Freshwater Sediment Standards

Reliability Statistics

As suggested by EPA, calculations of additional statistical measures for the freshwater SQVs have been conducted. The methods and results are described below.

1. Data Set

Each additional statistic was calculated for each individual bioassay endpoint at both the SQS and CSL levels. In addition, the statistics were calculated for a pooled dataset, using the proposed draft SQS and CSL values. This presented some difficulties, as not all stations in the data set have enough bioassays to meet the rule interpretation criteria for bioassays endpoints (3 acute and chronic endpoints). Therefore, only the stations that contained enough data to determine a hit/no-hit status as described in the SMS were included in the pooled analysis, as follows:

- Stations with 3 or more acute and chronic bioassays were retained
- Stations with 1 or 2 bioassays were retained if they were hit stations
- Stations with 1 or 2 bioassays were not retained if they were no-hit stations, since it is not known whether they would be hit stations if they had additional bioassays

Once the dataset was screened to the appropriate set of stations, the one-hit and two-hit rules were applied as described in the marine standards to identify pooled hit/no-hit status for each station, assuming that the same one-hit and two-hit decision framework will be applied to freshwater data. At that point, the analysis was performed exactly as for the individual endpoints, using the test row of the Criterion Worksheet in the FPMCalc.xlsx spreadsheet.

2. Statistical Measures

EPA suggested that a variety of statistical measures be used that would be less affected or not affected by the prevalence of hits and no-hits in the data set. The following additional statistical measures were agreed upon between EPA and Ecology:

- Bias
- Odds Ratio
- Hanssen-Kuipers Discriminant
- Post-Test Odds Ratio

2.1 Bias

Bias is defined as the number of samples predicted to be toxic divided by the number of samples that are actually toxic. Thus, bias provides a simple measure of how conservative a set of standards is:

- Bias > 1 indicates that the SQGs are conservative and over-predict toxicity
- Bias = 1 indicates that the SQGs are appropriately predictive
- Bias < 1 indicates that the SQGs are under conservative and under-predict toxicity

2.2 Odds Ratio

The odds ratio indicates the strength of a prediction of toxicity, either for a given chemical or for a group of chemicals. The odds ratio is calculated as the likelihood that a prediction of toxicity is correct over the likelihood that a prediction of toxicity is incorrect. Thus, an odds ratio of 5 indicates that if there is a prediction of toxicity, the sample is 5 times more likely to actually be toxic than not. A higher odds ratio indicates stronger predictive capability.

2.3 Hanssen-Kuipers Discriminant

The Hanssen-Kuipers Discriminant is used to evaluate the fit of a model, and is frequently used to evaluate logistic regression models (but can be used for any model). The Hanssen-Kuipers Discriminant is a less general version of the Kappa statistic, and is used in cases where the prevalence of hits and no-hits in the data set is skewed. This statistic is believed to be unaffected by prevalence and ranges from 0–1. Like r^2, a Hanssen-Kuipers Discriminant value closer to 1 represents a better fit to the data. The following framework has been proposed in the epidemiological literature for interpreting model fit; however, this classification scheme is somewhat arbitrary and may or may not translate well to environmental data:

- κ between .01-.20 = slight
- κ between .21-.40 = fair
- κ between .41.-60 = moderate
- κ between .61-.80 = substantial
- κ between .81-1 = nearly perfect

2.4 Post-Test Odds Ratio

Each of the above statistics provides a measure of how predictive the SQVs are and/or whether they are a good fit to the biological data. The Post-Test Odds Ratio allows a site manager, PLP, or consultant to evaluate chemical results at a station and predict the likelihood that the sample will pass or fail an individual bioassay test or a suite of bioassay tests, based on the specific chemical exceedances in the sample. The post-test odds ratio is calculated as follows:

$$\textit{Post-test odds ratio} = \textit{Pre-test odds ratio} \times LR_1 \times LR_2 \times \ldots \times LR_n$$

where:

- Pretest odds ratio = the value calculated in Section 2.2 for an individual endpoint or SQV set
- LR = positive likelihood ratio for each chemical's sediment quality benchmark (true positive rate divided by false positive rate for that chemical)

- n = number of chemicals at station whose sediment quality benchmarks are exceeded

Because this is a sample-specific calculation, we cannot evaluate it now, but can provide the equation for calculating it, as well as the Odds Ratios calculated in section 2.2 and the likelihood ratios for the individual chemicals. With this information, the post-test odds ratio can be calculated for any individual sample against any single bioassay endpoint or the pooled interpretive endpoint provided in the SMS. This information could be used to decide whether to conduct bioassay override testing, and may also provide a basis for prioritizing areas for testing or cleanup by giving a relative measure of their potential for benthic toxicity.

3. Results and Discussion

The following results were obtained for the individual and pooled data sets.

3.1 Bias

Table 1. Bias at the SQS Level for Individual and Pooled Endpoints

Endpoint	Correct Predicted Hits	Correct Predicted No-Hits	False Predicted Hits	False Predicted No-Hits	Bias
Chironomus 10-day growth	52	382	78	13	**2.0**
Chironomus 10-day mortality	68	380	103	17	**2.0**
Hyalella 10-day mortality	72	187	90	17	**1.8**
Hyalella 28-day growth	21	43	10	5	**1.2**
Hyalella 28-day mortality	38	232	33	9	**1.5**
Pooled	**179**	**173**	**191**	**39**	**1.7**

DRAFT

Table 2. Bias at the CSL Level for Individual and Pooled Endpoints

Endpoint	Correct Predicted Hits	Correct Predicted No-Hits	False Predicted Hits	False Predicted No-Hits	Bias
Chironomus 10-day growth	40	417	59	9	**2.0**
Chironomus 10-day mortality	54	375	126	13	**2.7**
Hyalella 10-day mortality	39	236	78	13	**2.3**
Hyalella 28-day growth	10	58	9	2	**1.6**
Hyalella 28-day mortality	22	273	12	5	**1.3**
Pooled	**47**	**352**	**74**	**60**	**1.1**

Bias ranged from 1.2 – 2.7 for all individual endpoints, and from 1.1 – 1.7 for pooled endpoints. It is interesting to note that all chronic and pooled endpoints had lower bias than all acute endpoints. The reason for this is not known, although it may be that the chronic tests have a larger number of true hits, which may lower the bias. The pooled data sets had an even larger percentage of true hits than the overall data set due to exclusion of no-hit stations with only one or two bioassays, which also likely lowered the bias.

Ranges of bias were comparable for the SQS and CSL levels among individual endpoints. However, for the proposed SQS and CSL standards evaluated using the pooled hit/no-hit definitions, the bias was more conservative at the SQS level and approximately = 1 at the CSL level. This suggests that we have combined endpoints appropriately in selecting the final criteria, erring on the conservative side for the SQS and achieving a good balance at the CSL level.

3.2 Odds Ratios

The number of correct and false hits and no-hits are the same as shown in Tables 1 and 2.

Table 3. Odds Ratios at the SQS and CSL Levels for Individual and Pooled Endpoints

Endpoint	SQS	CSL
Chironomus 10-day growth	20	31
Chironomus 10-day mortality	15	12
Hyalella 10-day mortality	8.8	9.1
Hyalella 28-day growth	18	32
Hyalella 28-day mortality	30	100
Pooled	**4.2**	**3.7**

DRAFT

The Odds Ratios range from 9 – 15 for acute mortality endpoints, and from 18 – 100 for chronic and growth endpoints, suggesting that exceedance of SQVs for acute mortality endpoints is less likely to result in actual measurable toxicity in that bioassay. However, these odds ratios are quite high for all individual endpoints, resulting in approximately a 1 – 10% chance of not seeing an effect.

The pooled endpoints, representing the promulgated SQVs, have somewhat lower odds ratios. This suggests that the SQVs proposed for promulgation are somewhat more conservative than the values for the individual bioassays. The SQVs for both SQS and CSL levels have odds ratios of roughly 4: 1, suggesting an 80% likelihood of exceeding the biological standards using the interpretive criteria in the rule given an SQV exceedance. These are still quite reasonable odds, in line with the policy goals used to calculate the guideline (80% overall accuracy, maximum of20% false negatives and 20% false positives).

3.3 Hanssen-Kuipers Discriminant

The number of correct and false hits and no-hits are the same as shown in Tables 1 and 2.

Table 4. Hanssen-Kuipers Discriminants at the SQS and CSL Levels for Individual and Pooled Endpoints

Endpoint	**SQS**	**CSL**
Chironomus 10-day growth	0.63	0.69
Chironomus 10-day mortality	0.59	0.55
Hyalella 10-day mortality	0.48	0.50
Hyalella 28-day growth	0.62	0.70
Hyalella 28-day mortality	0.68	0.77
Pooled	**0.30**	**0.27**

The results for the Hanssen-Kuipers Discriminants suggest models with “moderate” or “substantial” fits for each individual endpoint. The best fits are again for the growth and chronic endpoints, with slightly lower values for the acute toxicity endpoints.

The pooled endpoints have results suggestive of only “fair” fits. However, it is worth noting that these represent combined SQVs that were not calculated using the FPM model, but were instead developed by the agencies through selecting the lowest or second-lowest of the individual endpoint values. Having not been developed using a modeling process, it may not be reasonable to expect the higher degrees of fitness to the data that the individual endpoints show. In addition, not all the data were suitable for use in the pooled endpoint evaluation, as discussed in the Data Set section.

DRAFT

3.4 Post-Test Odds Ratio

As described in section 2.4, calculation of the Post-Test Odds Ratio for a given future sample with chemistry data requires knowing the pre-test odds ratio, as well as the likelihood ratios for individual chemicals. The pre-test odds ratios applicable to individual bioassays or to the combined SMS biological interpretive criteria are given in Section 3.2. This section will provide the chemical-specific likelihood ratios for each individual endpoint and the combined SMS biological interpretive criteria for use in the equation. The associated tables are currently not available, as they will take some time to complete and will be prepared in May/June once the criteria are entirely finalized.

DRAFT