PERSPECTIVE

Experimental design and analysis for consideration of sex as a biological variable

Clare M. Diester¹, Matthew L. Banks¹, Gretchen N. Neigh² and S. Stevens Negus¹*Neuropsychopharmacology* (2019) 44:2159–2162; <https://doi.org/10.1038/s41386-019-0458-9>

The National Institutes of Health (NIH) issued a policy in 2015 to promote inclusion of both male and female subjects in preclinical biomedical research [1]. As guidance to address this policy, NIH has articulated “The Four Cs of Studying Sex to Strengthen Science” (<https://orwh.od.nih.gov/sex-gender/nih-policy-sex-biological-variable>): (1) Consider (design studies that take sex into account, or explain why it is not incorporated), (2) Collect (tabulate sex-based data), (3) Characterize (analyze sex-based data), and (4) Communicate (report and publish sex-based data). This article describes a general, multi-step approach to address The Four Cs when investigation of sex differences is not the primary aim, and sample sizes required to detect sex differences are not known. We acknowledge that there are many instances where sex may be a top priority; however, as written, the NIH policy and associated “Four C’s” guidance recognize that this is not invariably the case. The factors that may contribute to an individual researcher’s prioritization of sex differences among all possible independent variables are beyond the scope of this brief article.

The data to illustrate this approach are shown in Fig. 1 and Table 1 for a study involving pain-related behavior in mice. The expression and treatment of pain constitute one domain in which sex differences are apparent in both humans and laboratory animals [2], and preclinical research is playing a major role in analgesic drug development in response to the NIH “Helping to End Addiction Long-term” (HEAL) initiative. Accordingly, this is one area where preclinical research is active, the new NIH policy is pertinent, and inclusion of both males and females is warranted. For this illustrative study, which used methods similar to those described previously [3], adult male and female ICR mice (Envigo, Frederick, MD) received an intraperitoneal injection of dilute lactic acid (IP acid) as an acute visceral pain stimulus immediately before being placed into 4-inch diameter plexiglass cylinders for 20-min periods of videotaped observation. Two observers blinded to experimental treatment scored each video by counting the number of pain-related stretches (defined as contraction of the abdomen followed by extension of the hind limbs). Observer scores for each session were averaged for subsequent analysis. The cyclooxygenase inhibitor ketoprofen was also evaluated for its effectiveness to block IP acid effects in a separate group [3]. The following steps in experimental design and analysis were incorporated to address The Four Cs.

(1) IP acid alone was tested in one group of 12 mice, ketoprofen + IP acid was tested in a separate group of 12 mice, and each group contained equal numbers of male and female mice ($N = 6$ /

sex). The total group size, which was modestly larger than the $N = 6$ –10 mice used in our previous study with only males [3], was selected based on the observation that confidence-interval widths for any normally distributed population decline with increasing sample size and approach an asymptote at approximately $N = 12$ [4]. We further chose to constitute this $N = 12$ group size with equal numbers of $N = 6$ mice per sex in part to comply with published guidance by the *British Journal of Pharmacology* for use of equal sample sizes of $N \geq 5$ across groups [5]. However, it is important to note that NIH guidance does not specify minimum group sizes or allocations by sex, and our approach could be used regardless of the numbers of males and females included in a study.

(2) Sex was not the primary independent variable of interest in this study, so the primary analysis pooled data from both sexes and used repeated-measures one-way ANOVA (Prism 8.0, Graph-Pad, La Jolla, CA) to evaluate effects of IP acid or ketoprofen dose. A significant ANOVA was followed by Dunnett’s post hoc test to compare acid or ketoprofen effects to vehicle effects. The criterion for significance for this and all other statistical tests was $p < 0.05$. IP acid produced a dose-dependent stimulation of stretching, and ketoprofen dose-dependently blocked IP acid effects.

(3) Secondary analyses to address sex as a biological variable proceeded in two steps. First, data in each panel of Fig. 1 were segregated by sex and analyzed by repeated-measures one-way ANOVA. IP acid stimulated stretching in both sexes; however, ketoprofen failed to significantly decrease IP acid effects in either sex. Second, data from males and females in each panel were also analyzed by two-way ANOVA (Prism 8.0), with sex as a between-subjects factor and IP acid or ketoprofen dose as a within-subjects factor. A significant interaction was followed by a Holm–Sidak post hoc test. For IP acid alone, there was a significant main effect of acid dose but not of sex. The acid dose $\times$ sex interaction was significant, but post hoc analysis did not reveal a sex difference at any acid dose. For ketoprofen + IP acid, there was a significant main effect of ketoprofen dose but not of sex, and the ketoprofen dose $\times$ sex interaction was also not significant.

(4) Lastly, all one-way and two-way ANOVA results were submitted to power analysis to calculate three values: (a) Cohen’s f effect size, (b) achieved power ($1 - \beta$), and (c) the total number of animals predicted as necessary to detect a significant effect given the empirically determined effect size and criterion levels of $\alpha = 0.05$ and power ($1 - \beta$) = 0.8 ($G \times \text{Power}$ [6], free and publicly available: <http://www.gpower.hhu.de>).

¹Department of Pharmacology and Toxicology, Virginia Commonwealth University, Richmond, VA 23298, USA and ²Department of Anatomy and Neurobiology, Virginia Commonwealth University, Richmond, VA 23298, USA

Correspondence: S. Stevens Negus (sidney.negus@vcuhealth.org)

Received: 23 April 2019 Revised: 23 June 2019 Accepted: 28 June 2019

Published online: 5 July 2019

Fig. 1 Effects of IP lactic acid administered alone or after pretreatment with ketoprofen on pain-related stretching behavior in male and female mice. Abscissae: concentration of lactic acid (left panel, diluted in sterile water and administered IP in a volume of 10 ml/kg) or dose of ketoprofen (right panel, administered subcutaneously as a 30-min pretreatment to 0.32% IP acid). Veh = vehicle for IP acid (left panel) or ketoprofen before 0.32% IP acid (right panel). Ordinates: number of stretches observed during a 20-min observation period. Points show mean $\pm$ SEM for all mice (black circles, $N = 12$), just males (blue squares, $N = 6$), and just females (pink diamonds, $N = 6$). Filled symbols indicate significantly different from Veh as determined by one-way ANOVA and Dunnett's post hoc test for pooled data, males, or females. \$ Indicates a significant sex $\times$ dose interaction as determined by two-way ANOVA, but follow-up analysis with the Holm-Sidak post hoc test did not reveal a significant effect of sex at any acid dose. The results of ANOVA and power analysis for data in each panel are shown in Table 1

Power analysis complements the ANOVA results in three ways. First, "effect size" provides a basis for comparing the magnitude of sex differences or other effects across studies [7]. In this data set, for example, all of the effect sizes for IP acid alone were greater than the effect sizes for ketoprofen antinociception.

Second, "power" provides a basis for confidence in drawing conclusions based on the ANOVA results. In particular, just as " α " values specify the probability of a Type 1 (false-positive) error, so the " β " values specify the probability of a Type 2 (false-negative) error. Moreover, just as convention accepts $\alpha \leq 0.05$ as an acceptable criterion for Type 1 errors in concluding that an effect is PRESENT, so convention also generally accepts $\beta \leq 0.2$ as an acceptable criterion for Type 2 errors in concluding that an effect is ABSENT [8]. Insofar as the term "power" represents $1 - \beta$, then this criterion is equivalent to power ≥ 0.8 . Given this criterion, it is appropriate to conclude that an effect of sex (or any other variable) is absent only if the experiment is sufficiently powered to reach that conclusion at power ≥ 0.8 . In this data set, most sex-based analyses failed to reach the criterion for statistical significance, and even with the significant dose $\times$ sex interaction for effects of IP acid alone, the post hoc analysis failed to reveal a significant sex difference at any acid dose. However, power analysis indicated that it would be inappropriate to conclude from these ANOVA results that a sex difference was absent for either IP acid or ketoprofen, because none of the analyses involving sex achieved power ≥ 0.8 . The same principle applies to interpretation of ketoprofen effects in males or females alone. Although ketoprofen failed to produce a significant decrease in IP acid-stimulated stretching in either males or females, power was well under 0.8 in both sexes. Consequently, it would be inappropriate to conclude from these ANOVA results that ketoprofen had no effect. Notably, power ≤ 0.8 is not problematic in the event that a significant effect is deemed to be present, because in these cases, the concern is with potential false-positive conclusions (addressed by α) and not with potential false-negative conclusions (addressed by β). For example, the pooled analysis of ketoprofen effects had power < 0.8 (0.752); however, the ketoprofen data met the criterion for a significant effect, so there was no need to address the potential of a false-negative conclusion.

Lastly, the "predicted N" can inform experimental design for future studies that might pursue evaluation of sex differences. Thus, the effect size observed for a treatment in an initial sample of subjects from some population can be used to predict the sample size required to achieve target statistical criteria (e.g., $\alpha \leq 0.05$, $\beta \leq$

0.2) for confidence in reaching positive or negative conclusions regarding that treatment effect in other subjects from that population. Importantly, this prediction is founded on the assumption that the effect size observed in the initial sample size is representative of the whole population, but this of course is an empirical question (see ref. [8] for commentary). In this study for example, power analysis predicts that total sample sizes of 56 (28/sex) and 16 (8/sex) would be required to adequately characterize the presence or absence of a main effect of sex and a dose $\times$ sex interaction, respectively, for two-way ANOVA of effects produced by IP acid alone. Similarly, samples sizes would need to be increased to 15 (males) or 10 (females) to adequately characterize effects of ketoprofen by one-way ANOVA in males or females alone.

In summary, this approach provides a strategy to address the Four Cs in NIH-funded preclinical research. Inclusion of both males and females is responsive to the "Consider" mandate for the experimental design that takes sex into account. The pooling of data across sex permits focus on the primary variable(s) of interest, while segregation of data by sex addresses the "Collect" mandate for tabulating sex-based data. The secondary analyses address the "Characterize" mandate by using ANOVAs to analyze sex-based differences with existing sample sizes and power analyses to guide both interpretation of the ANOVA results and design of any future studies that might focus on sex differences. Finally, the results of ANOVAs (F statistics) and power analyses (effect size, power, and predicted N s) provide a useful array of statistical outcome measures that fulfill the "Communicate" mandate for reporting and publishing sex-based data.

Ultimately, three categories of sex differences have been described: Type 1 ("sexual dimorphism"; qualitatively different phenotypes between sexes), Type 2 ("sex differences"; quantitatively different phenotypes), and Type 3 ("sex convergence and divergence"; similar phenotypes with different biological mechanisms) [9]. The analysis proposed here will be more sensitive to Type 1 than Type 2 differences, but in either case, it can provide preliminary information on the existence of phenotypic sex differences. By contrast, this approach will not detect Type 3 differences. As a result, even if adequately powered statistical analysis indicates that a phenotypic sex difference is absent, it remains possible that the convergent phenotypes could have sexually divergent underlying mechanisms. With these caveats in mind, the approach proposed here can serve as a strategy for preliminary analysis of sex effects in studies that include both sexes, but do not aim to focus on sex as a primary variable of

Table 1. Summary of power analysis results from one-way and two-way ANOVA of data shown in Fig. 1

Analysis	IP lactic acid alone		Ketoprofen + IP acid		
	F Statistic, p-value	Current effect size (Cohen's f)	Sample size for power ≥ 0.8	F statistic, p-value	Current effect size (Cohen's f)
Sample size for power ≥ 0.8					
From one-way ANOVA					
Pooled males + females	F (2.19, 24.07) = 15.29, $p < 0.0001$	1.180	0.999	F = (1.31, 14.33) = 3.84, $p = 0.0147$	0.752
Males only	F (1.44, 7.18) = 43.55, $p = 0.0002$	2.951	0.999	F (1.12, 5.61) = 3.07, $p = 0.1329$	0.320
Females only	F (2.09, 10.43) = 7.84, $p = 0.0079$	1.253	0.879	F (1.66, 8.29) = 3.84, $p = 0.0714$	0.494
From two-way ANOVA					
Main effect of dose	F (2.26, 22.64) = 28.10, $p < 0.0001$	1.674	1.0	F (1.30, 12.97) = 6.40, $p = 0.0194$	0.713
Main effect of sex*	F (1.10) = 4.34, $p = 0.0639$	0.383	0.225	F (1.10) = 0.139, $p = 0.892$	0.057
Dose x sex interact	F (3.30) = 3.06, $p = 0.0430$	0.553	0.657	F = (4, 40) = 0.28, $p = 0.8920$	0.104
<p>For predicted "Sample Size for Power ≥ 0.8," numbers show the total sample size as well as the allocation to males (M) and females (F) assuming equal representation by both sexes</p> <p>*To illustrate the approach using Prism 8.0 and G*Power, detailed step-by-step methods are described here for power analysis of the Main Effect of Sex for IP Acid Alone on Stretching</p> <p>(1) From the two-way ANOVA table in Prism 8.0, calculate partial η^2 from the sum of squares (SS) values for the main effect of sex (SS_{sex}) and the residual variance unaccounted for by either sex or dose ($SS_{residuals}$) as follows: [10]</p> $\text{partial } \eta^2 = \frac{SS_{sex}}{SS_{sex} + SS_{residuals}} = \frac{164.7}{164.7 + 1121} = 0.1281$ <p>(2) In G*Power select the following:</p> <ul style="list-style-type: none"> - Test family: F tests - Statistical test: ANOVA: fixed effects, special, main effects, and interactions - Type of power analysis: Post hoc: Compute achieved power - a given α, sample size, and effect size - Input Parameters: Select "Determine," enter partial η^2 from step 1 above, and select "Calculate effect size f and transfer to main window"; also enter 0.05 for "α err prob"; 12 for "total sample size"; 1 for "numerator df"; and 2 for "number of groups" - Output Parameters: Select "Calculate" to calculate output parameters that include "Denominator df" and "Power (1 - β err prob)." Select "X-Y plot for range of values" to generate a graph and associated table that relate changes in Power to changes in Total Sample Size and determine from the table the smallest sample size necessary to obtain the desired power. This study selected the sample size necessary to achieve Power ≥ 0.8 (56 total animals) - Report: "Effect size f" from Input Parameters (Cohen's f), "Power (1 - β err prob)" from Output Parameters, and "Total Sample Size" necessary for desired power from Table (3) To ensure all metrics were entered correctly, check that the degrees of freedom in the "Output parameters" ("Denominator df") of G*Power matches that from the ANOVA table in Prism. If they do not agree, data were not entered correctly 					

interest. In those instances where sex is not the primary variable of interest, we suggest that priority of analyses should be assigned to those primary variables that have been adequately powered based on initial study design. Additional analyses of secondary variables such as sex are important and can be addressed in appropriately designed follow-up studies.

FUNDING AND DISCLOSURE

This work was supported by grants T32DA007027 (CMD), R01DA037287 (MLB), R01NR014886 (GNN), and R01DA030404 (SSN). The authors declare no competing interests.

ADDITIONAL INFORMATION

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Miller LR, Marks C, Becker JB, Hurn PD, Chen WJ, Woodruff T, et al. Considering sex as a biological variable in preclinical research. *FASEB J*. 2017;31:29–34.
2. Greenspan JD, Craft RM, LeResche L, Arendt-Nielsen L, Berkley KJ, Fillingim RB, et al. Studying sex and gender differences in pain and analgesia: a consensus report. *Pain*. 2007;132(Suppl 1):S26–45.
3. Bagdas D, Muldoon PP, AlSharari S, Carroll FI, Negus SS, Damaj MI. Expression and pharmacological modulation of visceral pain-induced conditioned place aversion in mice. *Neuropharmacology*. 2016;102:236–43.
4. van Belle G. *Statistical rules of thumb*. Hoboken, NJ: John Wiley & Sons; 2008.
5. Curtis MJ, Bond RA, Spina D, Ahluwalia A, Alexander SP, Giembycz MA, et al. Experimental design and analysis and their reporting: new guidance for publication in *BJP*. *Br J Pharmacol*. 2015;172:3461–71.
6. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*. 2007;39:175–91.
7. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev Camb Philos Soc*. 2007;82:591–605.
8. Marino MJ. How often should we expect to be wrong? Statistical power, P values, and the expected prevalence of false discoveries. *Biochem Pharmacol*. 2018;151:226–33.
9. McCarthy MM, Arnold AP, Ball GF, Blaustein JD, De Vries GJ. Sex differences in the brain: the not so inconvenient truth. *J Neurosci*. 2012;32:2241–7.
10. Richardson JTE. Eta squared and partial eta squared as measures of effect size in educational research. *Educational Res Rev*. 2011;6:135–47.