells (caput, cauda, or corpus): grade 2–5
- Granulomatous inflammation: grade 2–4<sup>b</sup>
- Tubular necrosis: grade 2–5
- Bilateral grade 5 oligospermia or azoospermia

**Testis histopathology**

- Interstitial cell hyperplasia: grade 2–4<sup>b</sup>
- Tubular necrosis/mineralization: grade 2–5
- Tubular vacuolation/loss of germ cells: grade 2–4<sup>b</sup>
- Seminiferous tubule degeneration-atrophy/hypoplasia: grade 2–4
- Seminiferous tubule degeneration-atrophy/hypoplasia: grade 5 in one testis, grade 1–4 or not present in other testis
- Loss of seminiferous tubules: grade 2–4<sup>b</sup>
- Seminiferous tubular degeneration-atrophy/hypoplasia: grade 5 in both testes
Unless otherwise indicated, an animal was designated as having the level 1 endpoint if either side (left or right) had the endpoint.

No animals had grade 5 granulomatous inflammation, grade 2–5 interstitial fibrosis, or grade 5 interstitial cell hyperplasia, tubular vacuolation, or loss of seminiferous tubules.
Table 2

Incidence of ordinal PS among male offspring evaluated at PNW 11–12 for each dose group (Saillenfait et al., 2008).

| Dose (mg/kg-d<sup>a</sup>) | Total Litters Examined | Total Males Examined | Incidence<sup>b</sup> | Level 0 | Level 1 | Level 2 | Levels 1 & 2 |
|---------------------------|------------------------|----------------------|-----------------------|---------|---------|---------|--------------|
|                           |                        |                      |                       |         |         |         |              |
| 0                         | 12                     | 24                   | 24 (100%)             | 0 (0%)  | 0 (0%)  | 0 (0%)  |              |
| 125                       | 10                     | 20                   | 18 (90%)              | 1 (5%)  | 1 (5%)  | 2 (10%) |              |
| 250                       | 14                     | 28                   | 21 (75%)              | 6 (21%) | 1 (4%)  | 7 (75%) |              |
| 500                       | 11                     | 22                   | 5 (23%)               | 9 (41%) | 8 (36%) | 17 (77%)|              |
| 625                       | 10                     | 20                   | 0 (0%)                | 3 (15%) | 17 (85%)| 20 (100%)|              |

<sup>a</sup> “mg/kg-d” = mg/kg-day.

<sup>b</sup> Ordinal levels 0, 1, and 2 are described in Section 2.2.
Table 3

Parameter estimates and standard errors (SEs) from fitting model (1) to the ordinal PS dataset, excluding the litter effect.\(^a\)

| Parameter                                      | Estimate | Standard Error (SE) |
|------------------------------------------------|----------|---------------------|
| Level 1 background response rate (\(\gamma_1\)) | 0.0320   | 0.0370              |
| Level 2 background response rate (\(\gamma_2\)) | 0.0168   | 0.0180              |
| Slope (\(\beta\))                              | 4.59     | 1.11                |
| \(\log(\text{BMD}_{0.5})\) for level 1 (\(\theta_{0.5,1}\)) | 5.17     | 0.217               |
| \(\log(\text{BMD}_{0.1})\) for level 2 (\(\theta_{0.1,2}\)) | 5.23     | 0.250               |

\(^a\)The model fit adequately according to chi-square goodness-of-fit test: \(\chi^2_5 = 7.10\), p-value = 0.213 (details in Appendix A). In addition, scatterplots of cumulative incidence rates for levels 1 and 2, with the estimated curve over-layed on each (Fig. 2ab), and residual plots of the estimated curves for the two cumulative incidence rates (Fig. 2cd) indicate adequate model fit.

\(^b\)The slope was statistically significantly greater than zero (\(t = 4.15\); p-value = 0.009). In addition, an F-test on the equality of the slopes yielded a non-significant result (\(F_{1,4} = 2.86\); p-value = 0.166).
Table 4
The estimates, SEs, and 90% and 95% confidence intervals of the BMDs from the normal-based and bootstrap analyses.

| Parameter | Analysis | Estimate (mg/kg-d) | SE (mg/kg-d) | CI Type | Two-sided 90% CI (mg/kg-d) | Two-sided 95% CI (mg/kg-d) |
|-----------|----------|--------------------|--------------|---------|--------------------------|--------------------------|
| Level 1 BMD | Normal-based | 176 | 38.2 | Unadj. | (123, 251) | (115, 269) |
| | Bootstrap | 215 | 91.8 | Unadj. | (116, 425) | (102, 477) |
| Level 2 BMD | Normal-based | 188 | 46.9 | Unadj. | (125, 283) | (115, 306) |
| | Bootstrap | 234 | 93.0 | Unadj. | (121, 449) | (105, 478) |

Superscript:
a “Unadj.”: CI is not adjusted for multiplicity; “M&B”: Mandel & Betensky multiplicity adjustment is applied; “Bonf.”: Bonferroni multiplicity adjustment applied; the principal confidence intervals used are in boldface.

b The normal-based SEs were computed using the delta method.

c The normal-based confidence intervals are t-intervals.

d The bootstrap estimates and SEs were the means and standard deviations of the BMD estimates from the 10,000 bootstrap runs.
Table 5
Results from dichotomous dose-response analyses for estimating the “any PS” BMD$_{05}$ and “severe PS” BMD$_{01}$, with one-sided 95% BMDLs.

| Dichotomous Analysis$^a$ | BMR | BMD (mg/kg-day) | One-sided 95% BMDL$^a$ (mg/kg-day) |
|--------------------------|-----|----------------|----------------------------------|
| AnyPS                    | 5% ER | 125           | 77                               |
| Severe PS                | 1% ER | 344           | 180                              |

$^a$The “any PS” and “severe PS” analyses are described in Section 2.2.

$^b$The one-sided 95% BMDL corresponds to the lower bound of a two-sided 90% confidence interval; BMDS does not provide the BMDUs for nested dichotomous models, so these are not listed.
Table 6

BMD model results from dichotomous analysis of selected PS endpoints.

| PS Endpoint                        | BMR    | BMD (mg/kg-day) | BMDL (mg/kg-day) |
|-----------------------------------|--------|-----------------|------------------|
| Hypospadias                       | 1% ER  | 401             | 242              |
| Undescended testes                | 1% ER  | 342             | 194              |
| Exposed os penis                  | 1% ER  | 361             | 112              |
| Areola or nipple retention        | 5% ER  | 317             | 205              |
| Azoospermia or grade 2–5 oligospermia<sup>a</sup> | 5% ER  | 117             | 60               |
| Tubular degeneration<sup>b</sup>  | 5% ER  | 480             | 266              |
| Sloughed cells                    | 5% ER  | 112             | 67               |

<sup>a</sup> An animal was designated as having “azoospermia or grade 2–5 oligospermia” if either epididymis had either azoospermia or grade 2–5 oligospermia (or both).

<sup>b</sup> "Tubular degeneration" consisted of tubular degeneration-atrophy/hypoplasia, tubular vacuolation, and loss of seminiferous tubules, all at grades 2–5.