Received: 10 July 2020 | Revised: 2 January 2021 | Accepted: 6 January 2021

DOI: 10.1002/em.22422

Environmental and Molecular Mutagenesis | Environmental Mutagenesis and Genomics Society | WILEY

RESEARCH ARTICLE

The use of benchmark dose uncertainty measurements for robust comparative potency analyses

Ryan P. Wheeldon[1] | Stephen D. Dertinger[2] | Steven M. Bryce[2] | Jeffrey C. Bemis[2] | George E. Johnson[1]

[1]Institute of Life Science, Swansea University Medical School, Swansea University, Swansea, UK

[2]Litron Laboratories, Rochester, New York

Correspondence
Ryan P. Wheeldon, Swansea University Medical School, Swansea University, Swansea SA2 8PP, UK.
Email: r.wheeldon.702067@swansea.ac.uk

Funding information
National Institute of Environmental Health Sciences, Grant/Award Number: R44ES029014

Accepted by: P. White

Abstract

The Benchmark Dose (BMD) method is the favored approach for quantitative dose–response analysis where uncertainty measurements are delineated between the upper (BMDU) and lower (BMDL) confidence bounds, or confidence intervals (CIs). Little has been published on the accurate interpretation of uncertainty measurements for potency comparative analyses between different test conditions. We highlight this by revisiting a previously published comparative in vitro genotoxicity dataset for human lymphoblastoid TK6 cells that were exposed to each of 10 clastogens in the presence and absence (+/−) of low concentration (0.25%) S9, and scored for p53, γH2AX and Relative Nuclei Count (RNC) responses at two timepoints (Tian et al., 2020). The researchers utilized BMD point estimates in potency comparative analysis between S9 treatment conditions. Here we highlight a shortcoming that the use of BMD point estimates can mischaracterize potency differences between systems. We reanalyzed the dose responses by BMD modeling using PROAST v69.1. We used the resulting BMDL and BMDU metrics to calculate "S9 potency ratio confidence intervals" that compare the relative potency of compounds +/− S9 as more statistically robust metrics for comparative potency measurements compared to BMD point estimate ratios. We performed unsupervised hierarchical clustering that identified four S9-dependent groupings: high and low-level potentiation, no effect, and diminution. This work demonstrates the importance of using BMD uncertainty measurements in potency comparative analyses between test conditions. Irrespective of the source of the data, we propose a stepwise approach when performing BMD modeling in comparative potency analyses between test conditions.

KEYWORDS
BMD, comparative, DNA damage, potency, uncertainty

1 | INTRODUCTION

An appreciation of dose–response analysis of genotoxicity data has accompanied a shift from a hazard identification testing approach, toward quantitative assessment of genotoxicity for risk assessment purposes (Macgregor et al., 2015; Dearfield et al., 2017). Under the auspices of the Quantitative Analysis Workgroup (QAW) of the Health and Environmental Sciences Institute Genetic Toxicology Technical Committee (HESI GTTC), different statistical methods for assessing dose–response relationships in genetic toxicology studies have been evaluated (Gollapudi et al., 2013). An overall conclusion was that the benchmark dose (BMD) approach for analyzing dose–response data derived from genotoxicity

studies exhibits the most favorable characteristics. The BMD for continuous data is defined as the dose that results in a predetermined change, typically ranging between 1 and 10% in the response rate of an adverse effect relative to existing background incidence (Macgregor et al., 2015). Unlike other dose–response analysis methods such as the No Observed Genotoxic Effect Level (NOGEL) approach, the BMD method is advantageous since it is not restricted to the study concentration/dose selection. The BMD method evaluates the entire range of concentrations/doses within the dataset while providing measures of uncertainty such as confidence limits (Crump, 1984; Slob, 2002). The uncertainty in the estimation of the BMD is defined as the range delineated between the upper (BMDU) and the lower (BMDL) confidence bounds.

BMDs and their associated confidence intervals (CIs) can be interpreted in a manner that conveys the potency of the studied test article. For example, CIs have been plotted to illustrate compound potency ranking from in vivo carcinogenicity and genotoxicity studies, as well as providing empirical comparisons across endpoints. Additionally, combined analysis can be performed for multiple dose–response datasets for a shared endpoint differentiated by one or more covariates (Slob and Setzer, 2014). This is known as the BMD combined covariate approach. A prominent study by Wills and colleagues (Wills et al., 2016) applied the BMD combined covariate approach to in vitro genotoxicity data and demonstrated that the precision of the BMD increases when compound, or another condition that pertains to mode of action (MoA), serves as the covariate. The precision of the BMD is defined by the ratio of the BMDL to the BMDU and has potential implications for regulatory decision making when used to define a Point of Departure (PoD) for extrapolation to a human exposure limit (White et al., 2020).

Wills et al. (2016) states that as CIs represent the range in which the true BMD lies, potency differences are only statistically defendable when there is no apparent overlap between intervals. Thus, evaluation of the entire BMD CI (BMDL and BMDU) is imperative to drawing potency conclusions from BMD analysis of dose–response datasets. Researchers have used BMD CIs of chemical classes of interest to plot compound genotoxic and/or carcinogenic potency in rank order (Hernández et al., 2011; Soeteman-Hernández et al., 2015a; Soeteman-Hernández et al., 2015b; Wills et al., 2016, 2017). Some prominent examples where BMD CIs were used in comparative genotoxicity potency analyses include studies by Allemang et al. (2018) and Wheeldon et al. (2020). Allemang et al. (2018) performed BMD analyses to evaluate the relative genotoxic potency of 15 pyrrolizidine alkaloids (PAs) via in vitro micronuclei formation in HepaRG cells. Wheeldon et al. (2020) performed BMD analyses to evaluate the comparative genotoxic potency of 8 Topoisomerase II poisons studied in human lymphoblastoid TK6 cells using the MultiFlow DNA damage response assay. Both publications identified the utility of BMD CIs in compound comparative genotoxicity potency analyses to support read across and MoA determination for a limited number of compounds of interest. In all of the aforementioned studies, compound potency comparisons within and between genotoxicity endpoints were always performed by scrutinizing the shape and steepness of the underlying dose–response curve, and by graphically representing the BMD CIs. Potency comparisons and measures of correlation were consistently evaluated by CIs spanning orders of magnitude versus deriving numerical values for comparison purposes.

Herein we consider a previously published in vitro genotoxicity dataset where the authors performed BMD analyses and drew conclusions without considering the uncertainty associated with the BMD measurements. Tian et al. (2020) investigated the use of phenobarbital/β-napthoflavone-induced rat liver S9 at maximal non-cytotoxic concentration (0.25% vol/vol final) in a flow cytometry based multiplexed DNA damage response assay. The laboratory was able to maintain the S9 enzyme/co-factor mix with cells and test compound for the entire exposure period of 24 hr in the TK6-based assay. The investigators focused on 15 chemicals: 8 of which are clastogens that are known to require metabolic activation to maximize formation of DNA-reactive metabolites; 5 are cytotoxicants; and 2 are direct acting clastogens that do not require metabolic activation. In addition to determining compound MoA through the use of biomarker responses, the authors applied the BMD approach with the aim of calculating a numerical "S9 potentiation ratio" value, which served as a comparison metric obtained by division of the BMD value in the absence of S9 by the BMD value in the presence of S9. Upon further consideration of this approach, we believe that there are significant shortcomings in the use of solely a BMD point estimate in their comparative potency analysis. Using the same logic highlighted by Wills et al. (2016); it is irreconcilable to rely upon a BMD point estimate to robustly compare potencies, since the BMD CI represents the range in which the true BMD lies. To our knowledge, there are no published reports that critique the use of a BMD point estimate, and how this differs from use of BMD CIs in comparative analysis between experimental conditions. Herein, the Tian et al. (2020) data is re-analyzed to further inform use of the same in vitro genotoxicity/low concentration S9 dataset. Primarily, we stress that the use of a BMD point estimate value does not provide an accurate representation of the likely potency range of the test compound. Said another way, the BMD point estimates and associated "S9 potentiation ratio" comparison metrics did not convey information about the uncertainty of the measurements that are consistent with the BMD uncertainty measurement approach that is advocated in the scientific literature.

This current report focuses on reanalysis and augmentation of the previously published Tian et al. (2020) dataset. Specifically, we convey the BMD uncertainty measurements that relate to the relative genotoxic potency of the 10 clastogenic compounds. To this end, we calculate "S9 potency ratio CIs" using the BMD uncertainty measurements (BMDL and BMDU) between S9 exposure conditions (the presence and absence of S9) and utilize said ratios to derive robust potency conclusions as a follow up to the Tian et al. (2020) work. Readers are encouraged to refer to the original Tian and colleagues' article for further context (Tian et al., 2020).

2 | MATERIALS AND METHODS

2.1 | In vitro genotoxicity dataset

The data were derived from a previously published article in which 15 compounds were studied using the in vitro MultiFlow® DNA

Damage Assay in the presence and absence of low dose (0.25% vol/vol) S9 (Tian et al. 2020). The in vitro MultiFlow DNA Damage Assay multiplexes several biomarkers that are responsive to diverse forms of DNA damage into a single flow cytometric analysis. The multiplexed biomarkers include: (a) phosphorylation of H2AX at serine 139 (γH2AX) for the detection of DNA double strand breaks, (b) phosphorylation of histone H3 at serine 10 (p-H3) to identify mitotic cells, (c) nuclear p53 content as an indicator of p53 activation, (d) frequency of 8n+ cells to monitor polyploidization, and (e) relative nuclei counts (RNC) to provide information about treatment related cytotoxicity (Bryce et al., 2016). A detailed description of the assay falls out of scope of this article; thus, interested readers are encouraged to refer to several publications that describe the MultiFlow assay (Bryce et al., 2016; Bryce et al., 2017; Bryce et al., 2018; Dertinger et al., 2019). As previously mentioned, the Tian et al. (2020) data included 8 direct acting clastogens that require metabolic activation, 5 cytotoxicants and 2 direct acting clastogens that do not require metabolic activation. In this reanalysis, we focused on the 10 clastogens, since this is where Tian et al. (2020) suggested the greatest differences in potency between S9 exist. The raw data were provided by Litron Laboratories and included 4 and 24 hr p53, γH2AX, and 24 hr RNC responses for the 10 clastogens: 2-acetylaminofluorene, 2-aminoanthracene, 7,12-dimethylbenzanthracene, benzo[a]pyrene, cyclophosphamide, dibenzo[a,l]pyrene (also known as dibenzo[def,p]chrysene), diethylnitrosamine, mitomycin C, 2-amino-1-methyl-6-phenylimidazo [4,5-b]pyridine (PhIP), and resorcinol.

2.2 | BMD analyses

Tian et al. (2020), performed BMD analysis on individual compounds with the S9 exposure (with or without [+/−] 0.25% vol/vol S9) condition serving as the covariate. Hence, individual dose–response curves on a compound basis already existed for this dataset. Here, the individual dose–response curves were scrutinized so that compounds and/or endpoints with little to no evidence of a dose–response could be disqualified from reanalysis. Hence, we excluded 4 hr p53 dose responses for diethylnitrosamine; 24 hr p53 dose–responses for diethylnitrosamine; and 24 hr γH2AX dose responses for 7,12-dimethylbenzanthracene.

PROAST version 69.1 operating in R 4.0.2 was used to analyze the continuous dose–response data (http://www.proast.nl) for the 4 hr and 24 hr γH2AX, 4 hr and 24 hr p53, and 24 hr RNC endpoints. We applied the exponential model as a sequence of nested models with increasing number of parameters (EFSA 2009; Slob and Setzer 2014; EFSA 2017). In this reanalysis, compound was selected as the covariate—differing from Tian et al. (2020) where S9 condition was the covariate—since this includes more dose–responses in a single analysis to increase the precision of the BMDs. PROAST provided the option to select model 3 or 5 from the family of models. In doing so, the number of parameters were increased to test whether the associated log likelihood is significantly increased, thus informing if the additional parameters were required for describing the dose–response (Slob, 2002; EFSA, 2009; Slob and Setzer, 2014; EFSA, 2017). Either model 3 or 5 was finally automatically selected as the most appropriate model to describe the shape of the dose–response curve. Readers are advised to refer to Slob and Setzer (2014) for detailed information about the dose response models (including models 3 and 5) and their algorithms.

Consistent with the approach utilized by Tian et al. (2020), an arbitrary critical effect size (CES) of 0.3 was selected for this reanalysis. A CES of 0.3 represents a 30% change in response compared to the concomitant control. CES −0.3 for the RNC endpoint represents a 30% decrease in response for this endpoint analyses. There is a lack of consensus on the appropriate choice of CES reported in the literature for in vivo endpoints (White et al., 2020), thus a lengthy discussion on CES falls outside the scope of this report. In any event, we justify the use of CES 0.3 for these in vitro endpoints since the resulting BMDs do not lie in the extremities of the dose response curves where the associated uncertainties could be relatively high.

The value of the BMD is a point estimate with an associated level of precision (the ratio between the BMDL and BMDU), thus 90% BMD CIs were obtained for each compound and endpoint combination in the presence and absence of S9 and were graphically plotted.

2.3 | Unsupervised clustering

Lower and upper values of the "S9 potency ratio CIs" (algorithm described in the results and discussion section) were evaluated using JMP software's unsupervised clustering platform (JMP, v12.0.1). Lower and upper "S9 potency ratio CI" values associated with the following 5 biomarkers were used as variables: 4 hr p53, 24 hr p53, 4 hr γH2AX, 24 hr γH2AX, and 24 hr RNC. The analysis options were set as follows: clustering method = hierarchical; method for calculating distances between clusters = "Ward"; data as usual = "Standardize Robustly"; data visualization = "Dendrogram," with "two-way clustering."

3 | RESULTS AND DISCUSSION

3.1 | BMD confidence intervals

Values of the BMD, BMDL, and BMDU were estimated and collated for all compounds (+/− S9) and endpoints that were included in the BMD analysis. The BMD CIs were plotted for each compound with +/−S9 CIs side-by-side (Figures 1–10). The BMD point estimates are included in the comparative potency plots to aid in graphical representation of the BMD point estimate relative to the corresponding BMDL and BMDU. This is important since some individuals misinterpret the BMD to be at the midpoint (or geometric mean) between the BMDL and BMDU. This is a misconception as evident by the differing lengths of the CIs either side of each BMD point estimate displayed in Figures 1–10.

CIs that span a maximum of approximately 1–2 Log units (i.e., 1–2 orders of magnitude) are considered good quality and consistent

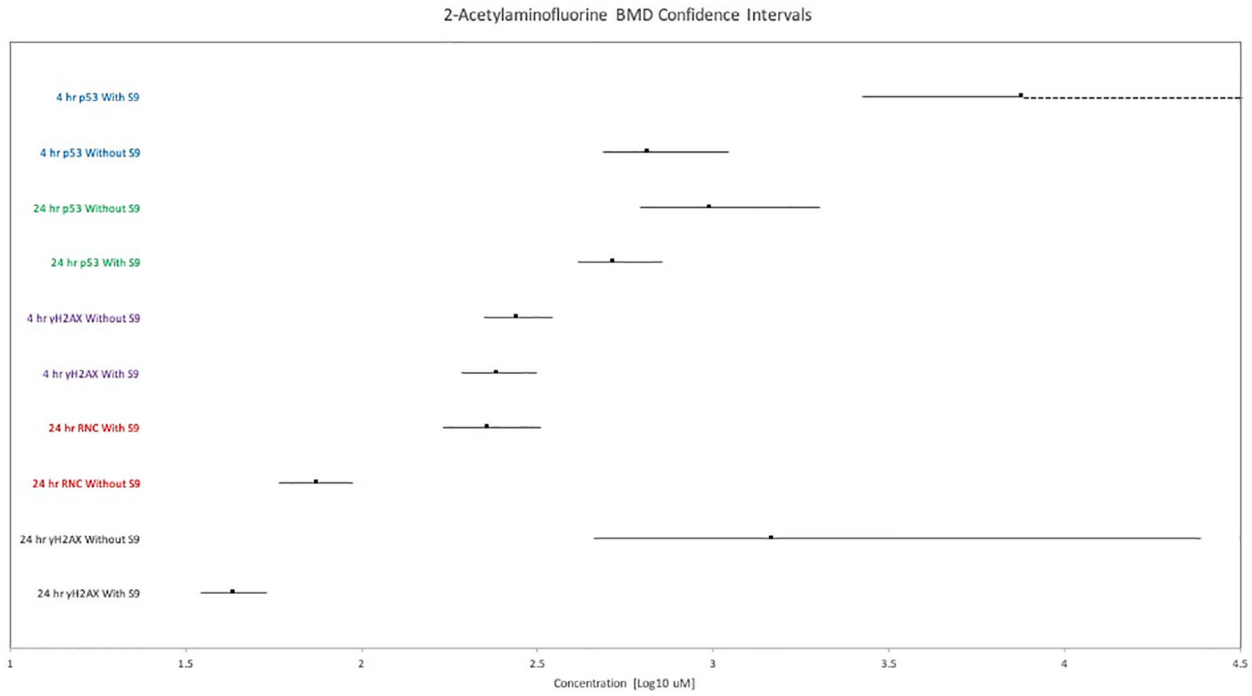

FIGURE 1 BMD confidence intervals for 2-acetylaminofluorene. 4 hr p53 endpoint with S9 infinite BMDU indicated with dashed positive direction lines. BMD point estimates are displayed as data points

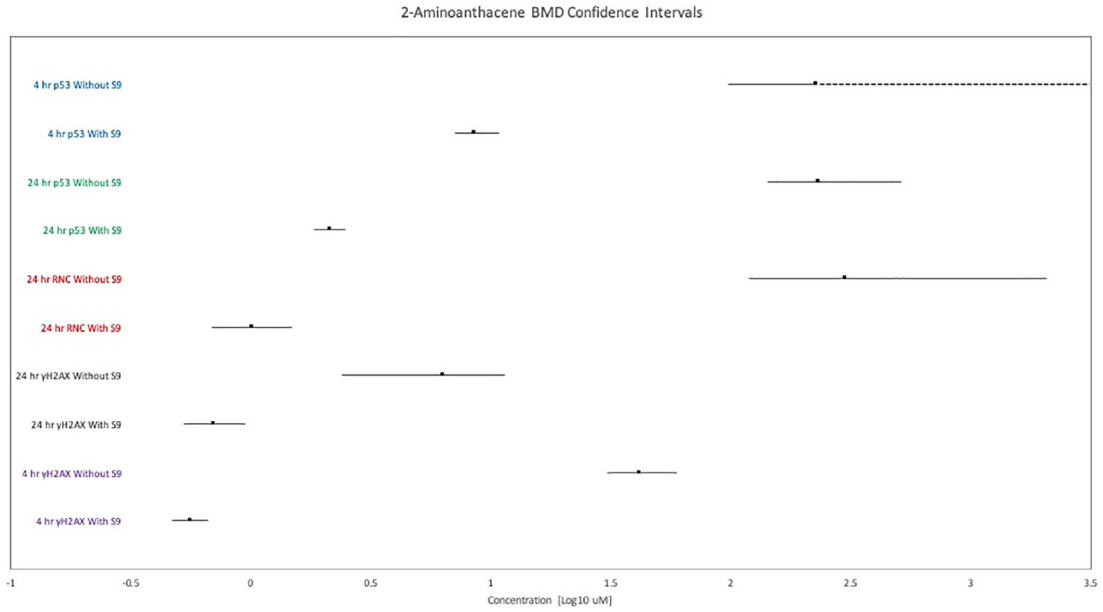

FIGURE 2 BMD confidence intervals for 2-aminoanthracene. 4 hr p53 endpoint without S9 infinite BMDU indicated with dashed positive direction lines. BMD point estimates are displayed as data points

with other BMD CI spans that are reported in the literature for other in vitro genotoxicity systems (Soeteman-Hernández et al., 2015a; Bemis et al., 2016; Wills et al., 2016; Allemang et al., 2018). While most compound's BMD analysis yielded tightly bound CIs, some displayed wide CIs or dose responses with unbound upper confidence limits. Specifically, the 4 hr p53 endpoint for 2-acetylaminofluorene –S9 yielded an infinite BMDU (BMD 3.88 Log10 μM, BMDL 3.42 Log10 μM, BMDU Infinite Log10 μM) (Figure 1). Second, both the 4 hr and 24 hr p53 (Figure 9) endpoints for PhIP -S9 yielded unbound BMDUs (BMD 4.39 Log10 μM, BMDL 3.03 Log10 μM, BMDU Infinite Log10 μM; BMD 5.20 Log10 μM, BMDL 3.22 Log10 μM, BMDU Infinite Log10 μM, respectively). The 4 hr γH2AX endpoint for 7,12-dimethylbenzanthracene returned a zero BMDL value

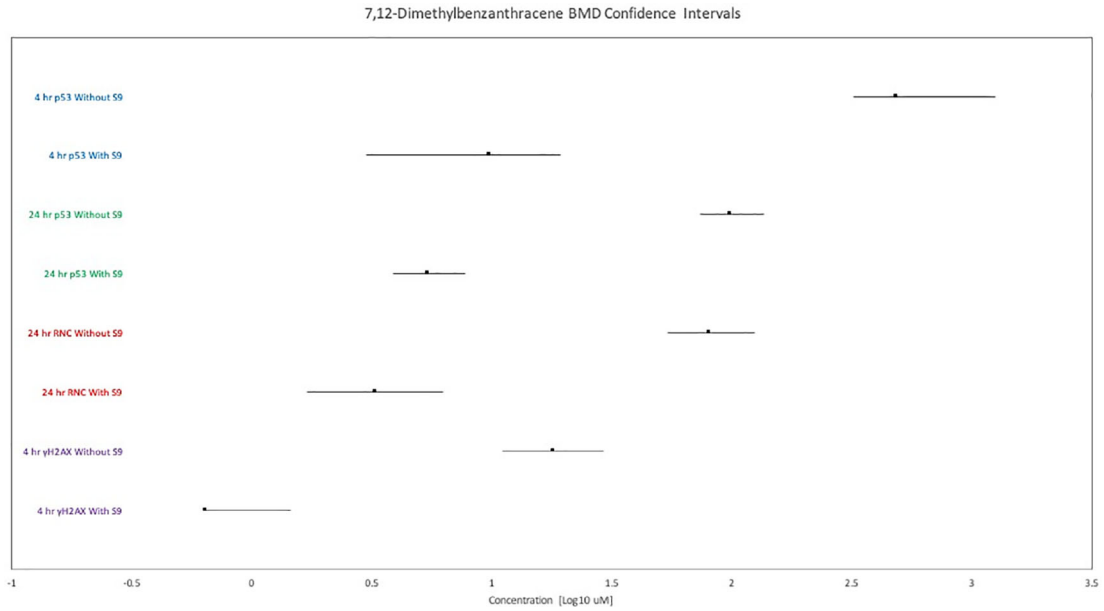

FIGURE 3 BMD confidence intervals for 7,12-dimethylbenzanthracene. 4 hr γH2AX endpoint with S9 yielded a zero value lower confidence bound. BMD point estimates are displayed as data points

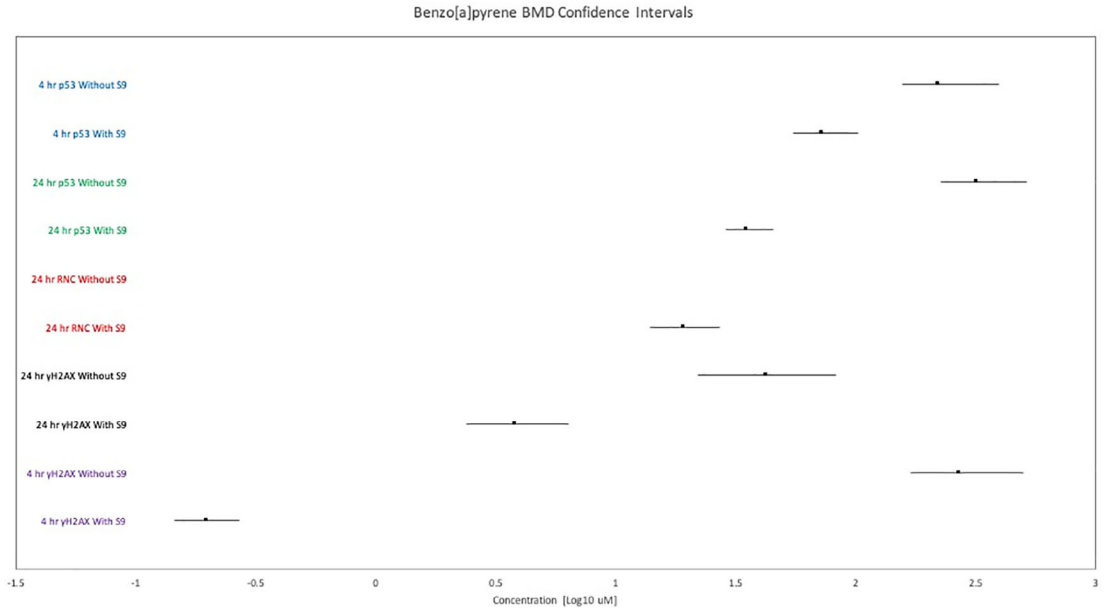

FIGURE 4 BMD confidence intervals for benzo[a]pyrene. 24 hr RNC endpoint without S9 yielded a disproportionately high BMD with infinite BMDU: indicated as a dashed line spanning the width of the plot. BMD point estimates are displayed as data points

indicating that the lower bound of the BMD is not significant from zero. We considered these limited instances where dose–response analyses yielded infinite BMDUs to have exhibited no evidence of dose–response relationships, and hence are unsuitable for drawing statistically robust conclusions from the analyses. In the single instance of a zero BMDL, the resulting BMDL to BMDU ratio is unobtainable (problems associated with division by 0), and hence justifies omission of this compound/endpoint combination from further analysis.

In addition to having unbound BMDUs, some compounds also displayed BMD estimates that are grossly disproportionate to the BMDs from other endpoints or S9 condition for the same compound. The BMD CIs for these instances were not plotted, since the relative disproportionality would make the other CIs appear comparatively small. These instances are indicated in the comparative potency plots with dashed lines spanning the entire graph range. Specifically, the 24 hr RNC endpoint –S9 for benzo[a]pyrene [BMD 12.33 Log10 μM] (Figure 4), the 24 hr RNC endpoint –S9 for dibenzo[a,l]pyrene [BMD

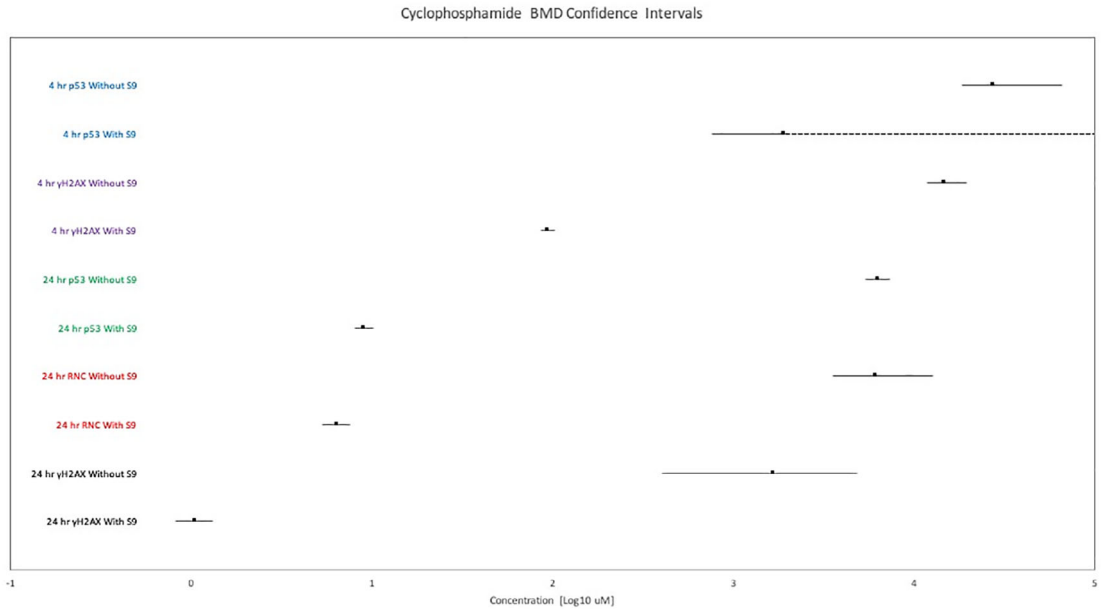

FIGURE 5 BMD confidence intervals for cyclophosphamide. Two-sided confidence intervals were obtained for all endpoint's BMD analyses. BMD point estimates are displayed as data points

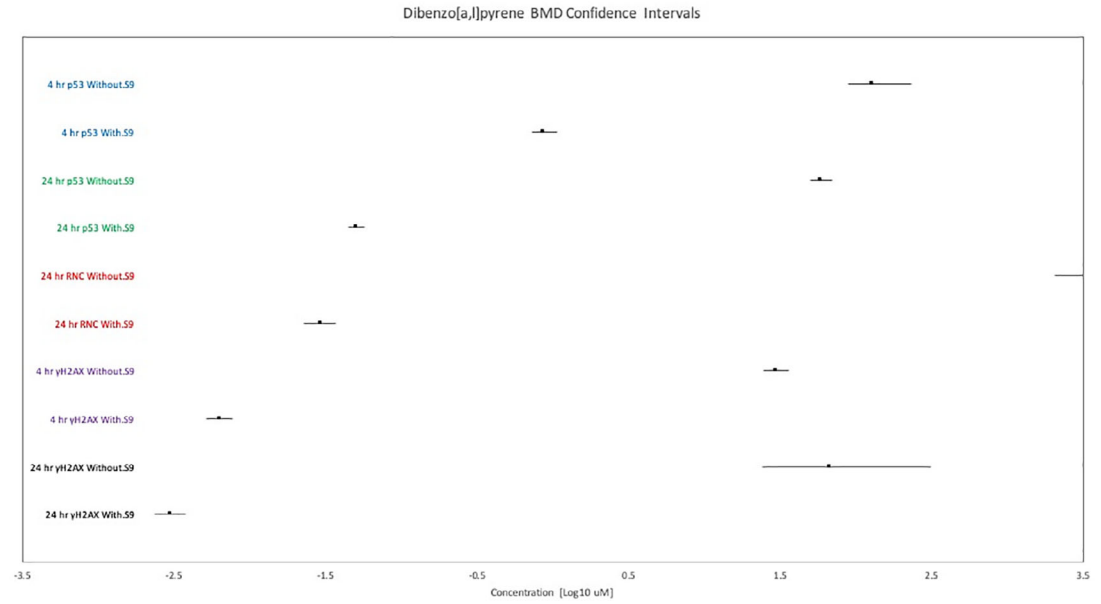

FIGURE 6 BMD confidence intervals for dibenzo[a,l]pyrene. 24 hr RNC endpoint without S9 yielded a disproportionately high BMD with infinite BMDU: BMD restricted and only displaying the BMDL. BMD point estimates are displayed as data points

11.99 Log10 μM] (Figure 6), and the 24 hr γH2AX endpoint −S9 for diethylnitrosamine [BMD 20.74 Log10 μM] (Figure 7). These endpoints also yielded infinite BMDUs, and consequently displayed no evidence of a dose response upon which one can then accurately define potency.

Visual scrutiny of each compound's BMD potency plot allows efficient visual discrimination of the potency difference of each endpoint that results from S9 exposure. Endpoints where the S9 exposure condition yielded BMD CIs that largely overlap results in potency differences that are statistically insignificant. Conversely, non-overlapping CIs are statistically significantly different and show that the S9 condition exerts an impact on the compounds potency in the same in vitro system.

For the most part, as one would expect, the tested clastogens that require metabolic activation to exert genotoxic effects show an increase in potency when exposed to the low concentration S9 system. This is evidenced by +S9 CIs that reside in more potent regions of the comparative potency plot by several orders of magnitude compared to −S9 CIs. An outlier is 2-acetylaminofluorene where all

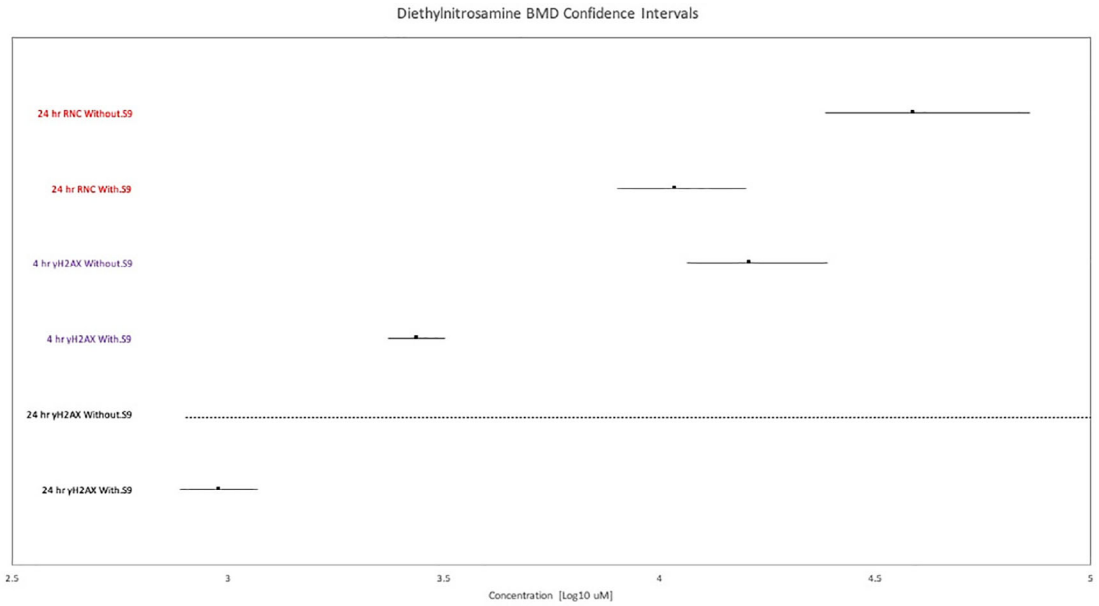

FIGURE 7 BMD confidence intervals for diethylnitrosamine. 24 hr γH2AX endpoint without S9 yielded a disproportionately high BMD with infinite BMDU: indicated as a dashed line spanning the width of the plot. BMD point estimates are displayed as data points

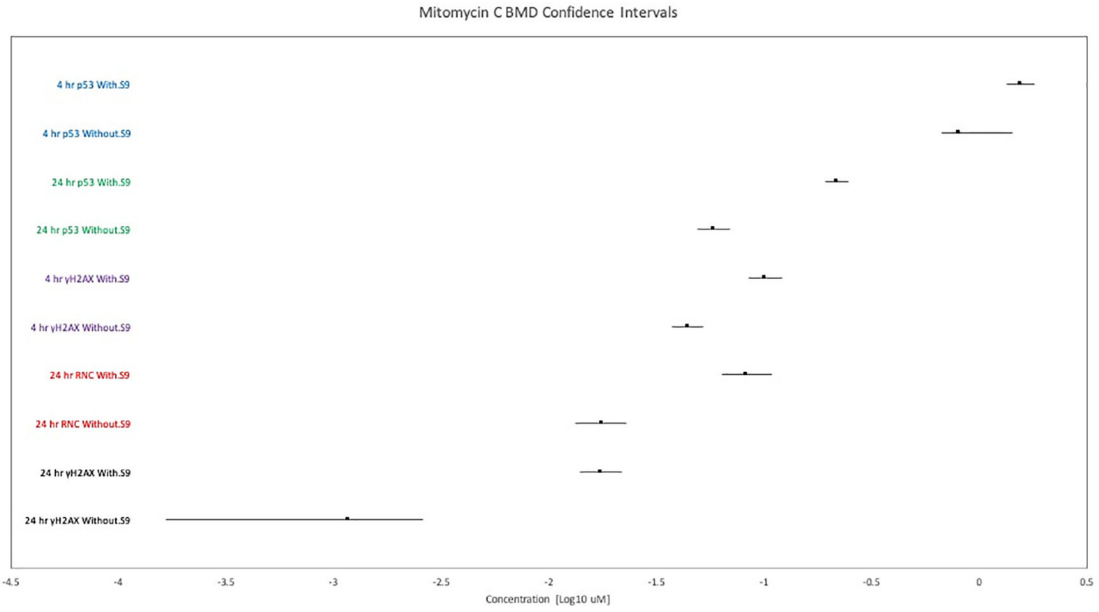

FIGURE 8 BMD confidence intervals for mitomycin C. Two-sided confidence intervals were obtained for all endpoint's BMD analyses. BMD point estimates are displayed as data points

endpoints except for 24 hr γH2AX and 24 hr RNC display CIs that reside in the same order of magnitude (Figure 1). This suggests that S9's potentiating effect was restricted to these endpoint/timepoint combinations. The comparative potency plots also show that the direct acting clastogens, mitomycin C and resorcinol, were not potentiated by S9—rather, to a certain extent, the presence of S9 slightly decreased their genotoxic potency.

3.2 | S9 potency ratio CIs

Tian et al. (2020) derived S9 potentiation ratios by comparative analysis of the BMD point estimates across the S9 exposure conditions for each compound and endpoint. There is a concern that the use of BMD point estimates can misrepresent the potency effect of S9 in this in vitro system since the BMD measure of uncertainty is disregarded.

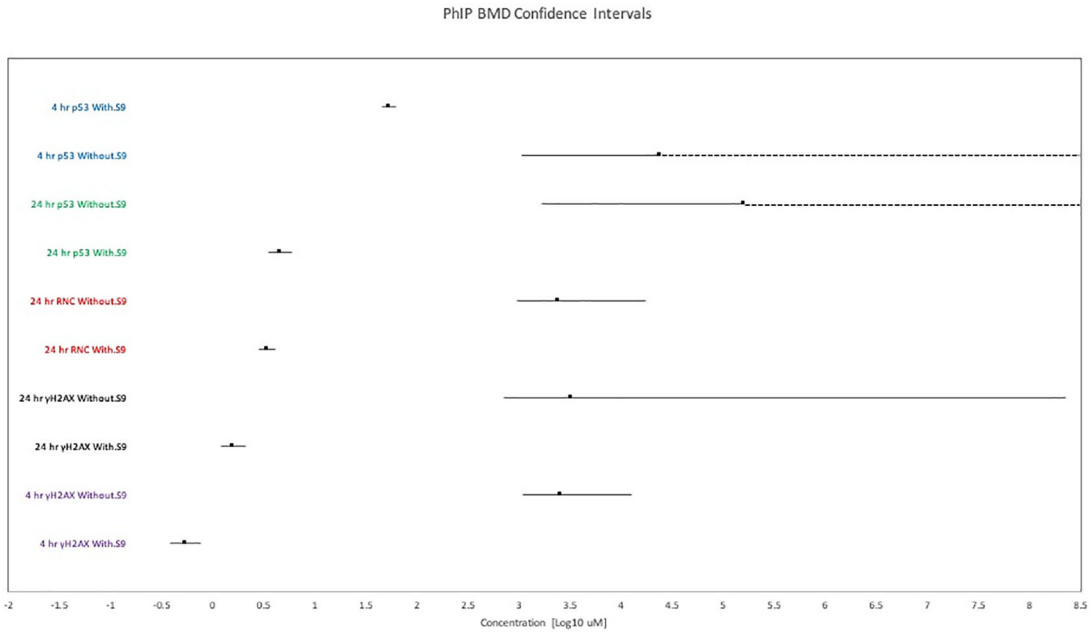

FIGURE 9 BMD confidence intervals for PhIP. 24 hr p53 endpoint without S9 and 4 hr p53 without S9 infinite BMDUs indicated with dashed positive direction lines. BMD point estimates are displayed as data points

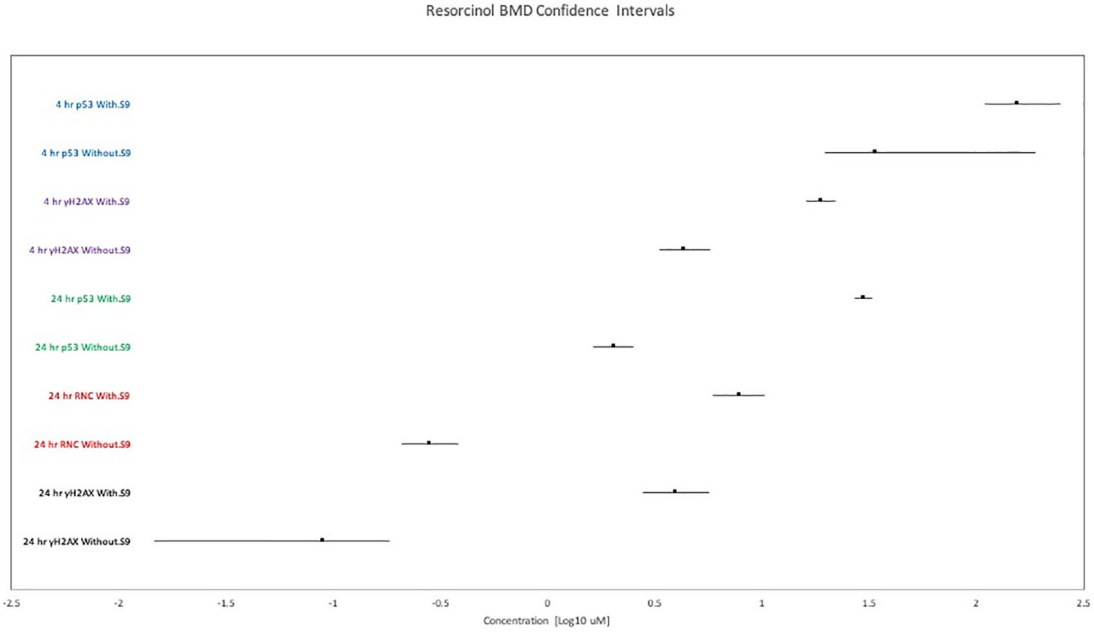

FIGURE 10 BMD confidence intervals for resorcinol. Two-sided confidence intervals were obtained for all endpoint's BMD analyses. BMD point estimates are displayed as data points

Tian et al. (2020) correctly pointed out that the BMD estimate can be beyond the top concentration used in the benchtop experiment. In these instances, the authors calculated the "S9 potentiation ratio" by restricting the BMD to the top concentration (denoted with a greater than [>] symbol in the tabulated results section of their article). However, these tight BMD restrictions led to mischaracterization of the true BMD ratio. We contend that one should only limit BMD values for graphical display where the value is disproportionately larger than other endpoints and S9 conditions for the same compound. In the BMD analysis reported in this re-analysis, disproportionately high BMD values also coincided with unbound BMDUs.

The BMD CIs obtained in this re-analysis were used to calculate an "S9 potency ratio CI" that is derived for each compound and endpoint BMD analysis that returned a dose response by comparing the

TABLE 1 Compound/endpoint combination BMD ratios and "S9 potency ratio CIs" (original and log scale) compared with the Tian et al. (2020) "S9 potentiation ratios"

Compound	Endpoint	S9 potentiation ratio from Tian et al. (2020) (μM)	BMD-BMD ratio[a]	S9 potency ratio CI range (μM)[b]	S9 potency ratio CI range (Log10 μM)[c]
2-Acetylaminofluorine	4 hr p53	NC	No DR	No DR	No DR
	24 hr p53	NC	1.89	0	0
	4 hr γH2AX	>2.6[d]	1.14	0	0
	24 hr γH2AX	>8.9	34.23	8.53–704.02	0.93 to 2.85
	24 hr RNC	0.53	0.33	0.18–0.55	−0.74 to −0.26
2-Aminoanthracene	4 hr p53	>7	No DR	No DR	No DR
	24 hr p53	>22	108.72	57.26–278.80	1.76 to 2.45
	4 hr γH2AX	167	74.55	46.03–126.81	1.66 to 2.10
	24 hr γH2AX	16.3	8.95	2.51–21.59	0.40 to 1.33
	24 hr RNC	>26.2	295.11	79.87–3,004.35	1.90 to 3.48
7,12-Dimethylbenzanthracene	4 hr p53	>32.9	49.90	16.63–418.60	1.22 to 2.62
	24 hr p53	23.7	18.06	9.52–34.87	0.98 to 1.54
	4 hr γH2AX	88.7	28.07	No DR	No DR
	24 hr γH2AX	NC	No DR	No DR	No DR
	24 hr RNC	21.8	24.43	8.59–73.37	0.93 to 1.87
Benzo[a]pyrene	4 hr p53	NC	3.06	1.54–7.23	0.19 to 0.86
	24 hr p53	>80	9.14	4.98–17.88	0.70 to 1.25
	4 hr γH2AX	>463	1,357.44	629.63–3,448.28	2.80 to 3.54
	24 hr γH2AX	1.8×10^6	11.18	3.47–34.73	0.54 to 1.54
	24 hr RNC	>31.4	No DR	No DR	No DR
Cyclophosphamide	4 hr p53	NC	No DR	No DR	No DR
	24 hr p53	870	702.02	535.64–923.65	2.93 to 2.97
	4 hr γH2AX	>101	158.03	115.69–227.54	2.06 to 2.36
	24 hr γH2AX	3,249	1,589.81	305.34–5,922.33	2.48 to 3.77
	24 hr RNC	797	963.44	466.40–2,405.30	2.67 to 3.38
Dibenzo[a,l]pyrene	4 hr p53	>56	148.74	88.94–316.87	1.95 to 2.50
	24 hr p53	>962	1,162.83	50.24–1,578.48	1.70 to 3.20
	4 hr γH2AX	22,727	4,666.876	24.49–6,967.37	1.39 to 3.84
	24 hr γH2AX	>38,462	22,167.08	6,325.46–130,672.27	3.80 to 5.12
	24 hr RNC	>1,220	No DR	No DR	No DR
Diethylnitrosamine	4 hr p53	NC	No DR	No DR	No DR
	24 hr p53	NC	No DR	No DR	No DR
	4 hr γH2AX	>3.5	5.90	3.64–10.47	0.56 to 1.02
	24 hr γH2AX	>10	No DR	No DR	No DR
	24 hr RNC	NC	3.58	1.53–9.10	0.18 to 0.96
Mitomycin C	4 hr p53	0.51	0.52	0.37–1.06	−0.43 to 0.03
	24 hr p53	0.73	0.27	0.20–0.36	−0.70 to −0.44
	4 hr γH2AX	0.82	0.44	0.31–0.62	−0.51 to −0.21
	24 hr γH2AX	0.25	0.07	0.01–0.19	−2.00 to −0.72
	24 hr RNC	0.31	0.21	0.12–0.36	−0.92 to −0.44
PhIP	4 hr p53	>3.1	No DR	No DR	No DR
	24 hr p53	>34.6	No DR	No DR	No DR
	4 hr γH2AX	>1,633	4,719.67	1,436.10–33,246.07	3.16 to 4.52
	24 hr γH2AX	>781	2067.12	336.19–189,075,630.25	2.53 to 8.28
	24 hr RNC	>18.2	703.43	230.83–6,142.75	2.36 to 3.79

(Continues)

TABLE 1 (Continued)

Compound	Endpoint	S9 potentiation ratio from Tian et al. (2020) (μM)	BMD-BMD ratio[a]	S9 potency ratio CI range (μM)[b]	S9 potency ratio CI range (Log10 μM)[c]
Resorcinol	4 hr p53	NC	0.22	0	0
	24 hr p53	0.11	0.07	0.0498–0.0937	−1.30 to −1.03
	4 hr γH2AX	0.26	0.23	0.1502–0.3594	−0.82 to −0.44
	24 hr γH2AX	0.05	0.03	0.0026–0.0659	−2.59 to −1.18
	24 hr RNC	0.047	0.04	0.0202–0.0645	−1.69 to −1.19

[a]The BMD-BMD ratio is presented to compare with the "S9 potentiation ratio" from Tian et al. (2020). In essence, the Tian et al. (2020) "S9 potentiation ratio" is a BMD-BMD ratio. The majority of BMD-BMD ratios from the re-analysis are similar to the Tian et al. (2020) "S9 potentiation ratios" showing that the BMD analyses do not differ significantly (same order of magnitude). One outlier is the analysis of the benzo[a]pyrene 24 hr γH2AX endpoint where the BMD-BMD ratio is significantly smaller than the "S9 potentiation ratio." This is likely due to increased precision resulting from combined covariate analysis of the clastogen compounds in the reanalysis of this endpoint. The significant difference between "S9 potentiation ratios" and the BMD-BMD ratios is the tight restrictions placed on the BMDs to calculate the "S9 potentiation ratio" denoted by the greater than symbol (>). This restriction resulted in mischaracterized comparison of potency.

[b]The S9 potency ratio CIs ranges are provided in the original scale to provide a like-for-like comparison with the Tian et al. (2020) "S9 potentiation ratio" presented in the original scale.

[c]The S9 potency ratio CI ranges are provided in the log scale to provide a like-for-like comparison with the comparative potency plots (Figures 1–10) presented in this article in the log scale.

[d]Greater than sign (>) from Tian et al. (2020) where tight restrictions were place on the BMD based on the upper concentration tested.

BMDU +S9 to BMDL −S9, to the BMDL +S9 to BMDU −S9. The "S9 potency ratio CI" is synonymous with the magnitude of the potency difference (range) exhibited for each compound/endpoint combination after exposure to S9. Although there were only 2 experimental conditions (+/− S9) included here, the same algorithm could apply to 2 or more experimental conditions where one potency is compared to another potency of interest. The "S9 potency ratio CIs" were calculated in the original scale so that the values can be compared to the "S9 potentiation ratios" values obtained from BMD-BMD ratios by Tian et al. (2020). Values greater than 1 represent increased compound potency for a particular endpoint after S9 exposure, whilst values less than 1 confer the inverse. 0 values represent CIs that overlap and hence potency differences are statistically indefensible. The BMD-BMD ratios from our analysis are also shown for comparative purposes. We have also displayed the S9 potency ratios in the Log scale for the sole purpose of aiding in demonstration of differences in orders of magnitude. The values are displayed in Table 1 in comparison with the S9 potentiation ratios from Tian et al. (2020).

Comparing our "S9 potency ratio CIs" with the S9 potentiation ratio obtained by Tian et al. (2020) shows that in almost all cases, the "S9 potentiation ratio" that was derived from BMD point estimates is either mischaracterized through tight restrictions on the BMD (where the researchers denoted these ratios with a > sign) or is in the upper range of our "S9 potency ratio CI" range, indicating a tendency for over estimation of the S9 bioactivation effect.

Calculating an "S9 potency ratio CI" for these experiments accurately conveys the uncertainty measurements which should be accounted for when assessing potency comparisons across conditions. We can summarize the "S9 potency ratio CIs" calculated here by stating that S9 increases the potency of the compounds ranging from approximately 1–2 orders of magnitude (Log10 μM) for lower potency differences, to approximately 3–5 orders of magnitude (Log10 μM) for higher potency differences. The effect of S9 is statistically insignificant for some compound/endpoint combinations where BMD CIs overlap. In other instances, S9 exposure decreased potency (expressed as negative values) of compounds by approximately 0.4 to 2.0 orders of magnitude.

3.3 | S9 potency ratio CI range

As previously mentioned, the "S9 potency ratio CI" values span several orders of magnitude. In order to objectively define the range of "S9 potency ratio CIs" obtained, we analyzed the lower and upper values (Log scale) of each "S9 potency ratio CI" via unsupervised hierarchical clustering. The clustering was based on squared Euclidean distance "Ward's method" (Ward and Hook 1963) between points. The resulting groups are presented in Figure 11 in the form of a two-dimensional dendrogram with accompanying heat map. The upper and lower values of the "S9 potency ratio CIs" associated with the in vitro biomarkers (4 hr p53, 24 hr p53, 4 hr γH2AX, 24 hr γH2AX, and 24 hr RNC) are displayed on the *X* axis. Compounds are plotted on the *Y* axis. The heat map represents the order of magnitude difference in "S9 potency ratio CIs." Compound/endpoint combinations where S9 exposure increased genotoxic potency are displayed with varying intensities of red on the heat map, whereas compounds and endpoints where S9 exposure decreased potency are displayed with varying intensities of blue on the heat map. Compound/endpoint combinations where the dose–response following S9 exposure was not statistically significantly different were plotted as zero values and displayed as gray-blue on the heatmap. Additionally, the clustering platform could not accommodate missing values for compounds/endpoint combinations that showed no dose–response. In these instances, a zero value was entered to accommodate the clustering method.

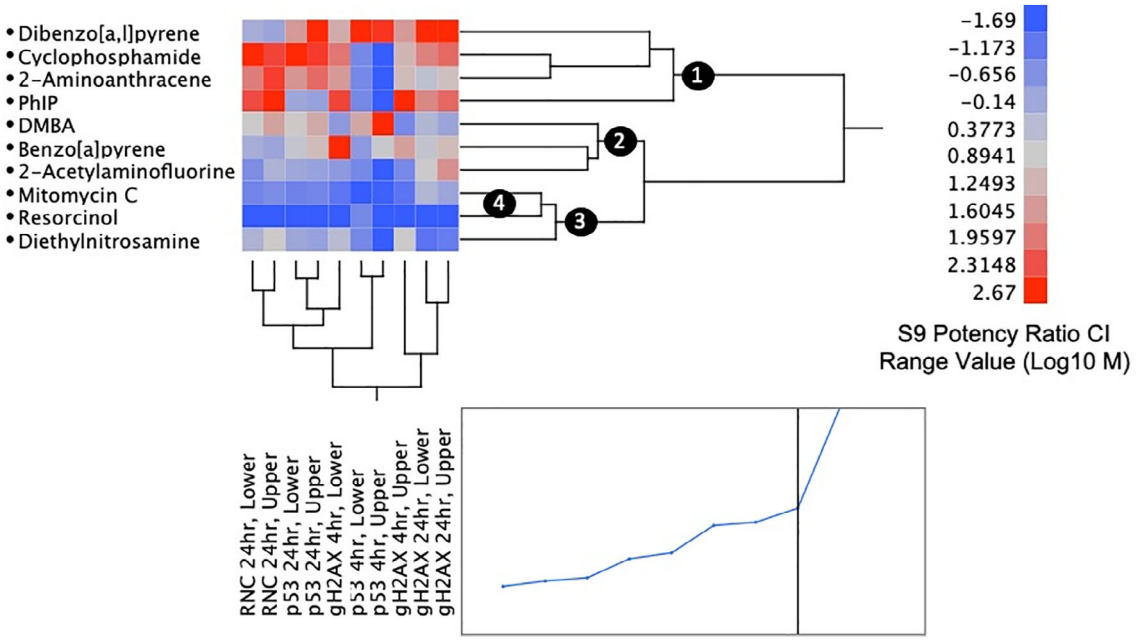

FIGURE 11 Unsupervised hierarchical clustering results are shown as a two-dimensional dendrogram with heatmap for the 10 clastogens "S9 potency ratio CIs." The values of the lower and upper confidence limits of the "S9 potency ratio CIs" were included in the analysis. The lower and upper bound of the "S9 potency ratio S9 CIs" are plotted on the *X*-axis per endpoint combination. Compounds are plotted on the *Y*-axis. Increasing intensities of red indicate a strong tendency for S9 exposure to increase the genotoxic potency of a compound for a specific endpoint. An increasing intensity of blue indicates the converse, where the presence of S9 decreases the genotoxic potency of a compound for a specific endpoint. Gray-blue represents compound/endpoint combinations where the dose–response following S9 exposure was not statistically significantly different (overlapping CIs), or where a zero value was included to accommodate the clustering method in instances where no S9 potency ratio CIs were obtained (infinite BMDUs). There are 4 distinct clades that group compounds into (1) high, (2) low, (3) zero, and (4) negative (as a subset of clade 3), effects on potency as a result of S9 exposure. Abbreviation: DMBA = 7,12-dimethylbenzanthracene

There are 4 distinct clades identified in the dendrogram: (a) Chemicals whose genotoxic potency was *dramatically* increased in the presence of S9 (dibenzo[a,l]pyrene, cyclophosphamide, 2-amino anthracene, and PhIP); (b) chemicals whose genotoxic potency was increased in the presence of S9 (dimethylbenzanthracene, benzo[a] pyrene, and 2-acetylaminofluorene); (c) chemicals whose genotoxic potency was not increased in the presence of S9 (mitomycin c, resorcinol, and diethylnitrosamine); and (d) a subset of clade 3, clade 4, whose genotoxic potency was reduced in the presence of S9 (mitomycin c, and resorcinol; indicated by a strong dark blue in the heat map).

The clustering results demonstrate that distinct groups exist that best describe the range of S9-dependent effects into high, low, zero and negative categories. The division of the S9 potency ratio CIs into groups follows application of quantitative metrics allowing objective interpretation in a hazard-based scale context. It is likely that when a more diverse set of clastogens are analyzed, it will be possible to detect further subdivisions of potency effect (e.g., low, medium, high, very high). The same clustering method could be applied to other dose–response comparative analyses. Several examples are envisioned, for instance the study of compound potency effects between different exposure durations (subchronic vs. chronic dosing regimens), between different target tissues (liver vs. bone marrow), or between different in vitro cell lines (liver HepaRG vs. HepG2, or V79 vs. CHO).

4 | CONCLUSIONS

The results of the data analysis presented here illustrate the necessity to utilize the full BMD CIs to draw conclusions from dose–response datasets. We have demonstrated that in comparative potency analysis, use of BMD point estimates can yield mischaracterized or overestimated potency ratios during instances where the width of the CIs varies considerably between conditions.

Interpretation of BMD CIs relies upon visual scrutiny by the evaluating scientist, and we speculate that one could be overwhelmed with the results from hundreds of compounds from in vitro screening experiments. Therefore, the evaluating scientist may wish to use statistical methods such as hierarchical clustering to aid in data interpretation. When objectively considering the range of "S9 potency ratio CIs" presented in our analysis, one is provided with a scale of high to low, zero, and negative S9 potency effects. While we illustrated the potency effect of S9 in an in vitro genotoxicity test system, the same approach can be applied to any comparative genotoxicity potency analysis. We propose that investigators apply the following stepwise approach when performing comparative potency analysis of 2 or more experimental conditions using the BMD methodology:

1. Utilize the results of combined covariate BMD analyses to plot BMD CIs that represent potency across the endpoints evaluated

under different dependent variables. For graphical purposes only, limit disproportionately high BMDs that display infinite BMDU values.

2. Derive the lower and upper values of "potency ratio CIs" between conditions by comparing BMDU condition 1 to BMDL condition 2, and BMDL condition 1 to BMDU condition 2 (ad infinitum), respectively.
3. Objectively identify groups in the data that best describe the magnitude of the difference in potency observed between conditions. While a scientist can successfully identify patterns in CI plots of a few experimental conditions (a small number of compounds, animal sex, limited number of cell lines, etc.) via visual techniques, we contend that hierarchical clustering may add significant value to the interpretation of large datasets, particularly those from in vitro screening experiments of large numbers of compounds and biomarker combinations.

ACKNOWLEDGMENTS

R. P. W. is an employee of Baxter Healthcare and was in receipt of PhD tuition fee funding from his employer at the time this manuscript was prepared. Benchtop work performed at Litron Laboratories was supported in part by a grant from the National Institute of Health/National Institute of Environmental Health Sciences (NIEHS; grant no. R44ES029014). The contents are solely the responsibility of the authors, and do not necessarily represent the official views of the NIEHS.

CONFLICT OF INTEREST

S. D. D., S. M. B., and J. C. B. are employed by Litron Laboratories where the benchtop work was conducted. Litron owns a patent covering the flow-cytometry based assay described in this manuscript and sells a commercial kit based on these procedures: MultiFlow® DNA Damage kit-p53, γH2AX, Phospho-Histone H3.

AUTHOR CONTRIBUTIONS

R. P. W. is the lead author of the manuscript and identified the issues with the S9 potentiation ratio published in Tian et al. (2020). R. P. W. performed the BMD analysis reported here as well as leadership on the results analysis and interpretation. S. D. D's laboratory provided the raw data. R. P. W. suggested the use of unsupervised hierarchical clustering to evaluate the S9 potency ratio CIs, and S. D. D. achieved this. R. P. W. provided the first draft of the manuscript, and all authors contributed to the revisions that followed.

ORCID

Ryan P. Wheeldon https://orcid.org/0000-0002-4893-246X
George E. Johnson https://orcid.org/0000-0001-5643-9942

REFERENCES

Allemang, A., Mahony, C., Lester, C. and Pfuhler, S. (2018) Relative potency of fifteen pyrrolizidine alkaloids to induce DNA damage as measured by micronucleus induction in HepaRG human liver cells. *Food and Chemical Toxicology*, 121, 72–81.

Bemis, J.C., Wills, J.W., Bryce, S.M., Torous, D.K., Dertinger, S.D. and Slob, W. (2016) Comparison of in vitro and in vivo clastogenic potency based on benchmark dose analysis of flow cytometric micronucleus data. *Mutagenesis*, 31, 277–285.

Bryce, S.M., Bernacki, D.T., Bemis, J.C., Spellman, R.A., Engel, M.E., Schuler, M., Lorge, E., Heikkinen, P.T., Hemmann, U., Dertinger, S.D., et al. (2017) Interlaboratory evaluation of a multiplexed high information content in vitro genotoxicity assay. *Environmental and Molecular Mutagenesis*, 58, 146–161.

Bryce, S.M., Bernacki, D.T., Smith-Roe, S.L., Witt, K.L., Bemis, J.C. and Dertinger, S.D. (2018) Investigating the generalizability of the MultiFlow® DNA damage assay and several companion machine learning models with a set of 103 diverse test chemicals. *Toxicological Sciences*, 162, 146–166.

Bryce, S.M., Bernacki, T., Bemis, J.C. and Dertinger, S.D. (2016) Genotoxic mode of action predictions from a multiplexed flow cytometric assay and a machine learning approach. *Environmental and Molecular Mutagenesis*, 57, 171–189.

Crump, K.S. (1984) A new method for determining allowable daily Intakes1. *Toxicological Sciences*, 4, 854–871.

Dearfield, K.L., Gollapudi, B.B., Bemis, J.C., Daniel Benz, R., Douglas, G.R., Elespuru, R.K., Johnson, G.E., Kirkland, D.K., LeBaron, M.J., Luijten, M., et al. (2017) Next generation testing strategy for assessment of genomic damage: a conceptual framework and considerations. *Environmental and Molecular Mutagenesis*, 58, 264–283.

Dertinger, S.D., Kraynak, A.R., Wheeldon, R.P., Bernacki, D.T., Bryce, S. M., Hall, N., Bemis, J.C., Galloway, S.M., Escobar, P.A. and Johnson, G.E. (2019) Predictions of genotoxic potential, mode of action, molecular targets, and potency via a tiered Multiflow® assay data analysis strategy. *Environmental and Molecular Mutagenesis*, 60, 513–533.

European Food Safety Agency (EFSA). (2009) European food safety authority. Guidance of the scientific committee on use of the benchmark dose approach in risk assessment. *EFSA Journal*, 1150, 1–72.

European Food Safety Agency (EFSA). (2017) Update: use of the benchmark dose approach in risk assessment. *EFSA Journal*, e04658, 15.

Gollapudi, B.B., Johnson, G.E., Hernandez, L.G., Pottenger, L.H., Dearfield, K.L., Jeffrey, A.M., Julien, E., Kim, J.H., Lovell, D.P., Thybaud, V., et al. (2013) Quantitative approaches for assessing dose–response relationships in genetic toxicology studies. *Environmental and Molecular Mutagenesis*, 54, 8–18.

Hernández, L.G., Slob, W., van Steeg, H. and van Benthem, J. (2011) Can carcinogenic potency be predicted from in vivo genotoxicity data? A meta-analysis of historical data. *Environmental and Molecular Mutagenesis*, 52, 518–528.

Macgregor, J.T., Frötschl, R., White, P.A., Crump, K.S., Eastmond, D.A., Fukushima, S., Guérard, M., Hayashi, M., Soeteman-Hernández, L.G. and Thybaud, V. (2015) IWGT report on quantitative approaches to genotoxicity risk assessment I. Methods and metrics for defining exposure–response relationships and points of departure (PoDs). *Mutation Research – Genetic Toxicology and Environmental Mutagenesis*, 783, 55–65.

Slob, W. (2002) Dose-response modeling of continuous endpoints. *Toxicological Sciences*, 66, 298–312.

Slob, W. and Setzer, R. (2014) Shape and steepness of toxicological dose-response relationships of continuous endpoints. *Critical Reviews in Toxicology*, 44, 270–297.

Soeteman-Hernández, L.G., Fellows, M.D., Johnson, G.E. and Slob, W. (2015a) Correlation of in vivo versus in vitro benchmark doses (BMDs) derived from micronucleus test data: a proof of concept study. *Toxicological Sciences*, 148, 355–367.

Soeteman-Hernández, L.G., Johnson, G.E. and Slob, W. (2015b) Estimating the carcinogenic potency of chemicals from the in vivo micronucleus test. *Mutagenesis*, 31, 347–358.

Tian, S., Cyr, A., Zeise, K., Bryce, S.M., Hall, N., Bemis, J.C. and Dertinger, S.D. (2020) 3Rs-friendly approach to exogenous metabolic activation that supports high-throughput genetic toxicology testing. *Environmental and Molecular Mutagenesis*, 61, 408–432.

Ward, J.H. and Hook, M.E. (1963) Application of an hierarchical grouping procedure to a problem of grouping profiles. *Educational and Psychological Measurement*, 23, 69–81.

Wheeldon, R.P., Bernacki, D.T., Dertinger, S.D., Bryce, S.M., Bemis, J.C. and Johnson, G.E. (2020) Benchmark dose analysis of DNA damage biomarker responses provides compound potency and adverse outcome pathway information for the topoisomerase II inhibitor class of compounds. *Environmental and Molecular Mutagenesis*, 61, 396–407.

White, P.A., Long, A.S. and Johnson, G.E. (2020) Quantitative interpretation of genetic toxicity dose-response data for risk assessment and regulatory decision-making: current status and emerging priorities. *Environmental and Molecular Mutagenesis*, 61, 66–83.

Wills, J.W., Johnson, G.E., Battaion, H.L., Slob, W. and White, P.A. (2017) Comparing BMD-derived genotoxic potency estimations across variants of the transgenic rodent gene mutation assay. *Environmental and Molecular Mutagenesis*, 58, 632–643.

Wills, J.W., Johnson, G.R., Doak, S.H., Soeteman-Hernández, L.G., Slob, W. and White, P.A. (2016) Empirical analysis of BMD metrics in genetic toxicology part I: in vitro analyses to provide robust potency rankings and support MOA determinations. *Mutagenesis*, 31, 255–263.

How to cite this article: Wheeldon RP, Dertinger SD, Bryce SM, Bemis JC, Johnson GE. The use of benchmark dose uncertainty measurements for robust comparative potency analyses. *Environ Mol Mutagen*. 2021;1–13. https://doi.org/10.1002/em.22422