

Question D.3 EPA re-implemented the model presented in the NRC(2001) in the language R as well as in EXCEL spreadsheet format. In addition, extensive testing of the resulting code was conducted. “Please comment on the precision and accuracy of the re-implementation of the model.”

Pre-meeting Comments/Clarifications on the Question

Question D-3 suggests that the estimation of the dose-response model and the hazard assessment were originally programmed in the R language. Page 63 of the Issue Paper indicates that the Poisson hazard model was originally estimated in the R language (optim routine) but the neither the main text of the paper nor its appendices provided any additional information. A clarifying question from the Panel through Tom Miller to the EPA staff resulted in the following response from Andrew Schulman through Jonathon Chen.

“The reference to the implementation in R in question D.3 is outdated, and should have been removed. This was an oversight on EPA's part. The model implementation in Excel is our implementation of record, and was used to prepare the results in the draft toxicological review. We would ask the Panel to please review and comment only on the implementation in Excel. (Background: EPA did originally implement its model in R. However we found that version to be not very transparent, and hard to debug. We then reimplemented the model in Excel, found and corrected some errors, and used that corrected version to prepare the tox review. While Excel may not be the best choice from the standpoint of numerical accuracy, it is greatly superior in the transparency of the implementation, and is powerful enough to perform the entire model calculation from start to finish, even including the nonlinear optimization. Once the Panel is satisfied that the implementation in Excel is correct and appropriate, then the model can be reimplemented in R or some other numerically superior language.)”

The Agency staff is to be commended for deciding to test its original R-language version of the model program through a separate implementation in EXCEL. The EXCEL version serves as a check of programming performed in alternative systems (e.g. R, S) and provides transparency for review by non-specialists. For the calculations required in this model of hazard and excess risk, the EXCEL computations should provide sufficient numerical accuracy. If the EPA returns to another model program, it should begin with the original model formulas and not simply transcribe the programming from the existing EXCEL version of the model. As a debugging and error-checking tool, comparisons of intermediate results from two model programs should be performed to verify the equivalence of the two model systems.

Overview of the EXCEL spreadsheet implementation of the model:

The EXCEL model implementation is described in Appendix B (pages 105-106) of the Issue Paper. The Issue Paper (page 65) referenced a URL,

www.epa.gov/waterscience.sab; however this proved to be not available. EPA staff notified the panel of the following correct address, <http://epa.gov/waterscience/sab/>. The Issue Paper suggests that a listing of the variable and parameter input fields is provided in Table B-3 but the current draft of the Issue Paper did not include this table. (The fields in the spreadsheet model were interpreted by the Panel based on the description provided in the text of the Issue Paper and general understanding of the model fitting procedure employed).

The spreadsheet model requires two Excel files and associated macros. The first of these is MCCancerfit.XLS. This workbook component of the model consists of eight worksheets in four pairs (e.g. fblad and MC fblad for female bladder cancer) that cover the two cancers of interest (lung and bladder) and gender (male, female). The initial worksheet (e.g. fblad) in each of the four cancer/gender pairs contains the input data for the fitting of the hazard model. The first step in the model fitting algorithm is to employ the EXCEL Solver to find initial values of a_1, a_2, a_3 and β (Cells G2:G5) that maximize the Poisson likelihood under the following model:

$$\lambda_{i,dose} = \exp(a_1 + a_2 \cdot age_i + a_3 \cdot age_i^2) \cdot (1 + \beta \cdot dose)$$

This is the model described by the EPA in the Issue Paper and is one of two models that appeared to provide the best fit to the data based on the Akaike Information Criterion (NRC, 2001).

The second worksheet in each the four disease/gender pairs (e.g. MC fblad) is used in conjunction with initial starting values generated by Solver for Cell N2 to simulate the empirical Bayes posterior distribution of the model parameters based on a set of 1000 random perturbations of the coefficient vector (a_1, a_2, a_3, β) about the maximum likelihood estimates produced by Solver. The perturbation involves independent, random (uniform) dispersion of the coefficient estimates in a relative range of +/- 10% about the point estimates generated by Solver. Parameter draws outside this range are not performed since the posterior likelihood takes on a near zero value outside the boundaries +/- 10% of MLE. The corresponding macro (e.g. mcfblad) is then invoked to apply the observed data and these perturbed coefficient values to establish the value of the posterior log-likelihood for each of the 1000 draws. The empirical Bayes estimate of the slope parameter and its lower confidence limit are then estimated based on the mean and standard deviation of the simulated posterior distribution:

$$\bar{b} = \frac{\sum_{j=1}^{1000} b_j \cdot \frac{L_j}{L_{\max}}}{\sum_{j=1}^{1000} \frac{L_j}{L_{\max}}}$$

$$sd(b) = \sqrt{\frac{1000}{999} \cdot \frac{\sum_{j=1}^{1000} \left[\frac{L_j}{L_{\max}} \cdot (b_j - \bar{b})^2 \right]}{\sum_{j=1}^{1000} \frac{L_j}{L_{\max}}}}$$

and,

$$UCL(b) = \bar{b} + 2 \cdot sd(b)$$

The estimated UCL(b) is then carried forward to the BIER.IV computation of the excess lifetime risk in the Bier.xls spreadsheet.

Based on its review, the Panel noted that for the given data inputs, the empirical Bayes estimation algorithm programmed in the MCCancerFit.xls spreadsheet does match the form of the model and the general description of the parameter fitting algorithm outlined in the Issue Paper.

As described in the Issue Paper, the EPA data inputs for at risk populations and cancer deaths agree with Morales, et al. (2000). In general, the panel recommends that all tables of inputs for these models be published in appendices to the Issue Report or final risk assessment so that reviewers can independently reference and verify the critical inputs to the hazard and excess risk analysis.

The MCCancerft.xls spreadsheet includes an adjustment of 50 µg/day of arsenic from food intake. Based on the formula provided on page 103 of the Issue Paper, the current model assumes a combined daily intake of 2 liters/day of cooking and drinking water. The issue paper suggests that the current analysis uses 30 µg/day. Although the Issue Paper notes the NRC(2001) finding that dietary intake had no significant effect on the estimated cancer slope factor, the apparent discrepancy between the value of 30 µg/day cited in the paper and the 50 µg/day value used in the spreadsheet model should be resolved. The model does not allocate a food input of arsenic to the control population. This is a decision that presumes food-based intake of arsenic originates through cooking water only.

The second EXCEL workbook in the risk assessment model employs the estimates of the dose response model parameter, β, and its upper bound to evaluate excess lifetime risk under the Bier-IV formula. The Bier.xls workbook includes four worksheets, one for each cancer type by gender combination (flung, mlung, fblad, mblad). The estimates of the linear dose response parameter and its estimated 95% UCL (see above) are manually pasted from the corresponding worksheet in MCCancerFit.xls.

The excess risk is computed in cell T15. Solver can be applied to the dose value in Cell T11 (not U10 as indicated on Page 105 in the Issue Paper) to establish the dose level required to produce a user-specified value of excess risk (i.e., ED₀₁).

The columns of each worksheet in the Bier.XLS spreadsheet incorporate data for a specific age range of the U.S. population. These columns are not labeled with the corresponding age range. Identifying labels should be applied to all rows and columns in these worksheets. By deduction, column 3 applies to individuals age 20-24, 4 to age 25-29, etc. If this is correct, the Panel recommends that the entry in cell B3 of each of the four Bier.XLS spreadsheets be verified. It appears that this mortality figure may apply to more than just the 20-24 year old population represented in Column 3. Referring to the data inputs for 20-24 year olds in the FLUNG spreadsheet in Bier.xls, the population value is 9,423,000, all deaths are 18,121 and the baseline hazard is .00192. Moving over one column to the 25-29 year olds, the population is nearly the same at 9,491,000, all deaths are 1580 and the baseline hazard is .00017—less than 1/10th that for the previous five year age group.

The Bier.xls spreadsheet implementation of the Bier.IV excess risk calculation includes a 3-fold divisor to transform the risk to a U.S. population base (assuming exposure per kg is 3-fold higher in the SW Taiwanese population). This scaling occurs in the calculation of the age-specific cancer hazard (Row 11). It should be documented and also should be a target for future sensitivity studies. Since this is a model parameter it should be identified as a distinct input on the spreadsheet instead of simply embedded in the calculations.

The notation for the Bier-IV formula on Page 102 in the Issue Paper does not distinguish between total survivorship (S_i) and survivorship adjusted for the added risk of cancer. However, the spreadsheet implementation of the model decomposes survival into the product of baseline survival and a survival factor that reflects excess cancer deaths due to the prior age's exposure to arsenic. Based on a version of the spreadsheet downloaded from Office of Water website, calculation of cancer-specific survival (Row 13) appears to incorporate mortality through time I, not time I-1 as it should. This should be checked. The calculation of baseline survival appears to be correct—the survival parameter at time I includes only mortality through the end of time period I-1. With this exception, calculation of Excess Risk follows the Bier IV formula.

Following the series of checks and corrections to the model listed above, the Panel encourages the Agency to extend its testing of the model's sensitivity to alternative models forms and model assumptions. Specific areas where the Panel felt additional sensitivity testing is warranted include:

- A Monte Carlo analysis in which the individual well concentrations for 22 villages with multiple wells are taken into account. The Panel recognizes the difficulties with this approach including the issue of how to allocate cases to wells within villages.

- MCCancerFit.xls :
 - A test of the sensitivity of the model to the choice of the reference population (SW Taiwan).
 - A test of the sensitivity of model results to the assumption that the reference population has 0 intake of arsenic via food.
 - A contrast of results for the linear dose model employed in this program to an alternative hazard model that has a dose contribution that is multiplicative and quadratic in form. This is the form of the model that NRC(2001) found to have best fit to the data based on the Akaike Information Criterion (AIC):

$$\lambda_{i,C} = \exp(a_1 + a_2 \cdot age_i + a_3 \cdot age_i^2) \cdot \exp(\beta_0 + \beta_1 \cdot dose + \beta_2 \cdot dose^2)$$

- Bier.xls
 - The Panel recommends a sensitivity analysis in which the age groupings used to estimate the baseline hazard and employed in the estimation of excess lifetime risk are altered. A logical choice is to test the sensitivity of the model results to using 10-year groupings (e.g. 20-29, 30-39) in both spreadsheets.
 - The exposure/kg parameter used to transfer the dose/response model from the original SW Taiwanese population to a U.S. general population is a major driver in the computation of excess lifetime risk. In preparing its final risk assessment, the EPA should conduct a sensitivity analysis to determine precisely how much the choice of a factor of 3 impacts the final estimates of excess lifetime risk.