ILL@UMKC

Rapid #: -22376207

CROSS REF ID: **547968**

LENDER: **T3M (Taipei Medical University) :: Main Library**

BORROWER: **UMK (University of Missouri-Kansas City) :: University Libraries**

TYPE:	Article CC:CCG
JOURNAL TITLE:	Neurotoxicology and teratology
USER JOURNAL TITLE:	Neurotoxicology and Teratology
ARTICLE TITLE:	Statistical modeling with litter as a random effect in mixed models to manage “intralitter likeness”
ARTICLE AUTHOR:	Golub, Mari S.,
VOLUME:	77
ISSUE:	
MONTH:	
YEAR:	2020
PAGES:	;106841-
ISSN:	0892-0362
OCLC #:	

Processed by RapidX: 4/8/2024 12:15:56 AM

Neurotoxicology and Teratology 77 (2020) 106841

Contents lists available at ScienceDirect

Neurotoxicology and Teratology

journal homepage: www.elsevier.com/locate/neutera

Statistical modeling with litter as a random effect in mixed models to manage "intralitter likeness"

Mari S. Golub[a], Christina A. Sobin[b,c,*]

[a] California National Primate Research Center, University of California, Davis, Davis, CA, United States of America
[b] College of Health Sciences, University of Texas EL Paso, El Paso, TX, United States of America
[c] Laboratory of Neuroendocrinology, The Rockefeller University, New York, NY, United States of America

ARTICLE INFO

Keywords:
Litter effects
Statistical models
Rodents
Pregnancy
Neurotoxicity

ABSTRACT

"Intralitter likeness," the possibility that the shared genetics and/or maternal environment in multiparous species causes strong similarity for outcome variables in littermates, violates a core statistical assumption, that of observation independence, when littermate outcomes are analyzed. Intralitter likeness has been of major concern to investigators for several decades. Despite consensus and guidance, many research reports in the rodent literature continue to ignore intralitter likeness. A historical review of the literature revealed that the long-preferred solution was to include litter as an effect in statistical models. Limitations in software development and computing capacity prior to 1990, however, appear to have led researchers and guidance authorities to endorse instead the method of using one value per litter. Here, the history of discussions regarding intralitter likeness in developmental neurotoxicological research is reviewed; growing knowledge regarding the biological bases and significance of intralitter likeness is discussed; principles underlying the use of litter as a random effect in mixed models are presented; statistical examples are provided illustrating the advantages and critical importance of including litter as a random effect in mixed models; and results using all data points (all pups from all litters) with litter as a random effect, are compared to results based on random selections of representative littermates. Mixed models with litter included as a random effect have distinct advantages for the analysis of clustered data. Modern computing capacity provides ready accessibility to mixed models for all researchers. Accessibility however does not preclude the need for appropriate expertise and consultation in the use of mixed (hierarchical) models.

1. Overview

Rodent models are a foundation of translational developmental neurotoxicological research. Differences in reproductive biology, however, can complicate use of this model. For developmental neurotoxicology, one issue has concerned statistical management of clustered data from rodent multiparous (multiple offspring) pregnancies. "Intralitter likeness," the possibility that the shared genetic and/or maternal environment causes strong similarity for outcome variables among littermates, violates the core theoretical assumption of observation independence.[1] Despite consensus and guidance, many research reports in the rodent literature continue to ignore intralitter likeness. For example, a 1997 review of 69 rodent developmental studies found that only 12% dealt with intralitter likeness (Zorrilla, 1997), and a 2013 review of 34 valproic acid studies in rodents reported that just 3 correctly addressed this issue (Lazic and Essioux, 2013).

For over 50 years, authorities in the fields of teratology and developmental neurotoxicology have discussed approaches for dealing with intralitter likeness. From at least the 1970's (Hughes, 1979), the optimal approach was a statistical solution. In this approach, data from all pups in all litters are analyzed, and the variable "litter" is included in the analysis as an additional model factor ("effect"). This approach has

* Corresponding author at: Department of Public Health Sciences, University of Texas at El Paso, 500 West University, El Paso, TX, 79968, United States of America.
E-mail address: casobin@utep.edu (C.A. Sobin).

[1] The logic of statistical tests begins with the notion that the axiomatic condition of nature is variability. Parametric statistical tests provide methods for calculating a single ratio value (e.g., the F- ratio) that indicates the amount of variability found between groups (attributable to the independent variable) as compared to the amount of variability observed across all groups ("normal" error variability). Because the analysis of variability is at the heart of parametric tests, the numeric value of one observation (data point) cannot be the result of some underlying factor that similarly influences the numeric values of other observations. When litter membership produces likeness among data points within the litter (cluster), the observations are no longer independent, confounding detection of exposure effects and yielding spurious results.

https://doi.org/10.1016/j.ntt.2019.106841
Received 20 March 2019; Received in revised form 25 October 2019; Accepted 31 October 2019
Available online 19 December 2019
0892-0362/

Table 1
Summary table comparing ANOVA results from models testing age of eruption of lower incisors in Wistar rats.[a]

	df	*F*	*p*	*est.* ω^2
Litter scores	3, 29	1.811	$0.2 > p > .1$	0.0687
Individual scores	3, 185	5.166	$0.005 > p > .001$	0.0620

[a] From (Tittmar, 1975), results from a study that included 4 treatment groups including untreated controls, sham controls, ethanol exposed and chloral hydrate exposed animals.

been long favored because it allows inclusion of all data points generated. Including all data points ensures that 1) whatever level of similarity or variability may be present within litters is separated from treatment effects; and 2) the variability from all pups in all litters is fully represented in the model results. In addition, this approach can provide a direct statistical test of the extent to which intralitter variability contributed to an outcome – information that might prove valuable in understanding the interaction of genetics and environment.

A practical historical detail, however, seems to have strongly influenced the course of research practice. When the notion of including litter as a factor in statistical models was first recommended in the toxicological literature during the 1970's, computing capacity was very limited (by current standards). Completing analyses that included litter as a model effect was not a unified process and required multi-step analyses and extensive programming knowledge that many researchers did not possess (Singer, 1998). Personal computers were not in widespread use until the 1990's, and even at that time, statistical software was just developing. During these same years, a computationally simpler statistical solution was adopted by some in which litter was used as a "nested" fixed effect (rather than as a random effect).

Perhaps because of earlier computational limitations and confusion regarding how to model litter statistically, the non-statistical alternative of using one value per litter was suggested (Haseman and Hogan, 1975; Staples and Haseman, 1974b). For example, it was recommended that one or two representative pups per litter be selected; alternatively, outcome values from all pups within each litter could be averaged to produce one value per litter. By reducing the overall sample size to the number of litters, these approaches very effectively reduced the chances of Type I (false positive) errors. This approach also simplified statistical models. At the same time, these approaches worked from the assumption that intralitter likeness was great enough to justify selecting only one or two pups as representative of the responses of all pups in the litter, or using litter means as representative of entire litters. For studies in which intralitter likeness was strong, these methods could produce valid findings. In cases where intralitter likeness was not strong however, or where the degree of clustering differed across litters, or across treatment conditions, these methods could yield results that would not in fact represent the "true" outcome for a given study. No guidance or objective criteria were suggested for determining the degree of intralitter likeness in a study. Nonetheless, the practice of using one value to represent the litter, became the standard approach to dealing with intralitter likeness in neurodevelopmental toxicological research.

Rapid and extensive advances in personal computing capacity and computing software capability, particularly over the past decade, now make completely feasible the long-preferred statistical approach of including all pups from all litters, and modeling litter as a random effect in statistical models. From a purely statistical perspective, ensuring that no data are excluded from analyses is consistent with research design and statistical "best practices" (Resnik, 2000). All of the major statistical software packages now offer the option for constructing the appropriate statistical models to address the issue of clustered data.

In the sections below, the history of discussions regarding intralitter likeness in developmental neurotoxicological research is briefly reviewed; and contemporary advances in reproductive biological issues are considered that may stimulate new interest in intralitter likeness as a variable in its own right. Principles underlying the use of random effects in mixed models for clustered data are discussed. Model examples are then presented to illustrate how ignoring litter can result in false positive results, and how adding litter as a random effect corrects the problem. Additional worked examples illustrate how selecting representative pups from each litter can yield misleading results that are not consistent with results obtained when all data from all litters are used, and litter is modeled as a random effect.

2. The history of using one value per litter in developmental neurotoxicology

2.1. Early papers

The thalidomide crisis (Vargesson, 2015) in the early 1960s was largely responsible for bringing developmental rodent models into the field of quantitative risk assessment. Clearly, for rodent risk assessment studies, the dam would be treated and the fetuses examined. Discussions of whether the dam or the fetus should be the unit of statistical analysis began in the early 1970's. For some outcomes (malformation, mortality), one value per litter could be generated as the proportion of the litter affected.

For non-binary outcomes however, further discussion was needed. For studies requiring several treatment conditions, and using species with large litter sizes, including all values from all pups produced large *F*-ratio values for effect sizes that were not meaningful.[2] The issue of inflated *F* values was illustrated in data provided by Tittmar from a study with 4 treatment groups (Tittmar, 1975). Using pups as the unit of analysis produced a statistically significant *p* value while using a single litter value (litter mean) did not. It is important to note that there was no change in effect size (*est.* ω^2); both approaches explained 6% of the total variance observed.

By 1975, influential authorities in teratology (Haseman and Hogan, 1975; Staples and Haseman, 1974a) had concluded that analyses based on the individual fetus would pose far greater risk of false positive results (Type I error) than analyses based on the litter: *"… if litter effects are present, a per-fetus analysis would be invalid and might seriously exaggerate the level of significance."* (Staples and Haseman, 1974a). The premise here must be noted. The statement implied that a singular choice had to be made between the fetus or the dam as the individual responder. It is important to note that in these early discussions (Abbey and Howard, 1973; Haseman and Hogan, 1975; Kalter, 1974; Staples and Haseman, 1974a), researchers repeatedly pointed out that fetuses *could* respond as individuals to a drug or toxicant treatment, and thus should not *always* be considered equivalent subunits (Becker, 1974). However, no suggestion was made as to how to test whether and to what extent intralitter likeness was present.

While statisticians took on the task of determining how best to represent the litter and the individual pup in data analysis of binary endpoints in teratology (e.g., death, malformation) (Allen et al., 1994; Kavlock et al., 1995; Ryan, 1992; Ryan and Molenberghs, 1999), developmental neurotoxicology researchers largely turned to managing intralitter likeness by avoiding use of the pup as the unit of analysis. Following on the then-current wisdom, the litter was to be the unit of analysis for outcomes evaluated across the lifespan. The validity of such an approach was never re-examined; and whether clustering of

[2] All other things being equal, as sample size increases, the denominator of the *F* ratio decreases, resulting in a necessarily larger, and thus "more likely to be statistically significant," *F* value. This is true for any study, and is the rationale behind interpreting a *p*-value only in the context of an effect size parameter. Statistical significance is only meaningful if "enough" of the variance is explained by the result, although what is considered "enough" varies across fields of study.

outcomes within litter continued across the lifespan was not and has not been systematically evaluated.

2.2. Intralitter and interlitter variability in contemporary reproductive science

2.2.1. The biology of within litter likeness

Intralitter likeness has most often been seen as simply a variable that limits our ability to detect “true” effects of toxicologic or neurotoxicologic agents. In fact, a rich literature from the field of reproductive biology helps to conceptualize a different perspective.

Each conceptus in a rodent litter is a genetically distinct individual. While all littermates have the same male and female parent, the selection of parental alleles for each individual varies according to population allele frequencies, as amply demonstrated in the breeding of transgenic mice. The assortment of available alleles may be smaller in inbred vs. outbred rodent strains, but these strains continue to accumulate spontaneous mutations and polygenic variability (Casellas, 2011). In addition each conceptus has a placenta which is genetically distinct. This genetic diversity can be assumed to influence response to toxicants in the fetus as demonstrated in gene-toxicant interactions that occur within litters in transgenic mouse research.

Each conceptus also has a distinct intrauterine environment. The rodent uterus is bicornate. Two horns (right and left) extend from each ovary to the common cervix, with fetuses implanted along the length of each horn. In the mouse, the blood flow comes from both ends of the horns (Vom Saal and Dhar, 1992). Each fetus experiences a slightly different concentration of blood components depending on its position along the uterine artery. With bidirectional flow, fetuses next to the arterial caudal and cranial entry to the circular uterine artery are larger, while the fetus in the middle of the horn is the lightest and smallest. Notably, detailed studies of variation in birthweight have identified and confirmed an effect of intrauterine position in rats, with fetal body weights, and brain weights, smallest in the middle of the uterine horn (McLaurin and Mactutus, 2015). This is presumably based on differential access to nutrients. Differential distribution of drugs and toxicants has also been found in some studies (McLaurin and Mactutus, 2015). The relevance of intrauterine position to treatment effects has been discussed in the bisphenol A literature (Vom Saal, 2016).

In addition, the development of a rodent fetus is affected by the sex of its neighbors. The fetus secretes gonadal hormones into the intrauterine lumen, which can subsequently be taken up by its neighbors. This provides one of the most studied aspects of intralitter variability. A multitude of sex-differentiated traits of the offspring can be altered by the sex of adjacent fetuses in uterine horn (Ryan and Vandenbergh, 2002). These traits include size, morphology, brain structure, behavior, and reproductive characteristics. Response to drugs and toxicants often differ between sexes, making these effects subject to modification by intrauterine position mediated by the sex of adjacent fetuses (Vom Saal, 2016). Most of the literature in this area used laboratory rodents, but the same phenomenon has been demonstrated in wild animals, domestic animals, and even in same-sex vs different-sex human twins (Ryan and Vandenbergh, 2002).

In addition to chemical influences, other characteristics of adjacent fetuses can be considered. A recent paper described the effect of motility of adjacent fetuses on fetal activity (Brumley et al., 2018).

This research suggested that intralitter variability should not be simply “controlled” in statistical analyses, but could be used to better understand the intra-uterine effects of toxicants. (This approach may not be feasible in standard developmental neurotoxicology studies however because it requires caesarean section delivery to identify the fetus's uterine position.)

2.2.2. Sources and significance of intralitter likeness and between litter variability

When the litter is considered the unit of analysis, the identity of the dam is the basis of the independent variable. A large number of variables known to influence outcome measures are carefully controlled to minimize interlitter error variability and prevent confounding. This includes genetic background (strain), age, size, reproductive experience (parity), cage size, location, cagemates, and time in lab. In studies that follow government guidelines, litter sizes are often normalized shortly after birth by culling (Suvorov and Vandenberg, 2016) or used as a covariate.

Identity of the male parent is another source of interlitter variability that is seldom considered. Commonly the identity of the male parent is determined by the breeding strategy, which might involve housing a small group of females with a given male and then assigning pregnancies from that male to different experimental groups, or housing each female with a different, individual male.

In addition to the genetic makeup of parents, we now know that epigenetic marks attained prior to conception can be transmitted to offspring making the physiological characteristics and pre-breeding experiences of both the male and female parent more relevant (Morgan et al., 2019; Skvortsova et al., 2018). In addition to epigenetic transmission through DNA, gametes can transfer information to the conceptus via RNA and other small molecules (Perez and Lehner, 2019). These sources of interlitter variability can inform interpretation of developmental neurotoxicology studies. They may also be a source of unexplained replication failures.

It is valuable to keep in mind that multiple births occur in most mammals and are within the reproductive capabilities of humans. Multiple births in humans are genetically influenced, have increased rapidly in incidence in recent decades and are an important public health consideration (Fell and Joseph, 2012). Fraternal twins, like littermates, are genetically distinct individuals conceived at the same time. Siblings, like littermates, have the same two parents and experience an intrauterine environment generated by the same mother. Understanding the sources of intralitter variability can help inform translation of toxicant effects to humans (Sauce et al., 2018) and will require the use of modern statistical techniques to quantify and report intralitter likeness.

2.3. Later revisiting of intralitter likeness

In the 1990's, approximately twenty years after the first discussions in the literature, the issue of intralitter likeness was revisited with specific emphasis on questions relevant to the field of developmental neurotoxicology. In a key paper, Holson and Pearce (Holson and Pearce, 1992) summarized possible approaches for managing intralitter likeness, and emphasized once again that the statistical solution, in which all littermates would be included in all analyses, and litter would be modeled as a random effect, was the preferred approach.

Even at this time, however, modeling litter as a random effect in regression models was simply not feasible for many investigators. A unified computing approach was only first available in the early 1990's, and the default options for these models were not appropriate for many applications (Singer, 1998), complicating their use. Many did not yet have access to the computing resources required for adopting the preferred statistical solution. Once again, the recommended non-statistical alternative was to use one value per litter, that is, each litter contributed one and only one value to the data analysis, either the average for a given outcome from each litter, or the value from one pup selected as representative of the litter (Holson et al., 2007; Holson and Pearce, 1992; Lazic and Essioux, 2013). This approach was broadly adopted for both investigator-initiated and regulatory studies (Holson et al., 2007). Current guidelines for regulatory developmental neurotoxicology studies (OECD, 2007, 2011; US EPA, 1998) continue to call for the selection of one male pup and one female pup per litter for behavioral testing.

During these years, an additional complication arose as sex-differentiation of postnatal outcomes came into focus. A single male was no

Table 2
Summary of approaches used to address "intralitter likeness" for continuous outcome data.

Approach	Advantages	Disadvantages
Fetus/pup as the unit of measure in statistical models		
Include all fetuses/pups in ANOVA/ANCOVA analysis, ignore possibility of intralitter likeness	No data are excluded; variability is fully represented	Intralitter likeness could exaggerate or suppress detection of group differences; provides no objective measure of intralitter likeness; large overall sample size may suggest statistical significance for non-meaningful effect size (Type I error)
Include all fetuses/pups in ANOVA/ANCOVA analysis and include litter as a fixed "nested" effect	No data are excluded; variability is fully represented; reduces Type I error	Exaggerates estimated influence of litter on other fixed effect outcomes, increasing probability of Type II error; provides no objective measure of intralitter likeness
Include all fetuses/pups in general linear regression analysis; ignore possibility of intralitter likeness	No data are excluded; variability is fully represented	Intralitter likeness could exaggerate or suppress beta estimates for fixed effects; provides no objective measure of intralitter likeness
Include all fetuses/pups in generalized linear "mixed" regression analysis of fixed effects (e.g., treatment, sex) ***with litter included as a random effect***	No data are excluded; variability within litters is fully represented; analyses provide objective estimate of the amount of intralitter likeness observed; models allow selection of appropriate distribution for outcome (e.g., poisson for count data)	Initial familiarization with modeling for "mixed" (hierarchical) regression models is needed
One value per litter in statistical models		
Determine and use one mean value per litter	Reduces sample size to the number of litters; avoids inadvertent exaggeration of treatment effects, or suppression of treatment effect detection	No representation of variability that may occur within litters; sex differences are confounded by overall mean values
Randomly select one representative fetus/pup per litter regardless of sex	Reduces sample size to the number of litters; avoids inadvertent exaggeration of treatment effects, or suppression of treatment effect detection	No representation of variability that may occur within litters; possible sex differences are not captured; does not allow for test of sex by treatment interaction
Randomly select one representative male and one representative female per litter, and analyze sexes separately	Reduces sample size to the number of litters; avoids inadvertent exaggeration of treatment effects, or suppression of treatment effect detection	No representation of variability that may occur within litters; two animals per litter may reintroduce intralitter likeness confound; does not allow for test of sex by treatment interaction

longer an appropriate representative of the entire litter. With regard to statistical solutions, because sex was an intralitter variable, it could not be used as a fixed effect, similar to treatment, in an interlitter (ANOVA) model. This twist led to the approach of selecting one male and one female from each litter, or obtaining different litter means for males and females, and analyzing the data for the sexes separately (Holson et al., 2007). It was pointed out, however, that selecting one male and one female pup per litter simply re-introduced the problem of litter if the outcome examined was not obviously influenced by sex (Elswick et al., 2000); importantly, this approach left no option for the statistical evaluation of treatment-sex interactions.

Table 2 below provides an overview of the advantages and disadvantages of different suggested approaches. As the table suggests, using mixed models and including litter as a random effect has several statistical advantages, perhaps the most critical of which is full representation of the variability associated with all pups from all litters. Modern-day computing capacity and sophisticated statistical software make mixed models readily accessible to all researchers who have appropriate statistical expertise and/or access to statistical consultation. In recent years, researchers in the area of, for example, developmental neuroscience (Aarts et al., 2015), veterinary science (Festing, 2006), developmental nutrition (Wainwright et al., 2007), and developmental psychobiology (Williams et al., 2017) have repeatedly demonstrated the many advantages of statistical models that include litter as a random effect.

3. Litter as a random effect in contemporary statistical models

Including a random effect in statistical models to account for clustered data is widely used in many fields of study and, particularly in the past ten years, has become highly accessible through modern laptop computing programs. In the sections below, we review basic concepts that are central to understanding mixed models, with a goal of building confidence in this approach among toxicological and neurotoxicological researchers. To make this approach and the results as transparent as possible, simple statistical examples are provided that include two main effects (exposure condition and sex), one interaction (exposure x sex) and, for the mixed model, one random effect (litter). We note that many studies will require substantially more complex statistical models than those here presented. The correct use of hierarchical models can become quickly challenging. It is critical for researchers to fully understand and have confidence in the basic logic of these tests. To ensure correct application of these models and interpretation of results, it is equally important for researchers to seek consultation with a statistician who has specific expertise in mixed (hierarchical) modeling.

3.1. Four key concepts for use of the random effect in statistical models

3.1.1. Clustering

As discussed above, rearing animals in natural litters creates a potential for correlated outcomes due to, for example, individual differences among dams in genetics, physiology (e.g. metabolism), and/or behavior, and/or other features of the rearing environment. These in turn may influence biological, physiological and/or behavioral outcomes in the pups, causing pup outcomes to be correlated (clustered data). There is also the possibility that clustering may occur in some litters and not in others, and could somehow be associated with experimental effects, sex, and or individual differences of the dam. Including all pups from all litters in statistical analyses, with litter as a random effect, allows the researcher to determine the extent to which clustering influenced the outcomes; additional model options allow the researcher to determine and test differences between clustering by litter which would be a first step for examining whether and how clustering interacts with model fixed effects.

3.1.2. Independent observations

While parametric statistical tests are considered "robust" to some underlying assumptions regarding characteristics of the data to be analyzed, the assumption of "independent observations" cannot be violated. This assumption refers to the nature of the variability in the data, such that data points are influenced only by the manipulated (independent) variable (with other possible influencing factors

controlled, such as age or sex). Data clustering violates this critical assumption. Within litters, pup outcomes may not be "independent," but instead may be the result of maternal and/or environmental characteristics (and are thus clustered). In group comparison models, group means are the foundation for determining "sum of squares between" and "sum of squares within," the parameters needed to calculate the final test statistic, for example, the *F*-ratio for an analysis of variance. When the assumption of independent observations is violated, the *F*-ratio becomes invalid. When group means are determined from clustered data within treatment groups, the effects of clustering cannot be distinguished from the treatment effect. When litter is simply ignored and all pups are included, "statistical significance" can be due simply to clustering, in other words, strong intralitter similarity that may or may not have been influenced by the independent variable (e.g., treatment effect). For this reason, the possible effects of litter cannot be ignored.

3.1.3. Fixed vs. random effects

In statistical models, whether the variability associated with an "effect" can be fully estimated fundamentally determines how an effect is mathematically included. The most commonly encountered models in the scientific literature are "single-level" models that include only fixed effects ("effect" here referring to a type of categorical predictor variable). When all categories of an effect are represented in a given dataset, the variability associated with the effect *can* be fully estimated. These types of effects are considered "fixed" in terms of the statistical model. Some examples of fixed effects are "malformation" (yes/no); "sex" (male/female); and "lifestage" (immature/mature/senescent), when all possible outcomes for the effects are included in the dataset.

While the most commonly reported models tend to include only fixed effects, a given categorical predictor variable can in fact be either a fixed effect or a random effect. A random effect is an effect for which not all possible levels or conditions have been sampled (not to be conflated with randomization or random assignment) (Eisenhart, 1947). If child subjects are selected from 5 of 12 possible elementary schools, "school" would be considered a random effect, because the variability associated with the effect "elementary school" is *not* fully represented. Similarly, the variable "litter" is a random effect because the set of litters in any given study cannot be assumed to fully represent the variability associated with a given outcome for "all litters" under the conditions tested. Because not all possible categories of the variable have been sampled, the data cannot be used to fully represent the variability associated with the effect, and the variability of the effect cannot be fully estimated in the model.[3]

3.1.4. Random effects in statistical models

While using litter as a random effect in statistical models was discussed for many years earlier, it was not until the early 1990's that SAS® software (SAS Institute: Cary, NC) was among the first to offer a unified method for including a random effect in a regression model (through the coding procedure "PROC MIXED"). Even six years after it was introduced, experts noted that most researchers were not familiar with this "newly" available analytic option and/or the conceptual basis for this approach (Singer, 1998), and excellent guidance papers were published (Singer, 1998). The approach has evolved and "next-gen" versions are now widely available in all standard statistical packages including for example, SAS, SPSS (®IBM Corp., Armonk: NY), and R (R Foundation for Statistical Computing, Vienna: Austria).

[3] As the reader has probably noted by now, it would not be difficult to develop a rationale for considering many variables as "random" rather than "fixed." There is a tendency in the literature for most predictor variables to be used as fixed effects although this may be more of a reflexive response rather than conscious decision. Not insignificantly, graduate students still tend to be taught the concepts of linear regression using only fixed effect models (the variability for which can be fully estimated).(Stroup, 2013)

Statistical models can include only fixed effects, only random effects, or both fixed and random effects. Models that include both fixed and random effects are referred to as "mixed" or "hierarchical" models because the fixed and random effects are conceptualized on two levels. Level 1 variables are fixed effects for which all subjects have at least one data point (typically the main outcome variable). For example, if the primary outcome in an exposure study is body weight, all subjects in the dataset will have one data point for the variable bodyweight. The Level 2 variables on the other hand are the effects for which not all levels are represented in the dataset, and each subject belongs to one and only one category of the level (e.g., litter).

How the random effect works in a statistical model is also important to consider. Simply stated, the random effect provides an alternate method for estimating variability for clustered data. When data are clustered, group means, used in the calculation of primary test statistics, are invalid because the variability around the group mean is restricted by characteristics of the cluster. Thus, for the purpose of calculating the primary test statistic (e.g., *F*-ratio), variability must be characterized in some way that does not depend on the cluster (litter) group means. Calculating a simple linear equation (slope) for each litter (for example, characterizing the change in bodyweight at a given level of contaminant exposure) produces a unique intercept for each litter. Unlike the mean, the outcome for each pup *will* vary freely around the intercept. In this way, the random effect quantifies the unique variability attributable to "litter" while controlling for its effects in the statistical model.

Importantly, calculation of the random effect is robust to differences in clustering across litters (for example, if a given exposure level interacts with litter effects resulting in strong clustering among one or more experimental groups and not in controls, or vice versa). Depending on the model options chosen within a given software package, parameters that indicate the amount of clustering within each litter can be requested, perhaps providing an additional perspective from which to understand mechanistic effects.

4. Comparison of results from simple fixed and mixed models using all pups ($N = 180$)

How variables are modeled in statistical tests fundamentally alters study results. Comparing the results from different models can be an efficient way of illustrating how different models estimate parameters, and how these differences determine study results and conclusions. We provide and discuss below alternative methods for analyzing data from a simulated study of developmental lead exposure.

The ranges of data for the database used in these examples were guided by actual data from previously published studies examining effects of developmental lead exposure (Basgen and Sobin, 2013; Flores-Montoya et al., 2015; Flores-Montoya et al., 2019; Flores-Montoya and Sobin, 2014; Sobin et al., 2017; Sobin et al., 2013) in brain, kidney and on behavior. The database included $N = 180$ pups in a balanced design of three exposure groups with equal numbers of males and females in each litter: controls (10 litters, 6 pups per litter, $n = 60$), low dose (30 ppm lead) (10 litters, 6 pups per litter, $n = 60$), and high dose (430 ppm lead) (10 litters, 6 pups per litter, $n = 60$). Exposures were delivered through dams' drinking water from birth to post-natal day (PND) 21, at which time body weights of all mice were measured and recorded.

Below we compare results from two models tested with the database described above ($N = 180$). The first analysis is a simple fixed model testing the main effects of exposure group, sex, and the interaction (exposure group × sex) on body weight at PND 21, with litter ignored. The second analysis is a mixed model testing the same effects with litter included as a random effect. Results from the simple fixed effects model suggested that the main effect (exposure) significantly influenced body weight. Results from the mixed effects model with litter included as a random effect showed that when litter clustering was included in the

Table 3
Comparison of model dimensions for simple model and mixed model.

FIXED EFFECTS MODEL		No. Levels	No. Parameters
Fixed Effects	Intercept	1	1
	Group	3	2
	Sex	2	1
	Group * Sex	6	2
Residual			1
Total		12	7

MIXED EFFECTS MODEL		No. Levels	Covariance Structure	No. Parameters	Subject Variables
Fixed Effects	Intercept	1		1	
	Group	3		2	
	Sex	2		1	
	Group * Sex	6		2	
Random Effects			Variance Components	1	Litter
Residual				1	
Total		12		8	

model, exposure group was not a significant predictor of body weight. That is, adding litter as a random effect effectively controlled the random effect of litter (clustering), "correcting" the false positive finding from the simple fixed effects model, while maximizing for the purpose of statistical estimation all of the variability observed in the study ($N = 180$).

To illustrate possible outcomes when the strategy of selecting 1 male and 1 female is used to eliminate the problem of litter clustering, four different unique databases of $N = 60$ each, were drawn from the full database of $N = 180$; each smaller dataset was created by selecting two (different) animals per litter. Each database was then tested using a simple fixed effects model (with litter accounted for by pup selection). Overall, none of the results from the smaller databases were consistent with results from the mixed model ($N = 180$) with litter as a random effect. Details are provided below.

For each analysis, the model parameters, the Type III sum of squares tests, and the parameter estimates are reviewed. Supplementary appendices provide the complete databases and software code (SPSS and SAS) used to generate the model results.

4.1. Simple fixed effects model with litter ignored vs. mixed model with litter as a random effect: comparison of model dimensions

The dimensions for the simple fixed and mixed models are nearly identical, with the addition of one parameter, litter, in the mixed effects model. Values in the parameters column show the number of parameters that will be calculated and displayed in the parameter estimates table (Table 3) for each model (number of variable categories minus 1, which is the reference group, in this case, the control group).

4.2. Simple fixed effects model with litter ignored vs. mixed model with litter as a random effect: comparison of Type III tests of effects

The Type III tests (Table 4 below) provide the overall test of significance for each fixed effect (two main effects and one interaction). In the simple model, the denominator *df* shows that all pups were included as independent observations (*Den df* for group = 180–6 = 174; *Den df* for sex and the interaction = 180–30–3 = 147). The overall test of the simple model suggested that differences in body weight were attributable to exposure effects and sex; the interaction between exposure group and sex was not statistically significant (exposure effects on body weight by group did not differ for males and females).

For the mixed model results that included litter as a random effect, also shown in Table 4, the Type III Test parameters are similar to the simple model results with one critical difference. The denominator *df* is 27 (*Den df* for group = 30–3 = 27; *Den df* for sex and the interaction = 180–30–3 = 147). This critically alters the calculation of the model. In this model, the overall test of the exposure group main effect is not statistically significant, suggesting that the statistical significance for exposure group in the simple model was due to inflation of group effects from data clusters in litters (within in exposure groups).

Calculating a mixed model with litter added as a random effect revealed a false positive result in the simple fixed model. The mixed model effectively controlled for the effects of clusters within litters while incorporating the variability from all pups in all litters. Also in this model, the main effect of sex was significant (as in the first model) and the interaction was not significant (as in the first model) suggesting that overall, regardless of exposure group, males weighed more than females. This was what would be expected if exposure did not alter expected physiological differences between males and females.

4.3. Simple fixed effects model with litter ignored vs. mixed model with litter as a random effect: comparison of parameter estimates

Whether any of the specific contrasts associated with each significant Type III test main effect reached statistical significance was determined by examining the model parameters shown in Table 5 below. All estimates in this table are unadjusted for other variables in the model.

The estimate for the intercept in this model is the mean bodyweight for the controls (reference group). The parameter estimate for the low and high dose groups is the amount of difference from the controls, in other words, the amount of difference between each experimental group and controls; or the amount by which males differ from females. (By default, SPSS uses the group coded with the highest numeric (arbitrary) value as the reference group; thus through coding choices, the researcher determines which group is used as the reference group.) The "estimate" for each reference group (that is, exposure group controls, or females) is "0".

As shown in Table 5 below, the 'estimate' values from the two models are identical, but, once again, the *df* differ reflecting the inclusion of litter as a random effect in the mixed model. Also, the

Table 4
Comparison of Type III tests of effects for simple fixed and mixed models.

Type III Tests, Simple Model with Fixed Effects only (N = 180)

Source	Num df	Den df	F	P
Intercept	1	174	9275.22	.000
Group	2	174	11.90	.000
Sex	1	147	8.40	.004
Group * Sex	2	147	0.13	.879

Type III Tests, Mixed Model with Litter as a Random Effect (N = 180)

Source	Num df	Den df	F	P
Intercept	1	27	2146.38	.000
Group	2	27	2.76	.082
Sex	1	147	21.55	.000
Group * Sex	2	147	0.33	.718

Table 5
Comparison of parameter estimates for simple and mixed models.

Simple Fixed Effects Model - Estimates of Fixed Effects

						95% CI	
Parameter	Estimate	Std Err	df	t	p	Lower	Upper
Intercept	12.40	0.30	174	40.70	.000	11.80	13.00
High Dose	-1.21	0.43	174	-2.81	.006	-2.06	-0.36
Low Dose	-1.11	0.43	174	-2.57	.011	-1.96	-0.26
Control	0	0	.	.	.	.	.
Male	0.89	0.43	174	2.07	.040	0.041	1.74
Female	0	0	.	.	.	.	.
High Dose*Male	-0.21	0.61	174	-0.34	.732	-1.41	0.99
High Dose*Female	0	0	.	.	.	.	.
Low Dose*Male	-0.30	0.61	174	-0.50	.620	-1.51	0.90
Low Dose*Female	0	0	.	.	.	.	.

Mixed Effects Model – Estimates of Fixed Effects

						95% CI	
Parameter	Estimate	Std Err	df	t	p	Lower	Upper
Intercept	12.40	0.47	27	26.50	.000	11.44	13.35
High Dose	-1.21	0.66	27	-1.80	.077	-2.56	0.14
Low Dose	-1.11	0.66	27	-1.67	.104	-2.45	0.24
Control	0	0	.	.	.	.	.
Male	0.89	0.27	147	3.32	.001	0.360	1.42
Female	0	0	.	.	.	.	.
High Dose*Male	-0.21	0.38	147	-0.55	.583	-0.96	0.54
High Dose*Female	0	0	.	.	.	.	.
Low Dose*Male	-0.30	0.38	147	-0.80	.427	-1.06	0.45
Low Dose*Female	0	0	.	.	.	.	.

calculation of the standard error, t, p and confidence intervals differ (the *t*-tests determined whether each estimate value differed significantly from zero). Thus, in both models, the mean (unadjusted) body weight of control mice was 12.40 g; high dose mice weighed 1.21 g less than controls; low dose mice weighed 1.11 g less than controls.

While the Type III tests tell the significance of a given factor in the entire model, the parameter estimates are evaluated to determine which comparisons differed and by how much. (In these examples we evaluate the significance of the parameter estimates; depending on the researcher's preference, post-hoc comparisons of the adjusted means can be requested and evaluated as an additional step in determining significance of effects.) Similar to the false positive results from the Type III tests, parameter estimates calculated from the simple fixed

Table 6
Marginal (adjusted) means for mixed effects model.

					95% confidence interval	
		Mean	SE	df	Lower	Upper
High dose		11.53	0.45	27	10.61	12.45
Low dose		11.58	0.45	27	10.67	12.50
Controls		12.84	0.45	27	11.92	13.76
Male		12.35	0.27	32	11.80	12.90
Female		11.62	0.27	32	11.10	12.18
High dose	Male	11.87	0.47	32	10.92	12.82
	Female	11.19	0.47	32	10.23	12.14
Low dose	Male	11.88	0.47	32	10.93	12.83
	Female	11.29	0.47	32	10.34	12.24
Controls	Male	13.29	0.47	32	12.34	14.24
	Female	12.40	0.47	32	11.44	13.35

model suggested that both exposure group differences were statistically significant. In the mixed model with litter included as a random effect, these differences were no longer statistically significant. Estimated marginal means (adjusted means) for these analyses generated by SPSS are provided in Table 6.

4.4. Simple fixed effects model with litter ignored vs. mixed model with litter as a random effect: comparison of variance explained (residuals)

To further compare the strength of these models, it is useful to consider parameters that indicate "goodness of fit" and amount of variance explained (not shown in the tables above). Restricted log likelihood values for the simple model (in "smaller-is-better" format) was 692.58, and for the mixed model was 593.67, representing modest model fit improvement with the mixed model. More impressive perhaps, the covariance residual (unexplained variance) was 2.79 g in the simple fixed model that ignored litter; for the mixed model the residual was 1.08 g, with litter accounting for approximately 1.7 g previously unexplained variance. In this way, the mixed model accounted for a substantially greater amount of variance, leaving approximately 1 g of variance unaccounted for in the mixed model (as compared with 2.79 g in the simple fixed model).

4.5. Using intraclass correlation (ICC) and design effect (DE) to examine the extent to which litter contributed to the model results

Using the same database ($N = 180$) to calculate two different models (a simple fixed effects model and a mixed model with litter included as a random effect) showed the critical importance of including litter as a random effect in a mixed model. Doing so approximately doubled the amount of variance explained, controlled for litter clustering, incorporated all of the variance in the study (all pups included), and of critical importance, revealed a false positive result produced by the simple fixed effects model with litter ignored. This provides a quantitative demonstration of the central importance of including litter as a random effect in analyses of developmental data from litters. At the same time, the full mixed model gives no quantification of the extent to which clustering was present within litters.

Quantitatively estimating the degree of litter clustering in a given dataset requires only two additional calculations. First, the intra-class correlation coefficient (*ICC*) for litter is determined (Müller and Büttner, 1994; Stanish and Taylor, 1983; West et al., 2007), then the Design Effect is calculated using the ICC. The Design Effect (DE) (Kish, 1965) was originally developed to provide a mathematical adjustment for statistical models with clustered data, conceptually similar perhaps to adjustments for multiple comparisons (Chen et al., 2017). In the current application, where litter is included as a random effect in a mixed model, calculating the design effect (*DE*) estimates whether the clustering in the data was strong enough to warrant using a multi-level mixed model (Kish, 1965; Muthen and Satorra, 1995). It has been suggested that when the *DE* is < 2, clustered data can be safely ignored (Muthen and Satorra, 1995). Thus, in addition to corroborating conclusions from the statistical model, these parameters can provide the researcher with rationale for either keeping the random effect of litter in the model (in cases where litter clustering contributes to the overall variability of the model), or re-calculating the model excluding litter (in cases where litter clustering did not influence results).[4]

The conceptual formula for the *ICC* is intuitive. The formula compares the variability attributable to litter clusters, to unexplained variability, ignoring all other factors in the model. The values of *ICC* range from −1.0 to +1.0 and larger positive *ICC* values indicate stronger association, in this case, greater clustering within litters. Unlike other forms of correlation that range between −1.0 and + 1.0, however, negative values of the *ICC* are not theoretically meaningful, and are simply interpreted as lack of within cluster likeness similar to 0 (Giraudeau, 1996).

Readers will find many discussions of and possibilities for calculation of the *ICC* in the statistical and research literature. We illustrate here one approach characterized specifically for litter-related associations (West et al., 2007).

$$ICC = \frac{\sigma^2 \text{ litter}}{(\sigma^2 \text{ litter} + \sigma^2)}$$

This formula compares the amount of variance attributable to litter as compared to total variance in the model. To determine the correct variance estimates for this formula, a simple separate mixed model is calculated predicting the outcome (e.g., bodyweight) from only one random predictor (e.g., litter). This yields covariance estimates for litter and the residual (unexplained) variance. The *DE* is then calculated using the following formula:

$$DE = 1 + (\textit{average cluster size} - 1) \times ICC$$

Thus, for the dataset of $N = 180$ used in the above analyses, the covariance parameters were: for litter, 2.05; and for the residual (all other variance in the model not attributable to litter) 1.23. Using these values in the above formula gives an $ICC = 0.63$ and $DE = 3.78$. Given the large difference in outcomes produced by simple fixed and mixed models, the relatively large *ICC* and *DE* are perhaps not surprising.

In mixed models that include data from all pups with litter as a random effect, the pup is the unit of analysis and clustering is statistically controlled. Mixed model results are influenced by the amount of clustering, but not the number of litters. For this reason, when an outcome is strongly clustered within all litters, and an adequate number of litters has been used (based on current literature for a given outcome, for example, between 6 and 10 litters per treatment group), all other things being equal, the statistical significance of group effects will be harder to detect. It is important to note that, in this case, adding more (and more) litters will not be useful. While the increased litter number could eventually produce a significant *p* value for a group effect, the effect size would not be expected to increase. Significant *p* values are only meaningful when a reasonable amount of variance is explained by the model. (Table 1, page 7, and Footnote 2, above). Note also that in epidemiological research, the ICC is used to determine sample size prior to experimentation (Killip et al., 2004), and could be used for this purpose in developmental neurotoxicology.

[4] While the principle of parsimony suggests that, all other things being equal, the simplest model that best explains the data is preferred, there is considerable debate among statisticians and researchers regarding whether or not a priori modeling decisions (e.g., including the use of litter as a random effect in a mixed model) should be revised following evaluation of model results. We offer the additional calculations as a methodologic option, with no specific recommendation for its use.

5. Consideration of results from analyses of selected mouse pairs (1 male/1 female per litter)

Current toxicological guidance recommends that researchers manage the potential problem of litter by selecting 1 male and 1 female from each litter as representative of the treatment effects. For a study of $N = 180$, with three treatment groups, 10 litters per treatment, each with 3 males and 3 females, 9 unique pairings of 1 male/1 female are possible for each litter. For 30 litters, 270 unique combinations of 1 male and 1 female from each litter (each of $N = 60$) are possible. While there is an attractive simplicity to randomly selecting 1 male and 1 female per litter to correct for possible litter effects, with 270 possible unique combinations, and litter clustering that nonetheless includes some variability (demonstrated in the above examples), it is logical to suggest that subsamples of data could differ in the extent to which they represent the entire sample (that is, results obtained from the mixed model of $N = 180$ with litter included as a random effect). To illustrate the possibilities, four unique databases of $N = 60$ each, were created by selecting different pairings of 1 male and 1 female per litter from the full sample of $N = 180$ used in the above model examples. We tested each unique dataset using a simple fixed effects model with two main effects (exposure group and sex) and one interaction (exposure group x sex). Because the mixed model shown above included all data from all litters, controlled for litter as a random effect, and substantially improved the amount of variance explained, for comparison purposes, we use the results obtained from the mixed model analysis as the closest approximation to "true" effects in the sample data. Table 6 below shows the range of significant and non-significant values obtained. To simplify the data presentation, the parameter estimates for the (non-significant) interaction are not shown.

5.1. Simple fixed effects models of N = 60 subsamples (1 m/1f) vs. mixed model with litter included as a random effect: comparison of results

The mixed model calculated above ($N = 180$) with litter included as a random effect suggested that exposure group did not significantly contribute to the variance in the data and, across all groups, males weighed more than females. Table 7 below shows results from 4 randomly selected

Table 7
Comparison of results from simple fixed effects model for four unique samplings of (1 male/1 female per litter) from the source database of $N = 180$.

Sampling A: N = 60, 1m/1f per litter

Source	Numerator df	Denominator df	F	p
Intercept	1	54	4432.68	.000
Group	2	54	18.47	.000
Sex	1	54	5.84	.019
Group * Sex	2	54	0.45	.639

						95% CI	
Parameter	Estimate	Std Err	df	t	p	Lower	Upper
Intercept	12.63	0.43	54	29.37	.000	11.78	13.49
High Dose	-1.95	0.61	54	-3.20	.002	-3.17	-0.73
Low Dose	-2.14	0.61	54	-3.53	.001	-3.36	-0.92
Control	0	0	.	.	.	.	.
Male	1.34	0.61	54	1.87	.068	-0.08	2.35
Female	0	0	.	.	.	.	.

Sampling B: N = 60, 1m/1f per litter

Type III Tests of a simple fixed effects

Source	Numerator df	Denominator df	F	p
Intercept	1	54	4007.46	.000
Group	2	54	22.73	.000
Sex	1	54	3.28	.076
Group * Sex	2	54	0.01	.994

						95% CI	
Parameter	Estimate	Std Err	df	t	p	Lower	Upper
Intercept	13.14	0.45	54	29.05	.000	12.24	14.05
High Dose	-2.65	0.64	54	-4.15	.000	-3.94	-1.37
Low Dose	-2.70	0.64	54	-4.22	.000	-3.98	-1.41
Control	0	0	.	.	.	.	.
Male	0.63	0.64	54	0.98	.332	-0.66	1.91
Female	0	0	.	.	.	.	.

(*continued on next page*)

Table 7 (*continued*)

Sampling C: N = 60, 1m/1f per litter

Type III Tests of a simple fixed effects				
Source	Numerator df	Denominator df	F	p
Intercept	1	54	3671.99	.000
Group	2	54	4.03	.023
Sex	1	54	1.95	.169
Group * Sex	2	54	0.31	.733

						95% CI	
Parameter	Estimate	Std Err	df	t	p	Lower	Upper
Intercept	13.18	0.52	54	25.35	.000	11.80	13.00
High Dose	-0.96	0.74	54	-1.30	.198	-2.43	0.52
Low Dose	-0.89	0.74	54	-1.21	.233	-2.36	0.59
Control	0	0	.	.	.	.	.
Male	1.06	0.74	54	1.45	.153	-0.41	2.54
Female	0	0	.	.	.	.	.

Sampling D: N = 60, 1m/1f per litter

Type III Tests of a simple fixed effects				
Source	Numerator df	Denominator df	F	p
Intercept	1	54	3497.45	.000
Group	2	54	3.38	.041
Sex	1	54	2.22	.142
Group * Sex	2	54	0.45	.640

						95% CI	
Parameter	Estimate	Std Err	df	t	p	Lower	Upper
Intercept	13.02	0.53	54	24.49	.000	11.95	14.08
High Dose	-0.80	0.75	54	-1.06	.295	-2.30	0.71
Low Dose	-0.72	0.75	54	-0.96	.340	-2.23	0.78
Control	0	0	.	.	.	.	.
Male	1.23	0.75	54	1.63	.208	-0.28	1.74
Female	0	0	.	.	.	.	.

subsamples of 1 male and 1 female (unique pairings) from the source database of $N = 180$. The results obtained varied broadly across the subsamples and none of the models agreed with the results obtained in the mixed model that included all pups with litter as a random effect.

With regard to the main effect of exposure group, results from Samplings A and B suggested that exposure group had a significant effect on body weight; in both samplings, the differences between the low and high dose group were estimated to be similar and both differed significantly from controls (false positive results). In Samplings C and D, the Type III tests suggested a possible significance of exposure group however tests of the parameter estimates proved to be non-significant. None of the sampling models detected the significantly higher body weights among males. The only point of agreement across all samplings and the mixed model results (using all animals from all litters with litter as a random effect), was the lack of an interaction effect. It should be noted that for these tests of data that included 1 male and 1 female per litter, calculating mixed models with litter as a random effect would not be a preferred approach. With only 1 male and 1 female selected per litter, the variability associated with litter is minimally represented. Thus, not surprisingly, calculating mixed models with litter as a random effect for the subsample databases that included 1 male and 1 female per litter did not alter the results.

6. Conclusions

For almost 50 years, there has been ongoing discussion of how best to handle statistical issues related to data clustering among multiple offspring in developmental neurotoxicologic and teratologic studies. Our understanding of the biological plausibility and perhaps new significance of intralitter likeness continues to expand. Early guidance recommended a statistical solution for data clustered by litter, which allowed researchers to include all littermates from all litters, and control for data clustering as a model effect. Doing so prior to the 1990's however was not a unified operation, leading perhaps to the widespread use of a different solution, that of determining one value per litter, either via calculation of a litter mean, or by selecting only representative littermates for analysis. To date, a majority of research studies have controlled for litter by using a litter mean, or one (or two) representative values per litter, perhaps impacting the extent to which the variability associated with all pups from all litters was represented in study results. Using data from all littermates in all litters and including litter as a random effect in mixed models, provides superior results as compared to using all littermates and ignoring the issue of litter, and as compared to using values from only representative littermates. Statistically modeling clustered data as a random effect is a preferred approach in many fields of study, and modern computing

capacity now provides ready accessibility for all researchers to this statistical solution for intralitter likeness. At the same time, accessibility does not preclude the need for appropriate expertise and consultation in the use of hierarchical (mixed) models.

Transparency document

The Transparency document associated this article can be found, in online version.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Supported by the California National Primate Research Center, National Institutes of Health, Office of Research Infrastructure (OD011107) (MSG); and the National Center for Research Resources, National Institutes of Health (5G12RR008124); Center for Clinical and Translational Science, The Rockefeller University; the J. Edward and Helen M.C. Stern Endowed Professorship in Neuroscience, University of Texas at El Paso (CAS).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ntt.2019.106841.

References

Aarts, E., Dolan, C.V., Verhage, M., van der Sluis, S., 2015. Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. BMC Neurosci. 16, 94. https://doi.org/10.1186/s12868-015-0228-5.

Abbey, H., Howard, E., 1973. Statistical procedure in developmental studies on species with multiple offspring. Dev. Psychobiol. 6 (4), 329–335. doi. https://doi.org/10.1002/dev.420060406.

Allen, B.C., Kavlock, R.J., Kimmel, C.A., Faustman, E.M., 1994. Dose-response assessment for developmental toxicity. III. Statistical models. Fundam. Appl. Toxicol. 23 (4), 496–509 (doi:S0272059084711341 [pii]).

Basgen, J.M., Sobin, C., 2013. Early chronic low-level lead exposure produces glomerular hypertrophy in young C57BL/6J mice. Toxicol. Lett. 225 (1), 48–56. https://doi.org/10.1016/j.toxlet.2013.1011.1031.

Becker, B.A., 1974. The statistics of teratology. Teratology 9 (3), 261–262. https://doi.org/10.1002/tera.1420090305.

Brumley, M.R., Hoagland, R., Truong, M., Robinson, S.R., 2018. Responsiveness of rat fetuses to sibling motor activity: communication in utero? Dev. Psychobiol. 60 (3), 265–277. https://doi.org/10.1002/dev.21615.

Casellas, J., 2011. Inbred mouse strains and genetic stability: a review. Animal 5 (1), 1–7. https://doi.org/10.1017/s1751731110001667.

Chen, S.-Y., Feng, Z., Yi, X., 2017. A general introduction to adjustment for multiple comparisons. J. Thorac. Dis. 9 (6), 1725–1729. https://doi.org/10.21037/jtd.2017.05.34.

Eisenhart, C., 1947. The assumptions underlying the analysis of variance. Biometrics 3 (1), 1–21. https://doi.org/10.2307/3001534.

Elswick, B.A., Welsch, F., Janszen, D.B., 2000. Effect of different sampling designs on outcome of endocrine disruptor studies. Reprod. Toxicol. 14 (4), 359–367. https://doi.org/10.1016/s0890-6238(00)00092-7.

Fell, D.B., Joseph, K., 2012. Temporal trends in the frequency of twins and higher-order multiple births in Canada and the United States. BMC Pregnancy Childbirth 12, 103. https://doi.org/10.1186/1471-2393-12-103.

Festing, M.F., 2006. Design and statistical methods in studies using animal models of development. ILAR J. 47 (1), 5–14.

Flores-Montoya, M.G., Sobin, C., 2014. Early chronic lead exposure reduces exploratory activity in young C57BL/6J mice. J. Appl. Toxicol. https://doi.org/10.1002/jat.3064. n/a-n/a.

Flores-Montoya, M.G., Alvarez, J.M., Sobin, C., 2015. Olfactory recognition memory is disrupted in young mice with chronic low-level lead exposure. Toxicol. Lett. 236 (1), 69–74. https://doi.org/10.1016/j.toxlet.2015.04.013.

Flores-Montoya, M.G., Bill, C.A., Vines, C.M., Sobin, C., 2019. Early chronic low-level lead exposure reduced C-C chemokine receptor 7 in hippocampal microglia. Toxicol. Lett. 314, 106–116. https://doi.org/10.1016/j.toxlet.2019.07.015.

Giraudeau, B., 1996. Negative values of the intraclass correlation coefficient are not theoretically possible. J. Clin. Epidemiol. 49 (10), 1205. https://doi.org/10.1016/0895-4356(96)00053-4.

Haseman, J.K., Hogan, M.D., 1975. Selection of the experimental unit in teratology studies. Teratology 12 (2), 165–171. doi. https://doi.org/10.1002/tera.1420120209.

Holson, R.R., Pearce, B., 1992. Principles and pitfalls in the analysis of prenatal treatment effects in multiparous species. Neurotoxicol. Teratol. 14 (3), 221–228 (doi:0892-0362(92)90020-B [pii]).

Holson, R.R., Freshwater, L., Maurissen, J.P., Moser, V.C., Phang, W., 2007. Statistical issues and techniques appropriate for developmental neurotoxicity testing a report from the ILSI risk science institute expert panel on neurodevelopmental endpoints. Neurotoxicol. Teratol. https://doi.org/10.1016/j.ntt.2007.06.001. doi:S0892-0362(07)00287-5 [pii].

Hughes, C.W., 1979. Outcome of early experience studies as affected by between-litter variance. J. Nutr. 109 (4), 642–645. https://doi.org/10.1093/jn/109.4.642.

Kalter, H., 1974. Editorial: choice of the number of sampling units in teratology. Teratology 9 (3), 257–258. doi. https://doi.org/10.1002/tera.1420090303.

Kavlock, R.J., Allen, B.C., Faustman, E.M., Kimmel, C.A., 1995. Dose-response assessments for developmental toxicity. IV. Benchmark doses for fetal weight changes. Fundam. Appl. Toxicol. 26 (2), 211–222 (doi:S0272059085710925 [pii]).

Killip, S., Mahfoud, Z., Pearce, K., 2004. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. Ann. Fam. Med. 2 (3), 204–208. https://doi.org/10.1370/afm.141.

Kish, L., 1965. Survey Sampling. John Wiley & Sons, Inc, New York.

Lazic, S.E., Essioux, L., 2013. Improving basic and translational science by accounting for litter-to-litter variation in animal models. BMC Neurosci. 14, 37. https://doi.org/10.1186/1471-2202-14-37.

McLaurin, K.A., Mactutus, C.F., 2015. Polytocus focus: uterine position effect is dependent upon horn size. Int. J. Dev. Neurosci. 40, 85–91. https://doi.org/10.1016/j.ijdevneu.2014.11.001.

Morgan, C.P., Chan, J.C., Bale, T.L., 2019. Driving the next generation: paternal lifetime experiences transmitted via extracellular vesicles and their small RNA cargo. Biol. Psychiatry 85 (2), 164–171. https://doi.org/10.1016/j.biopsych.2018.09.007.

Müller, R., Büttner, P., 1994. A critical discussion of intraclass correlation coefficients. Stat. Med. 13 (23–24), 2465–2476. https://doi.org/10.1002/sim.4780132310.

Muthen, B.O., Satorra, A., 1995. Complex sample data in structural equation modeling. Sociol. Methodol. 25, 267–316.

OECD, 2007. Test No. 426: Developmental Neurotoxicity Study.

OECD, 2011. Test No. 443: Extended One-generation Reproductive Toxicity Study.

Perez, M.F., Lehner, B., 2019. Intergenerational and transgenerational epigenetic inheritance in animals. Nat. Cell Biol. 21 (2), 143–151. https://doi.org/10.1038/s41556-018-0242-9.

Resnik, D.B., 2000. Statistics, ethics, and research: an agenda for education and reform. Account. Res. 8 (1–2), 163–188. https://doi.org/10.1080/08989620008573971.

Ryan, B.C., Vandenbergh, J.G., 2002. Intrauterine position effects. Neurosci. Biobehav. Rev. 26 (6), 665–678.

Ryan, L., 1992. The use of generalized estimating equations for risk assessment in developmental toxicity. Risk Anal. 12 (3), 439–447.

Ryan, L., Molenberghs, G., 1999. Statistical methods for developmental toxicity. Analysis of clustered multivariate binary data. Ann. N. Y. Acad. Sci. 895, 196–211.

Sauce, B., Bendrath, S., Herzfeld, M., Siegel, D., Style, C., Rab, S., ... Matzel, L.D., 2018. The impact of environmental interventions among mouse siblings on the heritability and malleability of general cognitive ability. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 373 (1756). https://doi.org/10.1098/rstb.2017.0289.

Singer, J.D., 1998. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. J. Educ. Behav. Stat. 23 (4), 323–355. https://doi.org/10.3102/10769986023004323.

Skvortsova, K., Iovino, N., Bogdanovic, O., 2018. Functions and mechanisms of epigenetic inheritance in animals. Nat. Rev. Mol. Cell Biol. 19 (12), 774–790. https://doi.org/10.1038/s41580-018-0074-2.

Sobin, C., Montoya, M.G., Parisi, N., Schaub, T., Cervantes, M., Armijos, R.X., 2013. Microglial disruption in young mice with early chronic lead exposure. Toxicol. Lett. 15 (13), 00151–00153.

Sobin, C., Flores-Montoya, M.G., Alvarez, J.M., 2017. Early chronic low-level Pb exposure alters global exploratory behaviors but does not impair spatial and object memory retrieval in an object-in-place task in pre-adolescent C57BL/6J mice. Neurotoxicol. Teratol.(Jan 12). https://doi.org/10.1016/j.ntt.2017.01.002. *[epub ahead of print]*.

Stanish, W.M., Taylor, N., 1983. Estimation of the Intraclass correlation coefficient for the analysis of covariance model. Am. Stat. 37 (3), 221–224. https://doi.org/10.2307/2683375.

Staples, R.E., Haseman, J.K., 1974a. Commentary: selection of appropriate experimental units in teratology. Teratology 9 (3), 259–260. doi. https://doi.org/10.1002/tera.1420090304.

Staples, R.E., Haseman, J.K., 1974b. Selection of appropriate experimental units in teratology. Teratology 9 (3), 259–260. https://doi.org/10.1002/tera.1420090304.

Stroup, W.W., 2013. Generalized Linear Mixed Models. CRC Press, Taylor Francis Group, Boca Raton, FL.

Suvorov, A., Vandenberg, L.N., 2016. To cull or not to cull? Considerations for studies of endocrine-disrupting chemicals. Endocrinology 157 (7), 2586–2594. https://doi.org/10.1210/en.2016-1145.

Tittmar, H.G., 1975. Letter: fetal units or litter units?—A possible solution. Teratology 12 (1), 89–90. https://doi.org/10.1002/tera.1420120113.

US EPA, 1998. Health Effects Guidelines OPPTS 870.6300 Developmental Neurotoxicity Study EPA/712/c-98/239. USEPA, Washington DC.

Vargesson, N., 2015. Thalidomide-induced teratogenesis: history and mechanisms. Birth Defects Res. C Embryo Today 105 (2), 140–156. https://doi.org/10.1002/bdrc.21096.

Vom Saal, F.S., 2016. TRIENNIAL REPRODUCTION SYMPOSIUM: environmental programming of reproduction during fetal life: effects of intrauterine position and the endocrine disrupting chemical bisphenol A. J. Anim. Sci. 94 (7), 2722–2736. https://doi.org/10.2527/jas.2015-0211.

Vom Saal, F.S., Dhar, M.G., 1992. Blood flow in the uterine loop artery and loop vein is bidirectional in the mouse: implications for transport of steroids between fetuses. Physiol. Behav. 52 (1), 163–171.

Wainwright, P.E., Leatherdale, S.T., Dubin, J.A., 2007. Advantages of mixed effects models over traditional ANOVA models in developmental studies: a worked example in a mouse model of fetal alcohol syndrome. Dev. Psychobiol. 49 (7), 664–674. https://doi.org/10.1002/dev.20245.

West, B., Welch, K., Galecki, A., 2007. Linear Mixed Models, A Practical Guide Using Statistical Software. CRC Press, Chapman & Hall.

Williams, D.R., Carlsson, R., Buerkner, P.-C., 2017. Between-litter variation in developmental studies of hormones and behavior: inflated false positives and diminished power. Front. Neuroendocrinol. 47, 154–166. https://doi.org/10.1016/j.yfrne.2017.08.003.

Zorrilla, E.P., 1997. Multiparous species present problems (and possibilities) to developmentalists. Dev. Psychobiol. 30 (2), 141–150. https://doi.org/10.1002/(SICI)1098-2302(199703)30:2<141::AID-DEV5>3.0.CO;2-Q. (pii).