External Peer Review

U. S. Environmental Protection Agency
BMDS Final Report

Final Compilation of Reviewer Comments
And Responses to Charge Questions

Prepared for
Integrated Risk Information System (IRIS) Program
Office of Research and Development
National Center for Environmental Assessment
U.S. Environmental Protection Agency

Prepared by
ORISE IRIS Technical Assistance Team
Oak Ridge Institute for Science and Education
Oak Ridge Associated Universities

June 2007

This document was prepared for the EPA by ORISE under interagency agreement No. DW-89939822-01-0 between EPA and the U.S. Department of Energy. ORISE is managed by Oak Ridge Associated Universities under a contract with DOE.

EPA BMDS Review

TABLE OF CONTENTS

EXTERNAL PEER REVIEWERS

Ralph L. Kodell, Ph.D.
University of Arkansas for Medical Sciences

Louise Ryan, Ph.D.
Harvard School of Public Health

R. Webster West, Ph.D.
Texas A&M University

The ORISE IRIS Technical Assistance Team has neither altered nor edited these comments for grammatical or other errors.

PEER REVIEW PROJECT

John Fox
Project Manager

Leslie Shapard, Peer Review Manager
Oak Ridge Institute for Science and Education

CHARGE TO EXTERNAL REVIEWERS FOR THE IRIS ASSESSMENT OF BMDS

Has the software been adequately tested with respect to functionality, providing correct results, and handling exceptions and errors?

Based on the record provided in the development (or methodology) and testing reports:

(a) Have the estimators been implemented correctly?

(b) Has the accuracy of estimates been verified, using alternative software or custom programs, for an appropriate range of data and parameter values?

(c) Is the record provided in the development and testing reports sufficient to document the algorithms used and results of software testing?

Is the model documentation and the reporting of results 'as good as' and consistent with that for existing BMDS quantal models? (i.e., will users of BMDS find it as easy to use these models and interpret the results as they do for existing BMDS quantal models?)

Is the user documentation ("Help File") clear and correct, and does it explain the application of the models well?

Are there any aspects of software development and testing, or model user documentation (BMDS Help files), or reporting of model results (user GUI and *.out file) that give you special cause for concern? If so, please describe your concerns and recommendations.

IMPORTANT NOTICE
(July 30, 2007, post-review)

Only seven of the models reviewed here will be released publicly with BMDS version 2 in 2007. Those are the new versions of quantal models having a background term additive to dose (the multistage, cancer, log-probit, gamma and Weibull models) and the new versions of the probit and logistic models having an explicit background response term. The MSW time to tumor model is undergoing revision (and, as stated in the review materials, is intended for internal use by EPA staff, not for public release, at this time). The log-logistic model with background additive to dose is not being released because of unresolved difficulties with convergence on solutions, as stated in the review materials.

RESPONSE FROM RALPH L. KODELL

Question 1

Has the software been adequately tested with respect to functionality, providing correct results, and handling exceptions and errors?

Answer 1

Unit testing was done on each individual functional unit. Results from modeling test data sets with the BMDS C programs were compared against published results, hand calculations, or results from other software packages. For the quantal models, SAS was the first choice as alternative software, while Mathematica was used for cases when SAS didn't converge. Also, Mathematica was used to verify confidence limit calculations for slopes of selected models. For the multistage Weibull model, TOX RISK was used for comparison. Generally, tests were considered successful if results matched to three decimal places.

Many different data sets having different features were tested in order to stress the code. A variety of interventions were implemented in the code to ensure reliable results for difficult cases. Most test data sets involved dose groups of size 10. As was clarified in the conference call May 22, 2007, a group size of 10 was thought to be reasonable for numerical testing, and the high-background problems described in the documentation were not related to sample size but rather to placement of doses. The limited range of dose-related response was also mentioned as a possible reason for the problems encountered in high-background cases. Some "off-line" testing of data sets with group sizes of 50 gave similar results to group sizes of 10. Some issues remain open for further investigation (e.g., scaling of doses). In a few extreme cases, errors in execution may still occur, in which cases error messages are produced. This is a good feature.

The Introduction and Background for Reviewers document indicates that the log-logistic model with background dose parameter is not included in the review because it is not yet suitable for public release. I agree that it needs further testing. In particular, the *ad hoc* upper bound constraint of 1 on the background dose parameter needs further justification. The Weibull model with background dose parameter proved difficult for SAS, so the validation of that case may be less complete than the other models.

System integration testing was done to verify that each module would fully integrate with the BMDS shell, except that the graphical user interface was not tested.

In addition, existing models in BMDS were tested before and after the new models were added to ensure that they worked the same as before.

I believe the software has been adequately tested.

Question 2

Based on the record provided in the development (or methodology) and testing reports:

(a) Have the estimators been implemented correctly?

(b) Has the accuracy of estimates been verified, using alternative software or custom programs, for an appropriate range of data and parameter values?

(c) Is the record provided in the development and testing reports sufficient to document the algorithms used and results of software testing?

Answer 2

(a) For all the quantal models, I believe the estimators have been implemented correctly. For the multistage Weibull model, I have some concerns, which I will discuss in detail in response to question 5.

(b) As indicated under Answer 1 above, the results of the quantal models have been tested against SAS and/or Mathematica for a variety of dose-response situations, including extreme cases. However, some extreme cases that were discussed during the May 22 conference call, such as 0% response at zero dose and 100% response at the highest dose, were tested "off-line" but not included in the documentation. For the cases considered, the accuracy of the estimates has been verified to three decimal places, with only a few exceptions. As was clarified in the conference call on May 22, the SAS results of Wheeler (SUGI, 2005) matched results of existing BMDS models, so his SAS code seemed like a natural place to start, and the SAS code supplied some degree of independence.

(c) I believe that the record provided in the development and testing reports is sufficient to document the algorithms used and results of software testing. The discussion of the modeling and results inspires confidence in the code.

Question 3

Is the model documentation and the reporting of results 'as good as' and consistent with that for existing BMDS quantal models? (i.e., will users of BMDS find it as easy to use these models and interpret the results as they do for existing BMDS quantal models?)

Answer 3

The model documentation and the reporting of results appear to be 'as good as' and consistent with that for existing BMDS quantal models. However, restrictions on model parameters seem not always to be stated. For example, I couldn't tell if the shape parameter of the quantal Weibull model is restricted to being greater than or equal to 1. Some restrictions are obvious, but others are not.

I did have a little trouble interpreting the results of the multistage and logistic models that give estimates of slopes at/near the BMD(BMDL). For the examples provided, the output provides an estimate of the "Slope at max:" Shouldn't this be labeled "Slope at BMD" or Slope at BMDL"?

Question 4

Is the user documentation ("Help File") clear and correct, and does it explain the application of the models well?

Answer 4

I disagree with the following statement on the second page of the Help file (Quantal Models with Background Dose Parameter): "The background parameter should not be interpreted literally or mechanistically as an internal equivalent to an applied dose (unless there is independent evidence to support a mechanistic interpretation)." The inclusion of a background dose parameter is done to reflect a specific dose-additive mechanism, and I think it should have that interpretation. I do agree with the qualifying statements on page 5 that follow a similar statement to that quoted above. I agree that, just because one can fit a background dose parameter doesn't mean that the underlying dose-response relationship is truly dose additive. This was discussed in the May 22 conference call and it was indicated that the language would be softened.

I suggested during the May 22 call that more explanation be provided for the logistic and probit models. Because a background response parameter has been added for these two models, with the "implied" background dose parameter left in, these 3-parameter models are really hybrids that include both background dose (implicitly) and background response (explicitly). I asked if the EPA had considered adding 2-parameter background response logistic and probit models where $\alpha=0$, if such models make sense and will fit data. The answer was yes, the models had been considered and some runs had been made but not included in the documentation. It was pointed out that users can restrict α to be zero in the 3-parameter models. I mentioned that it might be useful to highlight this in the help file, so that users can have 2-paremeter models of background dose and background response to compare. Subsequent to the call, the EPA Project Manager provided results of several examples using the 3-parameter logistic and probit models, and the 2-parameter background dose and background response logistic models. The results were very informative. In particular, it was pointed out that 2-parameter background response models are applicable only when the empirical response probability at zero dose is at least 0.5. The user help file will be modified to include guidance on the use of these models. This will be a valuable addition, and the manner of response by EPA to this particular question inspires additional confidence in the overall BMDS product.

The previous two paragraphs are related to an issue that came up during the May 22 call. That is, how are users to interpret model parameters that are not statistically significant? For example, should they fit a reduced model that doesn't include a nonsignificant parameter? I'm not sure. With models that contain both background dose and background response parameters (*logistic_bgr* and *probit_bgr*), a likelihood ratio testing approach might be used to determine if reduced models are adequate to describe the data. Thus, a user could possibly choose one or the other of the background models as the final model. However, one would still need to be cautions

about making definitive mechanistic interpretations without, as the document says, independent information about mechanisms. It was noted in the documentation for this review that a version of the multistage model with both types of background parameters was investigated, but was abandoned because of numerical complications. Furthermore, in the Introduction and Background document, it was stated that it was not the objective of EPA to enable users to estimate both parameters simultaneously for a number of reasons. Perhaps this area is something to consider for future releases of BMDS.

I had trouble distinguishing the properties of the multistage model and the 'cancer model.' It was clarified in the May 22 conference call that the parameters of the multistage model can be unrestricted, with the default being a non-negativity restriction, while those of the cancer model are restricted to be nonnegative. There are places in the some of the documentation where this is not clear. I had not realized, or perhaps had forgotten, that in the present BMDS version, one can elect not to restrict the multistage model's parameters.

Question 5

Are there any aspects of software development and testing, or model user documentation (BMDS Help files), or reporting of model results (user GUI and *.out file) that give you special cause for concern? If so, please describe your concerns and recommendations.

Answer 5

I have some questions and concerns about the multistage Weibull (MSW) model.

Has the MSW been tested for BMD calculation at any other time than the end of the study (e.g., 104 weeks)? It might not be a big deal, but it should work for any t, right?

I was confused when I questioned the restriction on the shape parameter, c, during the May 22 conference call. (I may have been thinking of the quantal Weibull model where c is the power of dose, not time.) Instead of allowing c to be less than 1 in the MSW as I mentioned during the call, I now think that even constraining c to be greater than or equal to 1 might not be strict enough. The documentation mentions some problems for c between 1 and 2. The more I think about it, I'm not surprised. I haven't thought about the MSW in a while, but it is really a version of the multistage model for continuous dosing where time is explicitly included, rather than being absorbed into the dose parameters because of the model's being evaluated at a single, specific time. With this interpretation, I believe that c is the total number of stages in the model, while the highest power of dose, k, is the number of those c stages that are dose-related. So, I believe that it would be reasonable to restrict $c \geq k$ (and maybe even have c an integer). Regardless of the implementation in TOX RISK, I suggest the developers consider offering the option to impose $c \geq k$. As noted in Appendix A of the methodology document, c has the biggest impact on the restrictions on the GEV parameters, because all are functions of c.

It might just be due to my naiveté, but I'm surprised that the (constrained) log-likelihood function for the MSW is not necessarily unimodal. This makes me want to question the likelihood contributions of the various types of observations, and especially the interpretation of

the location parameter, t_0. If t_0 is interpreted as the lag time between onset of and death from tumor and T_{DT} is the time to death from tumor, then the time to onset is T_{DT}-t_0. With this interpretation, it seems to me that

$$F(t,d) = 1 - \exp\{-(t - t_0)^c \sum_{i=0}^{k} \beta_i d^i\}$$

is the cumulative distribution function for time to onset of tumor, not time to death from tumor, where F=0 for $t<t_0$. It has been noted in the past that if one assumes a constant lag time between onset of and death from tumor, then a nonparametric test like Peto's fatal tumor test can be interpreted as a test that compares time-to-onset distributions. Could the MSW model with t_0 interpreted as the lag time between onset and death actually represent the time to onset distribution and not the time to death distribution? If so, I think it's ill advised; if not, then I think the interpretation of t_0 is wrong.

On the other hand, if t_0 were interpreted as the minimum time necessary for a tumor to develop (or the minimum time for a tumor to develop and become fatal), then it seems that the above cdf would indeed be the cdf for time to death from tumor. I note from the Introduction and Background for Reviewers document that the parameter t_0 often has its MLE at the value of the smallest observed time for Incidental, Fatal, and Unknown contexts. This was also discussed during the conference call May 22, and it was mentioned that TOX RISK behaves similarly in this respect. In the BMDS document it was stated that the software development team does not have confidence in the interpretation of t_0 as the lag time between onset and death from tumor. I share this lack of confidence. The data might be indicating that t_0 should not be interpreted as the lag time between onset and death, but rather as the minimum time to tumor (or minimum time to death from tumor). Of course, the likelihood contributions are dependent on the interpretation of t_0, and a different interpretation could (should) change the contributions.

I have not had time to think it through, but I wonder if the non-unimodal behavior of the likelihood function is related to the questionable interpretation of t_0. Even if not, I think the modeling experts at EPA/NCEA and Battelle should take a close look at how the MSW is implemented for time to death from tumor with respect to the interpretation of t_0, although this would not necessarily be a trivial undertaking. With the present formulations, there might be some mathematical inconsistencies that are leading to problems, or that might lead to incorrect interpretations. I urge caution in releasing the MSW module without additional investigation. At the very least, I recommend including the development team's concerns in the help file.

RESPONSE FROM LOUISE RYAN

Has the software been adequately tested with respect to functionality, providing correct results, and handling exceptions and errors?

In general, I do feel that the software has been well tested. There seem to be good QA systems in place for

1. Isolating existing code to make sure that the addition of new code does not interfere with existing code
2. For testing out distinct modules (unit testing)
3. For testing the overall suite of programs (system integration testing)
4. For double checking the results again those obtained with other programs.

Discussions about numerical precision and compiler issues appear to be very thorough. I do have a couple of specific comments however.

- The background report says that SAS and Mathematica were used to independently replicate the results obtained from the EPA software (also ToxRisk in the case of the MSW model). It would be useful to provide a little more detail about how this was accomplished. For example, SAS is a very diverse set of procedures. What procedure was used for model fitting? As indicated below, I also think that R is a very useful tool for checking and would encourage its use.

- As indicated below, I think there are aspects of the software that need to have some additional, deeper statistical thinking applied. In certain cases where there are convergence problems etc, my belief is that the issue is less numerical/computational that statistical (in the sense of thinking about what models are appropriate and/or fit well in certain sorts of settings). A good example is the discussion on page 4 of the Background report where it talks about the idea of including both background response and background dose parameters in the model. The discussion addresses this from a numerical stability issue whereas I think that the issue is really one of identifiability. It simply would not make sense from a statistical perspective to fit a model with both different types of background parameters included. The basic problem, I believe, is that some of the models can be very poorly identified, especially for certain data configurations. I think that it might be a good idea to invest some time and resources to address these identifiability issues more carefully and to build into the software some warnings that go beyond simply reporting on numerical issues, but perhaps offering advice that certain models are so poorly identified that the results may not be very trustworthy. In general, I feel that numerical aspects of the convergence issues have been well addressed (e.g. use of Safe Exponential and Log Functions; parameter scaling). One exception to this is the incorporation of the upper limit on the background dose parameter for the log-logistic. While I know that the intention is not to release this particular module at this time, I would like to say for the record that I think use of this upper limit is inappropriate. I also think that based on the theoretical structure of the models being considered, similar

considerations should apply for all the models that incorporate a background dose into the dose response model with logged dose, not just the logistic

- I also think that there are some problems related to the computation of confidence intervals etc, but again, more statistical than computational (see below).

- As raised in the conference call discussing this review, it is not clear to me that a wide enough variety of data scenarios has been developed to fully test out the models, especially at extremes (such as 0 responses in the control group etc).

- As suggested by another reviewer on the conference call, I think it would be extremely helpful to conduct some classical simulations where empirical data are repeatedly generated from a known true model and then the results of fitting various models compared to the expected true values. It would be useful to do this not only for settings where the fitted and data-generating models are the same, but also where they differ. In the latter case,

Based on the record provided in the development (or methodology) and testing reports:
(a) have the estimators been implemented correctly?

In a number of cases, I do not feel that the reports provide sufficient technical details for me to fully comment. For example, the Multistage Weibiull Model generally requires the specification of constraints that for the response rate to always lie between 0 and 1. In general, the programming required to enforce such constraints can be challenging. A simple and commonly used alternative is to force all the β coefficients to be positive. The document should provide more detail about which approach was used.

I have a number of concerns about the approaches taken to computing confidence limits (and relatedly, BMDL and BMDU values). When constraints are incorporated, then extra care is needed when doing inference. From a theoretical perspective, so long as the maximum likelihood solution is not on the boundary of the parameter space, then the usual inferential procedure (based on computation of the information matrix) can be used as the basis for computing confidence limits. In finite samples, however, standard inference will often break down for maximum likelihood involving constraints (see Self and Liang (Asymptotic Properties Of Maximum-Likelihood Estimators And Likelihood Ratio Tests Under Nonstandard Conditions, Journal of the American Statistical Association 82: 605-610 1987; also Molenberghs, Likelihood ratio, score, and Wald tests in a constrained parameter space, American Statistician 61: 22-27 2007). In more complicated settings, many people recommend the use of approaches such as the bootstrap. Of course this particular issue needs to be addressed not only for the current software expansions, but also the original BMDS.

I have concerns about the starting values. While the chosen approach seems reasonable, I always feel that it is a good idea to double check a model fit by starting the algorithms at several different starting points. It would not be too difficult to build this into the software.

In a number of places in the background, I feel that the discussion is rather confusing. In some places, it sounds as though the purpose of the extended BMDS models were to add in a background rate, either through the γ parameter (the so called background response model) or through the η parameter (the so called background dose model). However, table 1 of the document called "BG_Dose Quantal Model Development" seems to imply that in some cases, one or other of these background models is already incorporated in to the existing BMDS software. This needs to be clarified.

In certain places, the discussion goes beyond being simply confusing to become nonsensical and wrong. For example, page 9 of "BG_Dose Quantal Model Development" says "The new log likelihood function is now expressed in terms of parameters β, γ and BMD and is minimized to find the lower limit of BMD." Minimizing this re-parameterized likelihood will yield the mle of BMD, NOT its lower confidence limit. The following sentence makes no sense (in my opinion): "The issue with this approach is that the intercept parameter is eliminated from the model and depending on the shape of the likelihood function; the BMDL estimate might not be the true minimum that satisfies the constraints".

(b) has the accuracy of estimates been verified, using alternative software or custom programs, for an appropriate range of data and parameter values?

The background report says that SAS and Mathematica were used to independently replicate the results obtained from the EPA software. This is great. However, it would be useful to provide a little more detail about how this was accomplished. For example, SAS is a very diverse set of procedures. What procedure was used for model fitting? Also, I also think that R is a very useful tool for checking and would encourage its use. As indicated in the conference call, I think it could be useful to test the software with a broader range of datasets. Also some traditional simulations would be helpful.

(c) Is the record provided in the development and testing reports sufficient to document the algorithms used and results of software testing?

In general, documentation is fairly good, though there are some areas where improvement is needed. On some topics (especially the more computational aspects such as numerical precision), documentation is excellent. Documentation on issues that are more closely aligned to statistical concepts is often poorer. For example, discussion about the incorporation of parameter constraints is poor, often confusing and sometimes barely addressed. For several of the models (e.g. multistage and also the multistage Weibull), parameter constraints play and important role. The method used to incorporate the constraints should be used. Also, there needs to be considerably more thought and discussion on the issue of how parameter constraints are incorporated into confidence limit calculations.

Is the model documentation and the reporting of results 'as good as' and consistent with that for existing BMDS quantal models? (i.e., will users of BMDS find it as easy to use these models and interpret the results as they do for existing BMDS quantal models?)

I am afraid that I have not used BMDS much, so I cannot reliably comment on this point. I would say, however, that the example output shown for the MSW model does not appear as well laid out as it might be.

Is the user documentation ("Help File") clear and correct, and does it explain the application of the models well?

The help files are quite good, in general easy to read and reasonably detailed. I would suggest adding some discussion about the use of constraints for some of the models (e.g. multistage). I also think that some of the figures in the help files could be made more informative if based on full scale simulations (rather than the relatively adhoc procedure of fitting "data" that correspond to the expected number of adverse events at each dose group). I would like to see just a little more detail about how upper and lower limits were computed as well.

I don't think that the multistage Weibull model is described as well as it could be. On page 3 of the MSW Time to Tumor Methodology Report, the model is described by

$$F(t,d) = F(t,d,t_0,c,\beta_0,\beta_1,...,\beta_k) = 1 - \exp\left\{-(t-t_0)^c \sum_{i=0}^{k} \beta_i d^i\right\}$$

where model parameters satisfy the restrictions $c \geq 1$, $t > t_0 \geq 0$, and $\beta_i \geq 0$ ($I = 0, 1, ..., k$). I believe it is more precise to replace $(t-t_0)^c$ as $(t-t_0)_+^c$ where $(t-t_0)_+^c = (t-t_0)^c$ if $t > t_0$ and 0 otherwise. I don't think it is correct to simply say that there is a restriction that $t > t_0$, although it may end up to be effectively true for most if not all fitted models. The same applies in the other documents, e.g. MSW Time to Tumor model description for users. In this latter file, t_0 should be mentioned in the section about parameters.

In the model development and testing report, would like to suggest that the term F(.) be clearly defined as a cdf (mapping the real line to the interval [0,1]).

Are there any aspects of software development and testing, or model user documentation (BMDS Help files), or reporting of model results (user GUI and *.out file) that give you special cause for concern? If so, please describe your concerns and recommendations.

1. In the background document, you indicate some convergence problems for the log-logistic model with background dose parameter at this time, specifically for the case where the zero response is high (e.g., > 40%) or for settings with a very flat response curve (e.g., 9% rising to 20%). I am surprised by this, since I don't consider a 40% control rate to be particularly high, nor a rise from 9% to 20% to be particularly flat! In fact, I would think that the numerical procedures should perform better in settings where there is a reasonably high response rate among controls. The background response parameter should be well defined in such settings. Hence, I am particularly concerned by the report about convergence problems

in these settings. I realize that you are not considering the log-logistic to be ready for release yet. However, the described approaches to handling some of the convergence issues concern me. For example, the "BG_dose…" file says that incorporation of the constraint to keep the background dose less than or equal to 1 was "based on the observation that with the constraint in place the results were validated by the output of SAS." To me, this is not at all a sound argument. First, we don't know that SAS is correct. I wrote a small R program to estimate models for the following dataset given in the report. The code (see Appendix) fits the log-logistic model using the non-linear maximizer in R. Interestingly, it seems to give very similar answers to what you describe as coming from the BMDS code. I then constructed a modified version of the code that fixed the background dose parameter and then varied between the values of 1 and 500. For the example data (see below), the likelihood does continue to creep up, so there is not true mle value. But in the first panel, I plot the predicted curves for each of the fitted models. Notice how similar they are. The bottom panel shows the log-likelihood for various values of the background parameter. The bottom line, I believe, from this exercise, is that this is a model where there is a very flat likelihood over a broad range of the data. There is no particular problem with the software, it is just that the model is not a good choice for these data.

Data
0 2 8
0.5 2 8
1.0 2 8
2.0 6 4
4.0 9 1

2. There are several concerns related to the report "MSW Time to Tumor model description for users".

    a. The report states: "When the data represent a mixture of observations with fatal (*F*) and incidental (*I*) tumor contexts, it is meaningful and usually feasible to estimate t_0 when modeling death from tumor using the MSW model for fatal tumors". I am concerned about this statement since it is fairly ambiguous. Is the report suggesting that an analysis be applied only to fatal tumors? How are the non-fatal tumors handled in such a setting?

    b. I am concerned that there is no discussion about the complexities of inference on t_0 in settings where it is estimated to lie on the boundary (e.g. at the smallest observed time of death from tumor).

    c. The statement "Maximum likelihood estimation may not lead to efficient estimators of percentiles for either the Weibull distribution or BMD" is problematic, since "efficient" has a very specific technical interpretation in the context of maximum likelihood.

d. In the section describing the mle calculations, the term "subject group" is not defined. It is not clear where the likelihood contribution came from for incidental tumors. My guess is that the contribution is constructed under the assumption that there is an exact time t_0 between tumor onset and death. However, I do not think it is appropriate to make this assumption. Personally, I would rather have the contribution of an incidental tumor at time t_{js} simply be $(1-F(t_{js},d))$, implying that it occurred sometime before t_{js}. By the way, with the existing formula, I think the first term should be $F(t_{js}-t_0,d))$.

e. The section entitled BMDL computation is very terse and almost impossible for most readers to follow. As discussed above, there really should be discussion about the impact of parameter restrictions on inference.

One other small comment

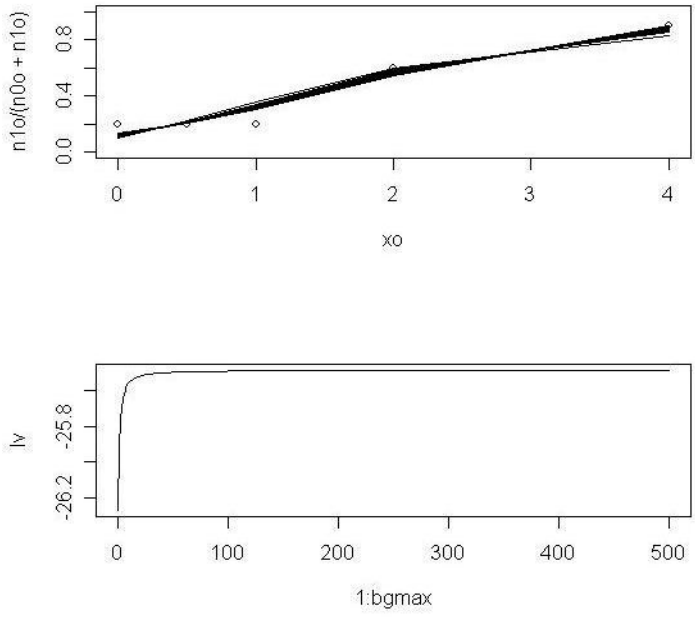
1. Why does the log-probit model have a constraint on the slope parameter to exceed one? (See Table 9 of background document)

Appendix:

```
#  R-Code to fit the log-logistic model as well as a modified version with
#  fixed power parameter

#### log-likelihood with no constraint on the parameters

llg=function(theta,x,n0,n1,prnt=F){
p=1/(1+exp(-(theta[1]+theta[2]*log(x+theta[3]))))
ll=sum(n0*log(1-p)+n1*log(p))
if (prnt) {print(theta)
print(p)
print(ll)
plot(x,n1/(n0+n1),ylim=c(0,1))
lines(x,p)}
-ll
}

#### log-likelihood with fixed value of background dose

llg.restricted=function(theta,bg,x,n0,n1,prnt=F){
p=1/(1+exp(-(theta[1]+theta[2]*log(x+bg))))
ll=sum(n0*log(1-p)+n1*log(p))
if (prnt) {print(theta)
print(p)
print(ll)
plot(x,n1/(n0+n1),ylim=c(0,1))
```

```
lines(x,p)}
-ll
}

###

theta0=c(-5, 2.37561,.5)
xo=c(0.0, 0.5, 1.0, 2.0, 4.0)
n0o=c(8, 8, 8, 4, 1)
n1o=c(2, 2, 2, 6, 9)

result.unrestricted=nlm(llg, p=theta0, x=xo,n0=n0o,n1=n1o,iterlim=1000)

par(mfrow=c(2,1))
##  run the restricted model over a range of values of the background
bgmax=500
lv=1:bgmax
plot(xo,n1o/(n0o+n1o),ylim=c(0,1))
for (bgr in 1:bgmax) {
result=nlm(llg.restricted, p=c(-5, 2.37561), bg=bgr, x=xo,n0=n0o,n1=n1o,iterlim=1000,prnt=F)
lv[bgr]=-result$minimum
prb=1/(1+exp(-(result$estimate[1]+result$estimate[2]*log(xo+bgr))))
lines(xo,prb)}
plot(1:bgmax,lv,type="l")
```

Figure 1



RESPONSE FROM R. WEBSTER WEST

For my review of the new additions to EPA's benchmark dose modeling software, I will loosely follow the general format of the charge to reviewers for each of the three areas up for review.

Has the software been adequately tested with respect to functionality, providing correct results, and handling exceptions and errors?

Based on the record provided in the development (or methodology) and testing reports:

(a) Have the estimators been implemented correctly?

(b) Has the accuracy of estimates been verified, using alternative software or custom programs, for an appropriate range of data and parameter values?

(c) Is the record provided in the development and testing reports sufficient to document the algorithms used and results of software testing?

1. Quantal Models with Background Parameter Additive to Dose or Additive to Response

On the whole, I would say yes to each of the above questions. Software testing is by no means a perfect process which is guaranteed to produce perfect code, but I feel the developers have set forth a reasonable testing plan and implemented it properly. However, I would like to add a few suggestions/comments here.

The software has currently been tested against output from SAS and Mathematica. SAS is pretty much an industry standard, so verifying output with SAS is a must. Also, I like the idea of implementing the routines in a more rudimentary language like Mathematica since this activity helps validate not only the output but also the process. However, I think additional testing should be considered because of the underlying properties of the likelihood functions used within BMDS. The likelihood surfaces for the dose response models to be fit are extremely bumpy with sometimes peculiar boundary behavior for the common designs used. For example, with the multistage model, there are frequently values along the β_1 axis (where β_2=0) and values along the β_2 axis (where $\beta_1 = 0$) that provide roughly the same value of the likelihood function. Because of this behavior, a grid search should be preformed to see if the optimizer achieves the true optimum value for the test cases considered. The large sample statistical properties of the resulting estimators are extremely dependent on achieving this true maximum. For the multistage model, one might do a grid search for all combinations of parameters over the range from 0 to 8 with a reasonably small step (assuming dose values have been divided by the largest dose). From the conference call, it appears that this may have been done informally. I only bring it up here to emphasize the point.

Given the changes to the underlying optimization code, I feel it is also important to test output against existing BMDS code when possible. For example, if one constrains the background dose

parameter to be 0 with the new optimization code, then the resulting estimators should be the same as those from the existing BMDS routines. A similar experiment can be done with the background response parameter. Once again, from the conference call, this may have been done, but it was not a part of the review document.

I also have some comments on the actual test sets that were considered. In my opinion, it would be a good idea to consider a few more extreme sets such as cases where the control has no response and the highest dose complete response (all developing cancer for example). These extremes can wreak havoc on the underlying likelihood function and they do occur often in practice. These extremes will also have an impact on many of the initial value routines. I assume the safe log function will work okay in these circumstances but this needs to be studied in more detail. I also feel it would be a good idea to test cases where the dose response data is actually inverted from the assumed risk relationship. The performance of the output in these situations may be useful in determining if the algorithms are working properly as well as serving as a test bed for error reporting. This may have been done but was not included as part of the review documents.

While some very large dose group sizes were considered among the test cases, it would also be a good idea to do a large scale simulation study where the true underlying parameters are known to verify the large sample properties of the estimates. Outliers from the simulation study may help determine potential problems with the optimization code that occur under specific conditions. Any potential biases in the optimization code might also stand out. Determining how large is "large" can be tricky with the models considered. In my experience, sample sizes of 100,000 or more may be required to reach the point where large sample behavior can be verified. Of course, these sample sizes would never occur in practice.

A few of the covariance estimates shown in Appendix A are distinctly different from the values given in SAS. In A.3 for example, the covariance between intercept and background dose is -0.5629 whereas the corresponding value from BMDS is -0.58. There may a problem with the values shown in this case as the BMDS output claims to be the covariance between background response and dose. Any problems in this area are likely to be caused by the numerical computation of the information matrix. I would not recommend its computation be done by numerical methods when analytical solutions are available. However, it is my understanding that this is the practice of the current BMDS system so this feature was maintained in the new routines.

Also, I feel I should comment on some of the convergence issues mentioned in the review documents when the response for controls was very high. My guess is that the reason for these issues is that this forces the overall dose response to be very shallow in the case of increasing response with dose. Therefore, the likelihood function will probably have multiple modes which will present problems for the optimization code. It will also make starting values much more critical.

2. Two Quantal Models Reporting Slope Of Dose-Response Function, With Confidence Interval

I realize that this is not the purpose of this review, but I must say that I think the inclusion of these methods may lead to a lower quality statistical practice. There are a number of serious

issues with this low dose linearization in my mind that are both practical and statistical. Regardless of my objections on statistical grounds, I found the tenor of the review document very uncertain for these methods. I found the detail of the discussion here to be quite interesting and written with a very honest tone, but these methods seem to be more at an exploratory stage rather than in final form. Therefore, I can not answer in the affirmative in terms of the charge questions above when it comes to these methods. The primary reason for this assessment is that the number of test cases considered is extremely small. I feel a great deal more validation is required before these methods go live. I suggest that a larger number of test cases be considered some of which should be extreme as discussed above. In this exploratory stage, I also feel that a simulation study is required to validate the coverage probabilities of the proposed methods since none is referenced in the literature. In short, I feel the majority of my comments from above apply doubly here. I also suggest that the developers consider other sources for validation such as the R programming language.

3. Multistage Weibull ("MSW") Time-to-Tumor Model

I should begin my review here by saying that I am not an expert in the time-to-tumor arena. I did, however, find the description of the methods in the review documents thorough enough that I feel I can make some comments. In my opinion, this model is overly complex for the majority of time-to-tumor data that I have seen. My guess is that the parameter t_0 is rarely estimable and when it is estimable that its interpretation is sketchy at best. In my opinion, it would be best to present the time-to tumor model first and then augment it with the fixed time t_0 to death. This way the real impact of the parameters makes more sense in that the βs are then connected to the time-to-tumor directly. This is quite important for the user guide. I suppose my biggest issue with the model is that the time to death from onset is fixed for each experimental unit. I am unaware of any researcher that believes this to generally be the case.

Assuming the mathematical development is correct and complete, I chose to focus here on the testing procedures. Testing is quite difficult here to due to the very complex nature of the model and the lack of commercially available software for fitting it. I do not consider ToxTools to be that much of standard, but I suppose it beats nothing. It would have been nice to see the data input formats along with the actual data for the models tested. Since only three data sets were considered and the matches to ToxTools were inconsistent in terms of benchmark dose, I do not feel that the testing is adequate for release. It appears that a number of issues must first be addressed. Once again, the model estimates needs to be compared to the true optima in a number of test cases where the truth is determined by a complete grid search of the parameter space. This is especially since there is no gold standard for comparison. After the comparison when the truth is known, the robustness of the software should be tested by applying it to extreme conditions. Since none of the data sets were provided, I can not tell if the sets tested were extreme in any way. Unfortunately, in this case extreme data may imply that the model is not estimable. However, this procedure could still be useful for error testing. Also, this is yet another situation where simulation should be conducted to test the software. While the real data sets are interesting, it is impossible to know the methods statistical properties without knowing the true values of the model parameters. Conducting a simulation study may provide a great deal of insights on the properties of the estimators provided by the software along with the benchmark dose lower and upper bounds.

There is sufficient documentation in the review materials to suggest that the intermittent component testing is adequate but I feel more should be done for the software as a whole. I also realize this is not an easy task.

My biggest concern here is that the software output may not be very useful even if it is correct. The rather bizarre nature of the profiles presented in the appendix show that instability of the method with the data tested. The optimization routine clearly has a preference towards the higher order coefficients. It is also interesting that t_0 is estimated to be zero in every single case. Clearly, more diversity in this regard is desirable. Indeed, the software may be working properly, but the model itself may have trouble fitting the observed data. I did notice that the profiles had stranger behavior for smaller risk values. This may be useful in diagnosing any issues.

Is the model documentation and the reporting of results 'as good as' and consistent with that for existing BMDS quantal models? (i.e., will users of BMDS find it as easy to use these models and interpret the results as they do for existing BMDS quantal models?)

I think this should not be a problem as the interface and output does not seem very different to me for any of the new methods.

Is the user documentation ("Help File") clear and correct, and does it explain the application of the models well?

Quantal Models With Background Dose Parameter

I think this document is sufficient. My only suggestion is that one might consider including a discussion of user options when either background dose or background response are estimated to be zero. While some statisticians may disagree, I personally feel that in most cases one should fit a reduced model because of a large reduction in the standard errors of the estimates. I am aware, however, that some may disagree. Perhaps the pros and cons in this regard should be added to encourage good statistical practice.

Examples of slope estimation with confidence limits

Clearly more documentation is required for the slope estimation procedure. A detailed example should be written up rather than just raw inputs and outputs as provided here. The documentation should also include some hints as to the suggested use and proper application of the procedure.

Multistage Weibull Time-to-Tumor Model Description

This document is clearly intended for a higher level user. I think it is a bit heavy on mathematical detail (it includes the likelihood function for example) and a bit light on implementation tips. What format should the data be in to apply this procedure? These sorts of

things should be included in the documentation along with detailed examples. This is even more important here as many people have trouble understanding censoring ideas correctly.

Are there any aspects of software development and testing, or model user documentation (BMDS Help files), or reporting of model results (user GUI and *.out file) that give you special cause for concern? If so, please describe your concerns and recommendations.

I think I have expressed most of my concerns above. However, I should say that I think the actual software itself needs to be extensively reviewed by an independent party before it is released. While here we are reviewing the testing procedures, a review of the software itself would be much more useful in answering many of the questions posed in the charge to reviewers. I strongly suggest contracting individuals to thoroughly test the software. In addition, the software might be submitted to the Journal of Statistical Software for review. In general, reviews from this journal can be quite extensive and helpful.

ATTACHMENT 1

HARVARD SCHOOL OF PUBLIC HEALTH

Department of Biostatistics
655 Huntington Avenue Boston, MA 02115

(617) 432-1056

FAX: (617) 432-5619
dept@biostat.harvard.edu

16 June 2007

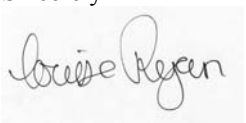
Margaret Lyday
Research Review Project Manager
ORAU/ORISE
P.O. Box 117, MS 17
Oak Ridge, TN 37831-0117
phone: 865-576-2922
fax: 865-241-3168
Margaret.Lyday@orise.orau.gov

Dear Margaret

Please find attached my report on the testing of the BMDS. While I feel that the software has a lot of strengths and seems numerically sound, I feel that there are a number of troubling aspects of the statistical theory underlying the software. I would advise EPA to address these in more depth before releasing the sofrware.

I hope this is helpful. I am happy to talk further, to add extra detail to my report and to generally do whatever I can to support this important effort.

Sincerely



Louise Ryan
Professor of Biostatistics