ADVANCES IN QUANTITATIVE BENCHMARK RISK ASSESSMENT FOR ENVIRONMENTAL HEALTH AND EDUCATIONAL SETTINGS

By

Mirjana Glisovic Bensa

A Dissertation Submitted to the Faculty of the

COLLEGE OF PUBLIC HEALTH

In Partial Fulfilment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2024

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: Mirjana Glisovic Bensa

titled: ADVANCES IN QUANTITATIVE BENCHMARK RISK ASSESSMENT FOR ENVIRONMENTAL HEALTH AND EDUCATIONAL SETTINGS

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Walter W Piegorsch
Walter Piegorsch — Date: Apr 22, 2024

Denise J Roe
Denise Roe — Date: Apr 22, 2024

KARey
Kelly A Reynolds — Date: Apr 22, 2024

Edward John Bedrick
Edward Bedrick — Date: Apr 23, 2024

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Walter W Piegorsch
Walter Piegorsch — Date: Apr 22, 2024
Mathematics

ACKNOWLEDGMENTS

First, I would like to express my gratitude to my advisor Dr. Walter W Piegorsch who invested numerous hours in my training. He helped me grow in my PhD and his teaching made this dissertation possible. His excitement, expertise, determination, and work ethic made this journey inspiring and enjoyable.

Secondly, I would like to thank Dr. Melanie Bell for encouraging me to apply and, when I struggled, to stay in the Ph.D program.

I would also like to thank my professors and members of my dissertation committee, Dr. Edward Bedrick, Dr. Denise Roe, and Dr. Kelly Reynolds for their dedication to truly teaching students and for their comments and edits of this dissertation manuscript.

I would like to thank my family, friends and classmates who supported me through this journey – Geronda Efraim, Geronda Paisious, Father Philaretos, Dr. Jones, Dr. Fathabad, Sayeh, Jovanka, Mariana, Milena, Ivan, Mihailo and Anastasia - for all their encouragement and prayers.

Finally, I would like to thank my husband Janko and daughters Emilija and Natalija for their love and support.

DEDICATION

In the loving memory of my mother Ljiljana and father Dorotheos.

TABLE OF CONTENTS

ABSTRACT

This dissertation is a unifying document of three related projects developed during my dissertation research. The projects involve comparison of existing methods for estimating the benchmark dose (BMD) and its lower confidence limit (BMDL), a Bayesian calculation for mixed-factor quantal data to estimate the BMD and BMDL, and an application of BMD methodology in educational setting.

Briefly, the first project compares the performance of current methods for BMD and BMDL calculation. We use 379 data sets from a curated Quantal Risk Assessment Database (QRAD) to compare performance of Bayesian model averaging (BMA), frequentist model averaging (FMA) and nonparametric regression modeling methods for BMD and BMDL calculations.

The second project applies Bayesian strategies to a mixed-factor setting with a secondary qualitative factor possessing two levels to derive two-factor Bayesian BMDs and BMDLs. We present reparametrized dose-response models that allow for explicit use of prior information on the target parameter of interest, the BMD. We also enhance our Bayesian estimation technique for BMD analysis by applying Bayesian model averaging to produce BMDs and BMDLs.

Lastly, the third project applies risk benchmarking methodology traditionally used in toxicology and environmental settings to educational data. We apply BMD and BMDL estimation techniques to students' placement exam scores in order to calculate the minimum placement exam score required to place in a class and successfully pass it.

Taken together, the three projects presented in this dissertation illustrate advances in quantitative benchmark risk assessment in environmental health and educational settings.

Statistical Advisor: Dr. Walter W. Piegorsch

CHAPTER 1
Introduction

1.1 Quantitative risk assessment

Quantitative risk assessment involves estimation of the likelihood and severity of an adverse response due to exposure to hazardous stimuli. Risk estimation is usually broken into four fundamental stages (Stern, 2002): hazard identification, dose-response assessment, exposure assessment and risk characterization. Quantitative considerations arise in each of the four stages with stage two, dose-response assessment, requiring the greatest statistical input. The following definitions provide a foundation for quantitative dose-response risk analysis.

*Definition 1.1***:** Risk, $R(x)$, is defined as the probability of some defined adverse effect, such as death, cancer, birth defect in a subject exposed to a particular stimulus at the dose level x. This definition assumes that risk changes with increasing exposure at dose level x to the environmental agent.

Definition 1.2: Background risk, $R(0)$, is defined as the risk that all subjects in a population would be expected to encounter in the absence of exposure.

To account for this spontaneous, background risk, the additional risk $R_A(x)$ and extra risk, $R_E(x)$ must be formally defined.

Definition 1.3: Additional risk or added risk, $R_A(x)$, is defined to be the risk beyond that of the control (i.e., background, spontaneous) level; that is:

$$R_A(x) = R(x) - R(0)$$

Definition 1.4: Extra risk, $R_E(x)$, is defined as the additional risk among subjects who would have not responded under control conditions:

$$R_E(x) = \frac{R_A(x)}{(1-R(0))} = \frac{R(x)-R(0)}{(1-R(0))}$$

Notice that in the absence of background risk, that is when $R(0) = 0$, the additional risk, $R_A(x)$, is the same as the extra risk $R_E(x)$ and when $R(0) = 1$, the extra risk, $R_E(x)$, is undefined.

If $0 \leq R(0) \leq 1, then\ R_E(x) \geq R_A(x)$; that is, extra risk can be viewed as an inflation of additional risk.

Together, $R_A(x)$ and $R_E(x)$ are known as forms of *excess risk.*

Definition 1.5: The reference dose (RfD) or reference concentration (RfC) is an ostensibly "safe" exposure level below which exposure to a toxin or other hazardous agent, little or no harm happens to an exposed subject.

Definition 1.6: The benchmark dose (BMD) is the dose that corresponds to a specified benchmark response (BMR) above which the excess risk is considered unacceptable.

1.2 Benchmark dose estimation

The estimation of minimum exposure levels, such as or based on Benchmark Doses (BMDs), that induce a pre-specified Benchmark Response (BMR), is a widely used tool in public health risk analysis (U.S. EPA, 2012). A central theme of my dissertation research concerns this benchmark dose estimation methodology.

In quantitative risk assessment, benchmark dose (BMD) calculations proceed as follows:

1. Identify the parent distribution of the data and model for the risk function $R(x)$. Select an appropriate form of excess risk function, $R_A(x)$ or $R_E(x)$, for use in the benchmark analysis. Excess risk is typically chosen based on existing conventions in risk literature.
2. Choose the benchmark response (BMR). Common values for BMR are: $0.1, 0.05, 0.01$ or even 10^{-6} depending on the nature of the adverse outcome, regulatory considerations, and the quality of the data. In many environmental studies, most common observed risks are between $0.05\ and\ 0.01$ and therefore most BMRs are often chosen to be in this range.
3. Acquire the data and fit the model in step 1 to them. Use this to estimate excess risk. If desired construct confidence limits on the excess risk.

4. Invert the relationship estimated in step 3 to estimate a BMD at the chosen BMR. For example, if extra risk, $R_E(x)$, is under study, solve the relationship $\widehat{R_E}(x) = BMR$ for x and use the result as $B\widehat{M}D$. If more than one solution results from the calculation, choose the smallest positive value.
5. Estimate a $1 - \alpha$ lower limit on the dose associated with the specified BMR. We call this lower limit a BMDL and use it as the basis for any future RfD calculations.

1.3 Article-based dissertation and content organization

This dissertation is a unifying document of three related projects developed during my dissertation research. We present 3 projects that build past the foundations set by Sans-Fuentes and Piegorsch (2021), Fang at al. (2015a) and Fang at al. (2015b).

The first project compares the performance of current methods for BMD and BMDL calculation. We use 379 data sets from a curated, online Quantal Risk Assessment Database (QRAD) to compare performance of Bayesian model averaging (BMA), frequentist model averaging (FMA) and nonparametric regression modeling methods for BMD and BMDL calculations. The second project applies Bayesian strategies to a mixed-factor setting with a secondary qualitative factor possessing two levels to derive two-factor Bayesian BMDs and BMDLs. It also enhances Bayesian estimation techniques for BMD analysis by applying Bayesian model averaging to produce BMDs and BMDLs. The third project applies risk benchmarking methodology traditionally used in toxicology and environmental setting to educational data.

The following chapters contain a summary of each project. The concluding chapter provides an overall summary. Following the references are appendices containing the complete projects that were summarized in preceding chapters.

CHAPTER 2

Bayesian Model Averaging, Frequentist Model Averaging and Nonparametric Regression Modeling Methods for BMD and BMDL Calculation

Theoretical features of the BMD and BMDL have been studied by a variety of researchers (Haber et al., 2018) and data scientists have worked to develop the BMD and BMDL concepts to address modern issues in quantitative risk assessment (Jensen et al., 2019). Missing, however, is an examination of the performance of the BMD on a "test bed" of high-quality data containing experimental results that span a range of possible dose-response patterns. Given the growing popularity of the benchmark analysis method, in the first paper (Glisovic et al., submitted 2024), we explored performance of current methods for BMD and BMDL calculations on 382 data sets from the Quantal Risk Assessment Database (QRAD). This is the first study that compares three popular methods for BMD and BMDL calculation on a large number of data sets.

The Quantal Risk Assessment Database (QRAD) consists of 733 separate quantal dose-response data sets assembled from a variety of existing toxicological data sources such as the U.S. National Toxicology Program (NTP), the Integrated Risk Information System (IRIS), the State of California EPA (CalEPA), and the U.S. Centers for Disease Control and Prevention's Agency for Toxic Substances and Disease Registry (ATSDR). To properly fit any statistical models, it was necessary to use data sets that contain at least 4 dose groups and exhibit a non-decreasing dose-response pattern (since variety of non-increasing patterns can lead to unstable computational results in the eventual point estimates). Out of 733 data sets available in QRAD, 382 data sets met this criterion.

A variety of models are available for fitting a non-decreasing dose-response relationship between a dose x of a chemical and a dose-response function $R(x)$, however, fundamental concern in benchmark analysis is potential uncertainty with specification of the dose–response function, $R(x)$. To mitigate problems associated with single model selection, model averaging, and nonparametric techniques have been developed. In the first paper (Glisovic et al., submitted 2024), we compared popular Bayesian model averaging (BMA), frequentists model averaging (FMA) and nonparametric regression modeling (NPAR) techniques for BMD and BMDL calculation. We calculated Bayesian model averaging (BMA) BMD and BMDL, frequentist model averaging (FMA) BMD and BMDL and nonparametric regression modeling (NPAR) BMD and BMDL across the pertinent data sets in QRAD.

We performed BMA and FMA across seven commonly used forms for $R(x)$ shown in the table below.

Model	**Dose-response function $R(x)$**
Logistic	$R(x) = \frac{1}{(1+e^{-\beta_0-\beta_1 x})}$
Probit	$R(x) = \Phi(\beta_0 + \beta_1 x)$
Quantal-linear	$R(x) = 1 - e^{-\beta_0-\beta_1 x}$
Two-stage	$R(x) = 1 - e^{-\beta_0-\beta_1 x-\beta_2 x^2}$
Log-logistic	$R(x) = \gamma_0 + \frac{1-\gamma_0}{(1+e^{-\beta_0-\beta_1 \log\{x\}})}$
Log-probit	$R(x) = \gamma_0 + (1-\gamma_0)\Phi(\beta_0 + \beta_1 \log\{x\})$
Weibull	$R(x) = \gamma_0 + (1-\gamma_0)\left(1 - \exp\left\{-e^{\beta_0} x^{\beta_1}\right\}\right)$

To calculate Bayesian model averaging (BMA) BMDs and BMDLs, we selected the ToxicR computer package (see https://niehs.github.io/ToxicR/) within the open-source R programming environment (R Development Core Team, 2019). ToxicR very efficiently conducts

Bayesian model averaging using methodologies as defined in Wheeler et al. (2020) and Wheeler et al. (2022). To calculate frequentist model averaging (FMA) BMDs and BMDLs, we again used ToxicR to obtain individual-model BMDs and combined these as described in section 3 of Piegorsch et al. (2013) to calculate BMDLs. To calculate nonparametric BMDs and BMDLs we follow methods presented in Piegorsch et al. (2014).

Applying these three methods on each of the 382 QRAD data sets we found that the BMA, FMA and nonparametric methods are not able to calculate BMD and BMDL values for every data set; however, when they did produce values for BMD and BMDL, they were highly correlated among all three methods (as we hoped they would). The BMA method was able to calculate BMDL values 99% of the time, the nonparametric method was able to calculate BMDL values 64% of the time, while the FMA method was able to calculate BMDL values only 54% of the time. This suggests that the most stable method for calculation of BMD and BMDL is the BMA method as described in Wheeler et al. (2020). Appendix A reproduces the complete Glisovic et. al (submitted, 2024) article.

CHAPTER 3

Bayesian BMD and BMDL Calculation for Mixed-Factor Quantal Data

Estimation of the benchmark dose (BMD) and of its lower confidence limit (BMDL) is well understood for the case of an adverse response to a single stimulus. In our first paper (Glisovic-Bensa, et al., submitted 2024), which was briefly described in chapter 2, we compared popular Bayesian model averaging (BMA), frequents model averaging (FMA) and nonparametric regression modeling (NPAR) techniques for BMD and BMDL calculation in the case of an adverse response to a single stimulus. In many environmental settings, however, one or more additional, secondary, qualitative factor(s) may collude to affect the adverse outcome, such that the risk changes with differential levels of the secondary factor Sans-Fuentes, 2021). In this two-factor setting the risk function $R(x,u)$ depends on two variables, x the dose and u some qualitative factor. Bayesian methods for estimation of BMD and BMDL have grown in popularity and in our first paper (Glisovic-Bensa, et al., submitted 2024) we have found that the BMA method outperforms other model-independent methods such as FMA and NPAR. However, to our knowledge, no Bayesian methodology has appeared that allows for incorporation of prior information in this mixed-factor setting. In the second paper (Glisovic-Bensa, et al., submitted 2023), shown in Appendix B, we apply Bayesian strategies to this mixed-factor setting with a secondary qualitative factor possessing two levels to derive two-factor Bayesian BMDs and BMDLs.

For $R(x,u)$ we consider five models often seen in toxicological and carcinogenic risk assessment. Table 1 lists formal expressions for each model.

***Table 1**. Popular quantal-response models in toxicological and carcinogenic risk assessment.*

Model	**Risk function**	**Form of risk function**
M_1	Logistic	$R(x,u) = 1/\left(1 + e^{-\beta_0-\beta_1 x-\delta_0\, u-\delta_1 xu}\right)$
M_2	Probit	$R(x,u) = \Phi(\beta_0 + \beta_1 x + \delta_0\, u + \delta_1 xu)$
M_3	Quantal-linear	$R(x,u) = 1 - e^{-\beta_0-\beta_1 x-\delta_0\, u-\delta_1 xu}$
M_4	Quantal-quadratic	$R(x,u) = 1 - e^{-\beta_0-\beta_1 x^2-\delta_0\, u-\delta_1 x^2 u}$
M_5	Complementary log-log	$R(x,u) = 1 - e^{-e^{\beta_0+\beta_1 x-\, u-\delta_1 xu}}$

where $u = \begin{cases} 0 \text{ if subject from group 1} \\ 1 \text{ if subject from group 2} \end{cases}$

To apply the Bayesian approach, we need to specify prior distributions for the unknown model parameters such as the β and δ-parameters in Table 1. With quantal data, it is unusual to have informative prior knowledge on parameters such as $\beta_0, \beta_1, \delta_0, \delta_1$. However, high-quality, historical control information is available for the background risk $R_j(0)$ which for notational simplicity we denote as γ_j, where γ_0 and γ_1 are the background risks for subjects from group 1 and group 2 respectively. For a Bayesian approach we specify prior distribution for the BMD (which again for notational simplicity we denote as ξ_j, where ξ_0 and $\xi_1$are the BMD for subjects from group 1 and group 2 respectively) and the background risks. To employ this prior information, we reparametrize $R(x,u)$ in terms of $\xi_j$and γ_j. For example, under the quantal-liner model

$R(x,u) = 1 - e^{-\beta_0 - \beta_1 x - \delta_0\, u - \delta_1 x u}$ from Table 2, while the reparametrized quantal-linear model becomes $R(x,u) = 1 - e^{\log(1-\gamma_0) - \frac{c}{\xi_0}x + u\log\left(\frac{1-\gamma_1}{1-\gamma_0}\right) - c\left(\frac{1}{\xi_1} - \frac{1}{\xi_0}\right)ux}$.

To apply a Bayesian analysis we assign a joint p.d.f. to the reparameterized vector

θ = [ξ_0 ξ_1 γ_0 γ_1]$^{\mathsf{T}}$. We write the prior generically as $\pi(\theta) = \pi(\xi_0, \xi_1, \gamma_0, \gamma_1)$ and we assume the unknown parameters enter into the prior independently, so that

$\pi(\xi_0, \xi_1, \gamma_0, \gamma_1) = \pi(\xi_0)\pi(\xi_1)\pi(\gamma_0)\pi(\gamma_1)$. We take $\xi_j \sim U(0, \max\{x_i\})$ and $\gamma_j \sim Beta(a_j^*, b_j^*)$ where *j* = 0,1. Using a binomial parent distribution for the data the corresponding joint posterior distribution for ξ_0, ξ_1, γ_0 and γ_1 is: $\pi(\xi_0, \xi_1, \gamma_0, \gamma_1 | \boldsymbol{Y}) =$

$$\frac{\prod_{i=1}^{n} \binom{N_i}{Y_i} R(x_i, u_i)^{\,Y_i}\{1 - R(x_i, u_i)\,\}^{N_i - Y_i}}{m(\boldsymbol{Y})} \times \frac{1}{\max\{x_0\}} \times \frac{1}{\max\{x_1\}}$$

$$\times \frac{\Gamma(a_0^* + b_0^*)}{\Gamma(a_0^*)\Gamma(b_0^*)} {\gamma_0}^{a_0^*}(1-\gamma_0)^{b_0^*-1} \times \frac{\Gamma(a_1^* + b_1^*)}{\Gamma(a_1^*)\Gamma(b_1^*)} {\gamma_1}^{a_1^*}(1-\gamma_1)^{b_1^*-1}$$

where

$$m(\boldsymbol{Y}) = \iiint \int \prod_{i=1}^{n} \binom{N_i}{Y_i} R(x_i, u_i)^{Y_i} \{1 - R(x_i, u_i)\}^{N_i - Y_i} \times \frac{1}{\max\{x_0\}} \times \frac{1}{\max\{x_1\}} \times$$

$$\frac{\Gamma(a_0^* + b_0^*)}{\Gamma(a_0^*)\Gamma(b_0^*)} {\gamma_0}^{a_0^*} (1 - \gamma_0)^{b_0^* - 1} \times \frac{\Gamma(a_1^* + b_1^*)}{\Gamma(a_1^*)\Gamma(b_1^*)} {\gamma_1}^{a_1^*} (1 - \gamma_1)^{b_1^* - 1} d\xi_0 d\xi_1 d\gamma_0 d\gamma_1$$

is the marginal distribution of **Y** in this Bayesian hierarchy.

Due to the non-conjugacy of the prior distributions here, for the posterior calculations we turn to Monte Carlo Markov Chain (MCMC) posterior approximations using the *R2OpenBUGS* package (Sturtz et al., 2005). For illustration purposes of this new Bayesian approach, we use goldenseal root powder carcinogenicity data (U.S. National Toxicology Program, 2010). For a given data ensemble (N, Y, x, u), we generated a tetravariate chain of 160,000 $(\xi_0, \xi_1, \gamma_0, \gamma_1)$ draws per model. Then, we estimate the BMDs in ξ_0 and ξ_1 by calculating the posterior sample means $\overline{\xi_0}$ and $\overline{\xi_1}$.

This paper (Glisovic-Bensa, et al., submitted 2023) shows how to apply Bayesian strategies to this mixed-factor setting, with a secondary qualitative factor possessing two levels, to derive two-factor Bayesian BMDs and BMDLs.

CHAPTER 4

Benchmark Risk Assessment for Mathematics Placement Scores

Students enrolled in college are often required to take at least one mathematics or statistics course. To ensure students success in such a class, universities and colleges have developed eligibility criteria and course prerequisites that students must meet before they can register for a given class. Often, students can satisfy required eligibility criteria by taking pertinent prerequisite courses, or by achieving the required score on a university-specific math placement exam or on standardized college entrance exams such as the SAT and ACT. At our home institution, the University of Arizona (UArizona), many incoming first-year students are required to take a precalculus course (MATH 120R). They can register for the course only if they meet at least one of the following criteria:

1. Earn at least 60 points on the UArizona implementation of the PPL Assessment Mathematics Placement Test.
2. Earned at least 640 points on the MATH-SAT exam.
3. Earned at least 26 points on the ACT exam.
4. Passed a College Algebra course with a grade of C or better.

At the end of each academic year, a team of faculty members from the UArizona mathematics department studies the passing rates of previous MATH 120R students with a PPL score at or near 60 points. If the passing rate is lower than usual, the required PPL score to place into MATH 120R is increased. In other words, the score required on the PPL assessment to place into MATH 120R is determined empirically on a rotating basis. To construct a somewhat more-objective target for the students to meet, in our 3rd Paper Glisovic et. al (in preparation, 2024) we apply risk-analytic techniques widely employed in toxicological and environmental applications (Piegorsch and Bailer, 2005, Sec. 4.3) to student mathematics placement score data. Using a transitional approach, we employ the benchmark risk analytic technique (Crump, 1984) to the placement question. As described above, benchmark risk assessment methodology uses components of the risk function to produce the benchmark dose, BMD, of the hazardous agent, at which the prespecified benchmark response, BMR, is obtained. We view PPL assessment score as the "dose" and benchmark *response* (BMR), as the prespecified passing rate in the mathematics/statistics course into which the students are placed.

This project illustrates the flexibility of benchmark risk assessment by extending its application outside of traditional toxicological and environmental applications to the educational setting. Appendix C reproduces the complete article (Glisovic et. al, in preparation, 2024) and Appendix D contains corresponding supplementary material.

CHAPTER 5

Conclusion

Theoretical features of the BMD and BMDL have been studied by a variety of researchers (Haber et al., 2018) and data scientists have worked to develop the BMD and BMDL concepts to address modern issues in quantitative risk assessment (Jensen et al., 2019). The overreaching goal in this collection of works is to build upon past existing knowledge and apply BMD and BMDL methodology to a new, novel assemblage of risk-analytic questions. The collective contribution of the three contained papers (see Appendices A-D) represent substantial progress towards meeting that goal.

In the first paper (Glisovic et. al, submitted 2024) we provide a thorough comparison between nonparametric, Bayesian model-averaged and frequentist-model averaged estimators of BMD and BMDL using a diverse, high-quality, curated data base (QRAD). The research presented in the first paper is informative to toxicological regulatory agencies and provides recommendations as well as advantages and disadvantages of these model-independent approaches. In paper one, we use ToxicR package to calculate Bayesian BMDs and BMDLs. Future studies could investigate the use of different packages such as BBMDS to determine if the results we present here translate well when other approaches are used.

In the second paper (Glisovic et. al, submitted 2023) we develop a Bayesian strategy to the mixed-factor setting, with a secondary qualitative factor possessing two levels, to derive two-factor Bayesian BMDs and BMDLs. Future directions here include developing more efficient methods to derive two-factor Bayesian BMDs and BMDLs that what was presented in the second paper (Glisovic et. al, submitted 2023). Indeed, for the posterior calculations we turned to Monte Carlo Markov Chain (MCMC) posterior approximations using the *R2OpenBUGS* package (Sturtz et al., 2005). This method is not as efficient as some other methods (Wheeler et al., 2020) and further study might determine if other posterior approximations could improve upon the results we achieved.

Finally, the third paper (Glisovic et. al, to be submitted 2024) extends application of the benchmark risk assessment methodology outside of traditional toxicological and environmental applications to the educational placement-score setting. In paper 3 we chose to employ a logit link function. Many other quantal-response models are possible for describing the sorts of mixed-factor data we studied in paper 3, however. Choices include the probit link, the complementary log link, the complementary log-log link, along with a variety of other quantal-response forms popular in environmental risk assessment (West et al., 2012; Wheeler et al., 2019). Future research is required to determine if the results we present here translate well to other forms of quantal-response model. Some form of *model averaging* amongst different quantal-response forms (Fang et al., 2015b; Piegorsch et al., 2013; Wheeler & Bailer, 2009; Glisovic et al., submitted 2023) as seen in papers 1 and 2 could be used to combine the different models as part of the benchmark analysis. Further study may reveal if this is a profitable avenue for conducting model-robust benchmark analytics with mixed-factor data in the educational setting.

REFERENCES

Crump, K. S. (1984). A new method for determining allowable daily intake. *Fundamental and Applied Toxicology* **4**, 854–871.

Fang, Q., Piegorsch, W.W., and Barnes, K.Y. (2015a). Bayesian benchmark dose analysis. *Environmetrics* 26, 373-382.

Fang, Q., Piegorsch, W. W., Simmons, S. J., Li, X., Chen, C., and Wang, Y. (2015b). Bayesian model-averaged benchmark dose analysis via reparametrized quantal-response models. *Biometrics*, 71(4), 1168-1175.

Haber, L. T., Dourson, M. L., Allen, B. C., Hertzberg, R. C., Parker, A., Vincent, M. J., Maier, A., & Boobis, A. R. (2018). Benchmark dose (BMD) modeling: current practice, issues, and challenges. *Critical Reviews in Toxicology* **48**, 387-415.

Jensen, S. M., Kluxen, F. M., & Ritz, C. (2019). A review of recent advances in benchmark dose methodology. *Risk Analysis* **39**, 2295-2315.

Piegorsch, W. W., and Bailer, A. J. (2005). *Analyzing Environmental Data*. Chichester: John Wiley & Sons.

Piegorsch, W. W., An, L., Wickens, A. A., West, R. W., Peña, E. A., & Wu, W. (2013). Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. *Environmetrics* **24**, 143-157.

Piegorsch, W. W., Xiong, H., Bhattacharya, R. N., & Lin, L. (2014). Benchmark dose analysis via nonparametric regression modeling. *Risk Analysis*, **34**(1), 135-151.

R Development Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sans-Fuentes, M. A., and Piegorsch, W. W. (2021). Benchmark dose risk analysis with mixed-factor quantal data in environmental risk assessment. *Environmetrics* 32, Article No. e2677.

Stern, A. (2002). Risk assessment, quantitative. In A. H. El-Shaarawi and W. W. Piegorsch (eds), Encyclopedia of Environmetrics, Vol. 3, pp. 1837-1843. Chichester: John Wiley & Sons.

Sturtz, S., Liggers, U., and Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12(3), 1-16.

U.S. EPA (2012). *Benchmark Dose Technical Guidance Document*. Technical Report number EPA/100/R-12/001. Washington, DC: U.S. Environmental Protection Agency.

U.S. National Toxicology Program (2010). Toxicology and carcinogenesis studies of goldenseal root powder (*Hydrastis Canadensis*) in F344/N Rats and B6C3F1 Mice (Feed studies). Technical

Report #562: U.S. Department of Health and Human Services, Public Health Service, Research Triangle Park, NC.

West, R. W., Piegorsch, W. W., Peña, E. A., An, L., Wu, W., Wickens, A. A., Xiong, H., & Chen, W. (2012). The impact of model uncertainty on benchmark dose estimation. *Environmetrics*, *23*(8), 706–716.

Wheeler, M. W., & Bailer, A. J. (2009). Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environmental and Ecological Statistics*, *16*(1), 37–51.

Wheeler, M. W., Piegorsch, W. W., & Bailer, A. J. (2019). Quantal risk assessment database: A database for exploring patterns in quantal dose-response data in risk assessment and its application to develop priors for Bayesian dose-response analysis. *Risk Analysis*, *39*(3), 616–629.

Wheeler, M.W., Blessinger, T., Shao, K., Allen, B.C., Olszyk, L., Davis, J.A. , Gift, J.S (2020). 408 Quantitative risk assessment: developing a Bayesian approach to dichotomous dose-response 409 uncertainty. *Risk Analysis* **40**(9), 1706–1722. 410

Wheeler, M.W., Cortinas Abrahantes, J., Aerts, M., Gift, J.S., Allen Davis, J. (2022). Continuous 411 model averaging for benchmark dose analysis: Averaging over distributional forms. 412 *Environmetrics* **33**(5), Article no. e2728.

APPENDIX A (Supplemental Material for Chapter 2)

Empirical operating characteristics of model-robust benchmark doses with quantal-response data

Mirjana Glisovic-Bensa [1]

Walter W. Piegorsch [2,3]

[1] Department of Epidemiology & Biostatistics, University of Arizona, Tucson, Arizona, USA

[2] Interdisciplinary Program in Statistics & Data Science, University of Arizona, Tucson, Arizona, USA

[3] BIO5 Institute, University of Arizona, Tucson, Arizona, USA

Correspondence

Walter W. Piegorsch, BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA. Email: piegorsch@math.arizona.edu;

Abstract

Public and environmental health officials studying the impacts of adverse environmental agents rely on scientific studies to assess risks to humans or animals. A popular modern, quantitative approach is the *benchmark dose* (BMD), the dose or exposure level of the adverse agent above which the damaging response exceeds acceptable levels. Theoretical features of the BMD have been studied by a variety of researchers; to expand on these efforts we provide an examination of the performance of the BMD on a "test bed" of high-quality data containing experimental results that span a range of possible dose-response patterns. We do so via the online Quantal Risk Assessment Database (QRAD). QRAD consists of 733 separate quantal dose-response datasets assembled from a variety of existing toxicological sources such as the U.S. National Toxicology Program (NTP), the U.S. Integrated Risk Information System (IRIS), the State of California EPA (CalEPA), and the U.S. Centers for Disease Control and Prevention's Agency for Toxic Substances and Disease Registry (ATSDR). We employ the QRAD data to examine performance of current model-robust methods for BMD analytics, including non-parametric statistical estimation strategies, frequentist model averaging over competing statistical models for the BMD and Bayesian model averaging over competing statistical models.

KEYWORDS

Benchmark analysis, BMDL, quantitative risk assessment, Bayesian model averaging, frequentist model averaging, nonparametric analysis.

1. INTRODUCTION

An important objective in quantitative risk assessment is characterization of the severity and likelihood of adverse outcomes caused by a hazardous agent (Stern, 2008). In this context, we define the risk function, $R(x)$, as the probability of some predefined adverse effect, such as death, cancer, mutation, birth defect, etc., in a subject exposed to a hazardous agent at dose or exposure level x. This definition assumes that risk changes with increasing exposure at dose x to the hazardous agent. Commonly, observations in such risk assessments are taken as the proportions of subjects exhibiting the adverse response associated with each x. This is called the *quantal response setting*. A contemporary approach to quantal-response risk estimation in this setting is known as *benchmark analysis*. Benchmark analysis was first introduced by Crump (1984) and it has since grown in popularity and flexibility (Jensen et al., 2019). Benchmarking employs components of $R(x)$ to produce the *benchmark dose* (BMD) of the hazardous agent at which a prespecified adverse response, the *benchmark* risk (BMR), is obtained. To account for any background risk, $R(0)$, that all subjects in a population would be expected to encounter in the absence of the hazardous agent, the risk function, $R(x)$, is refined into a form of *extra risk function, $R_E(x)$,* (Piegorsch & Bailer, 2005, Section 4.2). $R_E(x)$ is defined as the additional risk among subjects who would have not responded under control conditions: $R_E(x) = \frac{R(x)-R(0)}{1-R(0)}$. From this, the BMD is found by solving for x in $R_E(x) = \mathrm{BMR}$. The BMR often is taken in the range $0.01 \leq \mathrm{BMR} \leq 0.10$ (U.S. EPA, 2012). A statistical model is then used to produce a point estimator, $\widehat{\mathrm{BMD}}$.

To account for uncertainty associated with $\widehat{\mathrm{BMD}}$, $100(1-\alpha)\%$ confidence (or Bayesian credible) limits for the BMD are calculated. Driven by focus on risk reduction, only one-sided, lower confidence limits (BMDLs) are of interest (Crump, 1995). Theoretical features of the BMD have been studied by a variety of researchers (Haber et al., 2018) and data scientists have worked to develop the BMD concept to address modern issues in quantitative risk assessment (Jensen et al., 2019).

Given the growing popularity of benchmark analysis we felt it would be propitious to explore performance of modern methods for BMD calculation on a "sandbox" or "test bed" of high-quality data containing experimental results that span a range of possible dose-response patterns. For this we use the Quantal Risk Assessment Database (QRAD), a curated, online source of quantal dose-response data focusing on rodent carcinogenicity; see Section 2.5, below.

Strong evidence has appeared showing that specific, parametric, dose-response models — such as the quantal linear or the logistic — for building the BMDL can adversely affect the confidence limit (West et al., 2012; Ringblom et al., 2014). Therefore to analyze the QRAD data we appeal to *model-robust* formulations that counter problems with model dependency. An obvious choice is a nonparametric approach that makes no formal, parametric assumptions on the underlying dose response (Piegorsch et al. 2014). Alternatively, one can construct a wide class of parametric models over which the BMD is found and then average over all the models in the class. We study here two basic forms of model averaging: $\widehat{\mathrm{BMD}}$s averaged over frequentist operations to produce 100(1–α)% confidence limits for the unknown BMD (Piegorsch et al. 2013), or using Bayesian credible limits averaged over Bayesian posterior constructions (Wheeler et al. 2020). Our goal is to compare performance of Bayesian model averaging (BMA), frequentist model averaging (FMA), and nonparametric regression modeling (NPAR) for model-robust calculation of $\widehat{\mathrm{BMD}}$ and BMDL. Section 2 describes QRAD, along with the technical details of BMA, FMA, and NPAR in benchmark analysis. Section 3 presents the results of these three strategies when applied to the datasets from QRAD, while Section 4 ends with a discussion of our findings.

2. Methods

2.1. Model Averaging and Nonparametric Analysis

A fundamental concern in benchmark analysis is potential uncertainty with specification of the dose-response function, $R(x)$. In practice, many parametric forms for $R(x)$ operate well at higher doses near the range of the observed quantal outcomes but can produce wildly different estimates on the BMD at very small levels of risk (Faustman and Bartell, 1997; Kang et al., 2000). Some investigators use formal selection procedures such as the popular Akaike (1973) information criteria (AIC); however, AIC-guided selection has been shown to select incorrect models for BMD estimation almost as often as it selects correct models (West et al., 2012). Incorrect model selection can lead to wildly incorrect BMDLs (Ringblom et al, 2014). To mitigate problems associated with single-model selection, Bayesian model averaging (BMA), frequentist model averaging (FMA), and nonparametric regression techniques have been developed, as described in the following sections.

The model averaging approaches require specification of a class of possible models for the averaging calculations. We chose a class consisting of seven commonly used forms for $R(x)$:

Logistic:	$R(x) = \frac{1}{(1+e^{-\beta_0-\beta_1 x})}$
Probit:	$R(x) = \Phi(\beta_0 + \beta_1 x)$
Quantal − linear:	$R(x) = 1 - e^{-\beta_0-\beta_1 x}$
Two − stage:	$R(x) = 1 - e^{-\beta_0-\beta_1 x-\beta_2 x^2}$
Log − logistic:	$R(x) = \gamma_0 + \frac{1-\gamma_0}{(1+e^{-\beta_0-\beta_1 \log\{x\}})}$
Log − probit:	$R(x) = \gamma_0 + (1-\gamma_0)\Phi(\beta_0 + \beta_1 \log\{x\})$
Weibull:	$R(x) = \gamma_0 + (1-\gamma_0)\left(1 - \exp\left\{-e^{\beta_0} x^{\beta_1}\right\}\right)$

For all our calculations we set the benchmark risk, BMR, to 0.10 and for the one-sided lower BMDLs we set $\alpha = 0.05$.

2.2. Bayesian Model Averaging (BMA) for BMD and BMDL calculation

To perform the Bayesian model averaging (BMA) calculations we selected the ToxicR computer package (see https://niehs.github.io/ToxicR/) for the open-source R programming environment (R Development Core Team (2019). ToxicR very efficiently conducts BMA using algorithms presented in Wheeler et al. (2020) and Wheeler et al. (2022). For the larger Bayesian approach we referred to the modeling structure described in Wheeler et al. (2020, Section 2). The technical details are extensive and for space considerations we do not present them here. We refer readers to the presentation in Wheeler et al. (2020). There are, of course, many other BMA formulations (Fang et al., 2015; Shao & Shapiro, 2018; Simmons et al., 2015) often relying on Markov Chain Monte Carlo (MCMC) calculations to estimate the BMD. We chose to employ ToxicR as it uses accurate computational techniques that can analyze hundreds of dose-response curves very rapidly.

Within the Bayesian paradigm (Christensen, et al., 2011) it is necessary to specify prior information on each model's parameters from Section 2.1. Specification of the prior distributions for the Bayesian hierarchy under each model is built into ToxicR's various functions. This provides the analyst with great flexibility in the modeling process, although the prior specifications can be misappropriated in some cases. Luckily, ToxicR provides a default collection of carefully curated priors for its various parametric hierarchies. We appealed to these default priors in all our BMA computations.

Using the default priors as defined in Wheeler et al. (2020) on model parameters such as $\beta_0, \beta_1, \gamma_0$ in Section 2.1, we engaged the ToxicR function `single_dichotomous_fit()` to fit all seven models to each of the datasets we extracted from QRAD (see Section 2.5, below). ToxicR produced a $\widehat{\text{BMD}}$ and a corresponding BMDL for each of the seven models and a posterior probability $P(M \mid \mathbf{Y})$ estimating the strength of each model M for BMA analysis using the data $\mathbf{Y}$. Employing this information as described in Wheeler et al. (2020), ToxicR then calculates and reports Bayesian model averaged $\widehat{\text{BMD}}$s and BMDLs. Sample R code for these operations is provided in a Supplemental document.

2.3. Frequentist Model Averaging (FMA) for BMD and BMDL calculation

For the frequentist model averaging (FMA) approach we again collected the seven forms for $R(x)$ listed in Section 2.1 and applied the FMA formulation given in Piegorsch et al. (2013). The method requires weights w_q for each qth model (q = 1, ..., 7) that describe the strength of that model's contribution to the averaging process:

$$w_q = \frac{exp\{-\Delta_q/2\}}{\sum_{q=1}^{Q} exp\{-\Delta_q/2\}} \quad (2.1)$$

where $\Delta_q = AIC_q - min\{AIC_1, \ldots .AIC_Q\}$ and AIC_q is the AIC from the fit of the *q*th model M_q. With these, the FMA estimate of the underlying BMD is

$$\overline{BMD} = \sum_{q=1}^{Q} w_q \widehat{BMD}_q \qquad (2.2)$$

where $\widehat{\mathrm{BMD}}_{\mathrm{q}}$ is the estimated BMD under model *q* and Q=7. The corresponding standard error is

$$se(\overline{BMD}) \approx \sum_{q=1}^{Q} w_q \left\{var\left[\widehat{BMD}_q | M_q\right] + \left(\widehat{BMD}_q - \overline{BMD}\right)^2\right\}^{1/2} \qquad (2.3)$$

leading to an FMA BMDL:

$$BMDL = \sum_{q=1}^{Q} w_q \widehat{BMD}_q - z_\alpha \sum_{q=1}^{Q} w_q \left\{var\left[\widehat{BMD}_q | M_q\right] + \left(\widehat{BMD}_q - \overline{BMD}\right)^2\right\}^{1/2} \qquad (2.4)$$

Here again, we do not present an exhaustive description of the FMA method and instead refer readers to the presentation in Piegorsch et al. (2013).

To calculate the various individual-model quantities at each *q* we can again appeal to the ToxicR package and its `single_dichotomous_fit()` function. Then applying these in Equations (2.1)-(2.4) we obtain the FMA estimate $\overline{\mathrm{BMD}}$ and its FMA BMDL. Sample R code for these calculations is provided in a Supplemental document.

2.4. Nonparametric Regression Analysis for BMD and BMDL calculation

A third estimation strategy that can avoid issues with single-model misspecification, model-independent/nonparametric regression techniques have been developed (Piegorsch et al., 2014). For this nonparametric regression analysis (NPAR), specification of the risk function $R(x)$ is not required. All that is necessary is the assumption that *R*(*x*) is continuous with a non-negative first derivative, i.e., a non-decreasing dose response. The nonparametric approach then calculates estimates of *R*(*x*) and $R_E(x)$ as

$$\tilde{R}(x) = \begin{cases} \tilde{p}_i & if\ x = x_i \\ \tilde{p}_i + \frac{\tilde{p}_{i+1} - \tilde{p}_i}{x_{i+1} - x_i}(x - x_i) & if\ x_i \le x < x_{i+1} \end{cases} \qquad (2.5)$$

$$\tilde{R}_E(x) = \begin{cases} \frac{\tilde{p}_i - \tilde{p}_1}{1 - \tilde{p}_1} & if\ x = x_i \\ \frac{1}{1 - \tilde{p}_1}\left\{\tilde{p}_i - \tilde{p}_1 + \frac{\tilde{p}_{i+1} - \tilde{p}_i}{x_{i+1} - x_i}(x - x_i)\right\} & if\ x_i \le x < x_{i+1} \end{cases} \qquad (2.6)$$

The BMD is then estimated at the BMR using a linear interpolating spline for $\tilde{R}_E(x)$, producing

$$\widetilde{BMD} = \begin{cases} x_i + \frac{(1-\tilde{p}_1)BMR - (\tilde{p}_i - \tilde{p}_1)}{\tilde{p}_{i+1} - \tilde{p}_i}(x_{i+1} - x_i) & if\ \frac{\tilde{p}_i - \tilde{p}_1}{1 - \tilde{p}_1} < BMR < \frac{\tilde{p}_{i+1} - \tilde{p}_i}{1 - \tilde{p}_1}\ \ for\ i = 1, \ldots .., m-1 \\ x_i & if\quad BMR = \frac{\tilde{p}_i - \tilde{p}_1}{1 - \tilde{p}_1}\qquad for\ i = 2, \ldots .., m \end{cases} \qquad (2.7)$$

To calculate the BMDL using this NPAR analysis we turned to nonparametric bootstrapping (Davison & Hinkley, 1997), as described in Section 2.2 of Piegorsch et al. (2014). Briefly, pseudo data are simulated using the original dataset and then used to calculate a pseudo $\widetilde{BMD}_b$ (b = 1, ..., B), where B is the number of bootstrapped pseudo datasets. We took B = 2000. The BMDL is then found as the lower 100αth pseudo BMD among the ordered values of $\widetilde{BMD}_b$.

To perform the bootstrapping, Piegorsch et al. recommend that $\tilde{R}(x)$ be locally concave near the estimated benchmark point. They give a short algorithm for checking the concavity which, if it fails, invalidates the NPAR estimates; see Piegorsch et al. (2014, Section 3.3). Sample R code for these operations is provided in a Supplemental document. We do not present a more detailed description of this nonparametric regression method and refer readers to the presentation given in Piegorsch et al. (2014).

2.5. The Quantal Risk Assessment Database (QRAD)

QRAD, presented by Wheeler et al. (2019), consists of 733 separate quantal dose-response datasets assembled from a variety of existing toxicological data sources such as the U.S. National Toxicology Program (NTP), the Integrated Risk Information System (IRIS), the State of California EPA (CalEPA), and the U.S. Centers for Disease Control and Prevention's Agency for Toxic Substances and Disease Registry (ATSDR). Each quantal-response dataset contains m doses, x_i, with corresponding proportions, Y_i/N_i, of animals exhibiting a specific type of cancer (Y_i) out of the N_i animals tested (i = 1, ..., m). We assume a binomial parent distribution for Y_i: $Y_i \sim$ bin.($R(x_i)$, N_i). For all the datasets we standardized the dose, x, to be between 0 and 1.

To fit our statistical models properly to the QRAD data, we required datasets that contained at least four dose groups – i.e., $m \geq 4$ – and as mentioned above, that had a non-decreasing dose-response pattern. Out of 733 datasets, 405 had $m \geq 4$. Further, to determine which of these 405 datasets exhibited a non-decreasing dose-response pattern we appealed to the Cochran-Armitage trend test (Cochran, 1954; Armitage, 1955). If the test gave a one-sided upper p-value at or below 0.15, we viewed the dose-response as non-decreasing. Out of the 405 datasets, 382 had one-sided p-values less than or equal to 0.15 and were retained for our analysis. An additional three datasets were discarded due to failure of the ToxicR function `single_dichotomous_fit()` for at least one of the seven models in Section 2.1, leaving 379 datasets from QRAD for comparison of the model-robust methods.

3. Results

3.1. Algorithm Failure Rates

The BMA method was able to complete the BMDL calculations on all of our 379 QRAD datasets, i.e., 0.8% of the time BMA method was unable to calculate BMD and BMDL values. This is a remarkable success rate for a complicated estimation technology applied to a large, diverse knowledgebase. By contrast, the FMA method was unable to calculate $\overline{\mathrm{BMD}}$ via the `single_dichotomous_fit()` function, as one or mode of the model fits produced `NaN` (not a number) values. Of the remaining 338 datasets, a further 133 datasets produced a negative FMA BMDL which we labeled as NA.

The NPAR method produced $\widetilde{BMD}$s for all 379 QRAD datasets, but it was unable to produce a BMDL for 135 QRAD datasets as the requirement for the local concavity of $R(x)$ near the estimated benchmark point was not met. This gave NPAR BMDLs for only 244 of the QRAD datasets. A negative estimate for the BMD is not of practical use and we labeled these values as NaN (not a number). Thus, 11% of the time the FMA method was not able to produce BMD values and 46% of the time it was not able to produce meaningful BMDL values. The nonparametric method was able to calculate BMD values on all 379 datasets, however it was not able to calculate the BMDL in 135 out of 379 datasets. That is, the NPAR method failed to produce BMDL values 36% of the time. Table 1 summarizes the failure rates for all methods, coded as BMDn and BMDLn (for the NPAR method), BMDb and BMDLb (BMA method), and BMDfma and BMDLfma (FMA method). Focusing on the BMDLs, we see that the BMA method was far more resilient in calculating the benchmark lower limits.

Value	Percent
BMDn	0 %
BMDb	0.8%
BMDfma	11%
BMDLn	36%
BMDLb	0.8%
BMDLfma	46%

Table 1. Percent of time each algorithm failed or was otherwise unable to calculate meaningful BMD and BMDL estimates. See text for algorithm codes.

3.2. Graphical summaries

Frequency graphs in Figure 1 and boxplots in Figure 2 of the BMA BMD and BMDL (again labeled as BMDb and BMDLb respectively), FMA BMD and BMDL (BMDfma and BMDLfma), and NPAR BMD and BMDL (BMDn and BMDLn) all show positively skewed distributions for each of the six quantities. This is not unexpected: tested compounds from QRAD are suspected carcinogens or other toxins, for which BMD-based limits will likely be rather small. Very large, "safer" limits would then be less common. Further, these summary graphics suggest that NPAR provides more conservative (i.e., closer to zero) values for BMD than BMA and FMA, while the BMA and FMA methods provide more conservative values for the BMDL than NPAR.

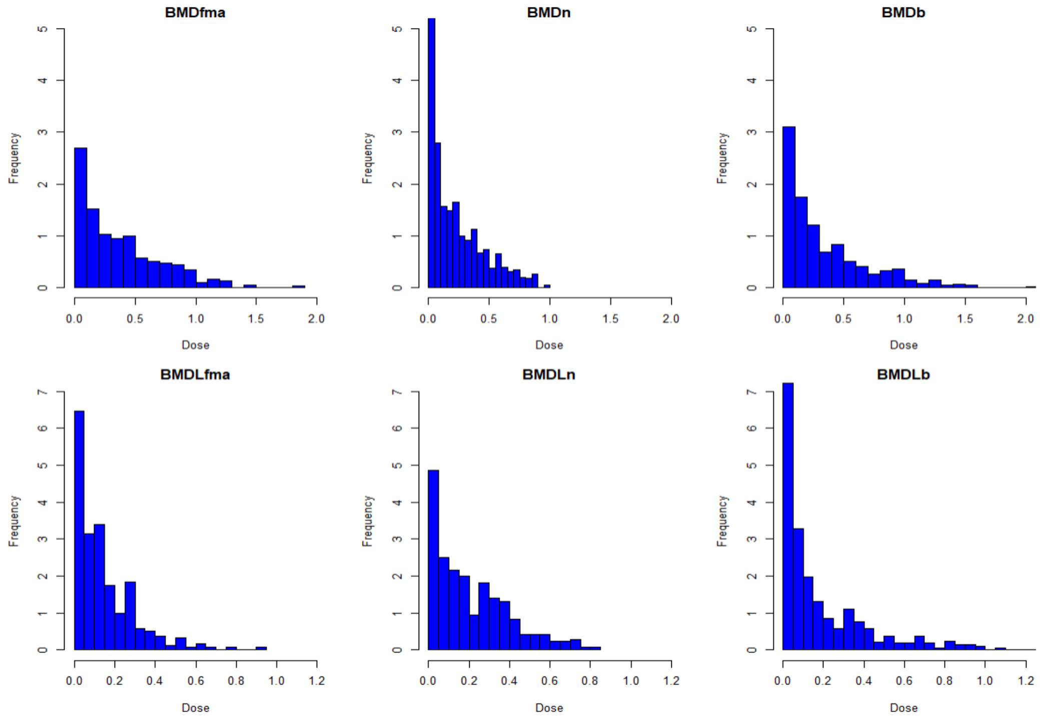

Fig. 1. Frequency of BMD and BMDL estimates from QRAD using each of the BMDfma, BMDn, BMDb, BMDLfma, BMDLn and BMDLb approaches. See text for algorithm codes.

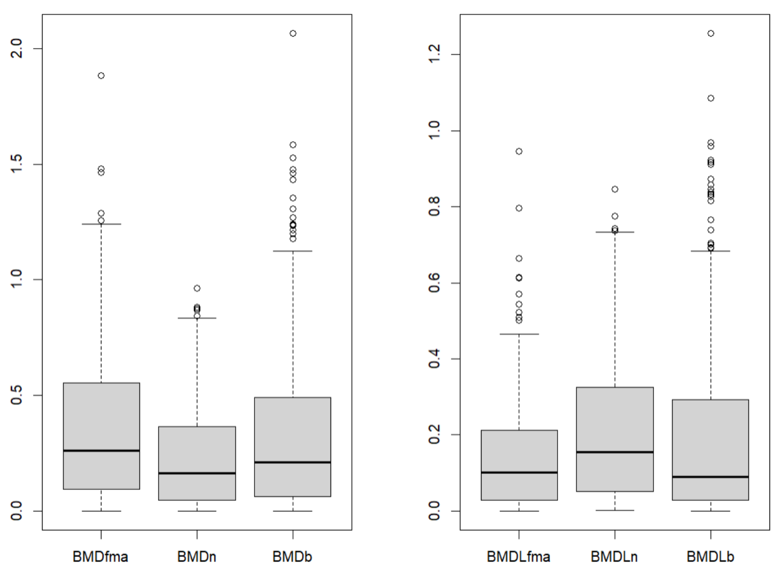

Fig. 2. Boxplots of BMD and BMDL estimates from QRAD using each of the BMDfma, BMDn, BMDb, BMDLfma, BMDLn and BMDLb approaches. See text for algorithm codes.

Scatterplots comparing (i) BMDfma vs. BMDn, (ii) BMDfma vs. BMDb, (iii) BMDn vs. BMDb, (iv) BMDLfma vs. BMDLn, (v) BMDLfma vs. BMDLb, and (vi) BMDLn vs. BMDLb appear in Figure 3. Compared to a 45 degree line which indicates perfect agreement (red lines in the figures), the agreement in the figures is generally good. The most stable relationship occurs for BMDfma vs. BMDb, i.e., the two model averaging methods, but this is not as strong for the corresponding BMDLs. Indeed, in all the other pairings there is bleed away from the 45 degree line. Those four plots are pairings of the model-averaging methods with NPAR and all indicate that when close-to-perfect agreement is disturbed, NPAR tends to give smaller values than BMA and FMA. This is consistent with indications in the frequency plots and boxplots, above.

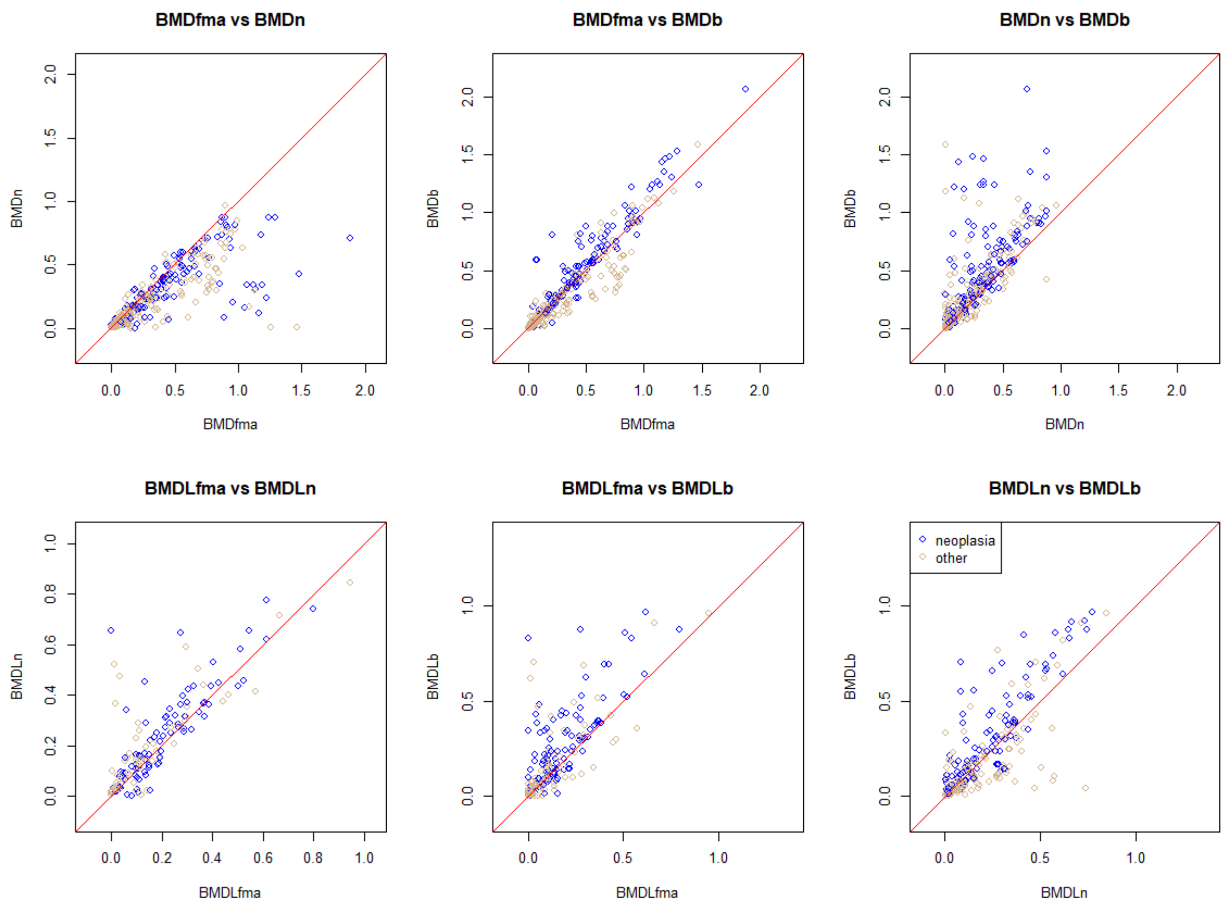

Fig. 3. Scatterplots from QRAD data of (i) BMDfma vs. BMDn, (ii) BMDfma vs. BMDb, (iii) BMDn vs. BMDb, (iv) BMDLfma vs. BMDLn, (v) BMDLfma vs. BMDLb, and (vi) BMDLn vs. BMDLb. Red lines are 45 degree lines marking perfect agreement. See text for algorithm codes.

3.3. Organ System Correlations

Within QRAD, each dataset is tied to a specific organ or organ system. This provides information on the organ system in which the particular toxicological or carcinogenic outcomes were observed. We investigated the relationships among the BMDfma, BMDb and BMDn values by examining the standard Pearson correlation coefficient, r, between each pairing for each organ; see Table 2. In all cases the correlations are positive. Across all points there appears to be a strong positive correlation between BMDn and BMDb (r = 0.74), a strong positive correlation between BMDfma and BMDn (r = 0.76) across all points, and a strong positive correlation between BMDfma and BMDb (r = 0.93). Across all points there is also a strong positive correlation between BMDLn and BMDLb (r = 0.79), between BMDLfma and

BMDLn (r = 0.83) and between BMDLfma and BMDLb (r = 0.75); see Table 2 for the BMD estimate correlations. Table 3 gives corresponding values for the BMDLs. The tables also give the correlations when stratified by a selection of high-interest organs/organ systems. High or very high positive correlations are exhibited for almost all the pairings, although we could not identify any particular across-organ pattern. A few cases appear where the correlation between FMA and NPAR is positive but close to zero; see, e.g., BMDfma vs. BMDn correlations in Developmental systems, or BMDLfma vs. BMDLn in the Bladder and Lung. The BMDLfma vs. BMDLb correlation in the Bladder is also marginally close to zero, although it moves very close to 1 for BMDLn vs. BMDLb. Viewing correlations close to 1 as indicative of greater within-method stability, this gives stronger support for the BMA and NPAR approaches.

Category	BMDfma vs. BMDn	BMDfma vs. BMDb	BMDn vs. BMDb
All Points	0.7638	0.9342	0.7414
Clinical Signs	0.9991	0.8994	0.8807
Developmental	0.1487	0.9953	0.3943
Mortality/Survival	0.6474	0.6404	0.9584
Neoplasia, all	0.7568	0.9554	0.7272
Nonneoplastic Histopathology	0.8703	0.9191	0.8649
Reproduction	0.4305	0.9943	0.5246
Bladder	0.7305	0.9935	0.7184
Kidneys	0.7803	0.9149	0.7823
Liver	0.8239	0.9367	0.7612
Lung	0.8989	0.7105	0.7503
Nasal Passages	0.6938	0.9401	0.6208
Stomach/Forestomach	0.7358	0.8820	0.6413

Table 2. Pearson correlation coefficients, r, between BMDfma, BMDb and BMDn values from QRAD data, stratified by organ/organ system. See text for algorithm codes.

In general, the comparison values in these figures and tables suggest that the three model-robust estimation and inferential methods give BMDs and BMDLs that are similar both across all points and, for the most part, in many specific organs/organ systems. This suggests a general level of good stability for any of these model-robust methods, as different methods lead to the same general inferences.

Category	BMDLfma vs. BMDLn	BMDLfma vs. BMDLb	BMDLn vs. BMDLb
All Points	0.8261	0.7496	0.7862
Clinical Signs	0	0	1
Development	NA	NA	0
Mortality/Survival	0.4756	0.1905	0.7110
Neoplasia, all	0.8360	0.7622	0.8694
Nonneoplastic Histopathology	0.7240	0.6256	0.6795
Reproduction	NA	NA	NA
Bladder	0.1394	0.2497	0.9252
Kidneys	0.9230	0.9096	0.8045
Liver	0.8513	0.7857	0.8481
Lung	0.1380	0.4354	0.6103
Nasal Passages	0.8900	0.8926	0.5321
Stomach/Forestomach	0.9294	0.9451	0.9438

Table 3. Pearson correlation coefficients, r, between BMDLfma, BMDLb and BMDLn values from QRAD data, stratified by organ/organ system. NA indicates that the correlation could not be calculated. See text for algorithm codes.

4. Conclusions and Discussion

In this study we have compared performance of Bayesian model averaging (BMA; Wheeler et al. 2020), frequentist model averaging (FMA; Piegorsch et al., 2013), and nonparametric regression modeling (NPAR; Piegorsch et al., 2014) for BMD and BMDL calculation on 382 datasets from Quantal Risk Assessment Database (QRAD). Our findings indicate that the three approaches perform similar to each other, both in general and in a variety of specific organs/organ systems for benchmark calculations with quantal carcinogenicity or other toxicological outcomes. For instance, all three methods exhibited generally high correlations along with reasonable, increasing, pairwise scatterplots for both BMD estimation and for BMDLs, and we did not notice any glaring discrepancies among their benchmark estimates. This is not altogether surprising, as all three were designed to estimate the same quantities and be resilient to model uncertainty and model misspecification.

Perhaps the most intriguing finding is the strong stability of BMA in terms of very low computational failures: over 99% of the available QRAD datasets produced BMD estimates and consequent BMDLs. For NPAR we expected a certain propensity to fail, as it comes with a local concavity requirement that appears to be nontrivial with real data. The contrasting failure rate for FMA, however, was a bit of a surprise. We did see some general similarities between the two model averaging methods; however, we expected the two to be more similar than they are different. Unfortunately, this expectation is now under question.

Of course, some caveats are in order, particularly for the model averaging approaches. Greatest among these is specification of the model averaging class. We chose a collection of seven parametric models (Section 2.1) which are very common in quantal-response risk analysis. Many more quantal-response function are available, however, and other analysts might wish to include some of these, and for that matter to remove some of our choices, when building their model averaging model class. Still, we expect small deviations in the list of considered models to make only a minor difference in the eventual BMDLs. Both BMA and FMA are designed to give lower weight(s) to models that fit poorly, giving them less impact on the benchmark calculations. Also, each of the three methods possesses requirements and assumptions to calculate BMD and BMDL estimates for any given dataset, as discussed in Section 2. If the requirements are not met, the benchmarking can fail or even worse, can produce BMDLs that are incorrect and possibly unsafe. We encountered the former event a number of times, leading to the failure rate issues mentioned above. As with any risk-analytic calculations, analysts and risk managers must be cognizant of the pitfalls, and the benefits, their methods possess.

Thus we are led to two basic recommendations. First, any of the three model-robust approaches we study can provide reasonable and consistent BMD estimates and BMDLs provided that the in-stream computations do produce actual values for risk benchmarking and do not fail. The possibility of the latter was made clear in our investigation, with the BMA method proving to be perhaps the most stable. Thus until further studies can explicate finer distinctions among model-robust approaches for BMD analysis, we encourage analysts to calculate BMD estimates and BMDLs via the BMA strategy as described in Wheeler et al. (2020).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, Petrov B. N. & Csaki B., (eds), 267–281. Akademiai Kiado, Budapest.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.

Christensen, R., Johnson, W. O., Branscum, A. J., & Hanson, T. E. (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Boca Raton, FL: Chapman & Hall/CRC Press.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10**, 417-451.

Crump, K. S. (1984). A new method for determining allowable daily intake. *Fundamental and Applied Toxicology* **4**, 854-871.

Crump, K. S. (1995). Calculation of benchmark doses from continuous data. *Risk Analysis* **15**, 79–89.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Fang, Q., Piegorsch, W. W., Simmons, S. J., Li, X., Chen, C., & Wang, Y. (2015). Bayesian model-averaged benchmark dose analysis via reparameterized quantal-response models. *Biometrics* **71**, 1168-1175.

Haber, L. T., Dourson, M. L., Allen, B. C., Hertzberg, R. C., Parker, A., Vincent, M. J., Maier, A., & Boobis, A. R. (2018). Benchmark dose (BMD) modeling: current practice, issues, and challenges. *Critical Reviews in Toxicology* **48**, 387-415.

Faustman EM, Bartell SM. (1997). Review of noncancer risk assessment: Applications of benchmark dose methods. *Human and Ecological Risk Assessment* **3**(5): 893–920.

Jensen, S. M., Kluxen, F. M., & Ritz, C. (2019). A review of recent advances in benchmark dose methodology. *Risk Analysis* **39**, 2295-2315.

Kang, S. H., Kodell, R. L., & Chen, J. J. (2000). Incorporating model uncertainties along with data uncertainties in microbial risk assessment. *Regulatory Toxicology and Pharmacology* **32**, 68–72.

Piegorsch, W. W., An, L., Wickens, A. A., West, R. W., Peña, E. A., & Wu, W. (2013). Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. *Environmetrics* **24**, 143-157.

Piegorsch, W. W., Xiong, H., Bhattacharya, R. N., & Lin, L. (2014). Benchmark dose analysis via nonparametric regression modeling. *Risk Analysis*, **34**(1), 135-151.

R Development Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ringblom, J., Johanson, G., & Öberg, M. (2014). Current modeling practice may lead to falsely high benchmark dose estimates. *Regulatory Toxicology and Pharmacology* **69**, 171–177.

Shao, K., & Shapiro, A. J. (2018). A web-based system for Bayesian benchmark dose estimation. *Environmental Health Perspectives*, **126**(1), 017002.

Simmons, S. J., Chen, C., Li, X., Wang, Y., Piegorsch, W. W., Fang, Q., & Dunn, G. E. (2015). Bayesian model averaging for benchmark dose estimation. *Environmental and Ecological Statistics* **22**(1), 5–16.

Stern, A. H. (2008). Environmental health risk assessment. In *Encyclopedia of Quantitative Risk Analysis and Assessment* **2**, Melnick, E. L. & Everitt, B. S. (eds.), 580–589.Chichester: John Wiley & Sons.

U.S. EPA (2012). *Benchmark Dose Technical Guidance Document*. Technical Report number EPA/100/R-12/001. Washington, DC: U.S. Environmental Protection Agency.

West, R. W., Piegorsch, W. W., Peña, E. A., An, L., Wu, W., Wickens, A. A., Xiong, H., & Chen, W. (2012). The impact of model uncertainty on benchmark dose estimation. *Environmetrics* **23**, 706–716.

Wheeler, M.W., Blessinger, T., Shao, K., Allen, B.C., Olszyk, L., Davis, J.A. , Gift, J.S (2020). Quantitative risk assessment: developing a Bayesian approach to dichotomous dose-response uncertainty. *Risk Analysis* **40**(9), 1706–1722.

Wheeler, M.W., Cortinas Abrahantes, J., Aerts, M., Gift, J.S., Allen Davis, J. (2022). Continuous model averaging for benchmark dose analysis: Averaging over distributional forms. *Environmetrics* **33**(5), Article no. e2728.

APPENDIX B (Supplemental Material for Chapter 3)

Bayesian benchmark dose risk assessment with mixed-factor quantal data

Mirjana Glisovic-Bensa [1]

Walter W. Piegorsch [2,3]

Edward J. Bedrick [1,2,3]

[1] Department of Epidemiology & Biostatistics, University of Arizona, Tucson, Arizona, USA

[2] Interdisciplinary Program in Statistics & Data Science, University of Arizona, Tucson, Arizona, USA

[3] BIO5 Institute, University of Arizona, Tucson, Arizona, USA

Correspondence

Walter W. Piegorsch, BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA. Email: piegorsch@math.arizona.edu

Abstract

Benchmark analysis is a general risk estimation strategy for identifying the benchmark dose (BMD) past which the risk of exhibiting an adverse environmental response exceeds a fixed, target value of benchmark response. Estimation of BMD and of its lower confidence limit (BMDL) is well understood for the case of an adverse response to a single stimulus. In many environmental settings, however, one or more additional, secondary, qualitative factor(s) may collude to affect the adverse outcome, such that the risk changes with differential levels of the secondary factor. Bayesian methods for estimation of the BMD and BMDL have grown in popularity, and a large variety of candidate dose-response models is available for applying these methods. This article applies Bayesian strategies to a mixed-factor setting with a secondary qualitative factor possessing two levels to derive two-factor Bayesian BMDs and BMDLs. We present reparameterized dose-response models that allow for explicit use of prior information on the target parameter of interest, the BMD. We also enhance our Bayesian estimation technique for BMD analysis by applying Bayesian model averaging to produce the BMDs and BMDLs, overcoming associated questions of model adequacy when multimodel uncertainty is present. An example from environmental carcinogenicity testing illustrates the calculations.

KEYWORDS

Bayesian inference, benchmark analysis, BMDL, environmental risk assessment, prior information, quantitative risk assessment

1 INTRODUCTION

1.1 Benchmark Risk Analysis

An important objective in environmental risk assessment is characterization of the severity and likelihood of adverse outcomes caused by a hazardous agent (Stern, 2008). In this context, the risk function, $R(x)$, is defined as the probability of some predefined adverse effect, such as death, cancer, mutation, birth defect, etc., in a subject exposed to a hazardous agent at dose or exposure level x. This definition assumes that risk changes with increasing exposure at dose x to the hazardous agent. Commonly, observations in such risk assessments are the proportions of subjects exhibiting the adverse response associated with each *x*. This is called the *quantal response setting* and is our focus here. A contemporary approach to quantal-response risk estimation is known as *benchmark analysis*. First introduced by Crump (1984), this method uses components of $R(x)$ to produce the benchmark dose, BMD, of the hazardous agent, at which a prespecified adverse response, BMR, is obtained. To account for the background risk, $R(0)$, that all subjects in a population would be expected to encounter in the absence of the hazardous agent, the risk function, $R(x)$, is refined into a form of *extra risk function,* $R_E(x)$. $R_E(x)$ is defined as the additional risk among subjects who would have not responded under control conditions: $R_E(x) = \frac{R(x)-R(0)}{1-R(0)}$.

From this, the benchmark dose, BMD, is found by solving for x in $R_E(x) = \text{BMR}$. The target BMR is often selected in the range $0.01 \leq \text{BMR} \leq 0.10$ (U.S. EPA, 2012).

To account for uncertainty associated with the point estimator, $\widehat{\text{BMD}}$, $100(1-\alpha)\%$ lower limits for the BMD are calculated. Driven by focus on risk reduction, only one-sided, lower limits (BMDLs) are of interest (Crump, 1995). The following example illustrates some of the considerations in a benchmark analysis.

1.2 Example: Risk analysis of goldenseal root powder (*Hydrastis canadensis*)

Goldenseal root powder (*Hydrastis canadensis*) is an herb in the buttercup family native to North America. Dried root of goldenseal is considered a natural antibiotic and is used in a variety of human activities. However, studies at the U.S. Environmental Protection Agency (EPA) raised questions about possible carcinogenic potential of this herb. With such relatively widespread usage, the U.S. National Toxicology Program (NTP) was called upon to conduct a mammalian carcinogenicity study of goldenseal root in which laboratory rodents were exposed to different concentrations of the powder (U.S. National Toxicology Program, 2010). F344/N male and female rats were exposed, through their feed, to 0, 3,000, 9,000 and 25,000 ppm of goldenseal root powder. Quantal-response data on induction of liver cancer were recorded for both males and females. Table 1 summarizes results of this study.

---Insert Table 1 here---

The data in Table 1 illustrate an example of a two-factor quantal-response model, where the goldenseal dose was the target factor and a discrete indicator (sex) served as a secondary factor. Simmons et al. (2015) analyzed these data from a Bayesian perspective, but only studied single-factor effects in the males. We will examine the full data with males *and* females in Section 4, below.

The data in Table 1 represent an unusual, mixed-factor design for benchmark risk analysis: Sans-Fuentes and Piegorsch (2021) adopted it and derived expressions for the BMD and BMDL using a frequentist approach. To our knowledge, no Bayesian methodology has appeared that allows for incorporation of prior information in this mixed-factor setting. A Bayesian approach was taken for the general case of one quantitative factor, the dose or exposure level, by Fang at al. (2015a). Those authors derived posterior expressions directly on the BMD and BMDL. In Fang et al. (2015b), they extended their Bayesian approach by employing Bayesian model averaging to account for different possible models for $R(x)$. Past the Sans-Fuentes and Piegorsch article, however, few works exist for broadly modifying single-factor benchmark analyses to accommodate secondary qualitative factors. Deutsch and Piegorsch (2012, 2013), following on work by Zhu (2005), expanded the risk benchmarking effort to the general problem of two continuous exposures. Wheeler and Bailer (2017) incorporated a secondary qualitative factor while focusing on a single continuous predictor; however, their models only considered a continuous outcome variable. Nonetheless, their approach was an early use of Bayesian techniques for benchmark risk assessment with multiple predictors. Boone et al. (2015) and Farhat et al. (2020) presented benchmarking strategies similar to Wheeler and Bailer (2017); they too incorporated multiple factors. In addition, their model expanded to multiple endpoints with both quantitative and quantal responses. As essentially none of these prior works – aside from Wheeler and Bailer (2017) – has focused on benchmarking with secondary qualitative factors, in this paper we blend Fang et al.'s reparametrized Bayesian strategies with the Sans-Fuentes and Piegorsch mixed-factor model to derive two-factor Bayesian BMDs and BMDLs in this mixed-factor setting. Section 2 describes a reparameterized Bayesian strategy similar to that in Fang et al. (2015a), allowing for direct operations on the BMD. To account for potential uncertainty in selection of a model for $R(x)$, Section 3 extends this to a Bayesian model averaging (BMA) strategy. Section 4 returns to the goldenseal root carcinogenicity data example in Table 1, while Section 5 ends with a brief discussion.

2 Bayesian benchmark analysis with quantal outcomes in the mixed-factor setting

2.1 Reparametrizing two-factor models under quantal response

With quantal data such as those in Table 1 we take Y_i as the number of subjects responding adversely out of N_i, the number of subjects tested at the dose i. Here we employ a binomial model and assume that the Y_is are independent such that $Y_i \sim Bin(N_i, R(x_i, u))$, where $R(x_i, , u)$ is the extended risk function for our two-factor setting. It gives the probability of responding adversely at dose x_i for a subject from group u, where

$$u = \begin{cases} 0 \text{ if male} \\ 1 \text{ if female} \end{cases}.$$

For $R(x, u)$ we consider five models often seen in toxicological and carcinogenic risk assessment: four discussed by Simmons et al. (2015) in their BMA study of single-factor BMDs – logistic, probit, quantal-linear, and quantal-quadratic – and an additional model popular with quantal data, the complementary-log-log. Table 2 lists formal expressions for each model.

---Insert Table 2 here---

For a Bayesian approach we specify prior distributions for the unknown model parameters such as the $\beta\ and\ \delta$-parameters in Table 2. With quantal data, it is unusual to have informative prior knowledge on parameters such as $\beta_0, \beta_1, \delta_0, \delta_1$, however, because their interpretation is so generic. More likely in practice is prior knowledge on quantities such as the benchmark dose, BMD, and the background risk, $R(0)$. For notational simplicity denote each BMD$_j$ by ξ_j and the background risk $R_j(0)$ by γ_j, where j =0 if male, or 1 if female. Notice in Table 2 that each model contains, in effect, a dose×group interaction. This can be viewed as modeling separate dose-response curves for each group and thus making the BMDs are necessarily group specific. Conveniently, high-quality, historical control information is available for the γ_js: the U.S. NTP publishes pertinent historical control rates for many of its carcinogenicity experiments (e.g., see the NTP technical report TR562, wherein Tables A3a and B3 provide historical data on goldenseal root carcinogenicity). Given the historical control means $\overline{\gamma}_j$ and standard deviations S_j, Method of Moment estimates of a_j and b_j in γ_j ~ independent Beta(a_j, b_j) can be found as $a_j^* = \left(\frac{\overline{\gamma}_j\left(1-\overline{\gamma}_j\right)}{S_j^*} - 1\right)\overline{\gamma}_j$ and $b_j^* = \left(\frac{\overline{\gamma}_j\left(1-\overline{\gamma}_j\right)}{S_j^*} - 1\right)\left(1 - \overline{\gamma}_j\right)$. If in addition prior information is available for ξ_j it can be incorporated as well. Otherwise, we set ξ_j ~ indep. U(0, max{x_i}), j=0,1, anticipating that a BMD would not exceed the largest dose at which a putative environmental carcinogen is studied.

We note in passing that typical applications of BMA often place separate priors on the regression effects for each model, but for convenience they may be set to be the same. - Of value with our model construct is that the same prior on interpretable effects essentially induces the prior on the regression coefficients for the different models (because models have same number of parameters). Thus different models have different priors on the regression parameters, but these different priors all map to the same prior on quantities that are easily interpreted. This may not occur with standard approaches that start by placing priors on the regression coefficients.

To employ the prior information we reparametrize $R(x,u)$ in terms of $\xi_j$and γ_j. For example, under the quantal-liner model $R(x,u) = 1 - e^{-\beta_0-\beta_1 x-\delta_0\, u-\delta_1 xu}$ from Table 2 the background risk for males is

$\gamma_0 = R(0,0) = 1 - e^{-\beta_0}$ and the background risk for females is $\gamma_1 = R(0,1) = 1 - e^{-\beta_0 - \delta_0}$. Clearly then, $\beta_0 = -\log\,(1-\gamma_0)$ and $\delta_0 = -\log\,((1-\gamma_1)/(1-\gamma_0))$. To find the ξs we solve for x in $R_E(x) =$ BMR. This gives

$$\xi_0 = -\frac{\log(1-BMR)}{\beta_1} = \frac{c}{\beta_1},$$

where c = –log(1 – BMR) and $\xi_1 = c/(\beta_1 - \delta_1)$. Then $\beta_1 = c/\xi_0$ and $\delta_1 = c\left(\frac{1}{\xi_1} - \frac{1}{\xi_0}\right)$. The reparametrized quantal-linear model becomes

$$R(x,u) = 1 - e^{\log(1-\gamma_0) - \frac{c}{\xi_0}x + u\log\left(\frac{1-\gamma_1}{1-\gamma_0}\right) - c\left(\frac{1}{\xi_1} - \frac{1}{\xi_0}\right)ux}.$$

We apply the same technique to reparametrize the other four models and obtain the results given in Table 3:

---Insert Table 3 here---

2.2 Bayesian benchmark analysis for reparametrized quantal response models.

To apply a Bayesian analysis under the hierarchy described in Sec. 2.1, we assign a joint p.d.f. to the reparameterized vector $\theta = [\xi_0\ \xi_1\ \gamma_0\ \gamma_1]^{\mathrm{T}}$. We write the prior generically as $\pi(\theta) = \pi(\xi_0, \xi_1, \gamma_0, \gamma_1)$.

Mimicking previous Bayesian models for benchmark analysis (Shao and Small, 2011; Shao, 2012), we assume the unknown parameters enter into the prior independently, so that $\pi(\xi_0, \xi_1, \gamma_0, \gamma_1). = \pi(\xi_0)\pi(\xi_1)\pi(\gamma_0)\pi(\gamma_1)$.

As proposed in Sec. 2.1, we take $\xi_j \sim U(0, \max\{x_i\})$ and $\gamma_j \sim Beta(a_j^*, b_j^*)$, j = 0,1. Using a binomial parent distribution for the data the corresponding joint posterior distribution for ξ_0, ξ_1, γ_0 and γ_1 is:

$$\pi(\xi_0, \xi_1, \gamma_0, \gamma_1|\boldsymbol{Y}) = \frac{\prod_{i=1}^{n}\binom{N_i}{Y_i}R(x_i,u_i)^{\ Y_i}\{1-R(x_i,u_i)\ \}^{N_i-Y_i}}{m(\boldsymbol{Y})} \times \frac{1}{\max\{x_0\}} \times \frac{1}{\max\{x_1\}} \times$$

$$\frac{\Gamma(a_0^* + b_0^*)}{\Gamma(a_0^*)\Gamma(b_0^*)}{\gamma_0}^{a_0^*}(1-\gamma_0)^{b_0^*-1} \times \frac{\Gamma(a_1^* + b_1^*)}{\Gamma(a_1^*)\Gamma(b_1^*)}{\gamma_1}^{a_1^*}(1-\gamma_1)^{b_1^*-1}$$

where

$$m(\boldsymbol{Y}) = \iiint\int \prod_{i=1}^{n}\binom{N_i}{Y_i}R(x_i,u_i)^{Y_i}\{1-R(x_i,u_i)\ \}^{N_i-Y_i} \times \frac{1}{\max\{x_0\}} \times \frac{1}{\max\{x_1\}} \times$$

$$\frac{\Gamma(a_0^*+b_0^*)}{\Gamma(a_0^*)\Gamma(b_0^*)}{\gamma_0}^{a_0^*}(1-\gamma_0)^{b_0^*-1} \times \frac{\Gamma(a_1^*+b_1^*)}{\Gamma(a_1^*)\Gamma(b_1^*)}{\gamma_1}^{a_1^*}(1-\gamma_1)^{b_1^*-1}d\xi_0 d\xi_1 d\gamma_0 d\gamma_1$$

is the marginal distribution of **Y** in this Bayesian hierarchy.

For historical data on the γ parameters we refer to values from the U.S. NTP (2010, Tables A3a and B3): the historical mean and standard deviation for hepatocellular adenomas or carcinomas in male rats are $\overline{\gamma}_0$ = 7/300 = 0.023 and S_0 = 0.023, while those for female rats are $\overline{\gamma}_1$ = 4/250 = 0.016 and S_1 = 0.022. This leads to the informative priors $\gamma_0 \sim Beta(0.95, 39.74)$, independent of $\gamma_1 \sim Beta(0.51, 31.29)$. No such prior information was available for the BMDs, hence we employed simple, independent, uniform priors between zero and the maximum dose for each sex: ξ_j ~ indep.U(0, 25000).

Due to the non-conjugacy of the prior distributions here, the resulting posteriors do not take convenient forms. Thus, for the posterior calculations we turn to Monte Carlo Markov Chain (MCMC) posterior approximations using the *R2OpenBUGS* package (Sturtz et al., 2005). For a given data ensemble (N, Y, x, u), we generated a tetravariate chain of $(\xi_0, \xi_1, \gamma_0, \gamma_1)$ draws. With these complex, multidimensional models our experience has shown that a longer chain of $K = 200{,}000$ draws can achieve reasonable convergence in the MCMC chains. For specifying the burn-in we applied Geweke's t-test approach (Geweke, 1992) via the `geweke.diag()` function in the *coda* **R** package (v0.19-4; Plummer et al., 2006).

We chose to apply the same burn-in for each component of the tetravariate chain, i.e., we took the largest recommended burn-in and applied it to all four chains. This produced burn-ins of 40,000 iterations for each chain, allowing us to operate with 160,000 draws per model.

To estimate the BMDs in ξ_0 and ξ_1 we calculated the posterior sample means $\overline{\xi_0}$ and $\overline{\xi_1}$ of the 160,000 post-burn-in draws. To correct for multiple inferences on the same chain we found corresponding Bonferroni multiplicity-adjusted 95% Bayesian BMDLs as the lower 2.5% quantiles of the post-burn-in chains.

Table 4 summarizes the results for the goldenseal root data given in Table 1. The table lists model-specific BMDs for males and females together with their corresponding 95% BMDLs and Deviance Information Criterion (DIC) values for each of the five models. DIC is a Bayesian model selection criterion based on the principle DIC = goodness of fit + complexity (Spiegelhalter et al. 2014), therefore lower values of DIC are prefered.

The complementary log-log model, M_5 had the lowest DIC, producing multiplicity-adjusted 95% lower limits $BMDL_0$ = 3,007 ppm for males and $BMDL_1$ = 5,118 ppm for females, while the two-parameter quantal-linear model, $M_{3,}$ had the highest DIC with 95% lower limits $BMDL_0$ = 887 ppm for males and $BMDL_1$ = 1,245 ppm for females. Further, consistently lower BMD and BMDL values for male mice across all five models indicate that male mice appear more susceptible to liver tumors after exposure to goldenseal powder root. The 95% BMDL values for model M_5 are, however, about four times higher than the 95% BMDL values for model M_3 which clearly shows that if an incorrect model is selected, BMD and BMDL values could be grossly underestimated. Due to issues associated with the single model approach or with unadjusted model selection procedures, model averaging is advocated by several authors for estimating both the BMD and the associated BMDL(s) (Bailer, Noble, and Wheeler, 2005; Faes, Aerts, Geys, and Molenberghs, 2007; Jensen and Ritz, 2015; Kang, Kodell, and Chen, 2000; Moon, Kim, Chen, and Kodell, 2005; Namata et al., 2008; Wheeler and Bailer, 2007, 2008). We discuss this in the next section.

---Insert Table 4 here---

3. Bayesian Model Averaging (BMA)

With single-factor model, many parametric forms for $R(x)$ operate well at higher doses near the range of the observed quantal outcomes but can produce wildly different estimates on the BMD at very small levels of risk (Faustman and Bartell, 1997; Kang et al., 2000). The Bayesian approach described in Section 2 assumes that model selection was made without error, i.e., the choice for the risk function is correct. Alternatively, some investigators use formal selection metrics such as the popular Akaike (1973) information criteria (AIC); however, AIC-guided selection has been shown to select incorrect models for BMD estimation almost as often as it selects correct models (West et al., 2012). Incorrect model selection can lead to wildly incorrect BMDLs (Ringblom et al. 2014).

To mitigate problems associated with model selection in our mixed-factor paradigm, we perform Bayesian model averaging across five commonly used forms for the risk function. Suppose we have a set of competing models for $R(x_i, u_i)$, say, an Uncertainty Class $U_H = \{M_1, M_2, M_3, M_4, M_5\}$, where each M_h represents a different dose-response model with the same mixed-model specification for the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_i + \delta_0 u_i + \delta_1 u_i x_i.$$

Possibilities include the four forms Simmons et al. (2015) employed, along with the popular complementary log-log model, giving H = 5

M_1 (logistic): $R = \frac{1}{1+e^{-\eta}}$, where $\xi_{01} = \beta_1^{-1}\log\left(\frac{1+BMRe^{-\beta_0}}{1-BMR}\right)$

$$\xi_{11} = (\beta_1 + \delta_1)^{-1}\log\left(\frac{1+BMRe^{-\beta_0-\delta_0}}{1-BMR}\right)$$

M_2 (probit): $R = \Phi(\eta)$, where $\xi_{02} = \beta_1^{-1}\{\Phi^{-1}[BMR(1-\Phi(\beta_0)) + \Phi(\beta_0)] - \beta_0\}$

$$\xi_{12} = (\beta_1 + \delta_1)^{-1}\{\Phi^{-1}[BMR(1-\Phi(\beta_0+\delta_0)) + \Phi(\beta_0+\delta_0)] - \beta_0 - \delta_0\}$$

M_3 (quantal-linear): $R = 1 - e^{-\eta}$, where $\xi_{03} = -\log(1-BMR)/\beta_1$

$$\xi_{13} = -\log(1-BMR)/(\beta_1+\delta_1)$$

M_4 (quantal-quadratic): $R = 1 - e^{-(\beta_0+\delta_0 u)-(\beta_1+\delta_1 u)x^2}$, where $\xi_{04} = \sqrt{-\log(1-BMR)/\beta_1}$

$$\xi_{14} = \sqrt{-\log(1-BMR)/(\beta_1+\delta_1)}$$

M_5 (complementary log-log): $R = 1 - e^{-e^{\eta}}$, where $\xi_{05} = \beta_1^{-1}(C_0 - 2\beta_0)$

$$\xi_{15} = (\beta_1+\delta_1)^{-1}(C_1 - 2(\beta_0+\delta_0))$$

with $C_0 = \log\left\{-log\left[1 - BMR\left(e^{e^{\beta_0}-1}\right)\right]\right\}$ and $C_1 = \log\left\{-log\left[1 - BMR\left(e^{e^{\beta_0+\delta_0}-1}\right)\right]\right\}$.

With these five forms, Bayesian Model Averaging (BMA) synthesizes the models into a weighted average of marginal posteriors: $\pi(\xi_u|\boldsymbol{Y}, U_5) = \sum_{h=1}^{H}\pi(\xi_u|\boldsymbol{Y}, M_h)\,P(M_h|\boldsymbol{Y})$ where the posterior model probabilities are obtained through Bayes' formula, via: $P(M_h|\boldsymbol{Y}) = m(\boldsymbol{Y}|M_h)P(M_h)/\sum_{r=1}^{H} m(\boldsymbol{Y}|M_r)P(M_r)$. The $P(M_h|\boldsymbol{Y})$ values are the per-model posterior probabilities and $m(\boldsymbol{Y}|M_h)$ is the marginal likelihood under each M_h model, i.e., the probability of the observed data given the model

M_h. If, as is common, the prior model probabilities $P(M_r)$ are assumed uniform, then $P(M_r) = 1/H$ and $P(M_h|\boldsymbol{Y})$ simplifies to $P(M_h|\boldsymbol{Y}) = m(\boldsymbol{Y}|M_h)/\sum_{h=1}^{H} m(\boldsymbol{Y}|M_h)$.

Not surprisingly, it is challenging to calculate each $P(M_h|\boldsymbol{Y})$. We estimated these quantities via the bridge sampling approach described in Fang et al. (2014a). Once the H chains were realized, we found each of the BMD point estimates as $\hat{\bar{\xi}}_u = \sum_{h=1}^{H} P(M_h|\boldsymbol{Y})\bar{\xi}_{uh}$ $(u = 0,1)$.

In practice, more important are the BMDLs. To obtain them, following Fang et al. (2015b), we solved the equation (Bonferroni-adjusted)

$$0.025 = \sum_{h=1}^{H} P(M_h|\mathbf{Y}) \cdot \frac{1}{m^*} \sum_{j=1}^{m^*} I_{(0,\xi_u)}(\xi_{uh}|M_h)$$

$u = (0,1)$, where m^* is the size of the sample from model M_h (after burn-in).

4. Example Revisited: Bayesian Model Averaging for BMD and BMDL Calculation

We return to the goldenseal data from Table 1 and extended the analysis in Section 2.2 using BMA approach. Our analysis was performed using BMR = 0.01 and the bridge sampling approach described in Fang et al. (2014a). The quantal-quadratic model had the highest posterior model probability (weight) of 0.5142 – despite having the second-highest DIC – followed by the probit, logistic, and complementary log-log models whose posterior model probabilities (weights) were 0.1861, 0.1447 and 0.1345 respectively. The quantal-linear model, which had the highest DIC value, gave, as expected, the lowest posterior model probability (weight) of 0.0204. With these posterior model probabilities, we calculated Bayesian model averaged (BMA) BMD's and 95% BMDLs. The Bayesian model averaged BMD for males was 5,163 ppm with 95% BMDL of 3,467 ppm, while the Bayesian model averaged BMD for females was 7,799 ppm with 95% BMDL of 4,723 ppm. These model averaged values are in the range of the individual model values and are closest to the higher-weight models. By comparison, Simmons at al. reported Bayesian model averaged BMD for males of 4,592 ppm with 95% BMDL of 2,862 ppm. Values that Simmons at al. reported rest in the similar range as the values we obtained.

---Insert Table 5 here---

This example illustrates consequence of selecting a wrong model. If a wrong model had been chosen, the consequent BMD's and BMDL's could have been, in this example, far too small.

5. DISCUSSION

Here, we extended the single-dose hierarchical Bayesian framework for estimating benchmark doses in quantitative risk analysis to the multiple mixed-factor setting by incorporating a secondary factor with two qualitative levels. (The model structure can be easily extended to multiple qualitative levels for the second factor.) Our approach estimates the BMDs from a class of five quantal-response models commonly employed in practice. We also account for potential model selection uncertainty using a Bayesian model averaging strategy (BMA).

We apply our methodology to goldenseal root carcinogenicity and find consistently lower BMD and BMDL values for male mice compared to female mice, across all five models. This indicates that male mice appear more susceptible than females to liver tumors after exposure to goldenseal powder root. However, the 95% BMDL values for model M_5 are, about four times higher than the 95% BMDL values for model M_3. This ilustrates that if an incorrect model is selected, BMD and BMDL values could be grossly underestimated. To eliminate issues associated with the single model approach and with the model selection procedures, we apply Bayesian model averaging to obtain BMA BMDs and BMDLs.

Of course, while we perform Bayesian model averaging across five popular forms for $R(x, u)$ (logistic, probit, quantal-linear, quantal-quadratic, and complementary log-log), BMA across a wider and/or different selection of models is also feasible. For the single-factor setting, Wheeler and Bailer (2007) showed that the statistical outcomes can differ depending on the number and type of models included in the average. From this, Wheeler et al. (2020) recommend the use of standard suite of nine models (quantal-linear, multistage, Weibull, gamma, dichotomous hill, logistic, log-logistic, probit, log-probit) in every risk assessment. Further research is required to determine if the results presented here translate well to cases when our methodology is applied across uncertainty classes with larger and more-diverse models.

Conflict of interest disclosure: The authors declare that no conflict of interest exists in this research.

Data availability statement: All data are contained in Table 1.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, Petrov, B.N. and Csaki, B. (eds), 267–281. Akademiai Kiado, Budapest.

Boone, E. L., Brooks, J.P., Nyirabahizi, E., and Chen, X. (2015). A Bayesian model-based approach for determining multivariate tolerable regions. *Journal of Environmental Statistics* **7**(6), 1–21.

Bailer, A.J., Noble, R.B., and Wheeler, M.W. (2005). Model uncertainty and risk estimation for experimental studies of quantal responses. *Risk Analysis*, *25*(2), 291–299.

Crump, K. S. (1984). A new method for determining allowable daily intake. *Fundamental and Applied Toxicology* **4**, 854–871.

Crump, K. S. (1995). Calculation of benchmark doses from continuous data. *Risk Analysis* **15**, 79–89.

Deutsch, R.C. and Piegorsch, W.W. (2012). Benchmark dose profiles for joint-action quantal data in quantitative risk assessment. *Biometrics* **68**(4), 1313–1322.

Deutsch, R.C. and Piegorsch, W.W. (2013). Benchmark dose profiles for joint-action continuous data in quantitative risk assessment. *Biometrical Journal* **55**(5), 741–754.

Fang, Q., Piegorsch, W.W., and Barnes, K.Y. (2015a). Bayesian benchmark dose analysis. *Environmetrics* **26**, 373-382.

Fang, Q., Piegorsch, W. W., Simmons, S. J., Li, X., Chen, C., and Wang, Y. (2015b). Bayesian model-averaged benchmark dose analysis via reparametrized quantal-response models. *Biometrics*, **71**(4), 1168-1175.

Farhat, N. J., Boone, E. L., and Edwards, D. J. (2020). A new method for determining the benchmark dose tolerable region and endpoint probabilities for toxicology experiments. *Journal of Applied Statistics* **47**(5), 775–803.

Faustman, E. M. and Bartell, S. M. (1997). Review of noncancer risk assessment: Applications of benchmark dose methods. *Human and Ecological Risk Assessment* **3**, 893–920.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4,* Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.), 169-193. Oxford: Oxford University Press.

Hamon NW (1990). Goldenseal. *Canadian Pharmacists Journal* **123**(11), 508–510.

Jensen, S.M., and Ritz, C. (2015). Simultaneous inference for model averaging of derived parameters. *Risk Analysis*, 35(1), 68–76.

Kang, S. H., Kodell, R. L., and Chen, J. J. (2000). Incorporating model uncertainties along with data uncertainties in microbial risk assessment. *Regulatory Toxicology and Pharmacology* **32**, 68–72.

Namata, H., Aerts, M., Faes, C., and Teunis, P. (2008). Model averaging in microbial risk assessment using fractional polynomials. *Risk Analysis*, *28*(4), 891–905.

Plummer M, Best N, Cowles K, Vines K (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, **6**(1), 7–11. https://journal.r-project.org/archive/.

Ringblom, J., Johanson, G., and Öberg, M. (2014). Current modeling practice may lead to falsely high benchmark dose estimates. *Regulatory Toxicology and Pharmacology* **69**, 171–177.

Sans-Fuentes, M. A., and Piegorsch, W. W. (2021). Benchmark dose risk analysis with mixed-factor quantal data in environmental risk assessment. *Environmetrics* **32**, Article No. e2677.

Simmons, S.J., Chen, C., Li, X., Wang, Y., Piegorsch, W.W., Fang, Q., Hu, B., and Dunn, G.E. (2015). Bayesian model averaging for benchmark dose estimation. *Environmental and Ecological Statistics* **22**, 5-16.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B* (*Statistical Methodology*) **76**, 485-493.

Stern, A. H. (2008). Environmental health risk assessment. In *Encyclopedia of Quantitative Risk Analysis and Assessment* **2**, Melnick, E. L. and Everitt, B. S., (eds.), 580–589. Chichester: John Wiley & Sons.

Sturtz, S., Liggers, U., and Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12(3), 1-16.

U.S. EPA (2012). *Benchmark Dose Technical Guidance Document*. Technical Report number EPA/100/R-12/001. Washington, DC: U.S. Environmental Protection Agency.

U.S. National Toxicology Program (2010). Toxicology and carcinogenesis studies of goldenseal root powder (*Hydrastis Canadensis*) in F344/N Rats and B6C3F1 Mice (Feed studies). Technical Report #562: U.S. Department of Health and Human Services, Public Health Service, Research Triangle Park, NC.

West, R. W., Pigorsch, W. W., Peña, E. A., An, L., Wu, W., Wickens, A. A., Xiong, H., and Chen, W. (2012). The impact of model uncertainty on benchmark dose estimation. *Environmetrics* **23**, 706–716.

Wheeler, M. W. and Bailer, A. J. (2007). Properties of model averaged BMDLs: A study of model averaging in dichotomous response risk estimation. *Risk Analysis*, *27*(3), 659–670.

Wheeler, M. W. and Bailer, A. J. (2008). Model averaging software for dichotomous dose response risk estimation. *Journal of Statistical Software*, *26*(5), 1–15.

Wheeler, M. W., Bailer, A. J., Cole, T., Park, R. M., and Shao, K. (2017). Bayesian quantile impairment threshold benchmark dose estimation for continuous endpoints. *Risk Analysis* **37**(11), 2107–2118.

Wheeler, M.W., Blessinger, T., Shao, K., Allen, B.C., Olszyk, L., Davis, J.A., Gift, J.S (2020). Quantitative risk assessment: developing a Bayesian approach to dichotomous dose-response uncertainty, Risk Analysis, 40 (9), 1706–1722.

Zhu, Y. (2005). Dose-time-response modeling of longitudinal measurements for neurotoxicity risk assessment. *Environmetrics* **16**(6), 603–617.

Table 1. Carcinogenicity data: hepatocellular adenomas or carcinomas in mice after exposure to goldenseal root powder. Data from U.S. NTP (2010).								
	Female F344/N Mice				Male F344/N Rats			
Exposure Concentration (ppm), x_i	0	3,000	9,000	25,000	0	3,000	9,000	25,000
Animals tested, N_i	50	50	50	50	50	50	50	50
Animals with tumors, Y_i	0	0	1	8	1	1	2	11

***Table 2**. Popular quantal-response models in toxicological and carcinogenic risk assessment.*		
Model	**Risk function**	**Form of risk function**
M_1	Logistic	$R(x,u) = 1/\left(1 + e^{-\beta_0 - \beta_1 x - \delta_0 u - \delta_1 x u}\right)$
M_2	Probit	$R(x,u) = \Phi(\beta_0 + \beta_1 x + \delta_0 u + \delta_1 x u)$
M_3	Quantal-linear	$R(x,u) = 1 - e^{-\beta_0 - \beta_1 x - \delta_0 u - \delta_1 x u}$
M_4	Quantal-quadratic	$R(x,u) = 1 - e^{-\beta_0 - \beta_1 x^2 - \delta_0 u - \delta_1 x^2 u}$
M_5	Complementary log-log	$R(x,u) = 1 - e^{-e^{\beta_0 + \beta_1 x - \delta_0 u - \delta_1 x u}}$

Table 3. *Common quantal-response models in toxicological and carcinogenic risk assessment in reparametrized forms. Model codes are from Table 2. Note: BMR∈(0,1) is the benchmark response, ξ_0 and ξ_1 are the BMDs ,and γ_0 and γ_1 are the background risks.*

Model	Reparametrized risk function $R(x,u)$
M_1	$R(x,u) = 1/\left(1+e^{-\beta_0^*-\beta_1^* x-\delta_0^* u-\delta_1^* xu}\right)$ where $\beta_0^* = \log\left(\gamma_0/(1-\gamma_0)\right)$ $\beta_1^* = \xi_0^{-1} \cdot \log\left[\frac{1+BMR\cdot\left(\frac{1-\gamma_0}{\gamma_0}\right)}{1-BMR}\right]$ $\delta_0^* = \log\left[\frac{\gamma_1(1-\gamma_0)}{\gamma_0(1-\gamma_1)}\right]$ $\delta_1^* = \xi_1^{-1} \cdot \log\left[\frac{1+BMR\cdot\left(\frac{1-\gamma_1}{\gamma_1}\right)}{1-BMR}\right] - \beta_1^*$
M_2	$R(x,u) = \Phi(\beta_0^* + \beta_1^* x + \delta_0^* u + \delta_1^* xu)$ where $\beta_0^* = \Phi^{-1}(\gamma_0)$ $\beta_1^* = [\Phi^{-1}(BMR \cdot (1-\Phi(\beta_0^*) + \Phi(\beta_0^*)) - \beta_0^*]/\xi_0$ $\delta_0^* = \Phi^{-1}(\gamma_1) - \beta_0^*$ $\delta_1^* = \frac{[\Phi^{-1}(BMR\cdot(1-\Phi(\beta_0^*+\delta_0^*))+\Phi(\beta_0^*+\delta_0^*))-(\beta_0^*+\delta_0^*)]}{\xi_1} - \beta_1^*$
M_3	$R(x,u) = 1 - e^{-\beta_0^*-\beta_1^* x-\delta_0^* u-\delta_1^* xu}$ where $\beta_0^* = -\log\left(1-\gamma_0\right)$ $\beta_1^* = \frac{c}{\xi_0}$ $\delta_0^* = -\log\left((1-\gamma_1)/(1-\gamma_0)\right)$ $\delta_1^* = c\left(\frac{1}{\xi_1} - \frac{1}{\xi_0}\right)$ $c = -\log\left(1-BMR\right)$
M_4	$R(x,u) = 1 - e^{-\beta_0^*+\beta_1^* x^2+\delta_0^* u+\delta_1^* x^2 u}$ where $\beta_0^* = -\log\left(1-\gamma_0\right)$ $\beta_1^* = \frac{c}{\xi_0^2}$ $\delta_0^* = -\log\left[\frac{(1-\gamma_1)}{(1-\gamma_0)}\right]$ $\delta_1^* = c(\xi_1^{-2} - \xi_0^{-2})$ $c = -\log\left(1-BMR\right)$
M_5	$R(x,u) = 1 - e^{-e^{\beta_0^*+\beta_1^* x+\delta_0^* u+\delta_1^* xu}}$ where $\beta_0^* = log[-log\left(1-\gamma_0\right)]$ $\beta_1^* = \frac{c_0-\beta_0^*}{\xi_0}$ $\delta_0^* = log\left[\frac{log(1-\gamma_1)}{log(1-\gamma_0)}\right]$ $\delta_1^* = \frac{c_1-\beta_0^*-\delta_0^*}{\xi_1} - \beta_1^*$ $c_0 = \log\left[e^{\beta_0^*} - \log(1-BMR)\right]$ $c_1 = \log\left[e^{\beta_0^*+\delta_0^*} - \log(1-BMR)\right]$

Table 4. *BMD estimates based on posterior means and 95% BMDLs from each reparametrized model in Table 3, together with the corresponding Deviance Information Criterion (DIC). A zero subscript indicates males and a unit subscript indicates females.*

Model	**Model Name**	**DIC**	**BMD_0**	**BMD_1**	**$BMDL_0$**	**$BMDL_1$**
M1	Logistic	20.90	4,853	9,425	2,922	4,913
M2	Probit	21.15	4,296	8,620	2,649	4,538
M3	Quantal Linear	27.87	1,993	2,830	887	1,245
M4	Quantal Quadratic	23.06	5,760	6,760	4,139	4,771
M5	Complementary Log-Log	20.63	4,897	9,638	3,007	5,118

Table 5. *BMD estimates based on posterior means and 95% BMDLs (in ppm) from each reparametrized model in Table 3, along with corresponding Bayesian model averaged (BMA) BMD and BMDL, for goldenseal root powder data given in Table 1. The BMR is set to 0.01.*

Model	**Model Name**	**BMD_0**	**BMD_1**	**$BMDL_0$**	**$BMDL_1$**	**Weight**
M1	Logistic	4,853	9,425	2,922	4,913	0.144711434
M2	Probit	4,296	8,620	2,649	4,538	0.186120986
M3	Quantal Linear	1,993	2,830	887	1,245	0.020431097
M4	Quantal Quadratic	5,760	6,760	4,139	4,771	0.514218064
M5	Complementary Log-Log	4,897	9,638	3,007	5,118	0.134518419
BMA		**5,163**	**7,799**	**3,467**	**4,723**	

APPENDIX C (Supplemental Material for Chapter 4)

Benchmark Risk Assessment for Mathematics Placement Scores

Mirjana Glisovic-Bensa [1]

Walter W. Piegorsch [2,3]

Tina Schuster [3]

[1] Department of Epidemiology & Biostatistics, University of Arizona, Tucson, Arizona, USA

[2] Interdisciplinary Program in Statistics & Data Science, University of Arizona, Tucson, Arizona, USA

[3] Department of Mathematics, University of Arizona, Tucson, Arizona, USA

Correspondence

Walter W. Piegorsch, BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA. Email: piegorsch@math.arizona.edu

Abstract

A critical operation conducted by college mathematics (and other) departments is placement of incoming undergraduates into appropriate first-year mathematics classes. Students are placed in a given mathematics course if they took prerequisite courses or if they achieved the required score on math placement exam or standardized college entrance exams such as SAT and ACT. Traditionally, the required score on the placement exam is calculated empirically. In this paper we apply quantitative "benchmark" risk analytics techniques, widely used in toxicological and environmental applications, to calculate the score on the placement exam students must achieve to register for a given mathematics class.

Keywords

Benchmark analysis, BMDL, quantitative risk assessment, math placement score.

1. Introduction

1.1 Overview

Students enrolled in a college, pursuing either a bachelor's or an associate degree, are almost always required to take at least one mathematics or statistics course. In order to ensure student success in such a class, it is important that the students have the necessary mathematical background and ability required to learn concepts presented in the course. To assess whether a student is ready to take and successfully pass a given math or statistics class, universities and colleges have developed eligibility criteria and course prerequisites that students must meet before they can register for a higher class. Often, students can satisfy required eligibility criteria by taking pertinent prerequisite courses, or by achieving the required score on a university-specific math placement exam or on standardized college entrance exams such as the SAT and ACT. Determination of such criteria is often campus-specific, and possibly also based on campus-specific targets – which are objectively imposed but still subjective in the sense that a campus/departmental coordinator or coordination team determines the actual targets – the students must achieve. To construct a somewhat-more-objective target for the students to meet, we apply here risk-analytic techniques for such constructs. We employ a "benchmark" risk analytic technique (Crump, 1984) to the placement question. As we envision here, the benchmark approach fixes a certain, prespecified level of *benchmark response* (BMR) on the placement exam past which students are predicted to succeed in the mathematics/statistics course into which the students are placed. It then uses statistical methods to find the *benchmark dose* (BMD) required to achieve that BMR. BMDs are widely applied in toxicological and environmental applications (Piegorsch and Bailer, 2005, Sec. 4.3) where the "dose" is an exposure level of a toxin. To our knowledge, however, this risk-analytic technique has never been applied in educational testing. To do so, we view "dose" as the score on a PPL assessment test that students must achieve to successfully pass a specified math class. Section 2 describes the generics of benchmark analysis from quantitative risk assessment. Section 3 introduces a mathematics placement example in which we argue that the benchmark paradigm can apply. In Section 4 we present a statistical model for translating the benchmark paradigm to the placement score setting, while in Section 5 we illustrate the application of this risk analytic, benchmark technology on the data from Section 3. We conclude with a brief discussion in Section 6.

2. Benchmark Analysis for Quantitative Risk Assessment

An important objective in traditional quantitative risk assessment is characterization of the severity and likelihood of adverse outcomes caused by a hazardous agent (Stern, 2008). In the toxicological context a *risk function*, $R(x)$, is defined as the probability of some predefined adverse effect – such as death, cancer, mutation, etc. – in a subject exposed to a hazardous agent at dose or exposure level x. This definition assumes that risk changes with increasing exposure at dose x to the hazardous agent. Commonly, observations in such risk analyses are proportions of subjects exhibiting the adverse response associated with each x. This is called the *quantal response setting*. A variety of models are available for fitting a non-decreasing dose-response relationship between a dose *x* of a toxin and a dose-response function $R(x)$. Often-seen quantal-response models for *R*(*x*) in toxicological and carcinogenic risk assessment include the logistic, probit, quantal-linear, quantal-quadratic and complementary log-log models (Piegorsch, et al., 2013). Benchmark risk analysis is a contemporary approach to quantal-response risk estimation. Benchmark analysis was first introduced by Crump (1984), and since then it has grown in popularity and flexibility. Benchmark analysis uses components of $R(x)$ to produce the benchmark dose, BMD, of the hazardous agent, at which the prespecified benchmark response, BMR, is obtained. The BMR is often taken in the range $0.01 \leq \mathrm{BMR} \leq 0.10$ (U.S. EPA, 2012), although this recommendation is based on applications in toxicological risk analysis. To account for the background risk, $R(0)$, that all subjects in a population would be expected to encounter in the absence of the hazardous agent, *R*(*x*) is refined into a form of *extra risk function,* $R_E(x)$, (Piegorsch & Bailer, 2005, Section 4.2). $R_E(x)$ is defined as the additional risk among subjects who would have not responded under control conditions: $R_E(x) = \frac{R(x)-R(0)}{1-R(0)}$. From this, the BMD is found by solving for x in $R_E(x) = \mathrm{BMR}$.

To account for uncertainty associated with a point estimator, $\widehat{\mathrm{BMD}}$, we calculate $100(1-\alpha)\%$ confidence (or credible) limits for the BMD. Driven by focus on risk reduction, only one-sided, lower confidence limits (BMDLs) are of interest (Crump, 1995). Theoretical features of the BMD have been studied by a variety of researchers (Haber et al., 2018) and data scientists have worked to develop the BMD concept to address modern issues in quantitative risk assessment (Jensen et al., 2019). Benchmark risk analytic techniques have, however, never been applied in educational settings to calculate placement exam scores. The following example describes this new risk-analytic application.

3. Example: Placing Students into a Precalculus Course at The University of Arizona

At our home institution, the University of Arizona (UArizona), many incoming first-year students are required to take a precalculus course (MATH 120R). They can register for the course only if they meet at least one of the following criteria:

5. Earn at least 60 points on the PPL Assessment Mathematics Placement Test.
6. Have earned at least 640 points on the MATH-SAT exam.
7. Have earned at least 26 points on the ACT exam.
8. Have passed a College Algebra course with a grade of C or better.

(The PPL assessment test is a mathematics placement vehicle UArizona has used since 2018. It is a 25 question adaptive test with topics that range from prealgebra through precalculus. See https://www.mheducation.com/highered/alekspplhtml.)

At the end of each academic year, a team of faculty members from the UArizona mathematics department studies the passing rates of previous MATH 120R students with a PPL score around 60 points. If the passing rate is lower than usual (for MATH 120R the usual passing rate is 50%) the required PPL score to place into MATH 120R is increased. In other words, the score required on PPL assessment to place into MATH 120R is determined empirically on a rotating basis. For simplicity, we focus here on only the PPL assessment score viewed as the "dose" for benchmarking purposes. The adverse outcome is defined as a student who achieves a grade of D, E, or W, while the complementary success is a grade of C or better. The following sections explore this setting from a risk-analytic perspective.

4. Mixed-Factor Risk Modeling Applied in Educational Testing

In addition to applying benchmark analysis for calculating the PPL score required for students to successfully pass UArizona's MATH 120R precalculus course, interest also exists in identifying whether there is a difference between male and female students as they apply in the risk analysis. To test for this difference, we construct a two-factor model with the quantitative PPL test score (*x*) and the student's qualitative sex, denoted here as *u* via

$$u = \begin{cases} 0 \text{ if student is male} \\ 1 \text{ if student is female} \end{cases}.$$

To apply a two-factor benchmark analysis in this educational testing setting we begin with a binomial assumption: $Y_i \sim \text{indep. } Bin(N_i, R(x_i, u_i))$, *i* = 1, ..., *n*, where Y_i is the number of students with PPL assessment score of x_i who failed the precalculus class (recall that our adverse event is a grade of D, E, or W), and N_i is the total number of students with PPL assessment score of x_i who took the precalculus class. If we take a_i as the standardized score on the PPL assessment math placement test; that is, a_i is the student's original PPL assessment score divided by the observed maximum PPL score). Then, we set $x_i = 1 - a_i$ in order to mimic the toxicological setting where higher values of x_i (the "dose") correspond to higher risk of the adverse event. For the risk function $R(x, u)$ we employ the ubiquitous logistic model with these various terms, as

$$R(x, u) = 1/\left(1 + e^{-\beta_0 - \beta_1 x - \delta_0 u - \delta_1 x u}\right).$$

As mentioned in Section 2, in benchmark analysis the BMD is found by solving for x in $R_E(x) = \text{BMR}$. We proceed with the benchmark calculations via the following steps:

For the risk function $R(x, u)$ we use the popular logistic model, given above, and we calculate $R_E(x)$ for use in the benchmark analysis:

$$R_E(x) = \frac{R(x) - R(0)}{1 - R(0)} = \frac{1 - e^{-\beta_1 x - \delta_1 x u}}{1 + e^{-\beta_0 - \beta_1 x - \delta_0 u - \delta_1 x u}}$$

We initially set the BMR to be 0.1, a traditional BMR used in toxicological benchmark analysis, although we also present results for larger BMR levels such as 0.25, 0.50, and 0.75 in order to study how the toxicological standard operates in this new, educational testing paradigm.

To calculate the BMD, which in our placement application represents one minus the PPL assessment score student must achieve, we solve: $R_E(x) = BMR$ for *x*; that is, we solve

$$\frac{1-e^{-\beta_1 x-\delta_1 xu}}{1+e^{-\beta_0-\beta_1 x-\delta_0 u-\delta_1 xu}} = \text{BMR}.$$

More specifically, we take BMD_0 (when u = 0) and BMD_1 (when u =1); that is, we calculate x = 1 – PPL placement score for males and females. Under our logistic model, these are:

$$BMD_0 = \beta_1^{-1} \log\left(\frac{1+\text{BMR}e^{-\beta_0}}{1-\text{BMR}}\right)$$

$$BMD_1 = (\beta_1+\delta_1)^{-1} \log\left(\frac{1+\text{BMR}e^{-\beta_0-\delta_0}}{1-\text{BMR}}\right)$$

We employ the core *glm()* function in the statistical computing environment **R** (**R** Development Core Team, 2019) to fit our logistic model to the PPL data and obtain $\hat{\beta} = \left[\widehat{\beta_0}\ \widehat{\beta_1}\ \widehat{\delta_0}\ \widehat{\delta_1}\right]^T$, the maximum likelihood estimates (MLEs) for the unknown parameter vector $\beta = [\beta_0\ \beta_1\ \delta_0\ \delta_1]^T$.
We appeal to the multivariate Delta method (Casella and Berger, 2002, Theorem 5.5.28) to approximate variances of the MLEs for BMD_0 and of BMD_1. This yields:

$$Var\left(\widehat{BMD_0}\right) = \frac{1}{\widehat{\beta_1}^2}\left[\left(\frac{BMR\cdot e^{-\widehat{\beta_0}}}{1+BMR\cdot e^{-\widehat{\beta_0}}}\right)^2 \cdot Var\left(\widehat{\beta_0}\right) + \widehat{BMD_0}^2 \cdot Var\left(\widehat{\beta_1}\right) + \frac{2BMR\cdot e^{-\widehat{\beta_0}}\cdot\widehat{BMD_0}}{1+BMR\cdot e^{-\widehat{\beta_0}}} \cdot Cov\left(\widehat{\beta_0},\widehat{\beta_1}\right)\right]$$

and

$$Var\left(\widehat{BMD_1}\right) = \frac{1}{\left(\widehat{\beta_1}+\widehat{\delta_0}\right)^2}\left[\left(\frac{BMR\cdot e^{-\widehat{\beta_0}-\widehat{\delta_1}}}{1+BMR\cdot e^{-\widehat{\beta_0}-\widehat{\delta_1}}}\right)^2 \left(Var\left(\widehat{\beta_0}\right) + Var\left(\widehat{\delta_1}\right) + 2Cov\left(\widehat{\beta_0},\widehat{\delta_1}\right)\right) + \right.$$

$$\widehat{BMD_1}^2 \left(Var\left(\widehat{\beta_1}\right) + Var\left(\widehat{\delta_0}\right) + 2Cov\left(\widehat{\beta_1},\widehat{\delta_0}\right)\right) + \frac{2BMR\cdot e^{-\widehat{\beta_0}-\widehat{\delta_1}}\cdot\widehat{BMD_1}}{1+BMR\cdot e^{-\widehat{\beta_0}-\widehat{\delta_1}}}\left(Cov\left(\widehat{\beta_0},\widehat{\beta_1}\right) + Cov\left(\widehat{\beta_1},\widehat{\delta_0}\right) + \right.$$

$$\left.\left. Cov\left(\widehat{\beta_1},\widehat{\delta_1}\right) + Cov\left(\widehat{\delta_0},\widehat{\delta_1}\right)\right)\right].$$

To account for uncertainty associated with the point estimator, $\widehat{\text{BMD}}$, we determine $100(1-\alpha)\%$ Wald-type confidence limits for the BMDs. Driven by our construction with students' failure in the course, we follow standard practice (Crump, 1995) and only calculate one-sided lower Wald confidence limits (BMDLs):

$$\widehat{BMDL}_0 = \widehat{BMD}_0 - z_{\alpha/2} se\left(\widehat{BMD}_0\right)$$

$$\widehat{BMDL}_1 = \widehat{BMD}_1 - z_{\alpha/2} se\left(\widehat{BMD}_1\right)$$

Finally, as the BMDLs here record lower limits above which possible failures (D, E, W) occur, we calculate $100(1-\alpha)\%$ upper confidence limits for the PPL score, PPLU, above which students' risks for success (C or better) are good. For males and females these are

$$\widehat{PPLU}_0 = 1 - \widehat{BMDL}_0$$

$$\widehat{PPLU}_1 = 1 - \widehat{BMDL}_1 \ .$$

Students who record a PPL score above these PPLU values are viewed as exhibiting extra risks past which success in MATH 120R is likely.

5. Example Revisited: Students Placing into Precalculus Course at The University of Arizona

To illustrate further application of this risk analytic technology, we provide a dataset of the form imagined in our model construction. Unfortunately, actual data on student placement and eventual grades cannot be released due to concerns with unapproved student identification. Instead, we build a hypothetical dataset similar in nature to the sort of placement data seen at UArizona. The full hypothetical dataset is given in a supplemental document; Table 1 shows a portion of the hypothetical data.

We fitted the logistic model to these data; sample **R** code is given in the Appendix. We took *x* to represent one minus standardized placement exam score *a*, and *u* to represent the sex (0 = Male, 1 = Female). Figure 1 displays a scatterplot of 1 – PPL score versus Y/N with the estimated extra risk functions $\widehat{R_E}(x,u)$ for males and females.

Table 1. *Hypothetical student data: selected counts of students (males and females) who passed a precalculus course and their corresponding PPL score.*

	male				female			
PPL assessment score, a_i	.59	.60	.61	.62	.59	.60	.61	.62
Total number of students with PPL assessment score of a_i, N_i	18	23	10	6	10	13	13	13
Total number of students with PPL assessment score of a_i who passed precalculus course, Y_i	12	11	5	3	2	2	5	5

The fitted model gave MLEs of $\widehat{\beta_0} = -2.4286$ (se = 0.2643, *z*-value =-9.219), $\widehat{\beta_1} = 5.3277$ (se = 0.5743, *z*-value = 9.276), $\widehat{\delta_0} = -1.0720$ (se = 0.4478, *z*-value = –2.394), and $\widehat{\delta_1} = 0.4176$ (se = 0.8816, *z*-value = 0.474). Two-sided p-values for $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\delta_0}$ are all < 0.01. This indicates that a significant interaction exists between the placement exam score and sex. That is, males and females differ significantly in their failure rates after adjusting for their PPL placement scores.

To incorporate a risk analysis into this model construct, at BMR = 0.10, the recommended placement exam limits are $\widehat{PPL_0} = 0.84$ for males and $\widehat{PPL_1} = 0.73$ for females with 95% confidence limit of $\widehat{PPLU_0} = 0.89$ for males and $\widehat{PPLU_1} = 0.79$ for females. These values indicate that females need lower placement exam scores than the males to successfully pass a class. This is corroborated by assessing the PPL patterns across differing BMRs. Table 2 presents the male (*u* = 0) and female (*u* = 1) PPLUs at a selection of BMRs over a wider range (BMR = 0.25, 0.50, 0.75). The pattern of higher PPLU values for males is seen to recur across all values; for a visual perspective, Figure 2 plots the same effect. This suggests that, from a risk-analytic perspective, males require higher PPLU scores than females to imply success when placed into MATH 120R.

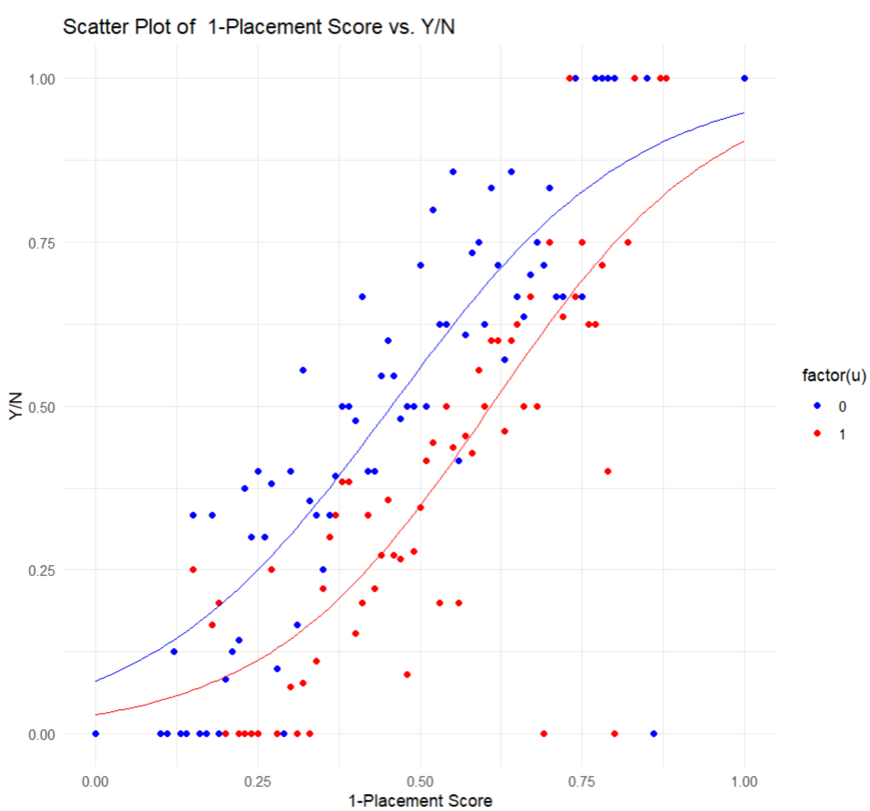

Figure 1: Scatter plot of 1 – PPL score versus observed proportions Y/N and estimated extra risk function $\widehat{R_E}(x, u)$. The estimated extra risk functions for students failing a precalculus course (MATH 120R) are shown for both male (*u* = 0, blue line) and female (*u* = 1, red line).

Table 2: The 95% upper confidence limits for PPL score at different BMR values for males and females.		
BMR	PPLU0	PPLU1
0.1	0.89	0.79
0.25	0.73	0.6
0.5	0.55	0.42
0.75	0.38	0.27

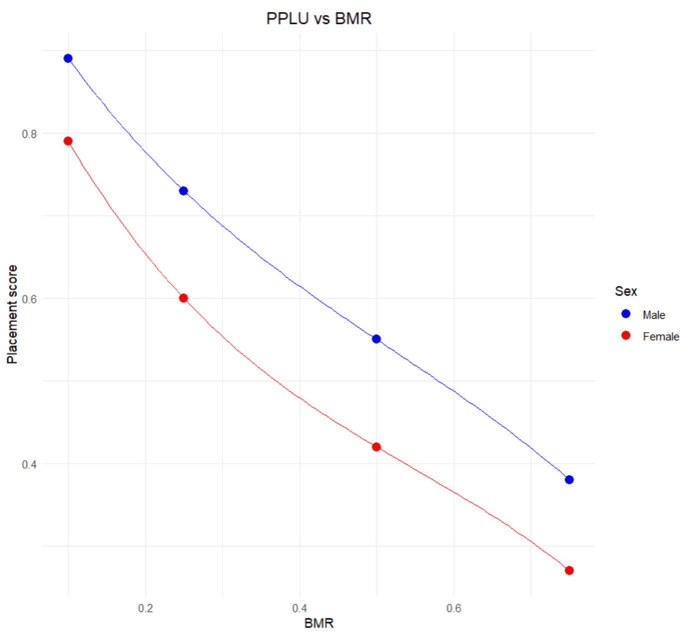

Figure 2: Plot of PPLU scores and BMR values for males and females.

6. Discussion

Herein, we have applied quantitative "benchmark" risk analytics techniques, widely used in toxicological and environmental applications, to calculate the placement exam score students must achieve in order to suggest success in a given mathematics class (UArizona's MATH 120R). We employed a mixed two-factor model with a logistic link function. Findings suggest that males and female responses differ significantly under our model, with males requiring higher PPL scores than females to imply success when placed into MATH 120R.

Of course, some caveats are in order. In particular, we have relied on the ubiquitous but still very specific logistic model. Many other quantal-response models are possible to describe the sort of mixed-factor data we studied here. Some of these popular in environmental risk assessment are the probit link, complementary-log link, complementary log-log link, multistage link, log-logistic link, log-probit link, Weibull link, among many others (Wheeler et al., 2020). Many investigator account for the consequent model uncertainty by applying some form of model selection. Even then, however, it has been seen that model selection strategies can adversely affect BMDL inferences (West, et al., 2012). Future research is needed to determine how the results we present here translate to other forms of quantal-response models.

Another possible remedy to the model uncertainty problem is to apply some form of model averaging among different quantal-response models. Previous authors (Fang et al., 2015; Piegorsch et al. 2013; Wheeler and Bailer, 2009) have shown that model averaging can be an effective strategy for overcoming model uncertainty, although here again, further research is required to assess whether these previous results can prove valuable in our new, risk-analytic placement testing scenario.

References

Casella, G., and Berger, R. L. (2002). *Statistical Inference*, 2nd Edn. Pacific Grove, CA: Duxbury Press.

Crump, K. S. (1984). A new method for determining allowable daily intake. *Fundamental and Applied Toxicology* **4**, 854–871.

Crump, K. S. (1995). Calculation of benchmark doses from continuous data. *Risk Analysis* **15**, 79–89.

Fang, Q., Piegorsch, W. W., Simmons, S. J., Li, X., Chen, C., & Wang, Y. (2015). Bayesian model-averaged benchmark dose analysis via reparameterized quantal-response models. *Biometrics* **71**, 1168–1175.

Haber, L. T., Dourson, M. L., Allen, B. C., Hertzberg, R. C., Parker, A., Vincent, M. J., Maier, A., & Boobis, A. R. (2018). Benchmark dose (BMD) modeling: current practice, issues, and challenges. *Critical Reviews in Toxicology* **48**, 387–415.

Jensen, S. M., Kluxen, F. M., & Ritz, C. (2019). A review of recent advances in benchmark dose methodology. *Risk Analysis* **39**, 2295–2315.

Piegorsch, W. W., and Bailer, A. J. (2005). *Analyzing Environmental Data*. Chichester: John Wiley & Sons.

Piegorsch, W. W., An, L., Wickens, A. A., West, R. W., Peña, E. A., & Wu, W. (2013). Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. *Environmetrics* **24**, 143–157.

R Development Core Team (2019). ***R**: A Language and Environment for Statistical Computing*. Vienna, Austria: **R** Foundation for Statistical Computing.

Stern, A. H. (2008). Environmental health risk assessment. In *Encyclopedia of Quantitative Risk Analysis and Assessment* **2**, Melnick, E. L. & Everitt, B. S. (eds.), 580–589.Chichester: John Wiley & Sons.

U.S. EPA (2012). *Benchmark Dose Technical Guidance Document*. Technical Report number EPA/100/R-12/001. Washington, DC: U.S. Environmental Protection Agency.

West, R. W., Piegorsch, W. W., Peña, E. A., An, L., Wu, W., Wickens, A. A., Xiong, H., and Chen, W. (2012). The impact of model uncertainty on benchmark dose estimation. *Environmetrics* **23**, 706–716.

Wheeler, M.W., Bailer, A.J. (2009). Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environmental and Ecological Statistics,* **16**, 37–51.

Wheeler, M.W., Blessinger, T., Shao, K., Allen, B.C., Olszyk, L., Davis, J.A. , Gift, J.S (2020). Quantitative risk assessment: developing a Bayesian approach to dichotomous dose-response uncertainty. *Risk Analysis* **40**(9), 1706–1722.

APPENDIX D (Supplemental Material for Appendix C)

APPENDIX A: SAMPLE R CODE

The following **R** code presents an approach to fit the logistic model and acquire benchmark values in our educational testing scenario. Note that this code is not intended to provide the most efficient approach for programming the benchmark calculations. Rather, the code illustrates the various steps taken for conducting the analysis in as straightforward a manner as possible. Users are welcome to develop their own code from the files here in order to increase efficiency.

```
#Code for Hypothetical data, February 2024

library(ggplot2)


FabData=read.csv(file = "FILE_NAME_HERE")

head(FabData)  # check data for proper upload


Y=FabData$Y

N=FabData$N

x=FabData$x

u=FabData$u


logitmodel= glm( cbind(Y,N-Y) ~ x + factor(u) +
I(x*u),family=binomial("logit"), maxit=99)


thetahat = coef(logitmodel)

VCV = vcov(logitmodel)

summary(logitmodel)


b0=thetahat[1]

b1=thetahat[2]

t=thetahat[3]

g=thetahat[4]
```

```
vb0=VCV[1,1]

vb1=VCV[2,2]

vt=VCV[3,3]

vg=VCV[4,4]

vb0b1=VCV[1,2]

vb0t=VCV[1,3]

vb0g=VCV[1,4]

vb1t=VCV[2,3]

vb1g=VCV[2,4]

vgt=VCV[3,4]


##############################################

BMR=0.10

#BMR=0.25

#BMR=0.50

#BMR=0.75


BMD0=log((1+BMR*exp(-b0))/(1-BMR))/b1

BMD1=log((1+BMR*exp(-b0-t))/(1-BMR))/(b1+g)

PPL0=1-BMD0

PPL1=1-BMD1


#Delta method calcultion for Variance of BMD0 and BMD1

db0u0=-BMR*exp(-b0)/((1+BMR*exp(-b0)*b1))          #b0 for u=0

db0u1=-BMR*exp(-b0-t)/((1+BMR*exp(-b0-t)*(b1+g))) #b0 for u=1


db1u0=-BMD0/b1     #b1 for u=0

db1u1=-BMD1/(b1+g) #b1 for u=1


dgu0=0           #g for u=0

dgu1=-BMD1/(b1+g) #g for u=1
```

```
dtu0=0                                        #t for u=0
dtu1=-BMR*exp(-b0-t)/((1+BMR*exp(-b0-t)*(b1+g))) #t for u=1

VBMD0=db0u0^2*vb0+db1u0^2*vb1+2*db0u0*db1u0*vb0b1
VBMD1=db0u1^2*vb0+db1u1^2*vb1+dgu1^2*vg+dtu1^2*vt+
  2*(db0u1*db1u1*vb0b1+db0u1*dgu1*vb0g+db0u1*dtu1*vb0t+
       db1u1*dgu1*vb1g+db1u1*dtu1*vb1t+dgu1*dtu1*vgt)

VBMD0
VBMD1
BMD0
BMD1
PPL0
PPL1

############################################################
#BMDL calculation
BMDL0=(BMD0-1.96*sqrt(VBMD0))
BMDL1=(BMD1-1.96*sqrt(VBMD1))

PPLU0=1-BMDL0
PPLU1=1-BMDL1

############################################################
# Create a scatter plot of Placement Score vs Y/N

# Define the function y = x
line0 <- function(x) 1/(1+exp(-b0-b1*x))

# Define the function y = x + x^2
line1<- function(x) 1/(1+exp(-b0-b1*x-t-g*x))
```

```
ggplot(FabData, aes(x = `x`, y = Y/N, color = factor(u))) +
  geom_point() +
  # Add the first line (y = x)
  geom_function(fun = line0, color = "blue") +
  # Add the second line (y = x + x^2)
  geom_function(fun = line1, color = "red") +
  labs(x = "1-Placement Score", y = "Y/N") +
  scale_x_continuous(limits = c(0, 1)) +
  scale_y_continuous(limits = c(0, 1)) +
  scale_color_manual(values = c("0" = "blue", "1" = "red")) +
  ggtitle("Scatter Plot of  1-Placement Score vs. Y/N") +
  theme_minimal()

################################################################
#Create BMR and PPLU values

# Create data frames for the blue and red points
blue_data <- data.frame(
  x = c(0.10, 0.25, 0.50, 0.75),
  y = c(0.89, 0.73, 0.55, 0.38),
  Sex = "Male"
)

red_data <- data.frame(
  x = c(0.10, 0.25, 0.50, 0.75),
  y = c(0.79, 0.60, 0.42, 0.27),
  Sex = "Female"
)

# Create a spline function for the blue and red points
blue_spline <- spline(blue_data$x, blue_data$y, n = 100)
red_spline <- spline(red_data$x, red_data$y, n = 100)
```

```
# Combine the blue and red data into one data frame
combined_data <- rbind(blue_data, red_data)

# Plot the points and splines
ggplot() +
  geom_point(data = combined_data, aes(x = x, y = y, color = Sex), size = 3)
+
  geom_path(data = data.frame(blue_spline), aes(x = x, y = y), color =
"blue") +
  geom_path(data = data.frame(red_spline), aes(x = x, y = y), color = "red")
+
  scale_color_manual(values = c("Male" = "blue", "Female" = "red")) +
  labs(x = "BMR", y = "Placement score") +
  theme_minimal() +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5)) +
  ggtitle("PPLU vs BMR")
```

Appendix B: Hypothetical Student Data.

a	x	u	N	Y
0	1	0	1	1
0.12	0.88	1	1	1
0.13	0.87	1	1	1
0.14	0.86	0	1	0
0.15	0.85	0	1	1
0.17	0.83	1	7	7
0.18	0.82	1	4	3
0.2	0.8	1	1	0

0.2	0.8	0	1	1
0.21	0.79	1	5	2
0.21	0.79	0	1	1
0.22	0.78	1	7	5
0.22	0.78	0	2	2
0.23	0.77	1	8	5
0.23	0.77	0	1	1
0.24	0.76	1	8	5
0.25	0.75	0	6	4
0.25	0.75	1	4	3
0.26	0.74	1	6	4
0.26	0.74	0	1	1
0.27	0.73	1	10	10
0.28	0.72	1	11	7
0.28	0.72	0	3	2
0.29	0.71	1	6	4
0.29	0.71	0	3	2
0.3	0.7	0	18	15
0.3	0.7	1	4	3
0.31	0.69	0	7	5
0.31	0.69	1	1	0
0.32	0.68	1	4	2
0.32	0.68	0	4	3
0.33	0.67	0	10	7
0.33	0.67	1	3	2
0.34	0.66	0	11	7
0.34	0.66	1	4	2
0.35	0.65	1	8	5
0.35	0.65	0	3	2

0.36	0.64	1	10	6
0.36	0.64	0	7	6
0.37	0.63	0	14	8
0.37	0.63	1	13	6
0.38	0.62	0	14	10
0.38	0.62	1	5	3
0.39	0.61	1	10	6
0.39	0.61	0	6	5
0.4	0.6	1	4	2
0.4	0.6	0	8	5
0.41	0.59	1	9	5
0.41	0.59	0	4	3
0.42	0.58	0	15	11
0.42	0.58	1	7	3
0.43	0.57	0	23	14
0.43	0.57	1	11	5
0.44	0.56	0	12	5
0.44	0.56	1	5	1
0.45	0.55	1	16	7
0.45	0.55	0	7	6
0.46	0.54	1	6	3
0.46	0.54	0	8	5
0.47	0.53	1	20	4
0.47	0.53	0	8	5
0.48	0.52	1	9	4
0.48	0.52	0	5	4
0.49	0.51	1	12	5
0.49	0.51	0	6	3
0.5	0.5	1	29	10

0.5	0.5	0	14	10
0.51	0.49	1	18	5
0.51	0.49	0	12	6
0.52	0.48	1	11	1
0.52	0.48	0	18	9
0.53	0.47	1	15	4
0.53	0.47	0	25	12
0.54	0.46	1	11	3
0.54	0.46	0	11	6
0.55	0.45	1	14	5
0.55	0.45	0	15	9
0.56	0.44	1	11	3
0.56	0.44	0	11	6
0.57	0.43	0	10	4
0.57	0.43	1	18	4
0.58	0.42	0	10	4
0.58	0.42	1	18	6
0.59	0.41	1	10	2
0.59	0.41	0	18	12
0.6	0.4	1	13	2
0.6	0.4	0	23	11
0.61	0.39	1	13	5
0.61	0.39	0	10	5
0.62	0.38	0	6	3
0.62	0.38	1	13	5
0.63	0.37	1	15	5
0.63	0.37	0	28	11
0.64	0.36	1	10	3
0.64	0.36	0	12	4

0.65	0.35	1	9	2
0.65	0.35	0	12	3
0.66	0.34	0	9	3
0.66	0.34	1	18	2
0.67	0.33	1	4	0
0.67	0.33	0	31	11
0.68	0.32	0	9	5
0.68	0.32	1	13	1
0.69	0.31	1	4	0
0.69	0.31	0	12	2
0.7	0.3	1	14	1
0.7	0.3	0	5	2
0.71	0.29	1	5	0
0.71	0.29	0	15	0
0.72	0.28	1	3	0
0.72	0.28	0	10	1
0.73	0.27	1	4	1
0.73	0.27	0	21	8
0.74	0.26	0	10	3
0.75	0.25	1	3	0
0.75	0.25	0	10	4
0.76	0.24	1	1	0
0.76	0.24	0	10	3
0.77	0.23	1	2	0
0.77	0.23	0	16	6
0.78	0.22	1	3	0
0.78	0.22	0	7	1
0.79	0.21	0	8	1
0.8	0.2	1	3	0

0.8	0.2	0	12	1
0.81	0.19	0	1	0
0.81	0.19	1	10	2
0.82	0.18	0	3	1
0.82	0.18	1	6	1
0.83	0.17	0	5	0
0.84	0.16	0	3	0
0.85	0.15	1	4	1
0.85	0.15	0	3	1
0.86	0.14	0	1	0
0.87	0.13	1	1	0
0.87	0.13	0	8	0
0.88	0.12	0	8	1
0.89	0.11	0	1	0
0.9	0.1	0	2	0
1	0	0	2	0