Boberg et al. (2011) - Corrigendum (2016): further significant modifications needed

Peter Morfeld 1,2, Min Chen3, Rainer Otter4, Christine Palermo5, Jessica Kemmerling6, Michela Mastrantonio7*

1 Institute for Occupational Epidemiology and Risk Assessment (IERA) of Evonik Industries, Germany, Rellinghauser Str. 1-11, 45128 Essen
2 Institute and Policlinic for Occupational Medicine, Environmental Medicine and Prevention Research of Cologne University, Germany, Kerpener Strasse 62, 50937 Köln
3 ExxonMobil Biomedical Sciences Inc., USA, 1545 Route 22 East, Annandale, New Jersey 08801
4 BASF SE, Germany, Carl-Bosch-Strasse 38, 67056 Ludwigshafen
5 ExxonMobil Biomedical Sciences Inc., Belgium, Hermeslaan 2, 1831 Machelen
6 Evonik Performance Materials GmbH, Germany, Paul-Baumann-Strasse 1, 45772 Marl
7 ECPI, Belgium, Avenue E. Van Nieuwenhuyse 4, B-1160 Brussels,
*Corresponding Author
October 18, 2016

Dear Editor,

We are writing to express our concerns regarding the Corrigendum [1] to ‘Reproductive and behavioral effects of diisononyl phthalate (DINP) in perinatally exposed rats” [2] as well as concerns relating to the original publication [2]. It is our view the changes to the statistical methods reflected in the Corrigendum are much more impactful to the results of the analysis and overall interpretation of data than the tone of the Corrigendum reflects. In the Corrigendum, Boberg et al. have made changes that reveal deviations from their originally published study protocol and clearly acknowledge their published results represent outcomes of non-standard statistical approaches. Additionally, the authors fail to address a number of other statistical and experimental errors that impact reproducibility. We consider it important to call attention to this matter and detail the discrepancies below as the statistical significance of the results is the central basis of the authors’ conclusions and is therefore not only relevant to the representation of results but also to the valid interpretation of these data. Boberg et al. [1], [2] should be significantly modified to improve the accuracy and transparency of reporting and importantly to reflect outcomes based on the originally published, standard statistical methods.

We had contacted Boberg et al. in May 2016 concerning difficulties in reproducing a number of their results based on the statistical methods as described in [2] prior to the recent publication of the Corrigendum[1]. Using the statistical methods as originally published in [2], we reanalyzed the raw data¹ that were provided for the following endpoints: testosterone, nipple retention, sperm motility, sperm/g cauda, percent progressive sperm and anogenital distance (AGD/AGDi) measurements. We were unable to reproduce the reported statistical significance for four of these six parameters as published in [2].

1. **Percent progressive sperm**: Boberg et al. [2] reported a significant difference in percent progressive sperm between the control and the 750 mg/kg/day exposure group. Our reanalysis concluded that there is **no significant difference between control animals and those treated with DINP at all doses**.
2. **Sperm/g cauda**: Boberg et al. [2] reported a significant increase in sperm/g cauda epididymis at the highest dose group (900 mg/kg/day). Our reanalysis concluded that there is **no significant difference between control animals and those treated with DINP at all doses**.
3. **Sperm motility**: Boberg et al. [2] reported a significant decrease in sperm motility between control and the 600, 750 and 900 mg/kg/day exposure groups. Our reanalysis confirmed statistical significance at the two highest doses (750 and 900 mg/kg/day); however, the reanalysis concluded that there is **no significant difference between control animals and those treated with 600 mg/kg/day DINP**.
4. **AGD and AGDi**: Boberg et al. [2] reported a statistically significant difference in both anogenital distance (ADG) and AGDi on PND 13 between the control and 900 mg/kg/day exposure group. These findings could not be reproduced according to the originally published methods and our analysis concluded that there is **no significant difference in AGD and AGDi between control animals and those treated with DINP**.

To address the discordance for the sperm parameters noted in # 1-3 above, the Corrigendum modified the Materials and Methods to reflect a change in the statistical approach as follows; “p-values for differences between LS means were reported without correction for multiple comparisons. This was only done in cases when the overall ANOVA had a p-value below 0.05.” By making this change, the authors eliminated the statistical correction for multiple comparisons. The explanation offered in the Corrigendum for this change is that “the authors regret that the Material and Methods section lacks some details regarding statistical analysis”. This improperly characterizes the nature of this correction. The “details” added change the statistical methodology applied, thereby aligning the

¹ The raw data as supplied by the study authors were provided to us via the Danish Environmental Protection Agency (with the agreement of Dr. Boberg) in February 2016 in the context of a discussion on a regulatory hazard classification proposal. At approximately the same time, the raw data were made publicly available from the US EPA HERO database in connection with the US EPA IRIS DINP review [3]. It should be noted that the discrepancies documented below could not have been identified at the time of peer review and publication of the paper, because the raw data were not made available at that time. It is recognized that during the peer review process, peer reviewers take the data at face value and understandably assume that researchers are following their stated methods and carefully documenting scientific results, consistent with standards of good scientific practice.
methods with the conclusions in the original 2011 publication. Of significant concern, Boberg et al. state that eliminating use of the Dunnett’s test to correct for multiple comparisons is “not considered appropriate by some statisticians.” This is not an accurate statement. The accepted statistical practice is to apply correction for multiple comparisons in all instances when multiple comparisons have been performed. As more statistical comparisons are done, the probability of incorrectly (i.e., making a Type I error) rejecting a null hypothesis (i.e. identifying a p<0.05) increases by chance [4]. This error should be controlled for by correcting for multiple comparisons. Therefore, the methods as written in the original publication reflect the appropriate statistical approach, and the results in the publication (and not the Materials and Methods) should be updated to reflect the Dunnett’s p-value corrected for multiple comparisons as described in #1-3 above. Without correction, readers may fail to note how this non-standard approach in the statistical analysis referred to in the Corrigendum may have affected the results and interpretation.

To address the discordance of AGD parameters noted in #4 above, the statistical approach was modified in the Corrigendum to include the date of AGD measurement as a random factor. The Corrigendum explains this statistical modification is warranted “due to unexpected matters”. The method as originally published of measuring AGD by the same technician blinded to exposure group is now replaced by the new method of “AGD was measured by different technicians on different days”. While this change facilitates reproducibility of the statistical significance reported in [2], there is, however, no proposed change to the Materials and Methods section of the manuscript to account for this different experimental approach, nor is there any explanation of the “unexpected matters” leading to this change. Furthermore, the variables introduced with this experimental approach should be addressed in the publication, e.g. how doses were randomized across days, whether the measurements were blinded, whether there was control for technician variability. We emphasize that the statistical methods chosen by the authors assume that randomization was performed efficiently. Thus, the authors should describe how they randomized while controlling for weight: “The dams were randomized into five groups of 16 with similar body weight distributions “[2, p. 201]. It should also be described how the rats were randomly allocated to the technicians who did the measurements.

The authors state in the Corrigendum that “[i]t should be emphasized that the conclusion that DINP reduces sperm motility is true also when considering results of the Dunnett’s test.” However, this conclusion hinges solely on statistical significance without consideration of biological significance. The low percentage of motile sperm in the control animals of 60% does not comply with the minimum quality requirements in OECD guidance [5, 6] where a percentage of at least 70 % motile sperm is defined as a standard requirement. Since sperm parameters are sensitive to sampling techniques, analyses techniques, and environmental conditions [5], interpretation of results is contingent upon experimental optimization. In addition, we extracted the historical control data from two studies performed in the same laboratory. The motile sperm control sets of Boberg et al. 2011 (mean=59.3%), Jarfelt et al. 2005 (mean=66.9%) and Taxvig et al. 2007 (mean=53.8%) did not meet the requirements of the OECD guidance. In addition, we found that the percentage of motile sperm varied statistically significantly between the controls in these three studies (F test: p-value = 1.3 %). Thus, the variance in motile sperm percentage cannot be explained by physiological (random) fluctuations and points to the conclusion that standardisation of this parameter failed. We note further that the changes reported for DINP are within the unexplained variance of the controls, and therefore do not differ substantially from the historical controls within the laboratory. The authors should account for the non-adherence to OECD standards and unexplained systematic variation in control sets from the same laboratory in their interpretation.

In addition to the above, the following have not been addressed in the Corrigendum: for nipple retention, Table 3 footnote c indicates ANOVA followed by the Dunnett’s test, whereas generalized linear models in combination with generalized estimating equations were described in the Materials and Methods section. The statistical significance reported in Table 1 (Testis histology) could not be reproduced upon adjustment for multiple comparisons. The Materials and Methods section reported the “ROM procedure” was used to adjust for multiple comparisons when analyses were done without ANOVA. We examined Table 1 that reports on results from many Fisher exact tests, the reported statistical significance could not be reproduced upon application of the ROM procedure or other multiple comparison adjustment procedures. The descriptive statistics reported in Table 3 for
the dose group of DINP 900 mg/kg/day, for AGD, ADGi and nipples in males could not be reproduced. The standard deviation of AGDi in males for the dose group DINP 600 mg/kg/day is reported as 1.20 in the raw data file, but reported as 0.20 in Table 3 in the publication. While most of these changes may seem insignificant, publications should accurately reflect the experimental outcomes.

As every published study contributes to scientific discovery and biological understanding, it is essential that data be transparently reported, verifiable and adhere to accepted scientific practices. To facilitate this, important methodological approaches and outcomes must be accurately and completely reported in [2] including clarity in statistical approaches, methods of measurement, and randomization. Furthermore, proper use of statistics and minimal experimental quality requirements - both of which have clear impacts on outcomes and interpretation - must be adhered to. Boberg et al [1] [2] should be significantly modified to reflect outcomes and conclusions based on the originally published and standard statistical methods and other errors corrected.

References


Acknowledgement

We would like to thank Dr. Boberg and the Danish EPA for making the raw data available to the European Council of Plasticisers and Intermediates