Summer 2020

Reporting and analysis of split plot designs in preclinical animal experiments

Pu Liu
Iowa State University

Reporting and analysis of split plot designs in preclinical animal experiments

Pu Liu[1], Chong Wang[1,2], Fangshu Ye[1], Max Morris [1], Annette M. O'Connor[2,3*],

1 Department of Statistics, Iowa State University, Ames, IA, United States of America.
2 Department of Veterinary Diagnostic and Production Animal Medicine, College of Veterinary Medicine, Iowa State University, Ames, IA, United States of America.
3 Department of Large Animal Clinical Sciences, College of Veterinary Medicine, Michigan State University, East Lansing, Michigan, United States of America.

Abstract

The split plot design (SPD) has at least two types of experimental units and at least two levels of complete random design. As a result of this SPD structure, a method of analysis that accounts for the different levels of experimental unit is required, which is commonly a mixed model or a split-plot ANOVA. The design is utilized when it is not feasible to randomize the multiple interventions to the same level. The classic example of a split plot arises from agronomy, and gives name to the design, where the effects of two irrigation methods (factor 1) that must be applied to the entire (whole) plot are investigated with the effect of two different fertilizer types (factor 2) that are applied to sub plots within the whole plot.

In toxicology and nutrition, the split plot design is also employed to investigate the impact of exposure to toxins or nutrients during pregnancy and after birth. In these split plot experiments, the whole plot is the dam and the offspring are the subplots. Our objective is to evaluate the impact of choice of statistical approaches on the type I and type II error rates in hypothesis testing of effects as well as the precision of the estimation. Firstly, we assessed the reporting of SPD of 20 rat research studies and 25 agricultural studies where anecdotally the design appears to be better recognized. For the second objective, we used simulation modelling to evaluate the influences of two analysis approaches on the statistical inference obtained. For the Three scenarios included I) empirical mean parameters, II) null whole-plot main effect mean parameters and III) null effects mean parameters. And two variance conditions were i) empirical variance of random effects from research data and ii) sequential variance magnitude pairs of random effects at whole-plot and split-plot levels. The simulation study shown that although the misusage of two-way ANOVA on SPD data would provided a higher power for hypothesis testing, it was meanwhile at a risk of greater type I error rate. Furthermore, type I error introduced by two-way ANOVA rose with the increase of ratio of variance of whole-plot random effect to split-plot random effect. On the contrast, split ANOVA offered a stable type I error which around 0.05. This is a solid evidence of necessary of correct application of mixed model and split ANOVA on split-plot data.

Introduction

The split-plot design is an experimental design that involves two or more different sizes of experimental unit in a factorial treatment structure. The design name was proposed by Fisher [1] to compare the variance of two classes as a result of two levels of experimental units in experimental field trails. In a split-plot design with two factors, factor A is applied to the whole plot experimental units following completed randomized design. Then each whole plot is

subdivided into split plots and serves as a block in split-plot level. Factor B is randomly applied to the split-plot units. [2].

In toxicology and nutrition, the split plot design is often employed to investigate the impact of exposure to toxins or nutrients during pregnancy and after birth. In these split plot experiments, the whole plot is the dam and the offspring are the subplots. The results of such experiments can lead it inferences about the impact of the maternal diet during pregnancy on children's health outcomes. Such topics are clearly important and therefore it is critical that the results from such experiments are reproducible and valid. However, in previous project on biomedical experiments [3], we observed that biomedical researchers did not use the term split-plot when they employed the design. We hypothesised that perhaps the researchers did not recognize when they employed a split plot design and as a consequence the approach to analysis may be inappropriate and the inferences questionable. It is this question we investigate in this study.

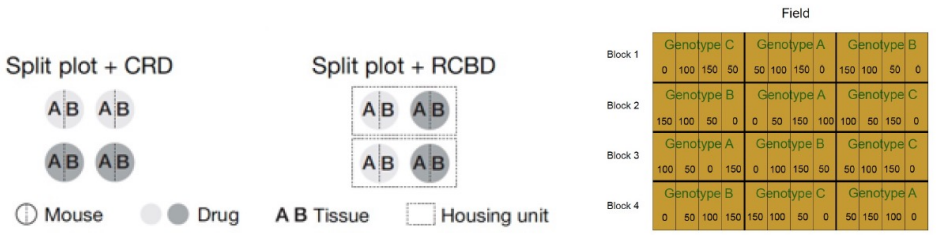
Our first objective in this study was to document approaches used by researchers to report and analyse the results of split plot experiments. Our approach to the first objective to assess the reporting of SPD in biomedical and compare it to agricultural studies where anecdotally the design appears to be better recognized. Our second objective was evaluate the impact statistical analysis approaches on the error rate in the studies, and the whole plot and sub plot effect size estimation estimation and the precision of the estimates. For the second objective, we used simulation modelling to evaluate the influences of two analysis approaches on the statistical inference obtained. Our third objective was to assess the impact of size of whole plot effects on the analysis results. We assessed how variance magnitude influenced the performance of the mixed model ANOVA on split-plot experiments using simulations of different effects at sequential variance magnitude pairs of random effects were used to assess Type I error and test power of main effect of whole-plot factor.

A reason of common practical application of SPD in agricultural field experiments is the difficulty that many agricultural treatments cannot be easily varied among small plots. [4] For example, in agricultural field experiments, it might be easier to change from one fertilizer level to another than planting one genotype to another. Hence, genotype could be randomly assigned to the larger plots as whole-plot factor. Within each plot, levels of fertilizer are randomly assigned to subplots in each whole plot as split-plot factor. A typical split-plot design in field study is a fertilizer $\times$ corn genotype experiment as shown in Fig. **??** b). A field is partitioned into four blocks. Each block is further partitioned into three plots. Three genotypes (A, B and C) can then be randomly assigned to plots within blocks. Each plot was partitioned into four subplots. Fertilizer amounts (0, 50, 100, 150 lbs. N / acre.) were randomly assigned to subplots. In this application of a split plot, randomization to group was employed in the whole-plot and split-plot levels.

Because of the hard-to-change factors and economic constraints, split-plot designs are also used industrial experimentations as well. [5] For example, if the research interest is to investigate the effects of calendering temperature, binder fibers and binder content on the strength of the fabrics, calendering temperature cannot be easily changed as the other two variables in the process. Then the calendering temperature is employed as the whole-plot factor in which each combination of the other two factors is applied while running at each temperature. [6] In all those cases, there are two levels of randomization corresponding to the two levels of experimental units. [2] [7] [4] [8]

In biomedical research, many research questions of interest which naturally involving two levels of the experimental units also need split-plot designs. For instance, scientific questions in reproductive toxicology, stress and nutrition are related to the interaction of effects of exposures during gestation and exposures after birth. For example, in human health questions about adipose tissue metabolism might be interested in diets during gestation and how they interact with diets during childhood. Similarly, in toxicology, the impact of exposure to compounds such as fluoride while in-utero then during childhood are of interest. Often, the first step in investigating these topics is to conduct experimental studies in animal models to determine these effects.

In animal research, the split-plot designs combines between-animal and within-animal testing as described in Fig **??** a) [10]. In this example, mice serve as the whole-plot experimental units while tissues served as split-plot experimental units. The mice are randomly assigned to diet and liver and kidney tissues from same study subject were randomly assigned to two drugs. In this design randomization on both two levels was satisfied. Although most of time, randomization in each level is obvious, it is notable when genotype can serves as split-plot factor. For instance, three rats, each of one strain, were held in one cage. Each cage is randomly assigned to receive either an enriched housing condition or standard housing condition. As enrichment housing condition was whole-plot treatment, genotype serves as the split-plot treatment. Unlike to field trail example, in which genotype was randomly assigned to split-plot fields, strain of mice is a natural characteristic cannot be manipulated. [11]



(a) (b)

Fig 1. Examples of different types of randomization. a)Randomization Examples in Animal Study. b)Randomization Example in Agronomy Study.

As there are at least two different sizes (or types /levels) of experimental unit in a split-plot design, the errors of estimates could be attributed to at least two sources of variances. [7] To separate the two random error effects from the sub-plot and whole-plot units, a linear mixed-model formulation is used for the split-plot design. The numerical calculation for the split-plot ANOVA elements are the same as for other balanced design. The key step is to identify the appropriate error terms for estimating the different effects in interest. [12] [13]. Practically, the model fitting and ANOVA analysis could be conducted in SAS and R by using **Proc Mixed** procedure and lme4 package provided the model is specified correctly.

Split plot designs are not only more statistically efficient by remaining the test power with less sample size than complete random designs and also able to provide more precise estimates through reducing the variance of effects in interest. [2] [14] The estimate of the whole plot main effect is based on the error in whole-plot level, while the split-plot effects and interactions are based on the error in split-plot level. however, failure to correctly separate and assign these sources of variability would result in incorrect estimation due to pooling the sums of squares of two errors. As a result, the p-values of F-test would be greater for whole-plot effects and smaller for split-plot effects and interactions than they really are. [7]

Materials and method

Eligible studies and study selection

To investigate reporting and analysis of split-plot experiments in preclinical animal studies, articles in mice studies were firstly be included in this study. The biomedical studies were

Based on our previous research [3], rat studies with descriptions of SPD and claiming employment of SPD were selected. Since there was only 1 paper claimed a split-plot design in 20 rat studies which actually conducted SPD, agronomy field studies with split-plot design were considered as well. Split-plot is originally developed from agricultural studies and it's still a

basic and popular method on experimental design in agronomy.

Assessment on the rationale and approach to the reported results of analysis

Assessment on the rationales

The articles are split into three groups. Group I and Group II were both articles in rat studies. However, articles in group I didn't include the term "split-plot", while the descriptions of experimental structure indicated the split-plot designs. Group II is composed by articles claimed "split-plot" language in their report, while their experiments were actually not split-plot designs. Group III constitutes split-plot studies in agronomy researches. To classify manuscripts to one of the groups, each article was assessed by two independent reviewers. Two criteria were used to determine whether a study is a split-plot design. From perspective of experimental structure, split-plot design is supposed to have at least one of those elements: two levels of experimental units (whole-plot and split-plot) [12] and complete randomly assignment of treatments to experimental units in each level [2]. Correspondingly, on the other hand, in the report of statistical analysis, at least one of those terms are supposed to be described, such as an employment of mixed model ; application of split-plot analysis or anova and/or assumption of random effect of whole-plot units. [7] [13] [14] The articles with split-plot structures were classified as group I. Statistical analysis sections from all three groups are assessed whether they reported those elements.

Assessment on statistical analysis and result reporting in split-plot experimental studies

In I and III article groups, the statistical analysis approach and result reporting of split-plot study were assessed by different elements. To determine whether the statistical analysis employed correct model, terms such as "split ANOVA" and/or "mixed model" are supposed be used in the statistical analysis section. For model fitting details, a random effect indicates variation introduced by whole-plot units ought to be included in the mixed model for split-plot experiments. In order to report a valid and informative statistical inference in a split-plot experiment, degree of freedom on each level is necessarily to be included in the results. Change between degree of freedoms attributed to whole-plot level residuals and split-plot level residuals is one of the characteristics of split-plot design. As a result, degree of freedoms and their change between levels are critical elements in result reporting of split-plot experiments.

Assessment on different analysis approaches to tests for treatment effects in split-plot designs

To explore the impact of different statistical analysis approaches on split-plot experiment data, settings of parameters (group means and variance of random effects) in the observed data points were required in the simulation datasets. These parameters were combined with other settings to generate multiple datasets, in which the treatment effects and variance magnitudes of random effects were varied. Since in most of the incorrect ANOVA results of split-plot experiments, two-way ANOVA was employed instead of split ANOVA. By applying both split ANOVA and two-way ANOVA on each simulation dataset, the impact of different statistical analysis approaches was assessed.

Parameter identification for simulation

In order to create an empirical simulation dataset, parameters were extracted directly from a experimental data of a mice study [11] (scenario I). In this study, a folded factorial experimental

design was employed. Observations were split to two partitions, one with split-plot design and the other with complete randomized design. Among the observations, the half with split-plot structure was an unbalanced design on the whole-plot level, with 8 cages in EE(enriched) diet treatment and 9 in S(standard) diet treatment. On the split-plot level, it's balanced that in each cage there are 3 mice in each strain. Therefore, EE is the whole-plot treatment. Strain is the split-plot treatment. And Cage is the whole-plot experimental unit. Rat is the split-plot experimental unit. Variable 'Growth' associated with growth of mouse during experiment (/grams) was chosen as response. With a total sample size 51, the simplified split ANOVA table is shown in supporting information Table . Note that, in the empirical dataset, both main effects of EE and Strain were significant, whereas the interaction is not significant.

On account of investigate impact of different statistical analysis approaches on the inference of whole-plot main effect, two more simulations were conducted for scenario II and III. In scenario II, null EE main effect model was applied, which equates to no whole-plot treatment main effect. The population mean parameter at each Strain level was the average of the two treatments at enriched EE treatment and standard EE treatment. In scenario III, null effects model was applied. Under null effects model, population mean for each treatment is the overall average across all treatment combinations. Population mean parameters for three scenarios were summarized in table 1.

Table 1. Population Mean Parameters for Three Simulation Scenarios

Enriched	Strain	I(empirical)	II(null EE)	III(null effect)
EE	BALB	14.1000	12.7833	14.5019
EE	C57	16.6625	16.0535	14.5019
EE	DBA	15.8375	14.6688	14.5019
SEE	BALB	11.46667	12.7833	14.5019
SEE	C57	15.4444	16.0535	14.5019
SEE	DBA	13.5000	14.6688	14.5019

Simulation procedure

Two sets of simulations were conducted based on parameters for three scenarios. The simulations were generated under two different conditions of random effect variances:

i) Empirical random effect variance from split ANOVA results.
Variance of Cage was 0.6930 and 3.8426 for Rat, which were calculated in R and confirmed in SAS.

ii) Sequential variance magnitude pairs of random effects.
First remain the total variance as (according to split ANOVA):

$$V_{total} = V_{Cage} + V_{Rat} = 0.693 + 3.843 \approx 4.5$$

Then generate simulation data at six choices of variance of random effects:

$$(V_{Cage}, V_{Rat}) = \{(0.7, 3.8), (1.4, 3.1), (2.1, 2.4), (2.8, 1.7), (3.5, 1), (4.2, 0.3)\}$$

In order to obtain a balanced structure, Cage number in each enrichment group was 9 in all simulations. Each simulation used multivariate normal random variables with corresponding mean and variance pair. Simulations under combinations $(3 \times (1 + 6))$ of population mean parameter scenarios and variance conditions were conducted in R with 10,000 times.

Analysis on simulated data

On $3 \times (1 + 6)$ simulated data outputs, split ANOVA and two-way ANOVA were applied:

i) Split ANOVA based on mixed model.

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{(i)l} + \epsilon_{ijl}$$

- fixed effect μ is associated with the intercept;
- fixed effect α is associated with the whole-plot factor Enrichment with two levels, Enriched (EE) and Standard (EES), corresponding to i = 1,2;
- fixed effect β is associated with the split-plot factor Strains with three levels, BALB, C57 and DBA, corresponding to j = 1,2,3;
- random effect γ is associated with whole-plot units Cage, which is nested in the whole-plot treatment factor Enrichment. There are 8 and 9 cages in each level of Enrichment treatment in original data, while there are both 9 cages in each level in simulation data, corresponding to (i)l = 1, ..., 8 or 9; where $\gamma \sim N(0, \sigma^2_\gamma)$;
- random effect ϵ is associated with split-plot experimental unit (rat), with a total number 54 in simulation data; where $\epsilon \sim N(0, \sigma^2_\epsilon)$.

ii) Two-way ANOVA based on Gauss-Markov model under normal error assumption.

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijl}$$

There was no random effect $\gamma_{(i)l}$ associated to variation attributed to Cage in the Gauss-Markov model. As a result, there was only one source of variability in this model, which was Rat.

Analysis of Variance i) and ii) were conducted in R correspondingly with 'lmer' function in 'lme4' package and 'lm' function.

To investigate effect of choice of different statistical analysis approaches on Type I error of tests and accuracy of point estimates of treatment effects, simulations (1×3) under variance condition I for three parameter scenarios were used. Then chance of rejection of treatment effects and features of confidence intervals at 5% level were summarized. Chance of rejection was calculated by proportion of p-values less than 0.05. Features of confidence interval comprise point estimate, width and coverage rate.

In order to evaluate influence of variance magnitudes ratio between two experimental units on performance of two types of ANOVA, simulations (6×3) under variance condition II for three parameter scenarios were used. Chances of rejection of treatment effects was summarized to presented as a long time frequency Type I error or test power approximation. For example, at null EE and null effects parameter settings, the significance report proportion in large simulation time converges to Type I error of EE main effect. Similarly, at original parameter settings, both EE main effect and Strain main effect were significant. The significance report proportion in large simulation time converges to test power of those main effects.

Results

Assessment of statistical analysis and results reporting among three article groups

Articles were split into three groups according to the criteria in Methods section. Between the two groups containing language that suggested split-plot designs (Group I and III), there were distinguishing differences of statistical analysis reporting.

As shown in Table 2, in rat studies in Group I, investigators just described the two levels of experimental units (dams and pups) and two factors assigned to each level without including exact "split-plot" terms. Only 1 study [15] in Group I stated a usage of both repeated-measurement design and repeated-measurement ANOVA. Another study [16] simply used

repeated-measurement ANOVA without statement of explicit study design. Those were merely 2/25 articles in Group I that imply split-plot designs with time treatment as split-plot level treatment. Interestingly, 1/20 study [17] without any indication of split-plot experimental design employed a mixed model in statistical analysis section. However, similar to other 19/20 articles in Group I, there was no assignment of random effect associated to whole-plot experimental units in the model.
On the other hand, in agronomy field studies in Group III, besides details on experimental designs, 25/25 articles used the terms such as "split-plot design" [18] or "randomized complete block split plot experiment" [19] as well. 7/25 [20] [21] [22] [23] [24] [25] [26] studies employed an mixed model in statistical analysis. Nevertheless, 4 [27] [28] [25] [26] of those 7 studies claimed the mixed model alone. The other 3 pinpointed the assginment of random effect associated to whole-plot experimental units in the model. Expect for those 3 studies, there were 2 studies included the random effect of whole-plot experimental unit but with incorrect model. One [29] made use of nonlinear model in SAS with PROC NLIN and the other [30] used a general linear model instead of the mixed model for split-plot experiment.

In Group I, there were 8/25 articles reported the degree of freedoms for F-test of effects. 6 [31] [32] [16] [15] [23] [24] of 8 reported different degree of freedoms of denominator for F-tests of effects in different stratum, which means that assessment of whole-plot main effect is associated to residuals in whole-plot level, whereas assessments of split-plot main effect and interaction were associated to residuals in split-plot level. The other 2 [33] [34] studies reported the same denominator degree of freedoms of F-tests for all effects.
In 5/25 studies in Group III that the degree of freedoms for F-tests of effects were reported, only 1 [29] stated en employment of mixed model. The other 4 studies [35] [36] [37] [38] didn't claim usage of mixed model, but reported degree of freedoms with changes between whole-plot level residuals and split-plot level residuals.

Table 2. The Reporting Characteristics of Articles in Group I and III

Sections	Key Terms	Group I (/20)	Group III (/25)
Analysis	Split design/anova	1	25
	Mixed model	1	7
	Random effect of wp units	0	5
Results	d.f. of wp effect	8	5
	d.f. of sp effect	8	5
	d.f. of interaction	8	5
	Changes among d.f.	6	5

Impact of different analysis approaches on Type I and Type II error and accuracy of point estimates of treatment effects

Impact of different analysis approaches on Type I and Type II error of treatment effects

The proportions of p-values less than 0.05 average across 10,000 simulation for three scenarios were shown in Table 3. According to the parameter settings in three scenarios, main effect of whole-plot factor EE was non-zero in scenario I. Main effect of split-plot factor Strain was non-zero in scenario II and III. Effect of interaction between whole-plot factor and split-plot factor was non-zero in scenario I. Since Type I error is known as "false positive" conclusion for hypothesis testing, chance of rejection of zero effects in population parameter settings corresponded to Type I error of hypothesis test of those effects. Similarly, chance of rejection of non-zero effects in population parameter settings was statistical power of effects.

In the whole-plot level, by using correct mixed model on split-plot data, the Type I error of EE main effect remained approximate 5% while achieving a fairly high statistical power (above 80%). The usage of incorrect two-way ANOVA would result in about twice Type I error rate as

mixed model.

In the split-plot level, mixed model provided higher power and kept the Type I error for split-plot factor effect and interaction around 5%. Statistical power by using two-way ANOVA was high as well, while the Type I errors were slightly lower than mixed model.

Table 3. Chance of Rejection on Tests of Treatment Effects by Split-plot ANOVA and Two-way ANOVA

Population Mean Parameters	EE		Strain		Interaction	
	Mixed	Two-way	Mixed	Two-way	Mixed	Two-way
scenario I (empirical)	82.7% *	90.8% *	99.4% *	99.0% *	15.7% *	11.0% *
scenario II (null EE)	5.2%	9.3%	99.3% *	98.9% *	5.1%	3.1%
scenario III (null effects)	4.9%	9.0%	5.3%	3.1%	5.1%	3.0%

* Non-zero effects in corresponding population parameter settings.

Impact of different analysis approaches on confidence intervals of whole-plot treatment main effect

Tabel 4 shown the impact of different statistical analysis approaches on 95% confidence intervals of EE main effect for three population mean parameter scenarios. True values were extracted directly from the population mean parameters. Point estimates from the mixed model (split-plot ANOVA) and general linear model (two-way ANOVA) were the same across three scenarios. And all point estimates obtained by two models were very close to the true values, which indicates that both two methods could provide valid point estimates. However, the mixed model provided wider confidence intervals and higher coverage rates than two-way ANOVA in all scenarios. It implied that mixed model increased the accuracy of estimates of whole-plot treatment main effect, which was consistent with the conclusion obtained from Table 3.

Table 4. Summary of 95% Confidence Intervals of EE Main Effect

Population Mean Parameters	True Value	Estimates		CI Width		Coverage Rate	
		Mixed	Two-way	Mixed	Two-way	Mixed	Two-way
scenario I (empirical)	2.0630	2.0483	2.0483	2.7799	2.3205	95.49%	91.81%
scenario II (null EE)	0	0.0027	0.0027	2.7731	2.3153	94.76%	90.73%
scenario III (null effects)	0	-0.0009	-0.0009	2.7755	2.3198	95.15%	91.03%

Impact of variances magnitude ratio on performance of analysis approaches

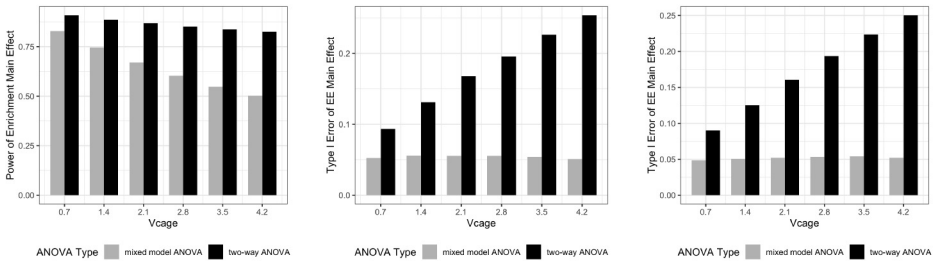
To evaluate the performance of analysis approaches, the chance of rejection of hypothesis test for three scenario with six variances magnitude ratios were plotted with respect to treatment effect.

Performances on split-plot ANOVA and two-way ANOVA of whole-plot treatment main effect

As shown in Fig 2, for scenario I (empirical population mean parameters), EE main effect is non-zero, as a result, the chance of rejection is corresponding to the power of hypothesis testing. With the variance proportion of Cage increasing, the power of EE main effect shown an obvious decrease by using split ANOVA, while the power decreased slightly by two-way ANOVA.

For scenario II (null EE main effect mean parameters) and III (null effects at all mean parameters), Type I error rate of EE main effect by using split ANOVA maintained around 0.05 across variance magnitude ratios. On the contrary, by two-way ANOVA, Type I error rate of EE main effect increased dramatically with the variance of Cage increasing.

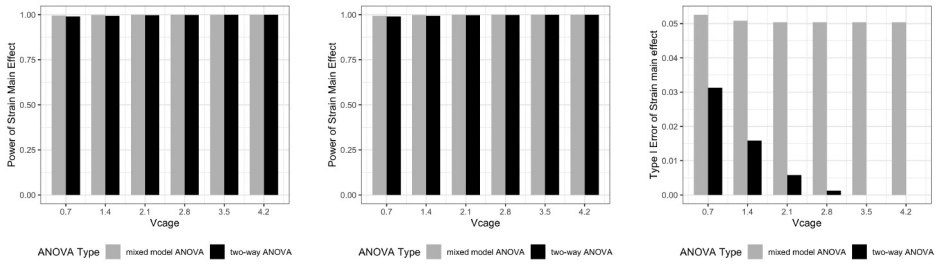
The Type I error rate and testing power of hypothesis testing for Strain main effect were shown in Fig 3. For scenario I (empirical population mean parameters) and II (null EE main effect mean parameters) Strain main effect is non-zero, as a result, the chance of rejection is corresponding to the power of hypothesis testing. The statistical powers were no obvious differences between using split ANOVA and two-way ANOVA. The statistical powers both remained high (greater than 80%) across the changing of variance proportion of Rat.



(a) (b) (c)

Fig 2. Power and Type I Error of Whole-plot Main Effect Comparison on Power and Type I Error of Whole-plot Main Effect (EE) between Split-plot ANOVA and Two-way ANOVA on Condition II Simulation for Three Scenarios. a):Power of EE Main Effect for Scenario I Simulation. b): Type I Error Rate of EE Main Effect for Scenario II Simulation. c): Type I Error Rate of EE Main Effect for Scenario III Simulation.

Performances on split-plot ANOVA and two-way ANOVA of split-plot treatment main effect



(a) (b) (c)

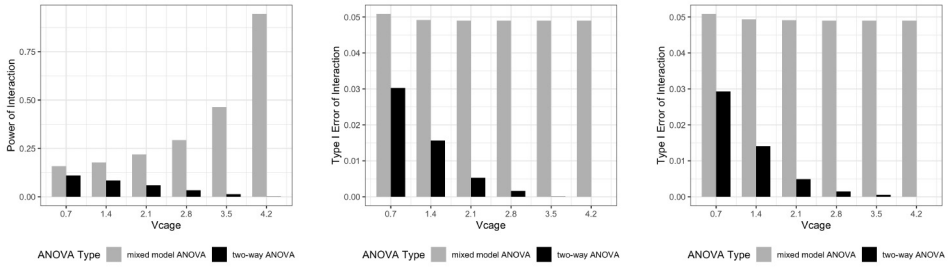
Fig 3. Power and Type I Error of Split-plot Main Effect Comparison on Power and Type I Error of Split-plot Main effect (Strain) between Split-plot ANOVA and Two-way ANOVA on Condition II Simulation for Three Scenarios. a): Power of Strain Main Effect on Original Parameters Simulation. b): Power of Strain Main Effect on Null EE Main Effect Parameters Simulation. c): Type I Error rate of Strain Main Effect on Null Effects Parameters Simulation.

For scenario III (null effects at all mean parameters), Type I error rate of Strain main effect by using incorrect two-way ANOVA decreased dramatically as variance magnitude of Rat decreasing. On the contrary, by split ANOVA, type I error rate of Strain main effect remained around 5%. Nevertheless, type I errors of both approaches were no larger than 5%

Performances on split-plot ANOVA and two-way ANOVA of interactions

For scenario I, the empirical mean parameters, all effects were non-zero. As shown in Fig 4 a), the statistical power of interaction by split ANOVA increased dramatically with the variance of

Rat decreasing, whereas it was decreased for two-way ANOVA. Since the interaction effects were zero for scenario II and III, as shown in Fig 4 b) and c), type I error rates of hypothesis testing for interaction between EE and Strain had same pattern across for scenario II and III. Type I error rates of interaction effect by using split ANOVA held around 5% while them displayed a sharp decrease pattern as the variance of random effect of Rat decreasing by two-way ANOVA.



(a) (b) (c)

Fig 4. Type I Error of Interaction Comparison on Type I Error of Interaction between Split-plot ANOVA and Two-way ANOVA on Condition II Simulation for Three Scenarios. a): Type I Error Rate of Interaction on Original Parameters Simulation. b): Type I Error Rate of Interaction on Null EE Main Effect Parameters Simulation. c): Type I Error Rate of Interaction on Null Effects Parameters Simulation.

Discussion

For the sake of animals' welfare, appropriate experimental design and a valid report are necessary for animal studies. Split-plot designs are used in animal experiments to reduce the use of animals and provide adequate statistical inference in the meanwhile. [39] The rational employment of split-plot designs requires two types of experimental units, which commonly refers to parents and offsprings in animal studies and plot and subplot in agronomy trails. However, with an aspiration of use split-plot design, animal studies and agronomy studies displayed different reporting performance.

25/25 agronomy pieces of research stated "split-plot" term while they conducted split-plot designs. Whereas in rat studies, some of the studies claimed a "split-plot design" without the use of two strata of experimental units. And only 5% (1/20) used the "split-plot" language to specify the form of experimental design. The lack of knowledge of split-plot designs might be attributed to there are more chance to end up with missing data in the rat study. Since adjustment for the degree of freedom is required for split-plot data with missing values. Considering that a consistent and concise system of notation could help readers understand seemingly difficult ideas [40], we suggest the researchers employ universal language to specify the experimental design of their studies.

In an experimental study, the foundation for the justification of the statistical approach is how data is collected. [41] In the description of how the experiments were conducted, treatment and corresponding experimental units were usually well demonstrated. But some details, such as randomization, were sometimes missing in the description. Since treatment efficacy would be overestimated as a result of missing randomization [42], a description of how to conduct the randomization in each stratum is important in reporting split-plot studies.

With respect to statistical analysis report, very few of studies in our article pools included a good manner of mixed model application. 0/20 studies in rat research with split-plot design provided report of both mixed model and random effect of whole-plot experimental unit. Not all 7/25 trial studies in agronomy research that applying mixed model assigned a random effect to whole-plot experimental unit. However, to analyze data from an explicit split-plot design, a mixed model with random effects associated to both whole-plot and split-plot experimental

units is imperative. Fail to identify the two sources of variation would result in overestimating significance of whole-plot treatment and underestimating significance of split-plot treatment and interaction. [7] Furthermore, although most of the researches conducted analysis of variance, 30% (6/20) rat studies reported F-test with different degree of freedoms in each stratum. The proportion is 20% (5/20) in agronomy trails. Different denominator degree of freedoms between F-test of whole-plot treatment effect and split-plot treatment effect is an insight of estimating the correct error. Therefore, to provide valid inference for split-plot designs, mixed model application, a random effect associated to whole-plot experimental units and degree of freedoms of F-test are necessary.

Based on a split-plot experiment, the employment of mixed model and split ANOVA is critical in statistical analysis. By analysis of the chance of rejection on hypothesis test of treatment effects on simulated data, it indicated that split ANOVA offered better statistical power when remaining significance level around 5%. In the hypothesis tests for whole-plot and split-plot factor main effects, the statistical power of split ANOVA and two-way ANOVA were both greater than 80%, while for interaction split ANOVA provided higher power than two-way ANOVA. Although the Type one error of tests for split-plot effects of two-way ANOVA was lower than split ANOVA, it could cause a lower significance level as well. The strengths of using the correct model and split ANOVA on split-plot designs were also shown in the accuracy of estimates of whole-plot treatment effects. The mixed model could always offer higher coverage in interval estimates. Which triggered a reflection on statistical reports on split-plot experiments with two-way ANOVA.

Resultingly most of the articles recognized split-plot designs as factorial designs with complete randomization, two-way ANOVA was inappropriately used on split-plot data. However, according to our analysis on simulation for three scenarios, although two-way ANOVA could provide higher test powers for treatment effects, it takes a risk of greater Type I error as well. With a fairly high test power (greater than 0.8) obtained by using mixed model, it's reasonable to use mixed model on split-plot data to keep both Type I and Type II error low.

The advantage of split-plot is that it separate the variation into two sources and use different error to make inference at each stratum. When the mixed model was employed to analyze split-plot data in our simulation, the split ANOVA table for simulation data sets was shown in Table 5. The MSE at whole-plot stratum was used to make inference for EE main effect. The MSE at whole-plot stratum was based on an effective sample size at Cage level, which was 16. On the other hand, if the two-way ANOVA was applied on the simulations, as shown in Table 6, the only MSE would be used to calculate the accuracy of estimates of EE main effect. And this MSE was based on an effective sample size of 48. With the increase of the effective sample size, the accuracy of the estimates would decrease, which explains the wider confidence interval occurred by using mixed model.

Table 5. Split ANOVA Table for Simulation Data

Stratum	Source	Degrees of Freedom	Sum Square
whole-plot	EE	2-1=1	$27\sum_{i=1}^{2}(\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot\cdot\cdot})^2$
	resid	17-1=16	$3\sum_{i=1}^{2}\sum_{l=1}^{9}(\bar{y}_{i\cdot l}-\bar{y}_{i\cdot\cdot})^2$
	corrected total	18-1=17	$3\sum_{i=1}^{2}\sum_{l=1}^{9}(\bar{y}_{i\cdot l}-\bar{y}_{\cdot\cdot\cdot})^2$
split-plot	Cage	18-1=17	$3\sum_{i=1}^{2}\sum_{l=1}^{9}(\bar{y}_{i\cdot l}-\bar{y}_{\cdot\cdot\cdot})^2$
	Strain	3-1=2	$18\sum_{i=j}^{3}(\bar{y}_{\cdot j\cdot}-\bar{y}_{\cdot\cdot\cdot})^2$
	EE*Strain	1*2=2	$9\sum_{i=1}^{2}\sum_{j=1}^{3}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y}_{\cdot\cdot\cdot})^2$
	resid	53-21=32	diff
	corrected total	54-1=53	$\sum_{ijl}(y_{ijl}-\bar{y}_{\cdot\cdot\cdot})^2$

Furthermore, to make inference on split-plot treatment main effect and interaction, different MSE were used in split ANOVA and two-way ANOVA. In split ANOVA, MSE at split-plot stratum with 32 effective sample size. And effective sample size in two-way ANOVA is 48.

Similar to the conclusions on confidence interval, with a smaller effective sample size indicating greater standard error, it's more difficult to reject the hypothesis test for significance of the treatment effects under the mixed model.

Table 6. Two-way ANOVA Table for Simulation Data

Source	Degrees of Freedom	Sum Square
EE	2-1=1	$27\sum_{i=1}^{2}(\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot\cdot\cdot})^2$
Strain	3-1=2	$18\sum_{i=j}^{3}(\bar{y}_{\cdot j\cdot}-\bar{y}_{\cdot\cdot\cdot})^2$
EE*Strain	1*2=2	$9\sum_{i=1}^{2}\sum_{j=1}^{3}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y}_{\cdot\cdot\cdot})^2$
resid	53-5=48	diff
corrected total	54-1=53	$\sum_{ijl}(y_{ijl}-\bar{y}_{\cdot\cdot\cdot})^2$

To explore the more general impact of different analysis approaches on statistical inference in split-plot designs, the proportion of variation attributed to whole-plot experimental units was reassigned in simulation under condition II for three scenarios. In the whole-plot level, as the variance attributed to whole-plot random effect increased, the power of whole-plot main effect of both approaches deceased and type I error of two-way ANOVA increased. This brought up an issue that if researchers use a general linear model to calculate the sample size under certain statistical power, the test power in their reports if the split-plot design was applied, would lower than their expectation. To draw the statistical inference of particular treatment effect, by using a mixed model, the s.e. to compute the power or type I error of hypothesis tests was the residual term in the corresponding level. On the other hand, the only residual term in two-way ANOVA is used as the denominator in every hypothesis test. When the variance associated to whole-plot random effect increases, the s.e. to compute the test statistics for whole-plot effects in split ANOVA increases, resulting in harder detection on significance of whole-plot effects. If the truth, which reflects by split ANOVA, is that the whole-plot effect is non-zero, the test power would decrease with s.e. increases. On the other hand, if the truth is null whole-plot effects, the type I error in two-way ANOVA would increase since the s.e. is smaller than it in split ANOVA and the differences in s.e. between two ANOVA increase as the variance of the whole-plot random effect increases.

To draw statistical inference for split-plot effects, split-plot factor and interaction between whole-plot and split-plot factors, by using split ANOVA, the error term is associated to residuals in the split-plot level, where as the two-way ANOVA uses the same error term with tests for whole-plot effects. When variance attributed to whole-plot random effect increases, the s.e. in split ANOVA to compute statistics for split-plot level hypothesis tests decreased while the s.e. in two-way didn't change. As a result, comparing to the truth reflected by split ANOVA, it was harder to detect the significance for split-plot effects by two-way ANOVA, since the s.e. were bigger than truth. Then the statistical reporting would end up with a false low type I error rate for the split-plot level effects. It is remarkable that the statistical power of interaction by split ANOVA reached 80% when variance of whole-plot random effect was dominant. Furthermore, when whole-plot treatment effect is zero, performance of either ANOVA approach on type I error of interaction was robust to split-plot treatment effect. The patterns in Fig 4 b) and c) were the same for non-zero or zero split-plot treatment main effect.

For application consideration, our initial thought is to evaluate the statistical reporting in rat studies that explore the impact of dam behaviors on their offspring's health conditions. Our simulation analysis showed that with the homogeneity of the dam decreasing, the type I error by an incorrect choice of two-way ANOVA increased, which leads to more false rejections to null whole-plot main effect hypothesis. As a result, we could mistakenly put too much stress on mothers for the pseudo impact of mothers' conduct during pregnancy on the health conditions of their children. Interestingly, when both main effects in whole-plot and split-plot level and interaction were all non-zero, the performances of two types ANOVA display opposite patterns as the variance of whole-plot random effect increasing. Which indicates that by using two-way

ANOVA on split data, the significance of interaction is harder to be detected. This might mislead the researchers to simpler conclusion than reality and ignoring the interaction between conducts of pregnant mothers and children in certain period.

Conclusion

With respect to validity of results of split-plot studies, the key findings in this study were summarized as:

- Although split-plot designs are commonly used in agronomy trails and preclinical animal studies, the terminology and methodology of split-plot design and mixed model were not pertinently used, especially in animal studies.
- By using mixed model on split-plot data, researchers could keep both Type I and Type II error fairly low for the hypothesis testing and achieve more accurate estimates of treatment effects.
- For whole-plot treatment main effect at higher stratum in split-plot design, hypothesis testing power decreased with the variance of whole-plot experimental units increasing. Testing power by using mixed model dropped faster than two-way ANOVA. However, although it seems that two-way ANOVA could provide powerful hypothesis testing, a risk of greater Type I error is notable.
- For split-plot effects, the performance of two types of ANOVA approaches on type I error rate of split-plot treatment main effect and interaction between whole-plot treatment and split-plot treatment were the same. The type I error by split ANOVA was around 5% while by two-way ANOVA, it was lower and decreasing with the variance of split-plot experimental units decreasing.
- The statistical power of split-plot treatment effect were almost the same by two types of ANOVA. Whereas the statistical power of interaction by split ANOVA increased as the variance of whole-plot random effect increased. On the contrary, the statistical power by two-way ANOVA decreased.

Supporting information

S1 Table. ANOVA Format for Two-Factor Split-plot Design in BMC Article

Stratum	Source	Degrees of Freedom
whole-plot	EE	2-1=1
	resid	16-1=15
	corrected total	17-1=16
split-plot	Cage(bolcks)	17-1=16
	Strain	3-1=2
	EE*Strain	1*2=2
	resid	50-20=30
	corrected total	51-1=50

S2 Table. Chance of Rejection of Hypothesis Testing of Treatment Effects on Simulations under Condition II for Scenario I

Effects	EE		Strain		Interaction	
(V(Cage),V(Rat))	Mixed	Two-way	Mixed	Two-way	Mixed	Two-way
(0.7,3.8)	82.9%	90.9%	99.5%	99.0%	15.8%	11.0%
(1.4,3.1)	74.5%	88.7%	99.9%	99.4%	17.7%	8.4%
(2.1,2.4)	67.0%	86.9%	100.0%	99.7%	21.9%	5.9%
(2.8,1.7)	60.4%	85.1%	100.0%	99.9%	29.3%	3.3%
(3.5,1.0)	54.8%	83.8%	100.0%	100.0%	46.4%	1.3%
(4.2,0.3)	50.3%	82.6%	100.0%	100.0%	94.6%	1.9%

S3 Table. Chance of Rejection of Hypothesis Testing of Treatment Effects on Simulations under Condition II for Scenario II

Effects	EE		Strain		Interaction	
(V(Cage),V(Rat))	Mixed	Two-way	Mixed	Two-way	Mixed	Two-way
(0.7,3.8)	5.2%	9.3%	9.30%	98.9%	5.1%	3.0%
(1.4,3.1)	5.6%	12.1%	99.9%	99.4%	4.9%	1.6%
(2.1,2.4)	5.6%	16.8%	100.0%	99.7%	4.9%	0.5%
(2.8,1.7)	5.5%	19.6%	100.0%	99.9%	4.9%	0.1%
(3.5,1.0)	5.4%	22.6%	100.0%	100.0%	4.9%	0%
(4.2,0.3)	5.1%	25.4%	100.0%	100.0%	4.9%	0%

S4 Table. Chance of Rejection of Hypothesis Testings of Treatment Effects on Simulations under Condition II for Scenario III

Effects	EE		Strain		Interaction	
(V(Cage),V(Rat))	Mixed	Two-way	Mixed	Two-way	Mixed	Two-way
(0.7,3.8)	5.2%	9.3%	9.30%	98.9%	5.1%	3.0%
(1.4,3.1)	5.6%	12.1%	99.9%	99.4%	4.9%	1.6%
(2.1,2.4)	5.6%	16.8%	100.0%	99.7%	4.9%	0.5%
(2.8,1.7)	5.5%	19.6%	100.0%	99.9%	4.9%	0.1%
(3.5,1.0)	5.4%	22.6%	100.0%	100.0%	4.9%	0%
(4.2,0.3)	0.5%	25.4%	100.0%	100.0%	4.9%	0%

References

1. Fisher RA. Statistical methods for research workers. Biological monographs and manuals. Edinburgh, London: Oliver and Boyd; 1925.
2. Jones B, Nachtsheim CJ. Split-plot designs: What, why, and how. Journal of quality technology. 2009;41(4):340–361.
3. O'Connor, Annette M, et al. The study design elements employed by researchers in preclinical animal experiments from two research domains and implications for automation of systematic reviews. PloS one; 2018.
4. Yates F. Complex experiments. Supplement to the Journal of the Royal Statistical Society. 1935;2(2):181–247.
5. Daniel C. Applications of statistics to industrial experimentation. vol. 124. John Wiley & Sons; 1976.
6. Morris M. Design of experiments: an introduction based on linear models. Chapman and Hall/CRC; 2010.

7. Christensen R. Plane answers to complex questions: the theory of linear models. Springer Science & Business Media; 2011.

8. Ramsey F, Schafer D. The statistical sleuth: a course in methods of data analysis. Cengage Learning; 2012.

9. Festing MF, Overend P, Das RG, Borja MC, Berdoy M. The design of animal experiments : reducing the use of animals in research through better experimental design / Michael FW Festing ... [et al.]. Laboratory animal handbooks. no. 14. London: Royal Society of Medicine; 2002.

10. Altman N, Krzywinski M. Points of significance: split plot design; 2015.

11. Walker M, Fureix C, Palme R, Newman JA, Dallaire JA, Mason G. Mixed-strain housing for female C57BL/6, DBA/2, and BALB/c mice: validating a split-plot design that promotes refinement and reduction. BMC medical research methodology. 2016;16(1):11.

12. John JA, Quenouille MH. Experiments: design and analysis. Griffin; 1977.

13. Kuehl RO, Kuehl R. Design of experiments: statistical principles of research design and analysis. Duxbury/Thomson Learning Pacific Grove, CA; 2000.

14. Box GE. Statistics for experimenters: design, innovation, and discovery. Hoboken, N.J. : Wiley-Interscience; 2005.

15. Bartos M, Gumilar F, Bras C, Gallegos CE, Giannuzzi L, Cancela LM, et al. Neurobehavioural effects of exposure to fluoride in the earliest stages of rat development. Physiology & behavior. 2015;147:205–212.

16. Paletz EM, Craig-Schmidt MC, Newland MC. Gestational exposure to methylmercury and n-3 fatty acids: Effects on high-and low-rate operant behavior in adulthood. Neurotoxicology and Teratology. 2006;28(1):59–73.

17. Cho CE, Sánchez-Hernández D, Reza-López SA, Huot PS, Kim YI, Anderson GH. Obesogenic phenotype of offspring of dams fed a high multivitamin diet is prevented by a post-weaning high multivitamin or high folate diet. International journal of obesity. 2013;37(9):1177–1182.

18. Sarlak S, Aghaalikhani M, Zand B. Effect of plant density and mixing ratio on crop yield in sweet corn/mungbean intercropping. Pakistan Journal of Biological Sciences. 2008;11(17):2128–2133.

19. Khan AU, Iqbal M, Islam K. Dairy manure and tillage effects on soil fertility and corn yields. Bioresource technology. 2007;98(10):1972–1979.

20. Hoffman P, Esser N, Shaver R, Coblentz W, Scott MP, Bodnar A, et al. Influence of ensiling time and inoculation on alteration of the starch-protein matrix in high-moisture corn. Journal of dairy science. 2011;94(5):2465–2474.

21. Amaducci S, Colauzzi M, Battini F, Fracasso A, Perego A. Effect of irrigation and nitrogen fertilization on the production of biogas from maize and sorghum in a water limited environment. European journal of agronomy. 2016;76:54–65.

22. Nazli MH, Halim RA, Abdullah AM, Hussin G, Samsudin AA. Potential of four corn varieties at different harvest stages for silage production in Malaysia. Asian-Australasian journal of animal sciences. 2019;32(2):224.

23. El-lethey HS, Kamel MM, Shaheed IB. Neurobehavioral toxicity produced by sodium fluoride in drinking water of laboratory rats. J Am Sci. 2010;6(5):54–63.

24. Carlin J, George R, Reyes TM. Methyl donor supplementation blocks the adverse effects of maternal high fat diet on offspring physiology. PloS one. 2013;8(5).

25. MacGuidwin A, Knuteson D, Connell T, Bland W, Bartelt K. Manipulating inoculum densities of Verticillium dahliae and Pratylenchus penetrans with green manure amendments and solarization influence potato yield. Phytopathology. 2012;102(5):519–527.

26. Mallowa SO, Esker PD, Paul PA, Bradley CA, Chapara VR, Conley SP, et al. Effect of maize hybrid and foliar fungicides on yield under low foliar disease severity conditions. Phytopathology. 2015;105(8):1080–1089.

27. Bao Y, Chen S, Vetsch J, Randall G. Soybean yield and Heterodera glycines responses to liquid swine manure in nematode suppressive soil and conducive soil. Journal of nematology. 2013;45(1):21.

28. Srisa-Ard K. Effects of crop residues of sunflower (Helianthus annuus), maize (Zea mays L.) and soybean (Glycine max) on growth and seed yields of sunflower. Pakistan Journal of Biological Sciences. 2007;10(8):1282–1287.

29. Leaf T, Ostlie KR. Nitrogen Rate Effects on Cry3Bb1 and Cry3Bb1+ Cry34/35Ab1 Expression in Transgenic Corn Roots, Resulting Root Injury, and Corn Rootworm Beetle Emergence. Journal of economic entomology. 2017;110(3):1243–1251.

30. Khan A, et al. Phosphorus and compost management influence maize (Zea mays) productivity under semiarid condition with and without phosphate solubilizing bacteria. Frontiers in plant science. 2015;6:1083.

31. Day JJ, Reed MN, Newland MC. Neuromotor deficits and mercury concentrations in rats exposed to methyl mercury and fish oil. Neurotoxicology and teratology. 2005;27(4):629–641.

32. Beyrouty P, Chan HM. Co-consumption of selenium and vitamin E altered the reproductive and developmental toxicity of methylmercury in rats. Neurotoxicology and Teratology. 2006;28(1):49–58.

33. Lucena GM, Porto FA, Campos ÉG, Azevedo MS, Cechinel-Filho V, Prediger RD, et al. Cipura paludosa attenuates long-term behavioral deficits in rats exposed to methylmercury during early development. Ecotoxicology and environmental safety. 2010;73(6):1150–1158.

34. Burdge GC, Lillycrop KA, Jackson AA, Gluckman PD, Hanson MA. The nature of the growth pattern and of the metabolic response to fasting in the rat are dependent upon the dietary protein and folic acid intakes of their pregnant dams and post-weaning fat consumption. British journal of nutrition. 2008;99(3):540–549.

35. Osborne S, Schepers JS, Francis D, Schlemmer MR. Use of spectral radiance to estimate in-season biomass and grain yield in nitrogen-and water-stressed corn. Crop Science. 2002;42(1):165–171.

36. Coelho C, Molin D, Joris HW, Caires E, Gardingo J, Matiello R. Selection of maize hybrids for tolerance to aluminum in minimal solution. Genetics and Molecular Research. 2015;14(1):134–144.

37. Jurado-Tovar A, Compton W. Intergenotypic competition studies in corn (Zea mays L.). Theoretical and Applied Genetics. 1974;45(5):205–210.

38. Khehra A, Bhalla S. Cytoplasmic effects on quantitative characters in maize (Zea mays L.). Theoretical and Applied Genetics. 1976;47(6):271–274.

39. Festing MFW, Overend P, Borja MC, Berdoy M. The Design of Animal Experiments: Reducing the Use of Animals in Research Through Better Experimental Design. Sage; 2016.

40. Kroese DP, Botev Z, Taimre T, Vaisman R. Data Science and Machine Learning: Mathematical and Statistical Methods. CRC Press; 2019.

41. Knight KL. Study/experimental/research design: much more than statistics. Journal of athletic training. 2010;45(1):98–100.

42. Liu E, Fan J. Fundamentals of Laboratory Animal Science. CRC Press; 2017.