MEMORANDUM

To:	Judy Facey, EPA; Timothy McMahon, EPA; Alexander Kliminsky, EPA; Deborah Burgin, EPA
From:	Jonathan Cohen, ICF; Rachel ONeal, ICF
Date:	September 5, 2022
Re:	Statistical Review of the Anderson and Molhave and Kulle et al Formaldehyde Inhalation Exposure Studies.

1. Introduction and Summary

This memorandum describes our statistical review of the Anderson and Molhave (1978, 1983) and Kulle et al (1987, 1993) formaldehyde inhalation exposure studies.

Anderson (1979) and Anderson and Molhave (1983) measured the effects of inhaled formaldehyde vapor on nasal mucociliary flow, nasal airflow resistance, forced expiratory vital capacity, and irritation threshold on sixteen human subjects. They reported a significant decrease in mucus flow rate at 0.3 mg/m^3 formaldehyde. They did not find statistically significant changes in nasal airflow resistance or forced expiratory vital capacity. They reported significant increases in the odor threshold for ethyl valeriate. Since the raw data were not provided, we were unable to verify those statistical analyses.

Anderson (1979) and Anderson and Molhave (1983) also reported the numbers of subjects at each dose that experienced discomfort at 2.5 or 5 hours or reported sensory irritation (conjunctive irritation and dryness of the nose and throat). Although the authors did not report a statistical analysis of those discomfort and sensory irritation rates, we used the available data to analyze those results and also compared our statistical dose-response and Benchmark Dose (BMD) analyses of the sensory irritation rates with the analyses presented in the IRIS report (EPA, 2022). We analyzed the discomfort and sensory irritation rates using Fisher exact tests of the differences in response rates at different dose levels and using Cochran-Armitage tests for trends in the response rates. These tests assume statistical independence between results at different doses; this is an assumption which may not be valid due to the fact that the same subjects were tested at multiple doses, but cannot be evaluated in the absence of the raw data for each subject. For discomfort at 2.5 hours and sensory irritation, we found statistically significant differences and trends in the response rates at the 5% level. For discomfort at 5 hours, we found statistically significant differences and trends in the response rates at the 10% level but not at the 5% level. We then used the current version of Benchmark Dose Software (BMDS Version 3.3rc10) to fit

and plot dose-response models, and to estimate the BMD as the dose at which there was a 10% extra risk above an assumed 0% risk for unexposed subjects. We also estimated the BMDL, defined as a one-sided 95% lower confidence limit for the BMD. The BMDL values were 0.151 mg/m^3 for discomfort at 2.5 hours, 0.250 mg/m^3 for discomfort at 5 hours, and 0.091 mg/m^3 for sensory irritation.

For Anderson (1979) and Anderson and Molhave (1983) we compared our BMD results for sensory irritation using BMDS Version 3.3rc10 with the results in the IRIS report (EPA, 2022) that used the earlier BMDS Version 2.2. In EPA (2022) they assumed either 0 or 3 observed cases at dose 0, even though Anderson (1979) and Anderson and Molhave (1983) did not report collecting observations at that dose. Our BMD, BMDL, and Akaike Information Criterion (AIC) values using BMDS Version 3.3rc10 agree (within 0.001) with the values reported in EPA (2022), although the reported p-values were very different. Since we were unable to obtain documentation or software for BMDS Version 2.2, we cannot explain the differences in the p-values. The BMDL values were 0.157 mg/m^3 for sensory irritation assuming 0 cases at dose 0, 0.219 mg/m^3 for sensory irritation assuming 3 cases at dose 0 using the log-logistic model selected in BMDS Version 2.2, and 0.312 mg/m^3 for sensory irritation assuming 3 cases at dose 0 using the Dichotomous Hill model selected in BMDS Version 3.3rc10 (not reported in Version 2.2).

Kulle et al (1987) and Kulle (1993) measured the effects of inhaled formaldehyde vapor on symptoms of odor sensation, nose/throat irritation, eye irritation, chest discomfort, cough, and headache, and on nasal resistance and pulmonary function. They collected data on 19 subjects at doses of 0, 1.and 2 ppm formaldehyde, 10 of the same subjects at a dose of 0.5 ppm formaldehyde, and 9 of the same subjects at a dose of 3 ppm formaldehyde. They reported mean symptom score differences (between the scores before and after exposure) and found significant linear trends for odor sensation and eye irritation. They also reported increases in nasal resistance at 2 ppm (not significant) and 3 ppm (significant), but no statistically significant effects on pulmonary function. Since the raw data were not provided, we were unable to verify those statistical analyses.

Kulle (1993) also used the data to analyze effects of inhaled formaldehyde vapor on the rates of symptoms of odor sensation, nose/throat irritation, and eye irritation at each dose. Kulle (1993) used McNemar tests and found significant differences in rates between clean air and formaldehyde exposures at 0.5, 2, or 3 ppm for odor sensation, and at 2 or 3 ppm for eye irritation. Since the raw data for each subject were not provided, we were unable to verify those statistical analyses. We used the available data to analyze the summarized symptom rates and also compared our statistical dose-response and Benchmark Dose (BMD) analyses of the odor sensation and sensory irritation rates with the analyses presented in the IRIS report (EPA, 2022). We analyzed the odor sensation and sensory irritation rates using Fisher exact tests of the differences in response rates at different dose levels and using Cochran-Armitage tests for trends in the response rates. These tests assume statistical independence between results at different doses; this is an assumption which may not be valid due to the fact that the same subjects were tested at multiple doses, but cannot be evaluated in the absence of the raw data for each subject. For odor sensation and eye irritation, we found statistically significant differences and trends in the response rates at the 5% level. For nose/throat irritation, we found no

statistically significant differences in the response rates at the 10% level, and statistically significant trends in the response rates at the 10% level. We then used the current version of Benchmark Dose Software (BMDS Version 3.3rc10) to fit and plot dose-response models, and to estimate the BMD as the dose at which there was a 10% extra risk above an assumed 0% risk for unexposed subjects. We also estimated the BMDL, defined as a one-sided 95% lower confidence limit for the BMD. The BMDL values were 0.182 ppm (0.224 mg/m^3) for odor sensation, 0.502 ppm (0.617 mg/m^3) for eye irritation, and 0.992 (1.220 mg/m^3) for nose/throat irritation.

For Kulle (1993) we compared our BMD results using BMDS Version 3.3rc10 for eye irritation with the results in the IRIS report (EPA, 2022) that used the earlier BMDS Version 2.2. Our BMD, BMDL, and Akaike Information Criterion (AIC) values using BMDS Version 3.3rc10 agree (within 0.001) with the values reported in EPA (2022), although the reported p-values were very different. Since we were unable to obtain documentation or software for BMDS Version 2.2, we cannot explain the differences in the p-values.

Some of the following text summarizing the Anderson and Molhave and Kulle et al papers was borrowed from EPA's draft DERs for these studies.

The attached files formdata.09522.xlsx, andersonkulle.sascode.090522.sas and andersonkulle.sascode.090522.lst contain the SAS code and listing file used for the Excel input file, Fisher exact tests, Cochran-Armitage tests, and sign test reported in this memorandum. The eight Excel output files containing the detailed BMDS Version 3.3rc10 dose-response analyses summarized in this memorandum are also attached.

2. Anderson (1979) and Anderson and Molhave (1983)

This study was described in two papers. Preliminary results were reported in Anderson (1979) and more detailed results were reported in a follow-on paper Anderson and Molhave (1983). The intent of this study was to determine the effect of inhaled formaldehyde vapor on nasal mucociliary flow, nasal airflow resistance, forced expiratory vital capacity, and irritation thresholds.

Sixteen human subjects (5 male, 11 female, age range 20-33 years; 5 smokers) were examined in groups of four. Each group underwent four different exposures on four consecutive days. Air concentrations of 0, 0.3, 0.5, 1.0, and 2.0 mg/m^3 formaldehyde were used in this study. Baseline measurements of nasal mucociliary flow, nasal airflow resistance, forced expiratory vital capacity, and odor threshold were made during a control period in which subjects were exposed to clean filtered air from outside for 2 hours. Following the control measurements, formaldehyde was added to the air, and after 'about one hour' a steady state concentration of formaldehyde was achieved and maintained for the duration of the experiment that day. Parameter measurements were then made after 2-3 hours of exposure to formaldehyde, and again after 4-5 hours of exposure to formaldehyde.

Formaldehyde concentration in the chamber air was measured by collection of 1h air samples and analysis by the chromotropic acid method. The variation was within ±20 percent from the target values.

During the exposures, subjects were also asked to perform tasks involving addition, multiplication, and card punching (whatever that is) of 15 minutes' duration. The addition test was performed during each exposure period, the multiplication test during the first exposure period only, and the card punching test in the control period and second exposure period.

RESULTS

Nasal Mucus Flow

Results of inhalation exposure to formaldehyde on nasal mucus flow are shown in the following Figure 1, copied from Figure 4 of the publication. As shown in the figure, for the anterior sections of the nasal cavity, inhalation exposure to formaldehyde vapor at 0.3 and 0.5 mg/m^2 resulted in a decrease in nasal mucus flow. Above the 0.5 mg/m^2 concentration, there was no further effect of formaldehyde on mucus flow. The posterior sections of the nasal cavity (slits 4-5 and 5-6 in the figure) were not significantly affected by formaldehyde inhalation exposure.

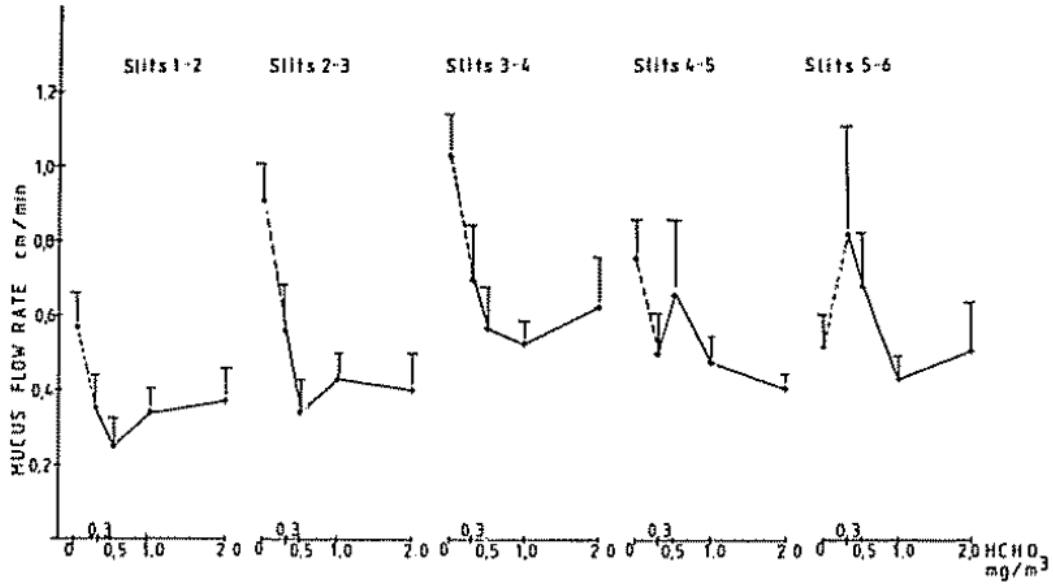

Figure 4 The average mucus flow rate in clean air (0) and after 4-5 h exposure to formaldehyde at four concentrations. Slits 1-3, 3-4, and 4-6 represent the anterior, middle, and posterior thirds of the ciliated part of the nose, respectively. One standard deviation of the mean is shown as a vertical bar.

Figure 1. Figure 4 from page 160 of Anderson and Molhave 1983. Mucus Flow Rates.

The statistical analyses described in the two papers consisted of non-parametric tests followed by an analysis of variance. Since the graphical results only show averages and the raw data were not provided, it is not possible to validate the statistical analyses.

Airway Resistance

The results of airway resistance experiments are shown in the following Figure 2, copied from Figure 6 from the publication. As noted in this figure, no significant changes in airway resistance measurements were observed at any concentration of formaldehyde tested in this study.

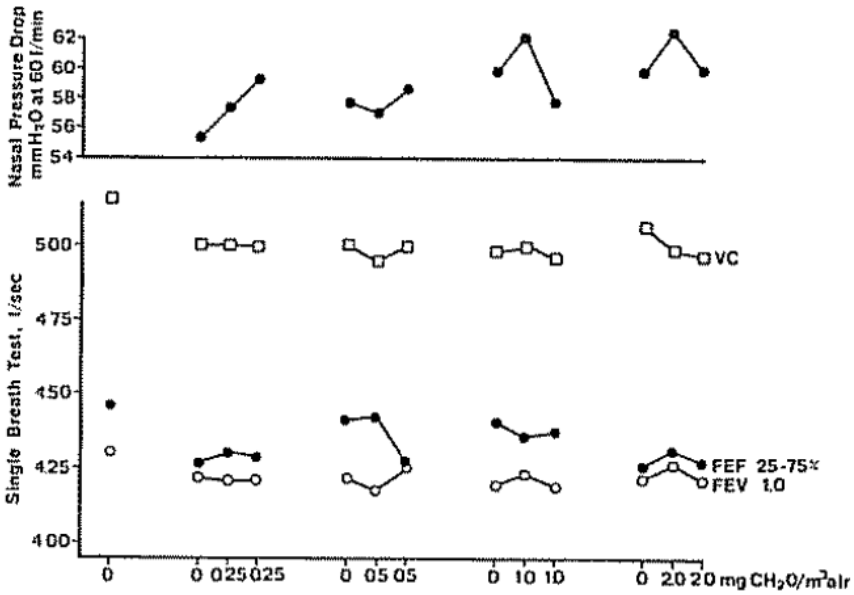

Figure 6 The variation with time of the averages of nasal pressure drop (upper part), of vital capacity (VC), forced expiratory flow ($FEF_{25-75\%}$), and forced expiratory volume during the first second of the expiration ($FEV_{1.0}$) (lower part). The interval between the measurements each day is approximately 2 h.

Figure 2. Figure 6 from page 161 of Anderson and Molhave 1983. Airway Resistance.

Statistical analyses described in the two papers consisted of non-parametric tests followed by an analysis of variance. Since the graphical results only show averages and the raw data were not provided, it is not possible to validate the statistical analyses.

Odor Threshold

The results of odor threshold experiments were reported to show a significant increase in the odor threshold for ethyl valeriate after 1-3 and 4-5 hours of exposure at 2 mg/cubic meter, but no significant changes in the odor threshold were found at lower concentrations.

The statistical analyses described in the two papers consisted of non-parametric tests followed by an analysis of variance. Since the raw data were not provided, it is not possible to validate the statistical analyses.

Discomfort/Irritation

Results of assessment of subjects for feelings of discomfort during formaldehyde inhalation exposures are shown in the following Figure 3, copied from Figure 7 from the publication. As noted in the figure, for the first 2 hours of inhalation exposure, there was no increase in subjective feelings of discomfort at the 0.3 and 0.5 mg/m^3 concentrations of formaldehyde. At the 1.0 and 2.0 mg/m^3 concentrations, discomfort was already reported during the first hour of exposure. After 2.5 hours of exposure, discomfort was reported in 3, 2, 7, and 10 subjects at the 0.3, 0.5, 1.0, and 2.0 mg/m^3 concentrations respectively. After 5 hours of exposure, discomfort was reported in 9, 3, 6, and 10 subjects at the 0.3, 0.5, 1.0, and 2.0 mg/m^3 concentrations respectively.

Following exposures, the subjects were asked to describe the symptoms experienced. Subjects reported the symptoms experienced as mainly conjunctival irritation and dryness of the nose and throat. The incidence of these sensory irritation symptoms reported was 3, 5, 15, and 15 subjects at the 0.3, 0.5, 1.0, and 2.0 mg/m^3 concentrations respectively.

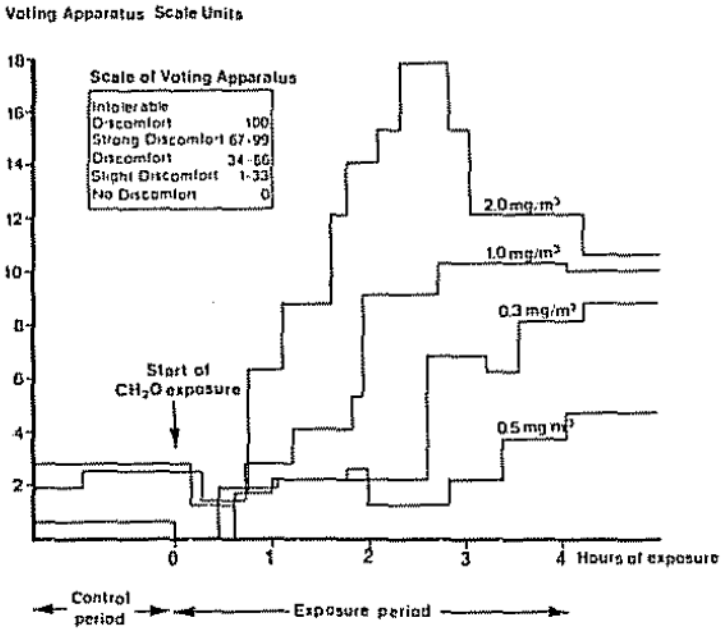

Figure 7 Variation with time of the mean discomfort vote in scale units at four different concentrations of formaldehyde In the control period, clean air was supplied to the subjects.

Figure 3. Figure 7 from page 162 of Anderson and Molhave 1983. Trends in Mean Discomfort Votes.

A statistical analysis of the discomfort and sensory irritation rates was not reported in the two papers. However, we used the available data for the following statistical analyses of the reported rates.

McNemar Test

A McNemar statistical test is often used to compare results of pairs of experiments carried out on a group of subjects, where for each subject the result of each experiment is either a positive result or a

negative result. The statistical test is based on the numbers of cases where the results of the two experiments disagree. If raw data were available for each individual subject, this test could have been applied to test whether the response rates at a pair of doses are significantly different. In the absence of the raw data, such a test cannot be carried out.

Fisher Exact Test

The results of the discomfort and sensory irritation experiments can be summarized in the following Table 1.

Table 1. Discomfort and Sensory Irritation Results from Anderson and Molhave (1983)

Dose mg/m^3	Discomfort at 2.5 hours Cases / Subjects 4 positive doses	Discomfort at 5 hours Cases / Subjects 4 positive doses	Sensory Irritation Cases / Subjects 4 positive doses	Sensory Irritation Cases / Subjects 4 positive doses Assume 0 cases at 0 mg/m^3 (unobserved)	Sensory Irritation Cases / Subjects 4 positive doses Assume 3 cases at 0 mg/m^3 (unobserved)
0	No data	No data	No data	0/16	3/16
0.3	3/16	9/16	3/16	3/16	3/16
0.5	2/16	3/16	5/16	5/16	5/16
1	7/16	6/16	15/16	15/16	15/16
2	10/16	10/16	15/16	15/16	15/16

The results in columns 2, 3, and 4 are the actual observed results. The results in columns 5 and 6 include hypothetical, but unobserved, results that at a dose of 0, there were either 0 or 3 cases. Those hypothetical results were assumed in the IRIS report (EPA, 2022) discussed below, although the IRIS report describes them as “Assumed response among controls = 0” or “Assumed response among controls = 3%”. We believe these descriptions in IRIS are incorrect. This is based on the fact that our calculated BMD values based on numbers of observed cases out of 16 at a dose of 0 match the IRIS BMD values that were reportedly based on percentages of cases. See the subsection on the dose/response and BMD modeling below for more details.

The Fisher exact test tests whether the response rates at different doses are equal. Application of this statistical test requires the assumption that the observed responses at different doses are statistically independent. It is questionable whether the independence assumption is valid since each subject was tested at multiple doses and it is very possible that their responses at different doses are correlated. However, in the absence of raw dose-specific data on each subject, the independence assumption cannot be statistically evaluated. The alternative to the independence assumption is clustering, such that there is dependence between repeated measures on the same subject.

The p-values for the Fisher exact tests are shown in the following Table 2. P-values of 5% or below can be treated as evidence that the response rates are different at different doses, under the independence assumption.

Table 2. Fisher Exact Tests for Discomfort and Sensory Irritation Results from Anderson and Molhave (1983)

Response	Discomfort at 2.5 hours Cases / Subjects 4 positive doses	Discomfort at 5 hours Cases / Subjects 4 positive doses	Sensory Irritation Cases / Subjects 4 positive doses	Sensory Irritation Cases / Subjects 4 positive doses Assume 0 cases at 0 mg/m^3 (unobserved)	Sensory Irritation Cases / Subjects 4 positive doses Assume 3 cases at 0 mg/m^3 (unobserved)
P-value	0.0117	0.0542	< 0.0001	< 0.0001	< 0.0001

The results of the Fisher exact tests show significantly different response rates at different doses at the 5% significance level for discomfort at 2.5 hours but not at 5 hours, and for sensory irritation using 4 positive doses with or without the assumed observed response rates at a dose of 0. However, the p-value for discomfort at 5 hours is 0.0542, which is almost significant at the 5% level.

Cochran-Armitage Trend Test

The Cochran-Armitage trend test also assumes independence of observed responses at difference doses. Under that assumption, this is a statistical test of the null hypothesis that the response rates are the same at every dose against the alternative one-sided hypothesis that the response rates increase with the dose. (It is unrealistic to assume that the true response rates would be lower at a higher formaldehyde dose.)

The p-values for the Cochran-Armitage trend tests using the data in Table 1 are shown in the following Table 3. P-values of 5% or below can be treated as evidence that the response rates increase with the dose, under the independence assumption. Two versions of the statistical tests are displayed. The asymptotic test uses a normal approximation of the test statistic to compute the p-value. The Monte Carlo version of the test displayed here used 10,000 Monte Carlo simulations of the test statistic to estimate a 99% upper confidence limit for the p-value.

Table 3. Cochran-Armitage Trend Tests for Discomfort and Sensory Irritation Results from Anderson and Molhave (1983)

Response	Discomfort at 2.5 hours Cases / Subjects 4 positive doses	Discomfort at 5 hours Cases / Subjects 4 positive doses	Sensory Irritation Cases / Subjects 4 positive doses	Sensory Irritation Cases / Subjects 4 positive doses Assume 0 cases at 0 mg/m^3 (unobserved)	Sensory Irritation Cases / Subjects 4 positive doses Assume 3 cases at 0 mg/m^3 (unobserved)
Asymptotic P-value (one-sided)	0.0007	0.0839	< 0.0001	< 0.0001	< 0.0001
P-value (one-sided) Monte Carlo 99% Upper Confidence Limit	0.0014	0.0975	< 0.0001	0.0005	0.0005

The results of the Cochran-Armitage trend tests show significantly increasing response rates at different doses at the 5% significance level for discomfort at 2.5 hours but not at 5 hours, and for sensory irritation using 4 positive doses with or without the assumed observed response rates at a dose of 0. The p-value for discomfort at 5 hours is 0.098, which is not significant at the 5% level but is significant at the 10% level.

Benchmark Dose Modeling Results Assuming Independence

The results of the discomfort and sensory irritation experiments shown in Table 1 were also analyzed by fitting statistical models for the probability of a response as a function of the dose. A variety of statistical models are fitted to the data and the best-fitting statistical model is selected. For these analyses we followed EPA's Benchmark Dose Modeling (BMD) method using BMD Software (BMDS) and, as in the IRIS draft report on formaldehyde (EPA, 2022), we chose the best model as the one with the lowest Akaike Information Criterion (AIC) statistic. (Note that the BMDS by default uses a different approach to select their recommended model.) The AIC is based on the log-likelihood of the fitted model, with an adjustment for the number of fitted parameters in the model, so that models with a large number of fitted parameters are penalized. The BMD software again assumes that repeated measures on the same subject are independent. The IRIS approach chosen to account for possible dependence is discussed below in the section "Benchmark Dose Modeling Results Adjusted for Dependence". We will refer to the Benchmark Dose approach here as BMD (for "Benchmark Dose"), since that is the name of the EPA software used, although the same approach is often called BMC (for "Benchmark Concentration") with the same statistical interpretation.

For the BMD modeling we assumed a Benchmark Response (BMR) of 10% extra risk above an assumed 0% risk for unexposed subjects (the control group). The BMD is defined as the estimated dose at which the response rate is the BMR. The BMDL is the benchmark dose lower confidence limit, defined as a one-sided 95% lower confidence limit for the BMD. The BMDL is often used by EPA as a point of departure (POD) for dose-response modeling. For the selected statistical model (based on the AIC), we used the BMDS to calculate the BMD, BMDL, AIC, p-value, statistical dose-response equation, and the parameter estimates. We also used the BMDS to display a graph of the fitted dose-response model.

In the IRIS report (EPA, 2022), they used BMDS Version 2.2 (see page B-700 of the IRIS report Appendix) to analyze the sensory irritation responses at the 5 doses shown in the last two columns of Table 1. In the report they quote these data sets as "assumed response among controls = 0" and "assumed response among controls = 3%". However, we were able to almost exactly reproduce the IRIS report BMD, BMDL, and AIC values, but not the p-values, using our analysis with the current BMDS Version 3.3rc10 where at a dose of 0, the "observed" response rate was assumed to be either 0 cases out of 16 (0%) or 3 cases out of 16 (18.75%) for the two data sets. We put quotation marks around "observed" since we do not believe that Anderson and Molhave collected data at the control dose of zero, based on the two papers. For this reason, we believe that the description of the BMD analyses in the IRIS report was misinterpreted. We could not find the older software BMDS Version 2.2. We were able to obtain and run a copy of BMDS Version 2.7 in EPA archives and obtained similar results to the current BMDS for the BMD, BMDL, and AIC values. However, the p-values from version 2.7 did not match the p-values from the new version and also did not match the p-values from IRIS. The p-values in the current BMDS version are consistent with the usual chi-square goodness-of-fit test. In the IRIS report they did not analyze the sensory irritation responses at the 4 positive doses only (column 4 of Table 1) and they also did not analyze the discomfort rate data.

We will present the dose response and BMD analyses separately for each of the five data sets tabulated in Table 1. Note that for the Anderson and Molhave papers the dose was reported as mg/m^3.

Discomfort at 2.5 hours

Table 4 presents the BMDS Version 3.3rc10 model summaries for the data on discomfort rates at 2.5 hours.

Table 4. BMDS Version 3.3rc10 Summary for Discomfort Rates at 2.5 Hours. Selected Model Based on Lowest AIC is Shown in Bold.

Model	BMD (mg/m^3)	BMDL (mg/m^3)	P-value	AIC
Dichotomous Hill	0.899	0.127	0.655	76.838
Gamma	0.461	0.153	0.348	77.849
Log-Logistic	0.492	0.115	0.364	77.758
Multistage Degree 3	0.293	0.152	0.331	77.933
Multistage Degree 2	0.293	0.152	0.331	77.933
Multistage Degree 1	**0.217**	**0.151**	**0.607**	**75.987**
Weibull	0.388	0.152	0.341	77.885
Logistic	0.495	0.364	0.588	76.155
Log-Probit	0.575	0.077	0.378	77.700
Probit	0.467	0.349	0.596	76.110
Quantal Linear	**0.217**	**0.151**	**0.607**	**75.987**

The selected model based on the AIC was the Multistage Degree 1 model, with the dose response equation: P(response) = g + (1-g)*[1-exp(-b1*dose^1). This is exactly the same model (with the same AIC) as the Quantal Linear model. The BMD and BMDL were 0.217 and 0.151 mg/m^3, respectively.

The estimated parameters are shown in Table 5. The selected dose-response model together with the BMD and BMDL values is plotted in Figure 4.

Table 5. BMDS Version 3.3rc10 Fitted Multistage Degree 1 Model for Discomfort Rates at 2.5 Hours with 95% Confidence Intervals.

Variable	Estimate	Standard Error	Lower Bound	Upper Bound
g	0.008	0.854	-1.665	1.682
b1	0.485	6.565	-12.382	13.353

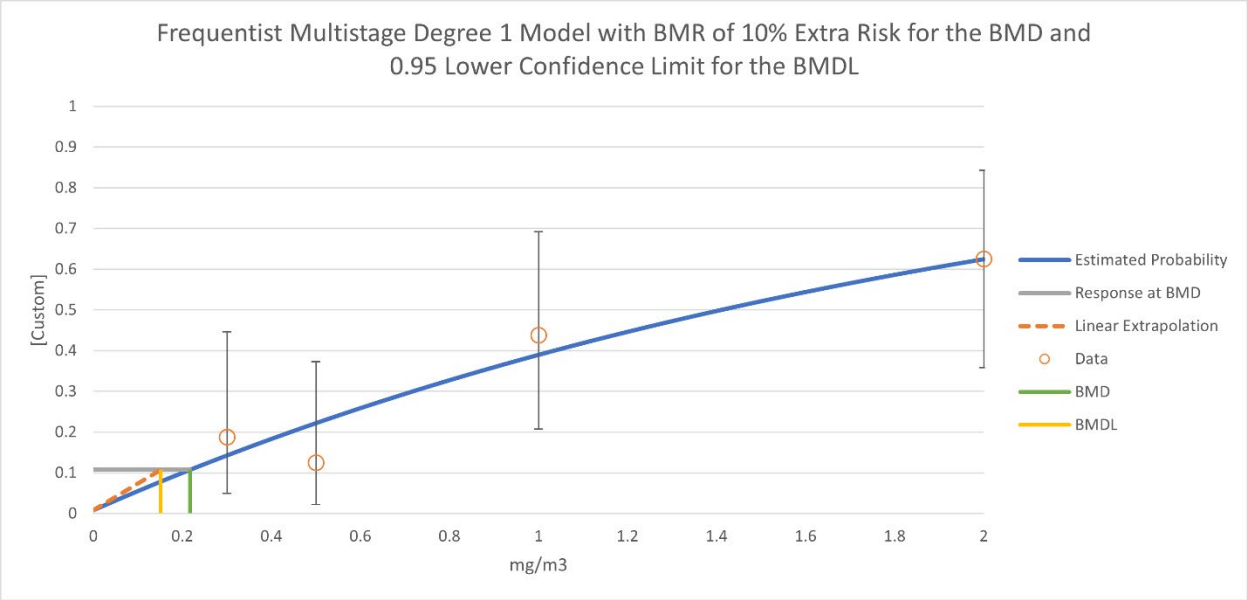

Figure 4. BMDS Version 3.3rc10 Fitted Multistage Degree 1 Dose-Response Model for Discomfort Rates at 2.5 Hours.

The wide confidence intervals for the parameters show a large uncertainty in the fitted model. The estimated response probability increases from about 0 at 0 mg/m^3 to about 0.6 at 2 mg/m^3.

Discomfort at 5 hours

Table 6 presents the BMDS Version 3.3rc10 model summaries for the data on discomfort rates at 5 hours.

Table 6. BMDS Version 3.3rc10 Summary for Discomfort Rates at 5 Hours. Selected Model Based on Lowest AIC is Shown in Bold.

Model	BMD (mg/m^3)	BMDL (mg/m^3)	P-value	AIC
Dichotomous Hill	1.810	0.001	NA	92.680
Gamma	**1.542**	**0.250**	**0.223**	**88.680**
Log-Logistic	1.807	0.195	0.083	90.680
Multistage Degree 3	1.185	0.242	0.204	88.848
Multistage Degree 2	0.944	0.227	0.182	89.227
Multistage Degree 1	0.502	0.202	0.151	90.072
Weibull	1.805	0.250	0.083	90.680
Logistic	0.564	0.328	0.154	89.818
Log-Probit	1.819	0.000	0.083	90.680
Probit	0.558	0.323	0.154	89.830
Quantal Linear	0.502	0.202	0.151	90.072

The selected model based on the AIC was the Gamma model, with the dose response equation: P(response) = g+(1-g)*CumGamma[b*dose, a]. The BMD and BMDL were 1.542 and 0.250 mg/m^3, respectively.

The estimated parameters are shown in Table 7. The selected dose-response model together with the BMD and BMDL values is plotted in Figure 5.

Table 7. BMDS Version 3.3rc10 Fitted Gamma Model for Discomfort Rates at 5 Hours with 95% Confidence Intervals.

Variable	Estimate	Standard Error	Lower Bound	Upper Bound
g	0.375	0.017	0.340	0.409
a	18 (Bounded)	NA	NA	NA
b	8.315	1.071	6.216	10.414

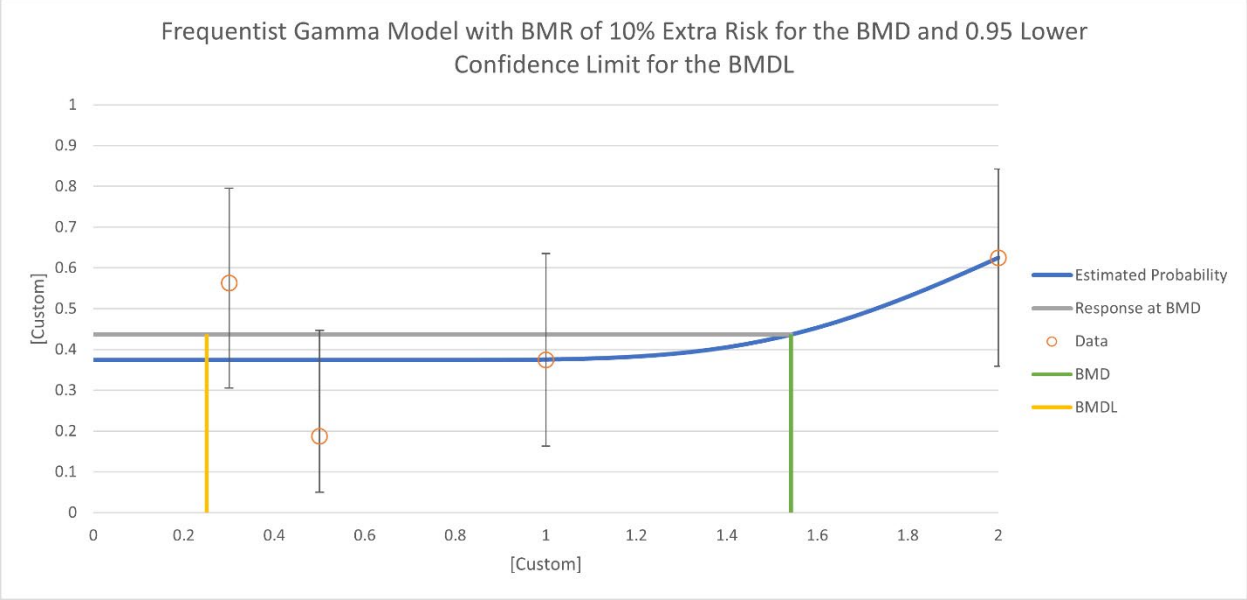

Figure 5. BMDS Version 3.3rc10 Fitted Gamma Dose-Response Model for Discomfort Rates at 5 Hours.

The fitted model shows an almost constant response rate at doses up to about 1.5 mg/m^3 after which the probability of discomfort at 5 hours increases rapidly.

Sensory Irritation Using Data at 4 Positive Doses Only.

Table 8 presents the BMDS Version 3.3rc10 model summaries for sensory irritation using the data at the 4 positive doses only. In this case, although the Dichotomous Hill model had the lowest AIC, the p-value for that model was not available and BMDS determined this model as questionable. Therefore, we excluded that model and selected the model with the second lowest AIC instead.

Table 8. BMDS Version 3.3rc10 Summary for Sensory Irritation Using the Data at the 4 Positive Doses Only. Selected Model Based on Lowest AIC, Excluding the Dichotomous Hill Model, is Shown in Bold.

Model	BMD (mg/m^3)	BMDL (mg/m^3)	P-value	AIC
Dichotomous Hill	0.483	0.256	NA	58.280
Gamma	**0.208**	**0.091**	**0.664**	**58.846**
Log-Logistic	0.325	0.161	0.497	59.104
Multistage Degree 3	0.138	0.068	0.531	60.321
Multistage Degree 2	0.138	0.068	0.531	60.321
Multistage Degree 1	0.080	0.060	0.420	60.261
Weibull	0.169	0.077	0.583	59.527
Logistic	0.224	0.155	0.681	60.115
Log-Probit	0.272	0.153	0.378	59.949
Probit	0.207	0.151	0.436	62.095
Quantal Linear	0.080	0.060	0.420	60.261

The selected model based on the AIC was the Gamma model, with the dose response equation: P(response) = g+(1-g)*CumGamma[b*dose, a]. The BMD and BMDL were 0.208 and 0.091 mg/m^3, respectively.

The estimated parameters are shown in Table 9. The selected dose-response model together with the BMD and BMDL values is plotted in Figure 6.

Table 9. BMDS Version 3.3rc10 Fitted Gamma Model for Sensory Irritation Using the Data at the 4 Positive Doses Only with 95% Confidence Intervals.

Variable	Estimate	Standard Error	Lower Bound	Upper Bound
g	1.523 E –8 (Bounded)	NA	NA	NA
a	2.275	0.904	0.503	4.047
b	3.255	1.448	0.418	6.092

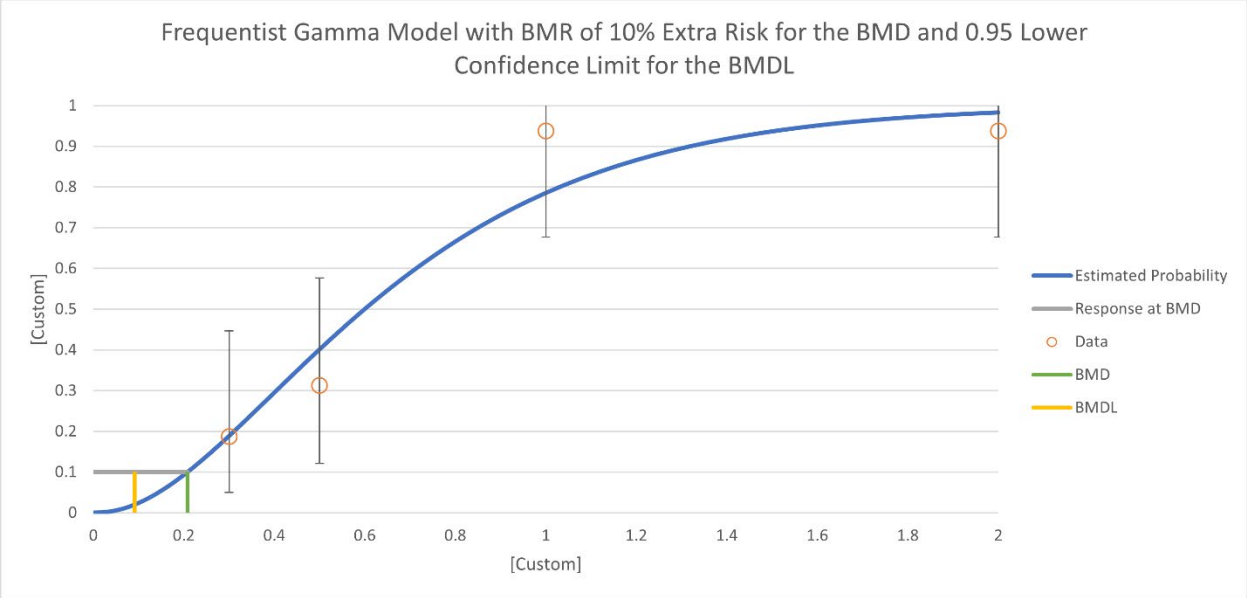

Figure 6. BMDS Version 3.3rc10 Fitted Gamma Dose-Response Model for Sensory Irritation Using the Data at the 4 Positive Doses Only.

The wide confidence intervals for the parameters show a large uncertainty in the fitted model. The fitted dose-response model shows that the response rate smoothly increases from 0 to 1 as the dose increases from 0 to 2 mg/m^3.

Sensory Irritation Using Data at 4 Positive Doses and Assuming 0 Responses at Dose 0.

Tables 10a and 10b present the BMDS model summaries for sensory irritation using the data at the 4 positive doses together with an assumed 0 responses out of 16 at dose 0. The results in Table 10a are from the IRIS report (EPA, 2022) that used the older BMDS Version 2.2. For comparison, the results in Table 10b are from the current BMDS Version 3.3rc10. Note that the IRIS report described their analyses

as “assumed response among controls = 0”. Also note that the IRIS report models do not include the Dichotomous Hill and Multistage Degree 1 models.

Table 10a. BMDS Version 2.2 Summary for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 0 Responses at Dose 0. Results From EPA (2022). Selected Model Based on Lowest AIC is Shown in Bold.

Model	**BMD (mg/m^3)**	**BMDL (mg/m^3)**	**P-value**	**AIC**
Gamma	0.209	0.091	0.049	58.847
Log-Logistic	**0.257**	**0.157**	**0.143**	**57.330**
Multistage Degree 3	0.137	0.068	0.016	60.321
Multistage Degree 2	0.137	0.068	0.016	60.321
Weibull	0.169	−0.077	0.040	59.527
Logistic	0.256	0.182	0.067	62.408
Log-Probit	0.249	0.153	0.111	57.965
Probit	0.239	0.175	0.047	65.167
Quantal Linear	0.080	0.060	0.025	60.262

Table 10b. BMDS Version 3.3rc10 Summary for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 0 Responses at Dose 0. Selected Model Based on Lowest AIC is Shown in Bold.

Model	BMD (mg/m^3)	BMDL (mg/m^3)	P-value	AIC
Dichotomous Hill	0.281	0.167	0.547	58.823
Gamma	0.208	0.091	0.845	58.846
Log-Logistic	**0.257**	**0.157**	**0.840**	**57.330**
Multistage Degree 3	0.138	0.068	0.737	60.321
Multistage Degree 2	0.138	0.068	0.737	60.321
Multistage Degree 1	0.080	0.060	0.588	60.262
Weibull	0.169	0.077	0.782	59.527
Logistic	0.256	0.182	0.729	62.408
Log-Probit	0.249	0.153	0.834	57.965
Probit	0.239	0.175	0.519	65.167
Quantal Linear	0.080	0.060	0.588	60.262

For both BMDS versions, the selected model based on the AIC was the Log-Logistic model, with the dose response equation: P(response) = g+(1-g)/[1+exp(-a-b*Log(dose))]. For both BMDS versions, the (rounded) BMD and BMDL were 0.257 and 0.157 mg/m^3, respectively. Except for the Weibull distribution BMDL, the BMD, BMDL, and AIC values for the two BMDS versions were all within 0.001 of each other, strongly suggesting that both versions used the same modeling formulations and data; the slight differences are likely due to differences in the convergence criteria. For the Weibull distribution, the BMDL values for the two versions were reported as –0.077 and 0.077, and it is reasonable to assume that the negative sign in the IRIS report was a typo.

The p-values for the two BMDS versions are extremely different. For example, the p-value for the selected model using BMDS Version 2.2 was 0.143 but the p-value for the selected model using BMDS Version 3.3rc10 was 0.840. Although we could not find documentation for the p-value calculations used

in BMDS Version 2.2, the values in BMDS Version 3.3rc10 agree with the usual p-value approach described on page 67 of the Benchmark Dose Technical Guidance (EPA, 2012): The scaled residuals for each dose (not shown here) are (O-E)/sqrt(E), where O and E are the observed and expected counts, the chi-squared statistic (0.841) is the sum of the squared scaled residuals, and the p-value (0.840) is indeed the probability that a chi-square value with 3 degrees of freedom exceeds 0.841.

The estimated parameters for BMD Version 3.3rc10 are shown in Table 11. (The corresponding estimates for BMD Version 2.2 are not reported in the IRIS report.) The selected dose-response model together with the BMD and BMDL values, is plotted in Figure 7.

Table 11. BMDS Version 3.3rc10 Fitted Log-Logistic Model for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 0 Responses at Dose 0 with 95% Confidence Intervals.

Variable	Estimate	Standard Error	Lower Bound	Upper Bound
g	1.523 E–8 (Bounded)	NA	NA	NA
a	1.630	0.510	0.630	2.630
b	2.819	0.667	1.511	4.127

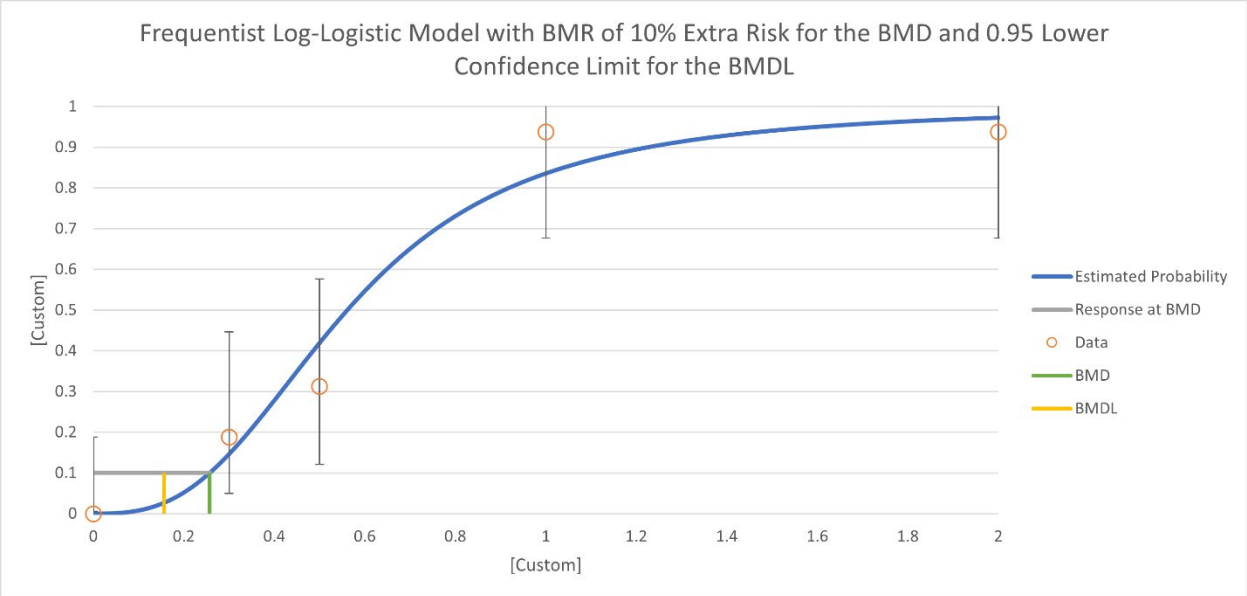

Figure 7. BMDS Version 3.3rc10 Fitted Log-Logistic Dose-Response Model for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 0 Responses at Dose 0.

The wide confidence intervals for the parameters show a large uncertainty in the fitted model. The fitted dose-response model shows that the response rate smoothly increases from 0 to 1 as the dose increases from 0 to 2 mg/m^3.

Sensory Irritation Using Data at 4 Positive Doses and Assuming 3 Responses at Dose 0.

Tables 12a and 12b present the BMDS model summaries for sensory irritation using the data at the 4 positive doses together with an assumed 3 responses out of 16 at dose 0. The results in Table 12a are from the IRIS report (EPA, 2022) that used the older BMDS Version 2.2. For comparison, the results in Table 12b are from the current BMDS Version 3.3rc10. Note that the IRIS report described their analyses as "assumed response among controls = 3%" although 3/16 = 18.75%. Also note that the IRIS report models do not include the Dichotomous Hill and Multistage Degree 1 models.

Table 12a. BMDS Version 2.2 Summary for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 3 Responses at Dose 0. Results From EPA (2022). Selected Model Based on Lowest AIC is Shown in Bold.

Model	**BMD (mg/m^3)**	**BMDL (mg/m^3)**	**P-value**	**AIC**
Gamma	0.304	0.142	0.195	77.217
Log-Logistic	**0.369**	**0.219**	**0.401**	**74.821**
Multistage Degree 3	0.262	0.091	0.115	79.039
Multistage Degree 2	0.262	0.091	0.115	79.039
Weibull	0.233	0.108	0.170	78.456
Logistic	0.201	0.148	0.0001	76.388
Log-Probit	0.350	0.208	0.320	75.800
Probit	0.196	0.149	0.0005	77.859
Quantal Linear	0.091	0.065	0.152	80.471

Table 12b. BMDS Version 3.3rc10 Summary for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 3 Responses at Dose 0. Selected Model Based on Lowest AIC is Shown in Bold.

Model	BMD (mg/m^3)	BMDL (mg/m^3)	P-value	AIC
Dichotomous Hill	**0.483**	**0.312**	**0.999**	**73.722**
Gamma	0.304	0.142	0.567	77.217
Log-Logistic	0.369	0.219	0.796	74.821
Multistage Degree 3	0.262	0.091	0.446	79.039
Multistage Degree 2	0.262	0.091	0.446	79.039
Multistage Degree 1	0.090	0.065	0.223	80.471
Weibull	0.233	0.107	0.395	78.456
Logistic	0.201	0.148	0.598	76.388
Log-Probit	0.350	0.208	0.688	75.800
Probit	0.196	0.149	0.528	77.859
Quantal Linear	0.090	0.065	0.223	80.471

For BMDS Version 2.2, the selected model based on the AIC was the Log-Logistic model, with the dose response equation: P(response) = g+(1-g)/[1+exp(-a-b*Log(dose))]. For BMDS Version 2.2, the (rounded) BMD and BMDL for the Log-Logistic model were 0.369 and 0.219 mg/m^3, respectively. For BMDS Version 3.3rc10, the selected model based on the AIC was the Dichotomous Hill model, not available in BMDS Version 2.2, with the dose response equation: P(response) = g +(v-v*g)/[1+exp(-a-b*Log(dose))] and the BMD and BMDL values were 0.483 and 0.312 mg/m^3, respectively . The BMD, BMDL, and AIC values for the two BMDS versions were all within 0.001 of each other, strongly suggesting that both versions used the same modeling formulations and data.

The p-values for the two BMDS versions are extremely different. For example, the p-value for the Log-Logistic model using BMDS Version 2.2 was 0.401, but the p-value for the same model using BMDS Version 3.3rc10 was 0.769. Although we could not find documentation for the p-value calculations used

in BMDS Version 2.2, the values in BMDS Version 3.3rc10 agree with the usual p-value approach described on page 67 of the Benchmark Dose Technical Guidance (EPA, 2012).

The estimated parameters for the Log-Logistic model selected using BMD Version 2.2 are shown in Table 13a (based on the IRIS report). The estimated parameters for the Dichotomous Hill model selected using BMD Version 3.3rc10 are shown in Table 13b. For a direct comparison between the IRIS report and the results using BMD Version 3.3rc10, the estimated parameters for the Log-Logistic model fitted using BMD Version 3.3rc10 are shown in Table 13c. The fitted dose-response models for the BMDS Versions 2.2 and 3.3rc10 together with the BMD and BMDL values, are plotted in Figures 8a (Log-Logistic model copied from the IRIS report), 8b (Dichotomous Hill model from the BMDS Version 3.3rc10 output) and 8c (Log-Logistic model from the BMDS Version 3.3rc10 output), respectively.

Table 13a. BMDS Version 2.2 Fitted Log-Logistic Model for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 3 Responses at Dose 0 with 95% Confidence Intervals. Based on Table B4 of the IRIS Report (EPA, 2022).

Variable	Estimate	Standard Error	Lower Bound	Upper Bound
g	0.160	0.072	0.020	0.301
a	1.462	0.610	0.267	2.657
b	3.668	1.129	1.456	5.881

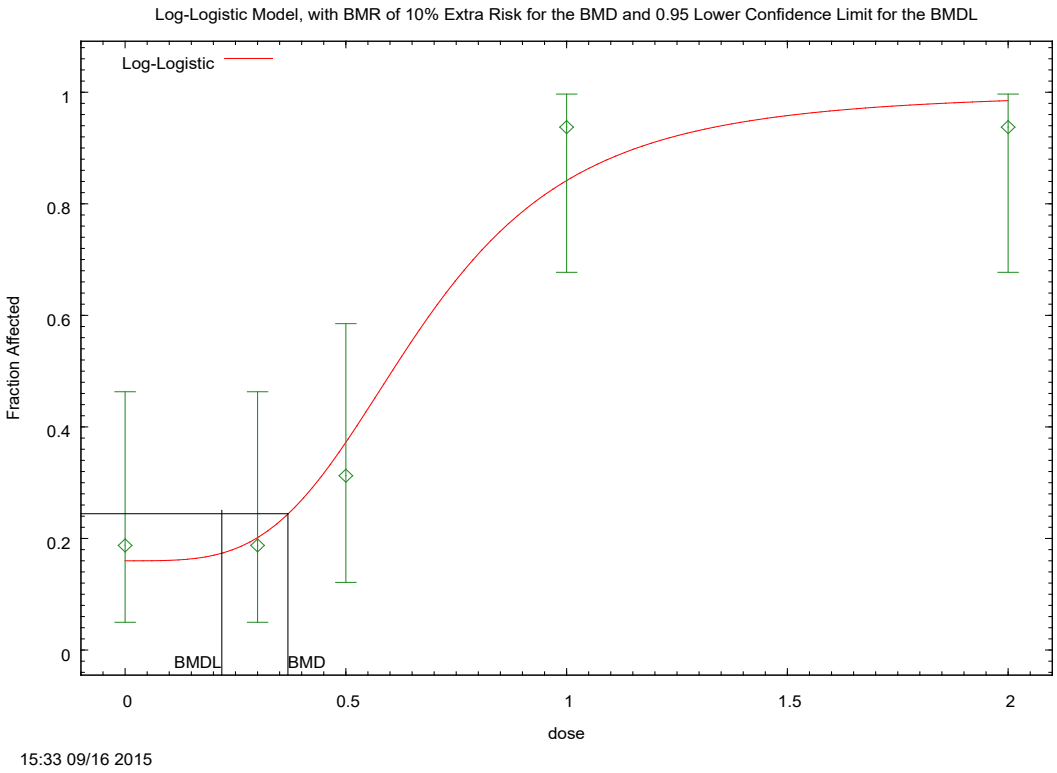

Figure 8a. BMDS Version 2.2 Fitted Log-Logistic Dose-Response Model for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 3 Responses at Dose 0. Copied from Figure B4 of the IRIS Report (EPA, 2022).

Table 13b. BMDS Version 3.3rc10 Fitted Dichotomous Hill Model for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 3 Responses at Dose 0 with 95% Confidence Intervals.

Variable	Estimate	Standard Error	Lower Bound	Upper Bound
g	0.187	0.028	0.133	0.242
v	0.923	0.031	0.862	0.984
a	8.195	0.622	6.976	9.415
b	14.145	0.897	12.386	15.904

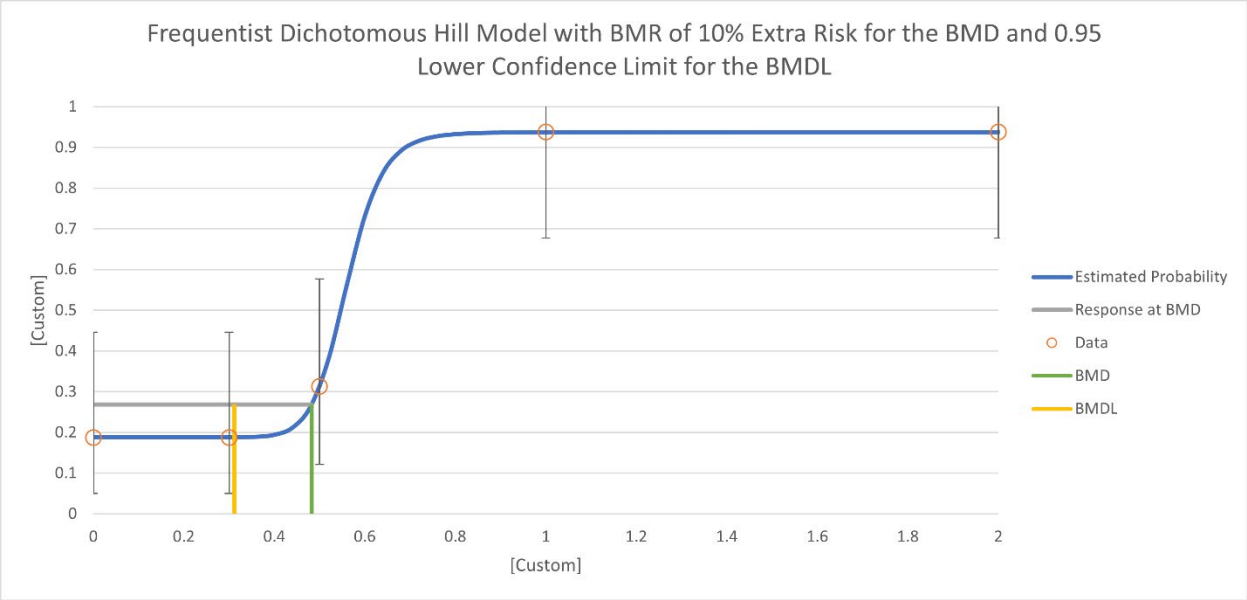

Figure 8b. BMDS Version 3.3rc10 Fitted Dichotomous Hill Dose-Response Model for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 3 Responses at Dose 0.

Table 13c. BMDS Version 3.3rc10 Fitted Log-Logistic Model for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 3 Responses at Dose 0 with 95% Confidence Intervals.

Variable	Estimate	Standard Error	Lower Bound	Upper Bound
g	0.160	0.033	0.096	0.225
a	1.462	0.609	0.267	2.657
b	3.668	1.129	1.456	5.880

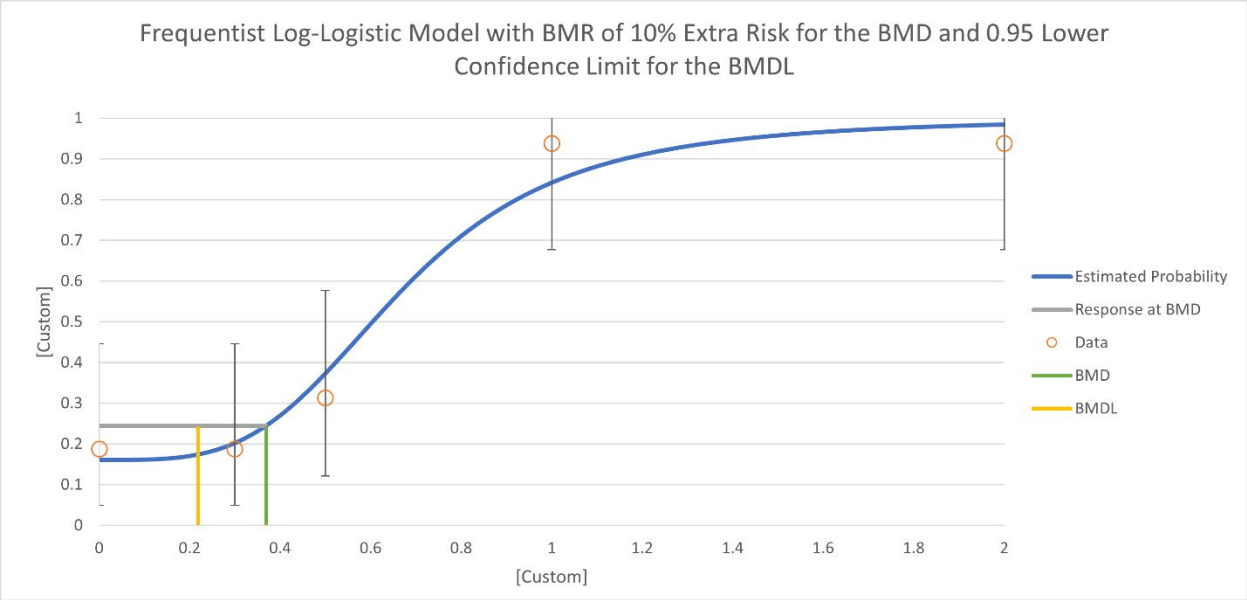

Figure 8c. BMDS Version 3.3rc10 Fitted Log-Logistic Dose-Response Model for Sensory Irritation Using the Data at the 4 Positive Doses and Assuming 3 Responses at Dose 0.

The log-logistic model parameter estimates and dose-response curves for BMDS Version 2.2 and BMDS Version 3.3rc10 (Tables 12a and 12c, Figures 8a and 8c) match almost exactly for the intercept and slope parameters and for the estimated value of the background parameter g, except that the standard errors and confidence intervals differ for the background parameter g. For that model the estimated response probability increases slowly from about 0.15 at 0 mg/m^3 to about 0.25 at 0.4 mg/m^3 then increases at a faster rate until it almost reaches 1 at 2 mg/m^3. For the better fitting Dichotomous Hill model (Table 12b and Figure 8b), there is a very rapid increase in the estimated response probability from about 0.25 at 0.5 mg/m^3 to about 0.9 at 0.7 mg/m^3.

Benchmark Dose Modeling Results Adjusted for Dependence

The above benchmark dose modeling results assume independence between the results at different doses. Since the available data did not include individual results, the IRIS report (EPA, 2022, page 2-9) chose to use BMD/2 (they refer to this as BMC/2) to approximate the BMDL after accounting for clustering, i.e., dependence between repeated measures at different doses. It is reasonable to assume that clustering will increase the width of the confidence intervals. However, the choice of a factor of 2 seems arbitrary and no rationale is provided for why this approximates the BMDL. Importantly, the true BMDL is a lower confidence limit for the dose at which the extra risk is 10%, and it therefore it needs to account both for the uncertainty (variance) in the parameter estimates **and** the extra uncertainty due to correlation between repeated measures. To demonstrate this point about the arbitrary nature of the factor 2, suppose that the best model for sensory irritation assuming 3 responses at dose 0 based on AIC had been the Multistage Degree 2 model and not the Log Logistic model. In that case the BMD and BMDL assuming independence would have been 0.262 mg/m^3 and 0.091 mg/m^3 and BMD/2 would have been 0.131 mg/m^{3}which is higher than the BMDL calculated assuming independence. Since clustering

generally decreases the effective sample size, and thus increases the parameter uncertainty, the true BMDL would be expected to be lower than the BMDL calculated assuming independence. At a minimum, we suggest that any adjustment to the BMDL to account for correlations should be applied to the BMDL from the model assuming independence instead of the BMD. One possibility might have been to divide the original BMDL by 2 instead of dividing the BMD by 2. However, without having good data from similar studies with repeated measures to use in a simulation study the numerical value of such an adjustment factor seems totally arbitrary. The following Table 14 summarizes the BMD and BMDL values for the selected models, with and without the adjustment factor. As noted above, the BMD and BMDL values are the same for BMDS Version 2.2 except for the Dichotomous Hill and Multistage Degree 1 models that are not reported in BMDS Version 2.2.

Table 14. BMDS Versions 2.2 and 3.3rc10 BMD, BMDL, BMD/2, and BMDL/2 Values from Anderson and Molhave (1983)

Outcome	Discomfort at 2.5 hours Cases / Subjects 4 positive doses BMDS 3.3rc10	Discomfort at 5 hours Cases / Subjects 4 positive doses BMDS 3.3rc10	Sensory Irritation Cases / Subjects 4 positive doses BMDS 3.3rc10	Sensory Irritation Cases / Subjects 4 positive doses Assume 0 cases at 0 mg/m^3 (unobserved) BMDS 2.2 and 3.3rc10	Sensory Irritation Cases / Subjects 4 positive doses Assume 3 cases at 0 mg/m^3 (unobserved) BMDS 2.2	Sensory Irritation Cases / Subjects 4 positive doses Assume 3 cases at 0 mg/m^3 (unobserved) BMDS 3.3rc10
BMD (mg/m^3)	0.217	1.542	0.208	0.257	0.369	0.483
BMDL (mg/m^3)	0.151	0.250	0.091	0.157	0.219	0.312
BMD/2 (mg/m^3)	0.109	0.771	0.104	0.174	0.184	0.241
BMDL/2 (mg/m^3)	0.076	0.125	0.046	0.078	0.110	0.156

3. Kulle et al (1987) and Kulle (1993)

This study was described in two papers. In the Kulle et al (1987) paper, the authors reported and analyzed the detailed results. In the follow-on paper Kulle (1993), the author used subject-specific data at pairs of doses to test if their symptomatic responses at the two doses were the same. The intent of this study was to determine the effects of acute exposures to inhaled formaldehyde vapor.

Nineteen healthy non-smoking human subjects (10 male, 9 female, average age 26 years; all non-smokers) participated in this study. Informed consent was obtained from all subjects, and subjects were financially compensated.

Each subject received 5 three-hour exposures to formaldehyde or control air with a week between exposures. Each subject served as their own control. The first group of 10 subjects (Group I) were exposed to 0.0, 0.5, 1.0, and 2.0 ppm formaldehyde (0.0, 0.62, 1.23, and 2.46 mg/m^3) at rest, with an additional 2.0 ppm exposure with exercise. The second group of 9 subjects (Group II) were exposed to 0.0, 1.0, 2.0 and 3.0 ppm (0.0, 1.23, 2.46 and 3.69 mg/m^3) at rest, and an additional 2.0 ppm with exercise. Spirometric measurements were made prior to and during exposures at 0, 30, 60, 90, 120, 150, and 180 minutes. On the day incorporating exercise, an 8-minute bicycle ride was performed every 30 minutes and minute ventilation was measured between the fourth and fifth minutes of each exercise stint. Spirometric measurements were completed 2 minutes after each exercise period. Post-exposure measurements were obtained 24 hours after the 3.0 ppm at-rest and 2.0 ppm with exercise HCHO exposures.

Airway resistance and thoracic gas volume were measured prior to exposure, at the completion of the 3-hour exposure and at 24 hours post-exposure. Non-specific airway reactivity was assessed at the end of the 3-hour exposure period. Nasal resistance was measured before and immediately following exposures to 2.0 and 3.0 ppm formaldehyde. Symptoms were scored by each subject prior to exposure, immediately following exposure, and 24 hours post-exposure. Incidence and severity of odor, nose/throat irritation, eye irritation, chest discomfort/tightness, cough, headache, heart palpitations, and double vision were recorded. The intent of this study was to determine the effect of inhaled formaldehyde vapor on nasal mucociliary flow, nasal airflow resistance, forced expiratory vital capacity, and irritation threshold.

RESULTS

Symptom Differences

Results of inhalation exposure to formaldehyde on reported symptoms for odor sensation, nose/throat irritation, eye irritation, chest discomfort, cough, and headache were summarized in Table II of the 1987 publication. Each response was assigned a score between 0 and 3. For the 9 subjects tested at all four doses 0, 1, 2, 3, and 4 ppm, Table 15 (copy of Table II of the 1987 publication) shows the mean symptom score difference (between the scores before and after exposure) at each dose together with its standard error. The last column of the table shows the results of a simple linear regression statistical test for a linear trend, showing significant linear trends (at the 5% significance level) for odor sensation and eye irritation. Figure 9 (copy of Figure 1 from the 1987 publication) shows the concentration-response for pooled Group I and Group II participants. Data from Group I participants were not reported in tabular format but were discussed within the paper.

Table II. Mean Symptom Differences ($t_{180} - t_0$) ± SE with formaldehyde exposure in Group II ($n = 9$).[a]

	HCHO concentration, ppm				Linear[b] dose significance
	0.0	1.0	2.0	3.0	
Odor sensation	0.00 ± 0.00	0.22 ± 0.15	0.44 ± 0.18	1.00 ± 0.29	$p < 0.0001$
Nose/throat irritation	0.00 ± 0.00	0.11 ± 0.11	0.33 ± 0.17	0.22 ± 0.15	$p = 0.054$
Eye irritation	0.00 ± 0.00	0.44 ± 0.24	0.89 ± 0.26	1.44 ± 0.18	$p < 0.0001$
Chest discomfort	0.00 ± 0.00	0.00 ± 0.00	0.11 ± 0.11	0.00 ± 0.00	$p = 0.62$
Cough	0.00 ± 0.00	0.11 ± 0.11	0.00 ± 0.00	0.00 ± 0.00	$p = 0.11$
Headache	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.11 ± 0.11	$p = 0.33$

[a] Presence and severity of symptoms were scored as: 0 = none; 1 = mild (present, but not annoying); 2 = moderate (annoying); 3 = severe (debilitating).
[b] No significant nonlinear trends were detected.

Table 15. Table II from page 921 of Kulle *et al* (1987). Mean symptom differences in Group II.

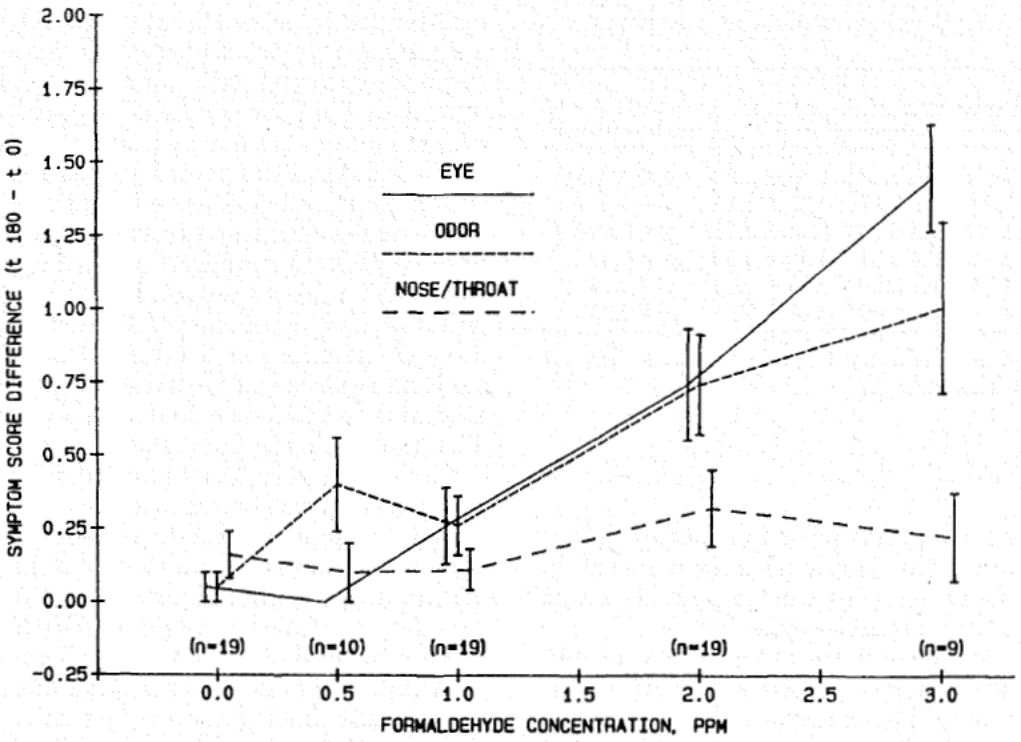

Figure 1. Mean sympton difference ±S.E. with formaldehyde exposure at rest.

Figure 9. Figure 1 from page 922 of Kulle (1987). Mean symptom differences for eye, odor, and nose / throat for all subjects.

Since the raw data for each subject were not provided, it is not possible to validate these statistical analyses.

Nasal Resistance

The paper presented graphs (Figure 3 of the 1987 publication) of individual nasal resistance percentage changes due to exposure to 2 ppm or 3 ppm formaldehyde. See Figure 10 below. As reported in the publication, "The mean increase in nasal resistance with at-rest exposure was not significant at 2 ppm HCHO (+10%, p =0.50), but was significant at 3.0 ppm (+27%, $p < 0.01$)". The statistical analysis used by Kulle et al (1987) was a t test. Since the results are presented in graphical format only, it would be difficult to properly validate these statistical analyses, although an approximate approach by digitizing the graphs is feasible.

We performed a non-parametric sign test to evaluate the nasal resistance. At 2 ppm formaldehyde, of the 13 subjects that showed a change (1 had no change), 6 of them showed an increase. Results for the remaining 5 subjects exposed to 2 ppm formaldehyde were not reported. Thus, the nasal resistance changes at 2 ppm formaldehyde were not statistically significant at the 5% level (p-value 0.5).
At 3 ppm formaldehyde, of the 6 subjects that showed a change (2 had no change), all 6 of them showed an increase. Results for the remaining 1 subject exposed to 3 ppm formaldehyde were not reported. Thus, the nasal resistance changes at 3 ppm formaldehyde were highly statistically significant at the 5 % level (p-value 0.000).

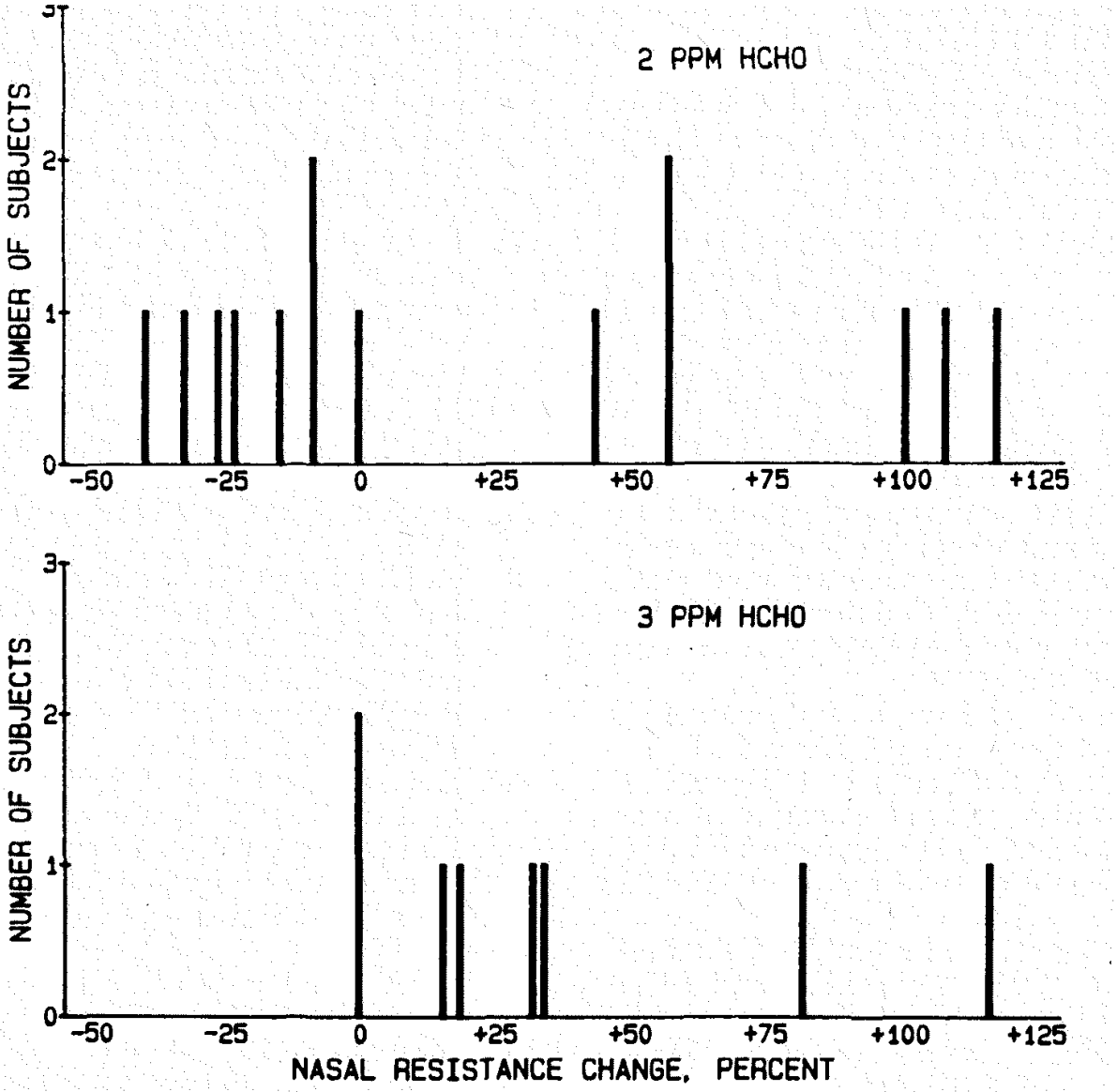

Figure 3. Subject nasal resistance change ($t_{180} - t_0$) with formaldehyde exposure at rest.

Figure 10. Figure 3 from page 923 of Kulle (1987). Individual subject nasal resistance changes at 2ppm and 3 ppm formaldehyde.

Pulmonary Function

As noted in the paper, "There were no significant decrements in pulmonary function (FVC, FEVi, FEF25-75%, SGaw) nor increases in bronchial reactivity to methacholine…as a result of acute 3-h exposures to 0.5-3.0 ppm HCHO at rest or with exercise (2.0 ppm HCHO), including 24-h postexposure." Numerical values were presented in Table III of the publication. The statistical analysis in the 1987 paper used an "analysis of variance for a randomized block design". Since the raw data for each subject were not provided, it is not possible to validate these statistical analyses.

Odor Sensation, Eye Irritation, and Nose/Throat Irritation

Summary tables for all dose levels for odor sensation (Table 2 of Kulle 1993), eye irritation (Table 3 of Kulle 1993), and nose/throat irritation (Table 4 of Kulle 1993) are available in tabular format from Kulle (1993) and presented below.

TABLE 2. Odor Sensation Reported by Subjects for Each Formaldehyde Concentration

HCHO (ppm)	Subjects (*n*)	Percent of Subjects Reporting Symptom[a]			
		None	Mild	Moderate	Severe
0.0	19	95 (10/8)	5 (0/1)	0 (0/0)	0 (0/0)
0.5	10	60 (1/5)	40 (3/1)	0 (0/0)	0 (0/0)
1.0	19	74 (8/6)	26 (2/3)	0 (0/0)	0 (0/0)
2.0	19	42 (4/4)	42 (4/4)	16 (2/1)	0 (0/0)
3.0	9	22 (2/0)	67 (3/3)	0 (0/0)	11 (1/0)

[a]Numbers in parentheses represent numbers (M/F) of males and females reporting symptom.

Table 16. Table 2 from page 327 of Kulle, 1993. Rates of Odor Sensation Symptoms.

TABLE 3. Eye Irritation Reported by Subjects for Each Formaldehyde Concentration

HCHO (ppm)	Subjects (*n*)	Percent of Subjects Reporting Symptom[a]		
		None	Mild	Moderate
0.0	19	95 (9/9)	5 (1/0)	0 (0/0)
0.5	10	100 (4/6)	0 (0/0)	0 (0/0)
1.0	19	74 (8/6)	21 (2/2)	5 (0/1)
2.0	19	47 (5/4)	32 (2/4)	21 (3/1)
3.0	9	0 (0/0)	56 (3/2)	44 (3/1)

[a]Numbers in parentheses represent numbers (M/F) of males and females reporting symptom.

Table 17. Table 3 from page 328 of Kulle, 1993. Rates of Eye Irritation Symptoms.

TABLE 4. Nose/Throat Irritation Reported by Subjects for Each Formaldehyde Concentration

HCHO (ppm)	Subjects (*n*)	Percent Reporting Symptom[a]	
		None	Mild
0.0	19	84 (9/7)	16 (1/2)
0.5	10	90 (4/5)	10 (0/1)
1.0	19	95 (9/9)	5 (1/0)
2.0	19	63 (8/4)	37 (2/5)
3.0	9	78 (6/1)	22 (0/2)

[a]Numbers in parentheses represent numbers (M/F) of males and females reporting symptom.

Table 18. Table 4 from page 329 of Kulle, 1993. Rates of Nose/Throat Irritation Symptoms.

As shown in Tables 16, 17, and 18, odor sensation (Table 16), eye irritation (Table 17), and nose/throat irritation (Table 18) were the most frequently reported symptoms from inhaled formaldehyde. A statistically significant linear dose-response at the 5% level (based on an analysis of variance) was observed for odor sensation and eye irritation ($p < 0.0001$), while nose/throat irritation almost reached the 0.05 level of significance ($p = 0.054$). As shown in Figure 9 and Table 15 "Eye irritation increased linearly from 0.5 to 3.0 ppm HCHO. At 2 ppm, 32 percent (6/19) of the subjects reported mild eye irritation and 21 percent (4/19) moderate eye irritation; at 3 ppm, all nine subjects experienced eye irritation, five at a mild and four at a moderate level." (Kulle et al, 1987)

For Group I participants exposed at rest, the paper stated that a "significant ($p< 0.05$ log-linear dose-response occurred with odor and eye irritation." Tests for non-linearity were not significant for Group I. The effect of exercise at the 2.0 ppm concentration did not have a significant effect on symptom reporting except for nose/throat irritation, which increased significantly at the 2.0 ppm concentration.

Since the raw data for each subject were not provided, it is not possible to validate any of the above statistical analyses of odor sensation, eye irritation, and nose/throat irritation.

McNemar Test

A McNemar statistical test is often used to compare results of pairs of experiments carried out on a group of subjects, where for each subject the result of each experiment is either a positive result or a negative result. The statistical test is based on the numbers of cases where the results of the two experiments disagree. For each of the three symptoms (odor sensation, eye irritation, and nose/throat irritation), Kulle (1993) reported the results of McNemar tests comparing the percentages of subjects reporting symptoms at each positive dose with the percentages of subjects reporting symptoms at the control dose. The McNemar test does not require data at different dose levels to be independent, so it is applicable even though a given subject was tested at more than one dose. For odor sensation, Kulle (1993) found no significant differences between clean air and 1.0 ppm formaldehyde (p-value 0.26) but

found significant differences between clean air and either 0.5, 2.0 or 3.0 ppm formaldehyde (p-values ≤ 0.01). For eye irritation, Kulle (1993) found no significant differences between clean air and 0.5 ppm formaldehyde or between clean air and 1 ppm formaldehyde (p-value 0.26) but found significant differences between clean air and either 2 ppm or 3 ppm formaldehyde (p-values ≤ 0.005). For nose/throat irritation, Kulle (1993) found no significant differences between clean air and 0.5 ppm formaldehyde (p-value 0.46), between clean air and 1 ppm formaldehyde (p-value 0.63), between clean air and 2 ppm formaldehyde (p-value 0.26), or between clean air and 1 ppm formaldehyde (p-value 0.19). Kulle (1993) did not discuss how the p-values were calculated. The standard McNemar test uses an asymptotic chi-square distribution with 1 degree of freedom for the test statistic Q_M, which can easily give poor estimates of the p-values with small samples. However, modern statistical software can employ an exact test that gives accurate results even for small samples.

Since raw data giving the numbers of subjects that had positive results at one dose and negative results at another dose were not provided, it is not possible to validate any of the McNemar test results for odor sensation, eye irritation, and nose/throat irritation.

Fisher Exact Test

The results of the odor sensation, eye irritation, and nose/throat irritation experiments can be summarized in the following Table 19, combining counts for both genders and counts for all symptom levels (mild, moderate, or severe).

Table 19. Odor Sensation, Eye Irritation, and Nose/Throat Irritation Results for Kulle et al

Dose ppm	Odor Sensation Cases / Subjects	Eye Irritation Cases / Subjects	Nose/Throat Irritation Cases / Subjects
0	1/19	1/19	3/19
0.5	4/10	0/10	1/10
1.0	5/19	5/19	1/19
2.0	11/19	10/19	7/19
3.0	7/9	9/9	2/9

The Fisher exact test tests whether the response rates at different doses are equal. Application of this statistical test requires the assumption that the observed responses at different doses are statistically independent. It is questionable whether the independence assumption is valid since each subject was tested at multiple doses and it is very possible that their responses at different doses are correlated. However, in the absence of raw dose-specific data on each subject, the independence assumption

cannot be statistically evaluated. The alternative to the independence assumption is clustering, such that there is dependence between repeated measures on the same subject.

The p-values for the Fisher exact tests are shown in the following Table 20. P-values of 5% or below can be treated as evidence that the response rates are different at different doses, under the independence assumption.

Table 20. Fisher Exact Tests for Odor Sensation, Eye Irritation, and Nose/Throat Irritation Results for Kulle et al

Response	Odor Sensation Cases / Subjects	Eye Irritation Cases / Subjects	Nose/Throat Irritation Cases / Subjects
P-value	0.0003	< 0.0001	0.1394

The results of the Fisher exact tests show significantly different response rates at different doses at the 5% significance level for odor sensation and eye irritation, but not for nose/throat irritation.

Cochran-Armitage Trend Test

The Cochran-Armitage trend test also assumes independence of observed responses at difference doses. Under that assumption, this is a statistical test of the null hypothesis that the response rates are the same at every dose against the alternative one-sided hypothesis that the response rates increase with the dose. (It is unrealistic to assume that the true response rates would be lower at a higher formaldehyde dose.)

The p-values for the Cochran-Armitage trend tests using the data in Table 19 are shown in the following Table 21. P-values of 5% or below can be treated as evidence that the response rates increase with the dose, under the independence assumption. Two versions of the statistical tests are displayed. The asymptotic test uses a normal approximation of the test statistic to compute the p-value. The Monte Carlo version of the test displayed here used 10,000 Monte Carlo simulations of the test statistic to estimate a 99% upper confidence limit for the p-value.

Table 21. Cochran-Armitage Trend Tests for Odor Sensation, Eye Irritation, and Nose/Throat Irritation Results for Kulle et al

Response	Odor Sensation Cases / Subjects	Eye Irritation Cases / Subjects	Nose/Throat Irritation Cases / Subjects
Asymptotic P-value (one-sided)	< 0.0001	< 0.0001	0.0632
P-value (one-sided) Monte Carlo 99% Upper Confidence Limit	0.0004	0.0005	0.0866

The results of the Cochran-Armitage trend tests show significantly increasing response rates at different doses at the 5% significance level for odor sensation and eye irritation, but not for nose/throat irritation The p-value for nose/throat irritation is 0.087 (Monte Carlo simulations), which is not significant at the 5% level but is significant at the 10% level.

Benchmark Dose Modeling Results Assuming Independence

The results of the discomfort and sensory irritation experiments shown in Table 19 were also analyzed by fitting statistical models for the probability of a response as a function of the dose. A variety of statistical models are fitted to be data and the best-fitting statistical model is selected. For these analyses we followed EPA's Benchmark Dose Modeling (BMD) method using BMD Software (BMDS) and, as in the IRIS draft report on formaldehyde (EPA, 2022), we chose the best model as the one with the lowest Akaike Information Criterion (AIC) statistic. (Note that the BMDS uses a different approach to select their recommended model.) The AIC is based on the log-likelihood of the fitted model, with an adjustment for the number of fitted parameters in the model, so that models with a large number of fitted parameters are penalized. The BMD software again assumes that repeated measures on the same subject are independent. The IRIS approach to account for possible dependence is discussed below in the section "Benchmark Dose Modeling Results Adjusted for Dependence". We will refer to the Benchmark Dose approach here as BMD (for "Benchmark Dose"), since that is the name of the software used, although the same approach is often called BMC (for "Benchmark Concentration") with the same statistical interpretation.

For the BMD modeling we assumed a Benchmark Response (BMR) of 10% extra risk above an assumed 0% risk for unexposed subjects (the control group). The BMD is defined as the estimated dose at which

the response rate is the BMR. The BMDL is the benchmark dose lower confidence limit, defined as a one-sided 95% lower confidence limit for the BMD. The BMDL is often used by EPA as a point of departure (POD) for dose-response modeling. For the selected statistical model (based on the AIC), we used the BMDS to calculate the BMD, BMDL, AIC, p-value, statistical dose-response equation, and the parameter estimates. We also used the BMDS to display a graph of the fitted dose-response model.

We will present the dose response and BMD analyses separately for each of the three data sets tabulated in Table 19. Note that for the Kulle et al papers the dose was reported as ppm.

Odor Sensation

Table 22 presents the BMDS Version 3.3rc10 model summaries for the data on odor sensation.

Table 22. BMDS Version 3.3rc10 Summary for Odor Sensation. Selected Model Based on Lowest AIC is Shown in Bold.

Model	**BMD (ppm)**	**BMDL (ppm)**	**P-value**	**AIC**
Dichotomous Hill	0.268	0.116	0.369	87.585
Gamma	0.260	0.182	0.597	85.187
Log-Logistic	0.268	0.116	0.369	87.585
Multistage Degree 3	0.303	0.184	0.366	87.013
Multistage Degree 2	0.296	0.182	0.364	87.121
Multistage Degree 1	**0.260**	**0.182**	**0.597**	**85.187**
Weibull	0.260	0.182	0.597	85.187
Logistic	0.575	0.432	0.396	86.010
Log-Probit	0.275	0.002	0.368	87.687
Probit	0.542	0.414	0.409	85.886
Quantal Linear	**0.260**	**0.182**	**0.597**	**85.187**

The selected model based on the AIC was the Multistage Degree 1 model, with the dose response equation: P(response) = g + (1-g)*[1-exp(-b1*dose^1). This is exactly the same model (with the same AIC) as the Quantal Linear model. Although the AIC values were also exactly the same 85.197 for the Gamma and Weibull models, we excluded those two models since they both had missing estimates for one of the parameters.

The BMD and BMDL were 0.260 and 0.182 ppm, respectively.

The estimated parameters are shown in Table 23. The selected dose-response model together with the BMD and BMDL values is plotted in Figure 11.

Table 23. BMDS Version 3.3rc10 Fitted Multistage Degree 1 Model for Odor Sensation with 95% Confidence Intervals.

Variable	Estimate	Standard Error	Lower Bound	Upper Bound
g	0.060	0.061	-0.059	0.179
b1	0.405	0.334	-0.250	1.059

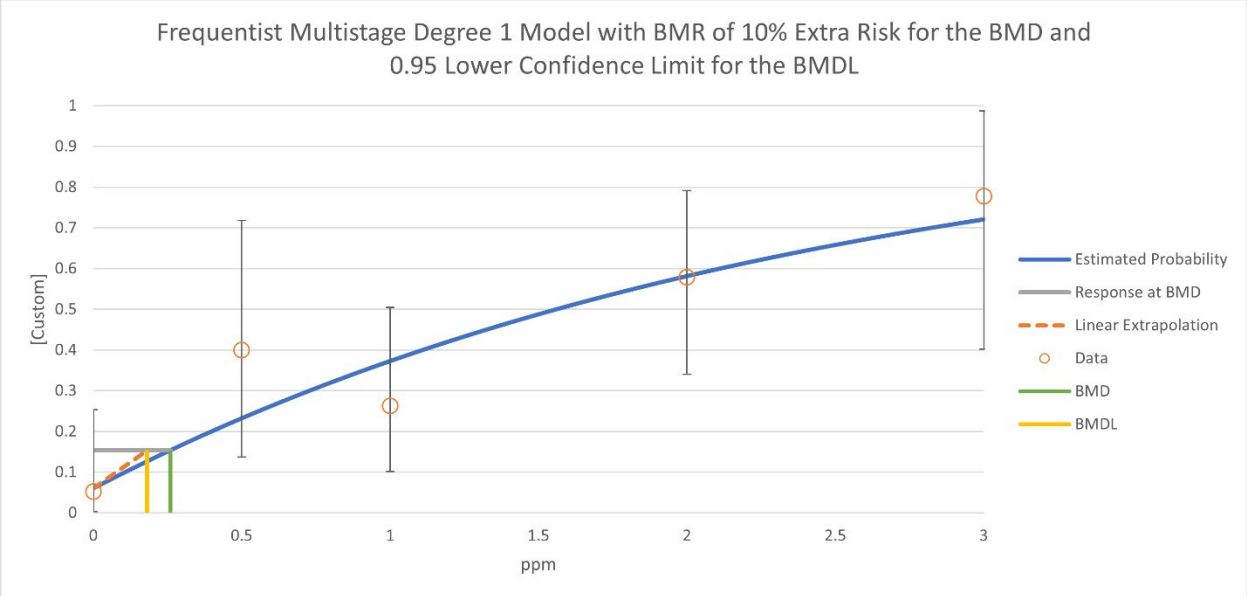

Figure 11. BMDS Version 3.3rc10 Fitted Multistage Degree 1 Dose-Response Model for Odor Sensation.

The wide confidence intervals for the parameters show a large uncertainty in the fitted model.

Eye Irritation.

Tables 24a and 24b present the BMDS model summaries for eye irritation. The results in Table 24a are from the IRIS report (EPA, 2022) that used the older BMDS Version 2.2. For comparison, the results in Table 24b are from the current BMDS Version 3.3rc10. Note that the IRIS report models do not include the Dichotomous Hill and Multistage Degree 1 models.

Table 24a. BMDS Version 2.2 Summary for Eye Irritation. Results From EPA (2022). Selected Model Based on Lowest AIC is Shown in Bold.

Model	BMD (ppm)	BMDL (ppm)	P-value	AIC
Gamma	0.853	0.497	0.182	66.839
Log-Logistic	0.852	0.510	0.147	67.596
Multistage Degree 3	0.863	0.369	0.226	66.134
Multistage Degree 2	0.676	0.395	0.373	65.090
Weibull	0.886	0.501	0.211	66.225
Logistic	0.760	0.546	0.364	64.737
Log-Probit	0.850	0.541	0.159	67.254
Probit	**0.694**	**0.502**	**0.369**	**64.645**
Quantal Linear	0.270	0.191	0.063	71.876

Table 24b. BMDS Version 3.3rc10 Summary for Eye Irritation. Selected Model Based on Lowest AIC is Shown in Bold.

Model	BMD (ppm)	BMDL (ppm)	P-value	AIC
Dichotomous Hill	0.852	0.510	0.415	67.596
Gamma	0.853	0.497	0.437	66.839
Log-Logistic	0.852	0.510	0.415	67.596
Multistage Degree 3	0.863	0.369	0.410	66.134
Multistage Degree 2	0.676	0.395	0.678	65.090
Multistage Degree 1	0.270	0.191	0.280	71.876
Weibull	0.886	0.501	0.395	66.225
Logistic	0.760	0.546	0.608	64.737
Log-Probit	0.850	0.541	0.452	67.254
Probit	**0.694**	**0.502**	**0.600**	**64.645**
Quantal Linear	0.270	0.191	0.280	71.876

For both BMDS versions, the selected model based on the AIC was the Probit model, with the dose response equation: P(response) = CumNorm(a+b*Dose). For both BMDS versions, the (rounded) BMD and BMDL were 0.694 and 0.502 ppm, respectively. The BMD, BMDL, and AIC values for the two BMDS versions were all within 0.001 of each other, strongly suggesting that both versions used the same modeling formulations and data; the slight differences are likely due to differences in the convergence criteria.

The p-values for the two BMDS versions are extremely different. For example, the p-value for the selected model using BMDS Version 2.2 was 0.369 but the p-value for the selected model using BMDS Version 3.3rc10 was 0.600. Although we could not find documentation for the p-value calculations used in BMDS Version 2.2, the values in BMDS Version 3.3rc10 agree with the usual p-value approach described on page 67 of the Benchmark Dose Technical Guidance (EPA, 2012): The scaled residuals for each dose (not shown here) are (O-E)/sqrt(E), where O and E are the observed and expected counts, the

chi-squared statistic (1.871) is the sum of the squared scaled residuals, and the p-value (0.600) is indeed the probability that a chi-square value with 3 degrees of freedom exceeds 1.871.

The estimated parameters for BMD Version 3.3rc10 are shown in Table 25a. The corresponding estimates for BMD Version 2.2 from the IRIS report are shown in Table 25b. The selected dose-response models together with the BMD and BMDL values, are plotted in Figures 12a and 12b. The numbers from the two BMDS versions match.

Table 25a. BMDS Version 2.2 Fitted Probit Model for Eye Irritation. Based on Table B-6 of the IRIS Report (EPA, 2022)

Variable	Estimate	Standard Error	Lower Bound	Upper Bound
a	-1.916	0.361	-2.624	-1.208
b	1.103	0.222	0.668	1.539

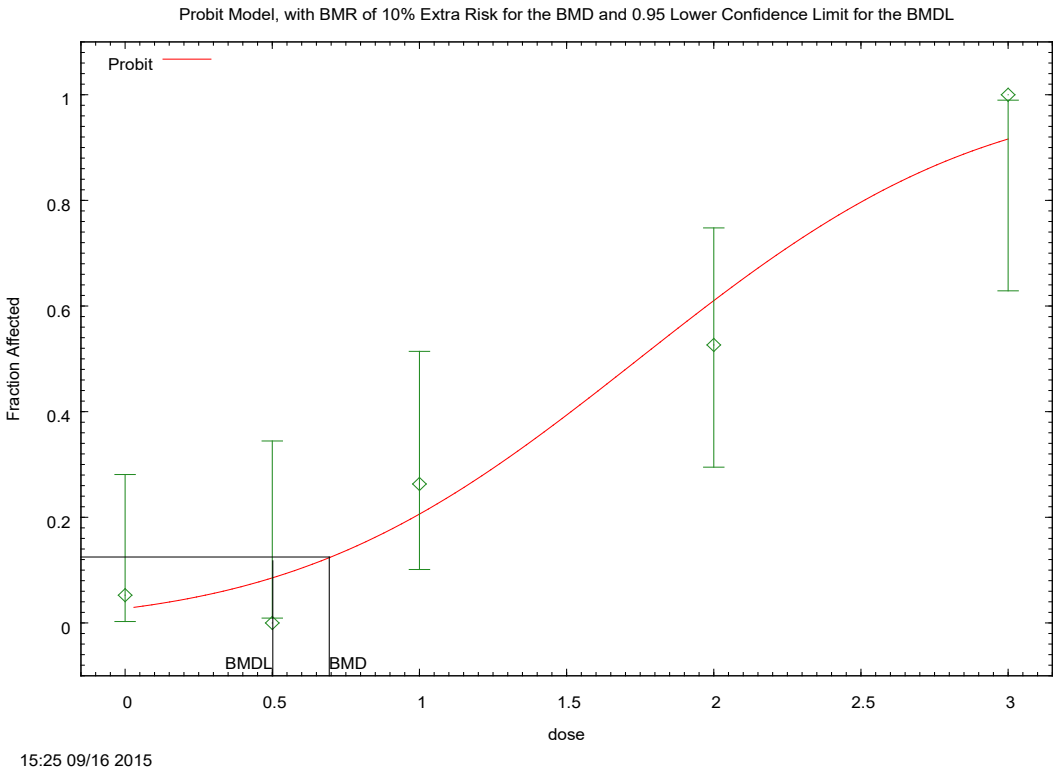

Figure 12a. BMDS Version 2.2 Fitted Probit Dose-Response Model for Eye Irritation. Copied from Figure B-5 of the IRIS Report (EPA, 2022)

Table 25b. BMDS Version 3.3rc10 Fitted Log-Logistic Model for Eye Irritation.

Variable	Estimate	Standard Error	Lower Bound	Upper Bound
a	-1.916	0.361	-2.624	-1.208
b	1.103	0.222	0.668	1.539

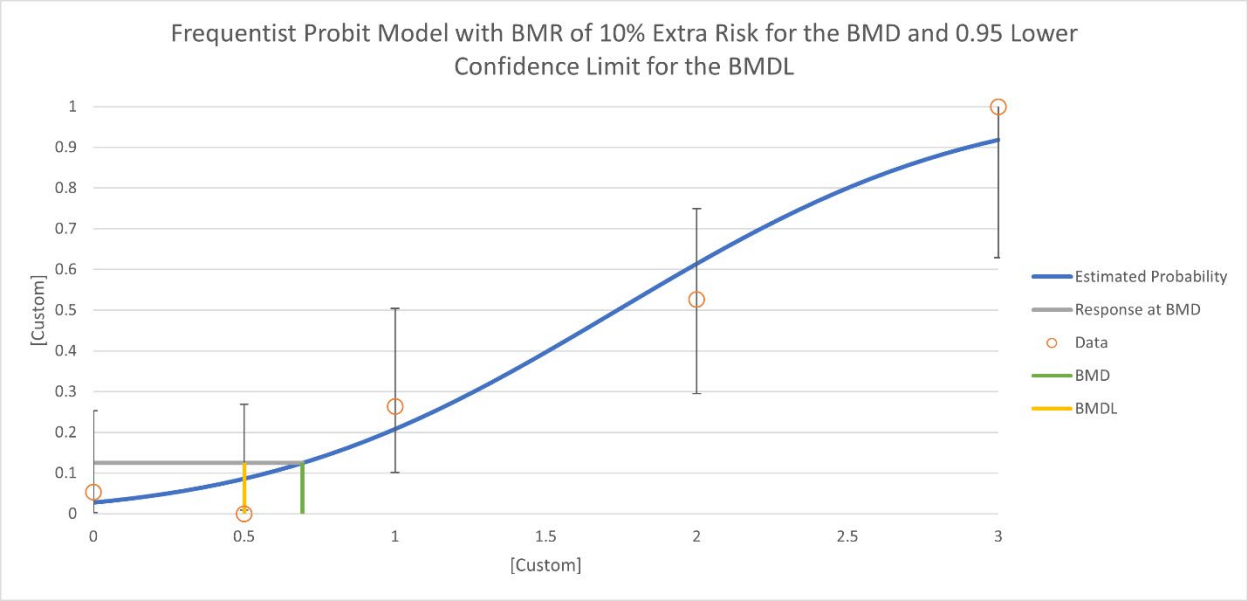

Figure 12b. BMDS Version 3.3rc10 Fitted Probit Dose-Response Model for Eye Irritation.

The wide confidence intervals for the parameters show a large uncertainty in the fitted model. The fitted dose-response model shows that the response rate smoothly increases from about 0 to 0.9 as the dose increases from 0 to 3 ppm.

Nose/Throat Irritation

Table 26 presents the BMDS Version 3.3rc10 model summaries for the data on nose/throat irritation.

Table 26. BMDS Version 3.3rc10 Summary for Nose/Throat Irritation. Selected Model Based on Lowest AIC is Shown in Bold.

Model	BMD (ppm)	BMDL (ppm)	P-value	AIC
Dichotomous Hill	**1.423**	**0.992**	**0.490**	**73.254**
Gamma	1.656	0.669	0.192	75.917
Log-Logistic	1.622	0.615	0.192	75.924
Multistage Degree 3	1.719	0.658	0.331	74.051
Multistage Degree 2	1.719	0.658	0.331	74.051
Multistage Degree 1	1.395	0.623	0.295	74.501
Weibull	1.646	0.660	0.183	76.025
Logistic	1.546	0.963	0.305	74.331
Log-Probit	1.618	0.001	0.212	75.684
Probit	1.506	0.912	0.305	74.336
Quantal Linear	1.395	0.623	0.295	74.501

The selected model based on the AIC was the Dichotomous Hill model, with the dose response equation: P(response) = g +(v-v*g)/[1+exp(-a-b*Log(dose))].

The BMD and BMDL were 1.423 and 0.992 ppm, respectively.

The estimated parameters are shown in Table 27. The selected dose-response model together with the BMD and BMDL values is plotted in Figure 13.

Table 27. BMDS Version 3.3rc10 Fitted Dichotomous Hill Model for Nose/Throat Irritation with 95% Confidence Intervals.

Variable	Estimate	Standard Error	Lower Bound	Upper Bound
g	0.104	0.029	0.047	0.162
v	0.243	0.035	0.174	0.311
a	-6.708	13.403	-32.977	19.561
b	18 (Bounded)	NA	NA	NA

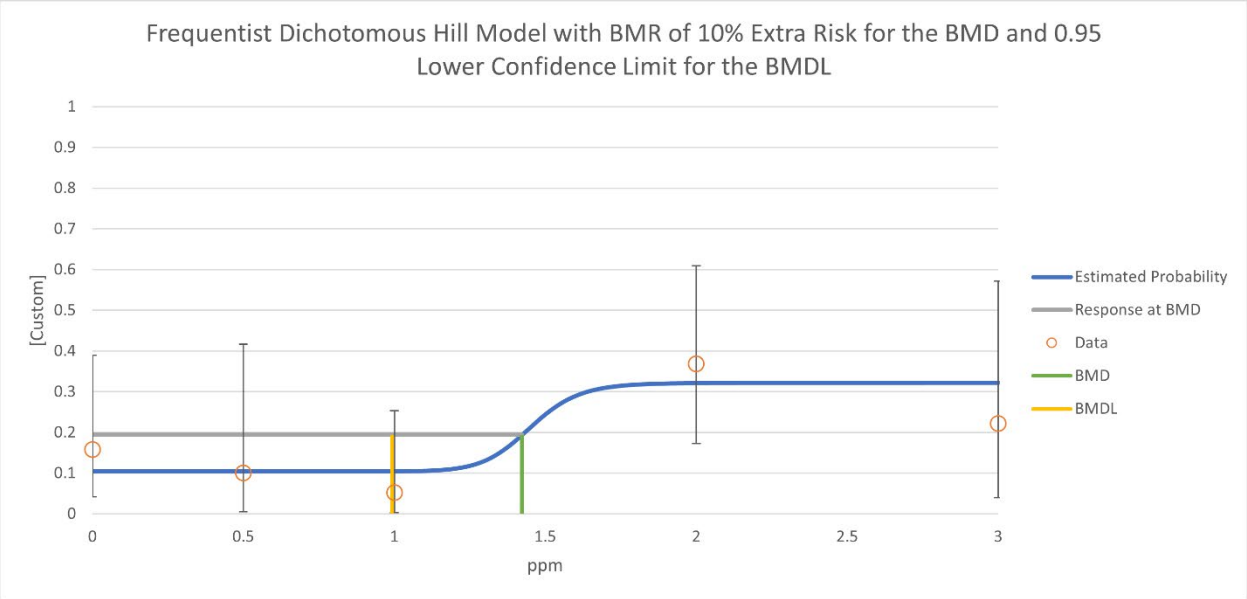

Figure 13. BMDS Version 3.3rc10 Fitted Dichotomous Hill Dose-Response Model for Nose/Throat Irritation.

The wide confidence intervals for the parameters show a large uncertainty in the fitted model. The dose-response curve increases from about 0.1 at 0 ppm to about 0.3 at 3 ppm.

Benchmark Dose Modeling Results Adjusted for Dependence

The above benchmark dose modeling results assume independence between the results at different doses. Since the available data did not include individual results, the IRIS report (EPA, 2022, page 2-9) chose to use BMD/2 (they refer to this as BMC/2) to approximate the BMDL to account for clustering, i.e., dependence between repeated measures at different doses. It is reasonable to assume that clustering will increase the width of the confidence intervals. However, the choice of a factor of 2 seems

arbitrary and no rationale is provided for why this approximates the BMDL. Importantly, the true BMDL is a lower confidence limit for the dose at which the extra risk is 10%, and it therefore it needs to account both for the uncertainty (variance) in the parameter estimates **and** the extra uncertainty due to correlation between repeated measures. To demonstrate this point about the arbitrary nature of the factor 2, suppose that the best model for eye irritation based on the AIC had been the Multistage Degree 3 model and not the Probit model. In that case the BMD and BMDL assuming independence would have been 0.863 and 0.369 and BMD/2 would have been 0.431 which is higher than the BMDL calculated assuming independence. Since clustering generally decreases the effective sample size, and thus increases the parameter uncertainty, the true BMDL should be lower than the BMDL calculated assuming independence. At a minimum, I suggest that any adjustment to the BMDL to account for correlations should be applied to the BMDL from the model assuming independence instead of the BMD. One possibility might have been to divide the original BMDL by 2 instead of dividing the BMD by 2. However, without having good data from similar studies with repeated measures to use in a simulation study the numerical value of such an adjustment factor seems totally arbitrary. The following Table 28 summarizes the BMD and BMDL values for the selected models based on BMDS Version 3.3rc10, with and without the adjustment factor. The unadjusted and adjusted BMD and BMDL values in this table have been converted from ppm to mg/m^3 units using a conversion factor for formaldehyde in air at 25 °C of 1 ppm = 1.23 mg/m^3. As noted above, the BMD and BMDL values are the same for BMDS Version 2.2, except that the BMD and BMDL for the Dichotomous Hill and Multistage Degree 1 models are not reported in Version 2.2.

Table 28. BMDS Version 3.3rc10 BMD, BMDL, BMD/2, and BMDL/2 Values from Kulle (1993)

Outcome	Odor Sensation	Eye Irritation	Nose/Throat Irritation
BMD (mg/m^3)	0.320	0.853	1.750
BMDL (mg/m^3)	0.224	0.617	1.220
BMD/2 (mg/m^3)	0.160	0.427	0.875
BMDL/2 (mg/m^3)	0.112	0.309	0.610

4. References

Anderson, I (1979). Formaldehyde in the Indoor Environment – Health Implications and the Setting of Standards. *In* Indoor Climate 5, edited by PO Fanger and O Valbjorn. Danish Building Research Institute.

Anderson and Molhave (1983). Controlled Human Studies with Formaldehyde. *In* Formaldehyde Toxicity, Chapter 14, James E Gibson (ed). Hemisphere Publishing Corp.

EPA (2012). Benchmark Dose Technical Guidance. Risk assessment Forum, US Environmental Protection Agency. EPA/100/R-12/001. June 2012.

EPA (2022). Toxicological Review of Formaldehyde – Inhalation. [CASRN 50-00-0]. External Review Draft. EPA/635/R-22/039a. April 2022.

Kulle T. J, Sauder L. R., Hebel J. R, Green, D. J, Chatham, M. D (1987). Formaldehyde Dose Response in Healthy Nonsmokers. *J. Air Pollut. Control Assoc.* 37:919-924.

Kulle T. J (1993) Acute Odor and Irritation Response in Healthy Nonsmokers with Formaldehyde Exposure. *Inhalation Toxicology* 5:3:323-332.