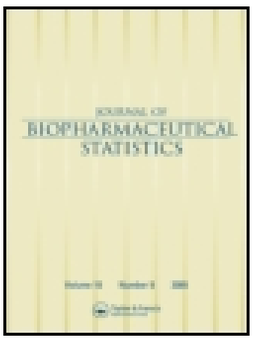
Identification of the minimum effective dose for normally distributed data using a Bayesian variable selection approach

OTAVA, Martin; SHKEDY, Ziv; Hothorn, Ludwig A.; TALLOEN, Willem; Gerhard, Daniel & KASIM, Adetayo (2017) Identification of the minimum effective dose for normally distributed data using a Bayesian variable selection approach. In: JOURNAL OF BIOPHARMACEUTICAL STATISTICS, 27(6), p. 1073-1088.

DOI: 10.1080/10543406.2017.1295247
Handle: http://hdl.handle.net/1942/26341



Journal of Biopharmaceutical Statistics

ISSN: 1054-3406 (Print) 1520-5711 (Online) Journal homepage: http://www.tandfonline.com/loi/lbps20

Identification of the Minimum Effective Dose for Normally Distributed Data Using a Bayesian Variable Selection Approach

Martin Otava, Ziv Shkedy, Ludwig A. Hothorn, Willem Talloen, Daniel Gerhard & Adetayo Kasim

To cite this article: Martin Otava, Ziv Shkedy, Ludwig A. Hothorn, Willem Talloen, Daniel Gerhard & Adetayo Kasim (2017): Identification of the Minimum Effective Dose for Normally Distributed Data Using a Bayesian Variable Selection Approach, Journal of Biopharmaceutical Statistics, DOI: 10.1080/10543406.2017.1295247

To link to this article: http://dx.doi.org/10.1080/10543406.2017.1295247

Accepted author version posted online: 16 Feb 2017.

Identification of the Minimum Effective Dose for Normally Distributed Data Using a Bayesian Variable Selection Approach

Martin Otava [*1], Ziv Shkedy[1], Ludwig A. Hothorn[2], Willem Talloen[3], Daniel Gerhard[4], and Adetayo Kasim[5]

[1]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Martelarenlaan 42, B-3500 Hasselt, Belgium

[2]Institute of Biostatistics, Leibniz University Hannover, Herrenhaeuserstr. 2, D-30419 Hannover, Germany

[3]Janssen, Pharmaceutical companies of Johnson & Johnson, Turnhoutseweg 30, B-2340 Beerse, Belgium

[4]School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, 8140 Christchurch, New Zealand

[5]Wolfson Research Institute for Health and Wellbeing, Durham University, Queen's Campus, University Boulevard, Stockton-on-Tees, TS17 6BH, United Kingdom

Acknowledgment: Martin Otava and Ziv Shkedy gratefully acknowledge the support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy). Martin Otava gratefully acknowledge the financial support of the Research Project BOF11DOC09 of Hasselt University. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government, department EWI.

Abstract

The identification of the minimum effective dose is of high importance in the drug development process. In early stage screening experiments, establishing the minimum effective dose can be translated into a model selection based on information criteria. The presented alternative, Bayesian variable selection approach, allows for selection of the minimum effective dose, while taking into account model uncertainty. The performance of Bayesian variable selection is compared with the generalized order restricted information criterion on two dose-response experiments and through the simulations study. Which method has perform better depends on the complexity of underlying model and effect size relative to noise.

Keywords: Model selection; Model uncertainty; Minimum effective dose; Order restricted models; Bayesian variable selection.

[*]Electronic address: martin.otava@uhasselt.be; Corresponding author

1 Introduction

The selection of the minimum effective dose (MED) is an important concept in the drug development process (European Medicines Agency, 2002 and Wang et al., 2011). It translates into the identification of the lowest dose that causes a desired effect or adverse events. The MED is often used in the context of the former case, while the latter is called the lowest observed adverse event level (LOAEL, Kodell, 2009) or the maximum safe dose (Hothorn and Hauschke, 2000). From a statistical point of view, there is no difference between these two concepts, only the interpretation of the response and the findings differ. An analogous framework arises when the determination of the maximum effective dose is of primary interest (Kong et al., 2014). In this manuscript, we restrict the discussion to the MED. In some cases, the clinical significance is included in the definition of the MED (Liu, 2010), while other cases are focused on statistical significance only (Kuiper et al., 2014). Note that clinical significance of the result can be included in stages following the analysis and treated separately.

The concept of the MED appears in multiple stages of drug development. If a large number of doses is used or prior knowledge about the shape of the dose-response profile exists, parametric methods can be applied (e.g. the four parameter logistic non-linear regression model, Hill's model, etc., Seber and Wild, 1989, Straetemans, 2012, Pramana et al., 2012). The MED is, in this case, based on a particular parametric model. Alternatively, methods can be used that combine model selection with parametric modelling, such as MCP-Mod (Bornkamp et al., 2009). In our framework, there are only few dose levels at which the response was measured and typically only limited knowledge about the dose-response relationship exists. Therefore, parametric modelling of the whole profile as a continuous function of dose is not suitable and an order restricted analysis of variance (ANOVA) is preferred. Typically, the monotonicity assumption is a reasonable choice, implying that a higher dose induces a stronger effect (positive or negative for upward or downward trend, respectively). Note that this assumption is often made in drug development studies (e.g. Bretz and Hothorn, 2003, Ohlssen and Racine, 2015).

The goal of the analysis is to determine the lowest active dose with significant difference to a control. For example, in an experiment with a placebo and three active doses, we would like to detect which of the three active doses is the MED. To achieve it, we need to be able to determine the probability of being the best model among the eight possible models (for each direction) shown in Table 1. As MED, we understand lowest dose that exhibits an effect, either increasing or decreasing, respectively to monotonocity assumption in place. Therefore, the $MED = 2$ for model g_2 in Table 1 for both Up and Down mean structure.

Within the frequentist framework, the MED can be viewed either in terms of inference of particular increments between consecutive doses or as model selection problem. The former approach is represented by multiple comparison procedures (Bretz and Hothorn, 2003), such as Dunnett's test (Dunnett, 1955). This approach may require to pool together some of the means in order to maintain a reasonable power, which does not provide complete information about the MED and can eventually lead to biased estimates (Hothorn and Hauschke, 2000). Multiple contrast tests are generally designed to preform an inference rather than to determine the MED (Bretz and Hothorn, 2003). Closed tests procedures can be applied instead, but they may lack overall power (Wang and Peng, 2015). Recently, Kuiper et al. (2014) suggested to focus on model selection methods and specifically on information criteria (IC) based approaches (e.g. Lin et al., 2009, Lin et al., 2012).

Within the IC approach, the weights of each one of the candidate models are estimated and used for the determination of the MED. It is crucial to realize that the MED cannot be established through a classical model selection process that focuses only on the best model (among a set of candidate models). Some of the competing models could have the same MED, i.e. the same dose that shows first significant effect compared to the mean of control dose. For example, the MED for all the models g_1, g_3, g_5 and g_7 in Table 1 is the first active dose. Although one model could have the highest model weights among all models, group of competing models with same MED could have higher weight when all pooled together. This reasoning suggests that IC is an appropriate approach, since IC based methods compare all candidate models and their IC values can be easily converted into weights. Such weights can be used to approximate posterior model probabilities (Burnham and Anderson, 2002) that can be pooled together for appropriate models (Kuiper et al., 2014).

Note that the set of models in Table 1 is based on strict inequalities. This is not typical set of models that would be used in context of IC methods that are designed to work with models containing inequalities rather than strict inequalities. The reason to approach the problem with strict inequalities models is that our main focus is on Bayesian variable selection described below and IC methods are used for comparison. This would lead to issues with fitting some strict inequalities models in Section 4. If the model with strict inequality between two given doses is fitted, but the observed means result in opposite direction, the model with equality is actually fitted, which is represented by another model in our strict inequalities based set (e.g. compare models g_2 and g_0). Therefore, model fitting need to be approached carefully and fitting of same models multiple times needs to be avoided. Note that for proper usage of IC with inequalities, smaller set of models would be used; but by extending the set of models and using only those that can be actually fitted, we only loose computational time.

Naturally, order restriction needs to be taken into account for IC based methods (Anraku, 1999) which leads to the generalized order restricted information criterion (GORIC, Kuiper et al., 2011). The advantage of the IC is that they provide the probability for a particular model being the best model, given the data, among all fitted models. Hence, multiple values of the MED can be computed together with their corresponding posterior model probabilities (Kuiper et al., 2014). The main disadvantage of this approach is that it requires to fit all the models under consideration. Total amount of possible models increases quickly with number of dose levels. For example, for an experiment with five or six dose levels, there are 16 or 32 order restricted one-way ANOVA models that may needs to be fitted, respectively. Ideally, scientific interest is limited only to the small set of models, but it may be the case that wide range of the models need to be explored. Procedures are available to reduce the number of models either by an efficient search in the model space (e.g. stepwise methods) or by reducing the model space itself (e.g. diversity index, Kim et al., 2014). However, they usually require additional input parameters or criteria specification and the resulting amount of models to be fitted can still remain prohibitive.

In such a case, Bayesian variable selection method (George and Mc-Culloch, 1993, O'Hara and Sillanpää, 2009) becomes an attractive alternative. In particular, for dose response experiments, the BVS approach (Kasim et al., 2012, Otava et al., 2014) allows fitting all models simultaneously and provides posterior probabilities for each of them, while computational time does not increase in a linear fashion as in case of the IC approach.

This manuscript is organized as follows. The methodological background for both the IC based methods and the BVS is summarized in Section 2. The two case studies analyzed in this paper are described in Section 3. The methods are applied for the two case studies in Section 4 and the results are evaluated. Further empirical comparison is investigated via simulation study and presented in Section 5. Finally, the findings are summarized and

discussed in Section 6.

2 Methodology

We consider a dose-response experiment with a control group and $K-1$ active dose levels. Denote the set of observations by

$$\mathbf{Y} = \left\{ Y_{ij},\ i = 0,\mathrm{K}\ , K-1,\ \ j = 1,\mathrm{K}\ , n_i \right\},$$

where n_i represents the number of observations of dose *i*. Our goal is to select the lowest dose *i* that shows a statistically significant difference compared to the control group. Such a dose is the MED. We denote such an event as $\mathrm{MED} = i$ and the probability that this event occurs as $P(\mathrm{MED} = i)$. Let $g_0, \mathrm{K}\ , g_R$ be a set of $R+1$ candidate models which are used to determine the MED. Based on the observed data and the models that are considered as plausible, the quantity of interest is the posterior probability of the particular value of the MED, $P(\mathrm{MED} = i \mid \mathrm{data}, g_0, \mathrm{K}\ , g_R)$. The determination of the MED can be translated into a model selection problem. For example, for $K = 4$ it translates to a selection of the best model among all models for given direction that are presented in Table 1. Note that multiple models induce the same MED, e.g. for $K = 4$ the probability that the MED is the second dose level is equal to $P(\mathrm{MED} = 2 \mid \mathrm{data}, g_0, \mathrm{K}\ , g_R) = P(g_2 \mid \mathrm{data}, g_0, \mathrm{K}\ , g_R) + P(g_6 \mid \mathrm{data}, g_0, \mathrm{K}\ , g_R)$, where $P(g_r \mid \mathrm{data}, g_0, \mathrm{K}\ , g_R)$ is the posterior probability of the model g_r, $r = 0,1,\mathrm{K}\ , R$. Therefore, the inference about the MED cannot be based on a single model only and our aim is to estimate $P(g_r \mid \mathrm{data}, g_0, \mathrm{K}\ , g_R)$ for all the suitable models. The posterior probabilities for the MED are obtained by summing appropriate posterior model probabilities. To simplify notation, from this point onwards, we denote $P(\mathrm{MED} = i \mid \mathrm{data}, g_0, \mathrm{K}\ , g_R)$ and $P(g_r \mid \mathrm{data}, g_0, \mathrm{K}\ , g_R)$ as $P(\mathrm{MED} = i \mid \mathrm{data})$ and $P(g_r \mid \mathrm{data})$, respectively.

2.1 Model averaging techniques

The likelihood based methodology addresses the problem of model selection through information criteria (IC) approaches (e.g. Akaike, 1974, Burnham and Anderson, 2002, Claeskens and Hjort, 2008, Lin et al., 2012, Kuiper et al., 2014). All candidate models are fitted and their corresponding IC values are computed. Based on the IC value, weights are calculated for each of the fitted models (as explained in detail below). The resulting weights can be considered as an approximation of posterior probabilities of the models being the best model among all fitted models, given the data (Burnham and Anderson, 2002).

As proposed by Burnham and Anderson (2002) and Claeskens and Hjort (2008), for set of models $g_0, g_1, \mathrm{K}\ , g_R$, we can select as the best model such that maximizes the posterior model probability given by

$$P(g_r \mid \mathrm{data}) = \frac{P(\mathrm{data} \mid g_r) P(g_r)}{\sum_{s=1}^{R} P(\mathrm{data} \mid g_s) P(g_s)}\ r = 0,\mathrm{K}\ , R. \qquad (1)$$

The term $P(\mathrm{data} \mid g_r)$ is the model likelihood (Burnham and Anderson, 2002) corrected with a penalization term and $P(g_r)$ is a prespecified prior probability of model g_r. In this section,

we consider a vague prior knowledge and so we use $P(g_r) = 1/(R+1)$ for all r. Different priors may be used in order to incorporate prior scientific knowledge, if available. The model likelihood $P(\text{data} \mid g_r)$ is approximated by

$$P_{IC}(\text{data} \mid g_r) = \exp(-\tfrac{1}{2}\Delta IC_r), \tag{2}$$

where $\Delta IC_r = IC_r - IC_{\min}$, with $IC_{\min} = \min_{r=0,\mathrm{K},R} IC_r$. Hence, combining equations (1) and (2) together and assuming equal prior probabilities, we get

$$w_r = P_{IC}(g_r \mid \text{data}) = \frac{\exp(-\tfrac{1}{2}\Delta IC_r)}{\sum_{s=0}^{R}\exp(-\tfrac{1}{2}\Delta IC_s)}. \tag{3}$$

The properties of this method depend on IC used.

An information criterion is a function of likelihood with a penalization term for model complexity given by

$$IC = -2\log L(\boldsymbol{\theta} \mid \text{data}) + \tau. \tag{4}$$

Here, $\boldsymbol{\theta}$ represents the model parameters, $L(\boldsymbol{\theta} \mid \text{data})$ maximum likelihood estimate for given model and τ is a penalization function. Note that if the order restricted model is considered, the likelihood will be computed under such restriction. As IC, Akaike's information criterion (AIC, Akaike, 1974) or Bayesian information criterion (BIC, Schwarz, 1978) can be applied. The AIC uses the penalty term $\tau = 2 \cdot A$, with A being number of distinct parameters in a model. The main criticism against the AIC is that it evaluates the goodness of fit without taking into account sample size (Burnham and Anderson, 2004). Small-sample size modification of the criterion was developed (Sugiura, 1978), but often the original version is used (Burnham and Anderson, 2004). The BIC uses the penalty term $\tau = A \cdot \log(B)$, where B is the number of observations. Hence, the BIC penalty is higher than for the AIC, if we have more than seven observations and the BIC favours simpler models as sample size increases. Although the criteria seem to be very similar, their motivation is grounded in very different principles. While the AIC arises from information theory and tries to find the model with the smallest distance to a complex true model, the BIC is related to an asymptotic Bayes factor and assumes that true model is contained in available set of models (Schwarz, 1978). However, as pointed out by Anraku (1999), none of these criteria is suitable in our framework, since they cannot properly evaluate order restrictions.

The order restricted information criterion (ORIC, Anraku, 1999) uses an order restricted likelihood in which the mean response at each dose level is estimated using isotonic regression (Barlow et al., 1972) and a penalty term is given by

$$\tau(ORIC) = 2 \cdot \sum_{\mathrm{l}=1}^{K} \mathrm{l}\, P(\mathrm{l}, K, \mathbf{v}). \tag{5}$$

The level probabilities (Robertson et al., 1988), $P(\mathrm{l}, K, \mathbf{v})$, are defined under the null model (of no dose effect, i.e. under g_0). We assume that there are K doses for an experiment with a control and $K-1$ dose levels. Then, $P(\mathrm{l}, K, \mathbf{v})$ represents the probability that number of distinct dose-specific means equals to l. The weights are given by $v_i = n_i / \sigma_i$ and they are constant for balanced experiment with equal variances. The generalized ORIC (GORIC, Kuiper et al., 2011) is an extension for more complicated profiles than simple order

restrictions. In our framework, for normally distributed data and monotonicity, GORIC reduces back to the ORIC.

The weights defined in Equation (3) can be used to estimate the dose-specific means as weighted average of the means estimated by the $R+1$ candidate models. This approach is closely related to model averaging techniques as discussed, in the context of dose-response modelling, in Bretz et al. (2005), Pinheiro et al. (2006), Whitney and Ryan (2009) and Lin et al. (2012).

Note that it is necessary to fit all candidate models $g_0,\mathrm{K}\,,g_R$ in order to compute the weights based on the IC described in this section. Therefore, with an increasing number of candidate models (e.g. when the number of dose levels increases), the number of fitted models increases as well. The number of candidate models can be significantly reduced if theory-based models only would be considered. However, in this manuscript, we focus on case when scientific knowledge to construct such framework is lacking and all the models need to be considered. Additionally, if the focus would be solely at MED and not the models themselves, set of models for ICs can be reduced to focus on MED only.

2.2 Order restricted estimation: hierarchical Bayesian approach

In this section we formulate a hierarchical Bayesian model in order to estimate the mean of the response at each dose level. The order constraints on the parameters are translated into order restrictions on the prior distributions (Klugkist and Mulder, 2008) which leads to a simplification of Markov chain Monte Carlo (MCMC) sampling (Gelfand et al., 1992 and Kasim et al., 2012). Following Klugkist and Mulder (2008) and Kasim et al. (2012), we assume a hierarchical one-way ANOVA model

$$Y_{ij} \sim N(\mu_i, \sigma^2), \tag{6}$$

where i represents the dose and $j = 1,\mathrm{K}\,,n_i$ the replicates within each dose. The dose-specific mean response at the dose level i is given by

$$E(Y_{ij}) = \mu_i = \begin{cases} \mu_0, & i = 0, \\ \mu_0 + \sum_{\mathrm{l}=1}^{i} \theta_{\mathrm{l}}, & i = 1,\mathrm{K}\,,K-1. \end{cases} \tag{7}$$

The constraints differ according to the direction of order restriction: $\theta_{\mathrm{l}} \geq 0$ for an upward trend or $\theta_{\mathrm{l}} \leq 0$ for a downward trend (Otava et al., 2014).

In order to ensure monotonicity among the means, the prior distributions of all components of vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \mathrm{K}\,, \theta_{K-1})$ are truncated (at zero) normal distributions. Note that $P(\theta_{\mathrm{l}} = 0) = 0$, a probability of any of the components to be exactly zero is zero. Hence, the parametrization in Equation (7) implies that the model with $K-1$ (ordered) parameters θ_{l}, is fitted (Dunson and Neelon, 2003). For example, for $K = 4$, only model g_7 can be fitted. Therefore, necessarily $\mathrm{MED} = 1$. However, all the other models $g_0,\mathrm{K}\,,g_{R-1}$ can be fitted by a slight modification of the parametrization of the mean structure, i.e. by fixing appropriate θ_{l} to be equal to zero. The deviance information criterion (DIC, Spiegelhalter et al., 2002), can be used to select the best model in the spirit of previous section. However, such approach shares the disadvantage of necessity to fit all models of interest separately and is not appropriate for inequalities. Instead, the model given above can be extended to BVS model, as will be shown in following section, by reformulating the mean structure in Equation (7) which allows to fit all candidate models $g_0,\mathrm{K}\,,g_R$ simultaneously.

2.3 BVS model formulation

The Bayesian variable selection (BVS) method is an extension of the Bayesian model in equation (7) that allows us to fit all candidate models at once (through one MCMC chain) via the internal variable selection procedure. The set of all candidate models is summarized by one mixture model and the mixture weights are estimated as additional parameters. Therefore, the BVS gains a clear advantage over any IC based method, where all the models need to be fitted separately. To incorporate all the models in the framework, the components of $\boldsymbol{\theta}$ has to be allowed to be equal to zero in the specification of the mean structure in Equation (7). Moreover, a specific selection of a subset of components from $\boldsymbol{\theta}$ to be equal to zero determines unambiguously which model of g_0, K, g_R is used. Hence, the BVS approach can be used to choose optimal models, given set of $R+1$ models.

Let z_l be an indicator variable, $l = 1, \mathrm{K}, K-1$ such as $z_l = 1$ if δ_l is included in the model and $z_l = 0$ otherwise and let $\theta_l = \delta_l \cdot z_l$. The dose-specific mean structure in Equation (7) can be expressed as a BVS model (O'Hara and Sillanpää, 2009) given by

$$E(Y_{ij}) = \mu_0 + \sum_{l=1}^{i} \theta_l = \mu_0 + \sum_{l=1}^{i} z_l \delta_l. \tag{8}$$

In order to specify prior distributions, we use notation $TN(\mu, \sigma^2, a, b)$ for a truncated normal distribution (where μ and σ^2 are the mean and variance parameters of the normal distribution and a, b are the limits of the truncation interval). The prior distributions of dose specific mean in control dose, the increments, the variance and all the hyperparameters are specified as

$$\begin{aligned}
&\mu_0 \sim TN(\eta_{\mu_0}, \tau_{\mu_0}^{-1}, 0, \infty), && \eta_{\mu_0} \sim N(0, 10^6), && \tau_{\delta_1} \sim \Gamma(1,1), \\
&\delta_l \sim TN(\eta_{\delta_k}, \tau_{\delta_k}^{-1}, 0, \infty), && \tau_{\mu_0} \sim \Gamma(1,1), && z_l \sim Bernoulli(\pi_i), \\
&\tau \sim \Gamma(10^{-3}, 10^{-3}), && \eta_{\delta_1} \sim N(0, 10^6), && \pi_1 \sim U(0,1),
\end{aligned} \tag{9}$$

where $l = 1, \mathrm{K}, K-1$. Further, following Kuo and Mallick (1998), we assume independence of z_l and δ_l, i.e. $P(\delta_l, z_l) = P(\delta_l) \cdot P(z_l)$. Detailed discussion on the model formulation and priors specification can be found in Kasim et al. (2012) and Otava et al. (2014).

The posterior mean of z_l (obtained through MCMC simulation) represents the posterior inclusion probability of δ_l in the model (O'Hara and Sillanpää, 2009). Due to the fact that the configuration of the vector ***z*** determines unambiguously a particular model, the posterior probability of a particular configuration of ***z*** translates into posterior probability of a particular model (Table 1). For example, in case of $K = 4$, posterior probability of model g_1 equals to

$$P(g_1 \mid \text{data}) = P\left[\mathbf{z} = c(1,0,0) \mid \text{data}\right]. \tag{10}$$

Note that $P(g_r \mid \text{data})$ is interpreted as posterior probability of model g_r, given the data, the priors and the set of all models. Naturally, prior specification can strongly influence the results of the analysis. In this way, prior information allows us to include information coming from scientific knowledge or previous experiments. Although we usually apply the BVS in case that all models are of interest (e.g. all models from Table 1), if a subset of the models is

a priori considered impossible, it can be easily omitted by setting its prior probabilities to zero. In case of lack of any prior information, non-informative priors can be used instead. Following Jeffreys (1961) and Kass and Wasserman (1996), we recommend to use the equal weights for all candidate models.

Analogously to the previous section, the MED can be obtained by summing the posterior probabilities of appropriate models. The resulting quantities represent the posterior distribution of the MED, i.e. to each possible value of the MED the posterior probability of being the true underlying MED is assigned. For example, for $K = 4$, $P(MED = 2|data) = P(g_2|data) + P(g_6|data)$. Hence, in terms of the inclusion vector $\boldsymbol{z}$, the posterior probability is given by

$$\bar{P}(MED = 2|data) = \bar{P}(\mathbf{z} = (0,1,0)|data) + \bar{P}(\mathbf{z} = (0,1,1)|data). \quad (11)$$

Note that the estimation of the mean vector $\boldsymbol{\mu}$ is computed as its posterior mean $\bar{\mu}$ of a MCMC chain of B iterations. It holds that $\bar{\mu} = 1/B\sum_{b=1}^{B}\hat{\mu}_b$, while in each iteration b, one model g_r is considered and estimate $\hat{\mu}_b$ is obtained. The model g_r is selected n_{g_r} times over all the B iterations. Therefore $\bar{\mu} = 1/B\sum_{r=0}^{R} n_{g_r}\hat{\mu}_{g_r}$, where $\hat{\mu}_{g_r}$ is the estimate of μ under model g_r. Since posterior probability $\bar{P}(g_r|\text{data}) = n_{g_r}/B$, i.e. it corresponds to proportion of selection of the model, the equation can be rewritten as $\bar{\mu} = \sum_{r=0}^{R}\bar{P}(g_r|\text{data})\hat{\mu}_{g_r}$. Therefore, mean estimates $\bar{\mu}$ are in fact model averaging based estimates, weighted by the posterior probabilities of the models.

There is very important difference between GORIC and BVS method in terms of underlying principles. GORIC (as well as AIC) arises from information theory and it estimates Kullback-Leibler divergence (Kullback and Leibler, 1951) between the true model and models under consideration. Therefore, it does not assume that the true model is necessarily among the candidate models. Candidate models represent potential approximation of complex underlying model. However, BVS model simply selects the best model among the candidate models, assuming that one of them is really the true underlying model. It fits the candidate models to the data in Bayesian framework and evaluates which of the models is the most likely to be the best one. Therefore, we do not expect similar performance of the methods, since they are constructed within very different frameworks.

3 Data sets

Two data sets, presented in Kuiper et al. (2014), are used for illustration of the BVS method and for the comparison between the BVS and the IC based approaches. Both data sets are displayed in Figure 1.

The Angina data set represents dose-response study of a drug to treat angina pectoris. The response is the duration (in minutes) of pain-free walking after treatment relative to the values before treatment. Four active doses were used together with a control dose (placebo only). Ten patients per dose were examined. Large values indicate positive effects on patients. The data were taken from Westfall et al. (1999, p. 164) and are available under the name angina in the package mratios (Djira et al., 2012) of the R software (R Core Team, 2014).

The Toxicity data set was introduced by Yanagawa and Kikuchi (2001, p. 320). It represents results of a chronic toxicity study on Mosapride Citrate (Fitzhugh et al., 1964).

Liver weight relative to the body weight was measured for 24 dogs. Three active doses of Mosapride Citrate were used and a control dose was added, six dogs were treated in each group. An increasing response suggests an increasing toxicity of the drug.

4 Results

We apply the BVS model, GORIC, the AIC and the BIC methods for the Toxicity and the Angina data sets described in Section 3. The attention is given to the comparison between the BVS and GORIC, since they are both taking into account order constraints within the estimation procedure of the MED. The model weights based on the IC are interpreted, in terms of Equation (3), as posterior model probabilities. In order to distinguish between the results of the methods, we denote posterior probabilities as $\bar{P}_{GORIC}$ and $\bar{P}_{BVS}$ for respective method. The analysis for all methods was done using the R software (R Core Team, 2014) version 3.1.1. For the BVS model, the MCMC was run using the package runjags (Denwood, In Review) together with the JAGS software (Plummer, 2003).

The results for the BVS model are shown in Figure 2 and Figure 3 for the Angina data and the Toxicity data, respectively. The left panels show the data, the BVS weighted average of mean estimates (solid line) and the best model selected by BVS (dashed line). For both case studies, the effect of model averaging is clearly seen. The right panels of both figures show the posterior model probabilities. While there is much clearer candidate for the best model for Toxicity data, g_1 with $\bar{P}_{BVS}(g_1|data) = 0.38$, the result for Angina data supports nearly equally two models, g_9 ($\bar{P}_{BVS}(g_9|data) = 0.249$) and g_{10} ($\bar{P}_{BVS}(g_{10}|data) = 0.269$). Note that the results are conditional on specification of priors (for details see Section 2.3).

The posterior model probabilities obtained for the BVS and GORIC for the Angina data set are shown in left panel of Figure 4. For both methods, the highest posterior probabilities were obtained for models with an increment between the last two doses. However, GORIC tends to prefer more complex models with smaller increments across multiple doses (g_{13}, g_{15}), while the BVS selects models with just few larger increments (g_9, g_{10}). The posterior probabilities of the MED are shown in the right panel of Figure 4. Both GORIC and the BVS assigned the highest posterior probability of being MED to the first dose. However, there is a difference between the two methods. Since GORIC method selects models with more parameters, it gives higher probability to models with increment already between first and second dose and therefore $P(\text{MED} = 1|\text{data})$ is estimated with large posterior probability, $\bar{P}_{GORIC}(\text{MED} = 1|\text{data}) = 0.741$. It also assigns nearly zero probability to $\bar{P}_{GORIC}(\text{MED} = 4|\text{data}) = 0.002$. In contrast, the BVS method gives much lower posterior probability to $\bar{P}_{BVS}(\text{MED} = 1|\text{data}) = 0.490$ and the posterior distribution of the MED is more equally spread over all doses, i.e. $\bar{P}_{BVS}(\text{MED} = 2|\text{data}) = 0.325$ and $\bar{P}_{BVS}(\text{MED} = 4|\text{data}) = 0.041$. The complete results are presented in Table 2 (the mean structure of the models is shown in Table S1 of the supplementary appendix). We can see that the results obtained for the AIC and BIC methods lie between the results obtained for GORIC and the BVS methods. Note that the results for the BIC are much closer to results of the BVS.

Similar pattern can be seen for the Toxicity data in Figure 5. While GORIC prefers a more complex model g_5 (having three different means) with $\text{MED} = 1$, the BVS suggests that the best model is g_1, while giving much higher posterior probabilities to other models, such as g_0, g_4 and g_5. Once again, both methods estimated the highest posterior probability

of being the MED for the same dose level, with GORIC estimate $\bar{P}_{GORIC}(\text{MED}=1|\text{data})=0.833$ and the BVS estimate $\bar{P}_{BVS}(\text{MED}=1|\text{data})=0.644$. Similarly to the Angina data, GORIC assigns very high posterior probability to $\text{MED}=1$ (see right panel of Figure 5), while BVS spread probability more equally, estimating relatively high posterior probabilities for other doses. Note that in Table 3 not all models were fitted for GORIC, AIC and BIC. That is caused by the violation of monotonicity assumption in the observed means between dose 2 and dose 3 (see Figure 3). As mentioned above, isotonic regression was used to estimate the order restricted means. While we incorporate the order restrictions for maximum likelihood estimation, the models with increase between dose 2 and dose 3 reduced to models that have a flat mean profile between dose 2 and dose 3 (e.g. model g_2 will reduce to model g_0). Therefore, only a subset of models g_0, g_1, g_4, g_5 with no increment between the dose 2 and dose 3 can be actually fitted and estimated. This property does not apply to the BVS model, because it does not use isotonic regression for the estimation of the means.

In both data sets, GORIC seems to support models with less equalities (i.e. more complex models) compared to the BVS and therefore estimates higher posterior probabilities for the lower values of the MED. Both methods tend to select similar patterns, but small differences between consecutive doses are treated as flat by the BVS but as increments by GORIC. The cause of this difference is due to the fact that the penalty of GORIC is rather low when additional parameters are added to the model. Hence, GORIC supports more complex models and results in much higher $\bar{P}_{GORIC}(\text{MED}=1|\text{data})$. On the other hand, the results for the BVS suggest that a model reduction step is addressed automatically within the procedure and a relatively large difference among doses is needed to include the increment in the model. As a consequence, the distribution of $\bar{P}_{BVS}(\text{MED}=i|\text{data})$ is spread more equally across the doses. The AIC and BIC are somewhere between the other two methods, AIC being closer to GORIC and BIC closer to BVS. This is expected since compared to the AIC, the BIC has a tendency to select less complex models due to a high penalty term.

As expected, the choice of the criterion determines the posterior distribution of MED. Although the MED with the highest posterior probability could be the same for different methods, substantial differences can be observed in the underlying posterior distribution that quantifies the uncertainty in the choice of MED. On the other hand, the choice of the criterion can incorporate our preference for a more or less complex model in the process of the estimation of the posterior probabilities.

5 Simulation study

5.1 Simulation setting

Considering the findings in Section 4, we conducted a simulation study to explore the performance of various methods according to a true underlying model. The simulation setting represents an experiment with $K=4$ dose levels with $n=3$ observations per dose. The configuration for the mean structure $\mu_0, \mu_1, \mu_2, \mu_3$ followed the specification given by Marcus (1976) (details are given in Section S4 in the supplementary appendix for the manuscript). Data were generated according to an order restricted model defined in Equation (6), $Y_{ij} \sim N(\lambda\mu_i, \sigma^2)$, with $\sigma^2=1$, for each of the models g_0, K, g_7. The values of $\lambda=1,2,3$ were used, representing different magnitudes of true effect. In total, $N=1000$ data sets were generated for each combination of a specific model and λ (i.e. in total 22 combinations were

simulated, 7×3 for g_1,K , g_7 and one for g_0, each 1000 times).

For all the methods, an assumption of a non-decreasing trend was made. As explained in the previous section, not all the models can be fitted for the IC methods in each simulated data set (when violation of monotonicity in simulated means occurs), while the BVS provided posterior probability for all the models in each simulated data set. The posterior model probabilities, $\bar{P}(g_r \mid \text{data})$, were computed according to the BVS, AIC, BIC and GORIC methods. The posterior probabilities for the MED, $\bar{P}(\text{MED} = i \mid \text{data})$, were derived by summation of appropriate posterior model probabilities. The methods were evaluated based on two criteria: the correct identification of the true underlying model and the correct identification of the true underlying MED. Additionally, the setting when the best model and the second best model are considered for evaluation is briefly discussed in Section 6 and the full results are shown in Section S4 of the supplementary appendix for the manuscript.

5.2 Simulation results

As shown in Table 4, performance according to model complexity is profound in simulation study results. While the BVS clearly performs better for simple models with only one or two different mean levels (g_0, g_1, g_2 and g_4), GORIC achieves better results for complex models (g_3, g_6, g_7). The result for model g_5 highlights another interesting point. While the magnitude of the difference is getting higher, GORIC seems to prefer more complex models (splitting high increment among more dose levels). Therefore, if $\lambda = 3$, the BVS overtakes GORIC in terms of correct selection of the model g_5 and reduces the difference for models g_3 and g_6. Clearly, GORIC is better method for the detection of model g_7. On the other hand, it shows the worst performance for the simplest model g_0 that can be of profound interest, representing absence of dose-response relationship. Interestingly, the AIC method performs well. While being always between BVS and GORIC, it shows good performance, except for model g_7. Performance of BIC is rather poor, being among the worst methods for all the possible models (and except g_0, being always worse than AIC). The complexity of the models selected by a specific method depends on the penalty term of that method. Typically, it holds that penalty of GORIC is smaller than penalty of the AIC that is (for $n > 7$) smaller than penalty of the BIC. Therefore, the AIC and GORIC may select more complex models. As was pointed out in Section 2, the AIC and GORIC methods do not assume that the true model is necessarily among the candidate models and they try to approximate it, while BVS model assumes that the true model is among the candidate models. Additional results for varying number of doses ($K = 4,5$) and replicates within dose ($n = 3,4,5,10$) indicate the same patterns and are presented in Section S4 in the supplementary appendix for the manuscript.

The main goal of the analysis is to estimate the MED. The selection was done after summing up posterior probabilities of the models with respective MED. The evaluation of methods based on correct identification of the MED, presented in Table 5, leads to different conclusions than correct model selection based analysis. We can see an overall improvement in the correct identification rate. This is due to the fact that if the true model is not selected, the methods tend to select the model with the same MED. The clearest improvement occurs for GORIC, especially for model g_1. The magnitude of the increment, represented by λ, seems to be an important factor for a correct MED determination. Clearly, GORIC performs better for $\lambda = 1$ for most of the models, while the BVS outperforms GORIC for nearly all of

the models if $\lambda = 3$. The model complexity factor stays clearly visible only for model g_4 (increment only in last dose) and g_7 (increment in all doses). The AIC seems very suitable for MED selection. It has never been the best method, but it has never had worse performance than both BVS and GORIC simultaneously. The BIC does not provide good results, in some cases it performed slightly better than other methods, but it is often the worst method with rather poor overall performance. Similar results for additional settings are presented in Section S4 in the supplementary appendix for the manuscript.

6 Discussion

The manuscript discusses the Bayesian variable selection method for the model selection and the estimation of the minimum effective dose. A comparison with competing methods based on information criteria GORIC, AIC and BIC was conducted in both case studies and simulation study. The AIC and BIC are not designed for comparison of order restricted models, but their performance was not entirely poor. In most of the cases they have been behaving in between GORIC and BVS, while being rather close to worse of them, especially in case of MED selection. Therefore, focus is put on comparison of BVS and GORIC.

General advantage of BVS compared to IC based methods is its unified framework for inference, estimation and model selection. While posterior probabilities $\bar{P}(g_r \mid \text{data}, g_0, \text{K}, g_R)$ can be used as a model selection tool, the dose-specific means estimates are based on the weighted average of model-specific estimates according to the posterior model probabilities. Therefore, the BVS model provides estimates for the dose-specific means while taking model uncertainty into account. Similarly, the model averaged estimates can be obtained for IC methods by using model-specific maximum likelihood estimates weighted by appropriate model weights.

Additionally, the BVS fits all the models simultaneously. In contrast with the IC based methods, the number of fitted models does not increase with increasing number of dose levels. For $K > 5$, the amount of models to be fitted can become prohibitive for IC based methods if fitting many candidate models is required. Ideally, the set of candidate models can be reduced by focusing on informative hypothesis framework, but multiple models always need to be considered. If the set of hypotheses is based on strict inequalities, as in our case, the careful interpretation of IC based models is needed. As we have seen in the simulation study and case studies, some of the models in the set could reduce to single model in fitting stage, which may lead to underestimation of certain posterior model probabilities, if violation of monotonicity in the dose-specific means is present. Such issue does not occur for BVS. However, note that this is not general property of IC based methods and it arises from fact that strict inequality based set is used in order to compare the Ic methods with BVS model.

There is clear pattern when BVS and GORIC are compared. The BVS model outperforms GORIC in case of less complex underlying models or higher magnitude of overall difference (λ), mainly for models g_2 and g_4. However, in case of small overall differences, it tends to oversimplify the models, especially for the most complex model g_7. In contrast, GORIC method prefers complex models, performing best for g_7 and often for g_5. However, this leads to its poor performance in case of high magnitude of difference and simplest models as g_0 or g_1.

While taking into account not only the best model, but also the second best model (with respect to posterior probability), the BVS model performs much better, relatively to GORIC (details are presented in Section S4 in the supplementary appendix).

Regarding the MED selection, the performance for higher magnitudes (λ) is of main

interest from an application point of view. As mentioned in the Section 1, the MED is typically related to the clinical significance as well as to the statistical significance. Therefore, cases of small overall effects are not of imminent interest. The bigger the overall dose effect is, the higher is the chance that MED would be relevant and its correct estimate is needed.

In summary, we have seen that both BVS and GORIC have its strong points. BVS performs better for the less complex models from the candidates set, while GORIC is able to select correctly models that are more complex. Therefore, GORIC will tend to select as MED lower doses, while BVS will tend to select higher doses, because GORIC will tend to split the overall effect across multiple doses (more complex model), while BVS will keep it rather at one particular dose (less complex models). Due to its insensitivity to model g_0, GORIC should be only used after initial filtering step, in case that the model selection is would be interpreted in terms of inference.

Based on our findings, the choice of methodology depends on the scope of the particular project and interpretation. Our main recommendation to the readers is to keep this fact in mind and not to select any of these methods as automatic preference, but always carefully evaluate, which set of models is considered, if strict inequalities are assumed and if set of candidate models can be a priori reduced based on scientific knowledge.

References

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Anraku, K. (1999), "An Information Criterion for Parameters Under a Simple Order Restriction," *Biometrika*, 86, 141–152.

Barlow, R. E., Bartholomew, D. J., Bremner, M. J., and Brunk, H. D. (1972), *Statistical Inference under Order Restriction*, New York: John Wiley & Sons.

Bornkamp, B., Pinheiro, J. C., and Bretz, F. (2009), "MCPMod - An R Package for the Design and Analysis of Dose-Finding Studies," *Journal of Statistical Software*, 29, 1–23.

Bretz, F. and Hothorn, L. A. (2003), "Statistical Analysis of Monotone or Non-monotone Dose-Response Data from *In Vitro* Toxicological Assays," *Alternatives to Lab Animals*, 31, 81–96.

Bretz, F., Pinheiro, J. C., and Branson, M. (2005), "Combining multiple comparisons and modeling techniques in dose-response studies," *Biometrics*, 61, 738–748.

Burnham, K. P. and Anderson, D. R. (2002), *Model Selection and Multimodel Inference: A Practical Information - Theoretic Approach*, New York: Springer.

— (2004), "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261–304.

Claeskens, G. and Hjort, N. L. (2008), *Model Selection and Model Averaging*, Cambridge University Press.

Denwood, M. J. (In Review), "runjags: An R Package Providing Interface Utilities, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS," *Journal of Statistical Software*.

Djira, G. D., Hasler, M., Gerhard, D., and Schaarschmidt, F. (2012), *mratios: Inferences for Ratios of Coefficients in the General Linear Model*, R package version 1.3.17.

Dunnett, C. W. (1955), "A Multiple Comparison Procedure for Comparing Several Treatments with a Control," *Journal of the American Statistical Association*, 50, 1096–1121.

Dunson, D. B. and Neelon, B. (2003), "Bayesian Inference on Order Constrained Parameters in Generalized Linear Models," *Biometrics*, 59, 286–295.

European Medicines Agency (2002), *Points to Consider on Multiplicity Issues in Clinical Trials*, no. CPMP/EWP/908/99, London.

Fitzhugh, O. G., Nelson, A. A., and Quaife, M. L. (1964), "Chronic Oral Toxicity of Aldrin and Dieldrin in Rats and Dogs," *Food Cosmetic Toxicology*, 2, 551–562.

Gelfand, A. E., Smith, A. F. M., and Lee, T.-M. (1992), "Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling," *Journal of the American Statistical Association*, 87, 523–532.

George, E. I. and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Hothorn, L. A. and Hauschke, D. (2000), "Identifying the Maximum Safe Dose: A Multiple Testing Approach," *Journal of Biopharmaceutical Statistics*, 10, 15–30.

Jeffreys, H. (1961), *Theory of Probability*, London: Oxford University Press, 3rd ed.

Kasim, A., Shkedy, Z., and Kato, B. S. (2012), "Estimation and Inference Under Simple Order Restrictions: Hierarchical Bayesian Approach," in *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R*, eds. Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., and Bijnens, L., Springer, Berlin, pp. 193–214.

Kass, R. E. and Wasserman, L. (1996), "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 91, 1343–1370.

Kim, S. B., Kodell, R. L., and Moon, H. (2014), "A Diversity Index for Model Space Selection in the Estimation of Benchmark and Infectious Doses via Model Averaging," *Risk Analysis*, 34, 453–464.

Klugkist, I. and Mulder, J. (2008), "Bayesian Estimation for Inequality Constrained Analysis of Variance," in *Bayesian Evaluation of Informative Hypotheses*, eds. Hoijtink, H., Klugkist, I., and Boelen, P. A., New York: Springer, pp. 27–52.

Kodell, R. L. (2009), "Replace the NOAEL and LOAEL with the $BMDL_{01}$ and $BMDL_{10}$," *Environmental and Ecological Statistics*, 16, 9–12.

Kong, M., Rai, S. N., and Bolli, R. (2014), "Statistical Methods for Selecting Maximum Effective Dose and Evaluating Treatment Effect When Dose-Response is Monotonic," *Statistics in Biopharmaceutical Research*, 6, 16–29.

Kuiper, R. M., Gerhard, D., and Hothorn, L. A. (2014), "Identification of the Minimum Effective Dose for Normally Distributed Endpoints Using a Model Selection Approach," *Statistics in Biopharmaceutical Research*, 6, 55–66.

Kuiper, R. M., Hoijtink, H., and Silvapulle, M. J. (2011), "An Akaike-type Information Criterion for Model Selection Under Inequality Constraints," *Biometrika*, 98, 495–501.

Kullback, S. and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86.

Kuo, L. and Mallick, B. (1998), "Variable Selection for Regression Models," *The Indian Journal of Statistics*, 60, 65–81.

Lin, D., Shkedy, Z., and Aerts, M. (2012), "Classification of Monotone Gene Profiles Using Information Theory Selection Methods," in *Modeling Dose-response*

Microarray Data in Early Drug Development Experiments Using R, eds. Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., and Bijnens, L., Springer, Berlin, pp. 151–164.

Lin, D., Shkedy, Z., Burzykowski, T., Aerts, M., Göhlmann, H. W. H., De Bondt, A., Perera, T., Geerts, T., Van den Wyngaert, I., and Bijnens, L. (2009), "Classification of Trends in Dose-Response Microarray Experiments Using Information Theory Selection Methods," *The Open Applied Informatics Journal*, 3, 34–43.

Liu, J. (2010), "Minimum Effective Dose," in *Encyclopedia of Biopharmaceutical Statistics*, ed. Chow, S., Taylor & Francis, 3rd ed., pp. 799–800.

Marcus, R. (1976), "The Powers of Some Tests of the Equality of Normal Means against an Ordered Alternative," *Biometrika*, 63, 177–183.

O'Hara, R. B. and Sillanpää, M. J. (2009), "Review of Bayesian Variable Selection Methods: What, How and Which," *Bayesian Analysis*, 4, 85–118.

Ohlssen, D. and Racine, A. (2015), "A Flexible Bayesian Approach for Modeling Monotonic Dose-Response Relationships in Drug Development Trials," *Journal of Biopharmaceutical Statistics*, 25, 137–156.

Otava, M., Shkedy, Z., Lin, D., Göhlmann, H. W. H., Bijnens, L., Talloen, W., and Kasim, A. (2014), "Dose-Response Modeling Under Simple Order Restrictions Using Bayesian Variable Selection Methods," *Statistics in Biopharmaceutical Research*, 6, 252–262.

Pinheiro, J., Bretz, F., and Branson, M. (2006), "Analysis of Dose-Response Studies: Modeling Approaches," in *Dose finding in drug development*, ed. Ting, N., Springer, New York, pp. 146–171.

Plummer, M. (2003), "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Pramana, S., Shkedy, Z., Göhlmann, H. W. H., Talloen, W., De Bondt, A., Straetemans, R., Lin, D., and Pinheiro, J. (2012), "Model-Based Approaches," in *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R*, eds. Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., and Bijnens, L., Springer, pp. 215–232.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R

Foundation for Statistical Computing, Vienna, Austria.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference*, John Wiley & Sons Ltd.

Schwarz, G. (1978), "Estimating the Dimension of a Model." *Annals of Statistics*, 6, 461–464.

Seber, G. A. F. and Wild, C. J. (1989), *Nonlinear Regression*, New York: Wiley & Sons.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit (with Discussion)," *Journal of the Royal Statistical Society, Series B*, 64, 583–639.

Straetemans, R. (2012), "Nonlinear Modeling of Dose-Response Data," in *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R*, eds. Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., and Bijnens, L., Springer, pp. 43–66.

Sugiura, N. (1978), "Further Analysts of the Data by Akaike's Information Criterion and the Finite Corrections," *Communications in Statistics - Theory and Methods*, 7, 13–26.

Wang, S.-J., Hung, H. M. J., and O'Neill, R. (2011), "Regulatory Perspectives on Multiplicity in Adaptive Design Clinical Trials throughout a Drug Development Program," *Journal of Biopharmaceutical Statistics*, 21, 846–859.

Wang, W. and Peng, J. (2015), "A Step-Up Test Procedure to Find the Minimum Effective Dose," *Journal of Biopharmaceutical Statistics*, 25, 525–538.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., and Hochberg, Y. (1999), *Multiple Comparisons and Multiple Tests Using the SAS System*, Cary, NC: SAS Institute Inc.

Whitney, M. and Ryan, L. (2009), "Quantifying Dose-Response Uncertainty Using Bayesian Model Averaging," in *Uncertainty Modeling in Dose Response: Bench Testing Environmental Toxicity*, ed. Cooke, R. C., John Wiley & Sons, Inc., pp. 165–179.

Yanagawa, T. and Kikuchi, Y. (2001), "Statistical Issues on the Determination of the No-Observed-Adverse-Effect Levels in Toxicology," *Environmetrics*, 12, 319–325.

Figure 1: *The two case studies. Crosses represent dose-specific means. Left panel: the Angina data set. Right panel: the Toxicity data set.*

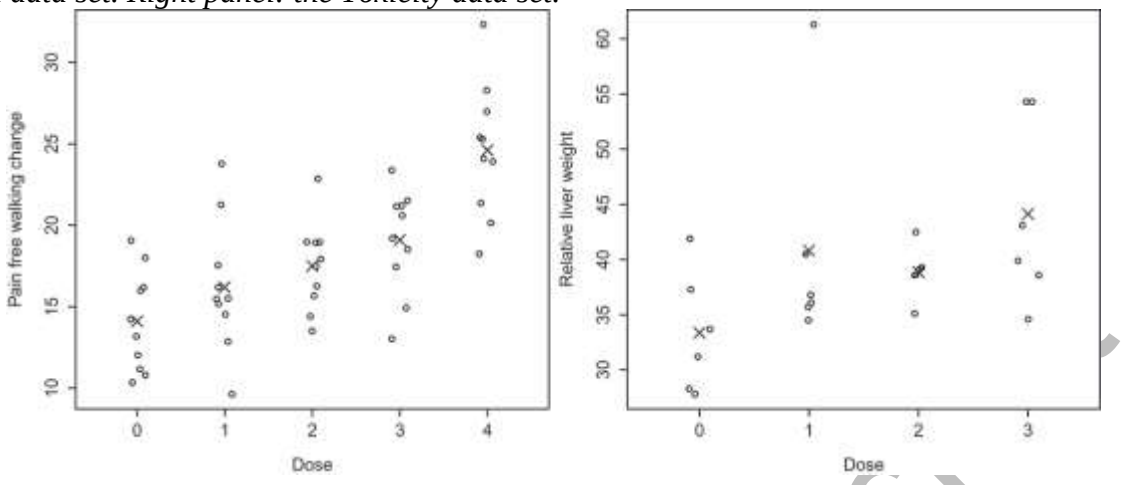


Figure 2: *The Angina data. Left panel: Observed data, sample means (crosses) and posterior means of the BVS model (solid line) and model* g_{10} *(dashed line). Right panel: Posterior probability for* g_r, $r = 0,\mathrm{K},15$. *Notation corresponds to the model numbers presented in Table 1, extended respectively for* $K = 5$ *(see Table S1 in a supplementary appendix of the manuscript).*

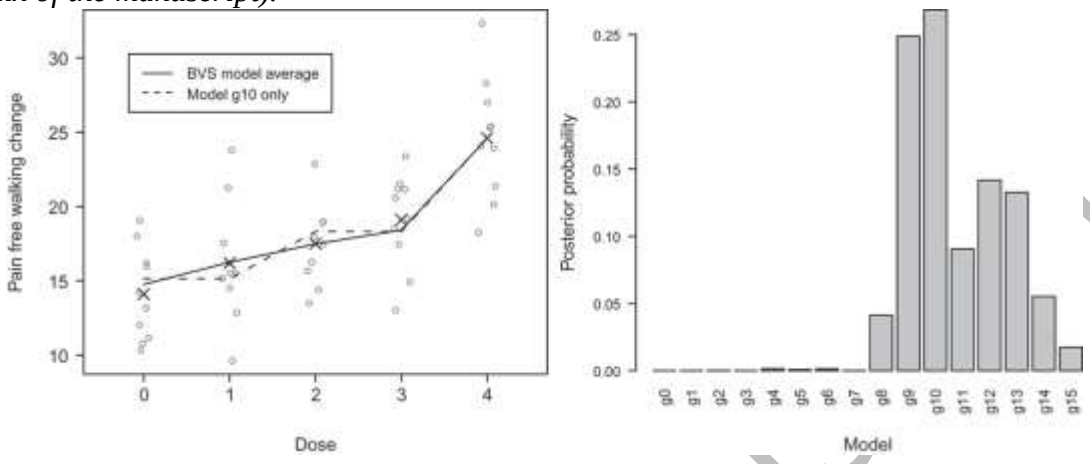


Figure 3: *The Toxicity data. Left panel: Observed data, sample means (crosses) and posterior means of the BVS model (solid line) and model* g_1 *(dashed line). Right panel: Posterior probability for* g_r, $r = 0,\mathrm{K},7$. *Notation corresponds to the model numbers presented in Table 1.*

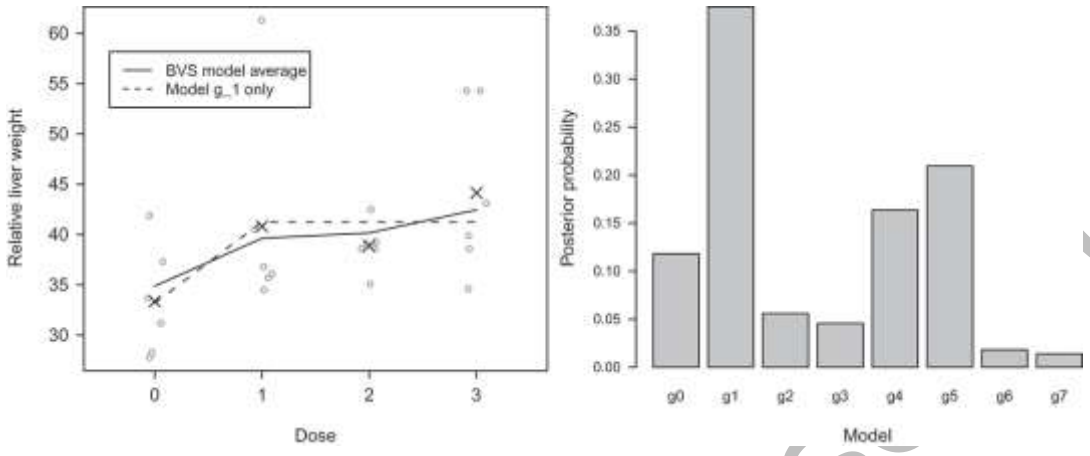


Figure 4: *The Angina data. The BVS results (black) and GORIC results (grey) comparison. Left panel: Posterior probability for* g_r, $r = 0,\text{K},15$. *Right panel: Posterior probability for the MED.*

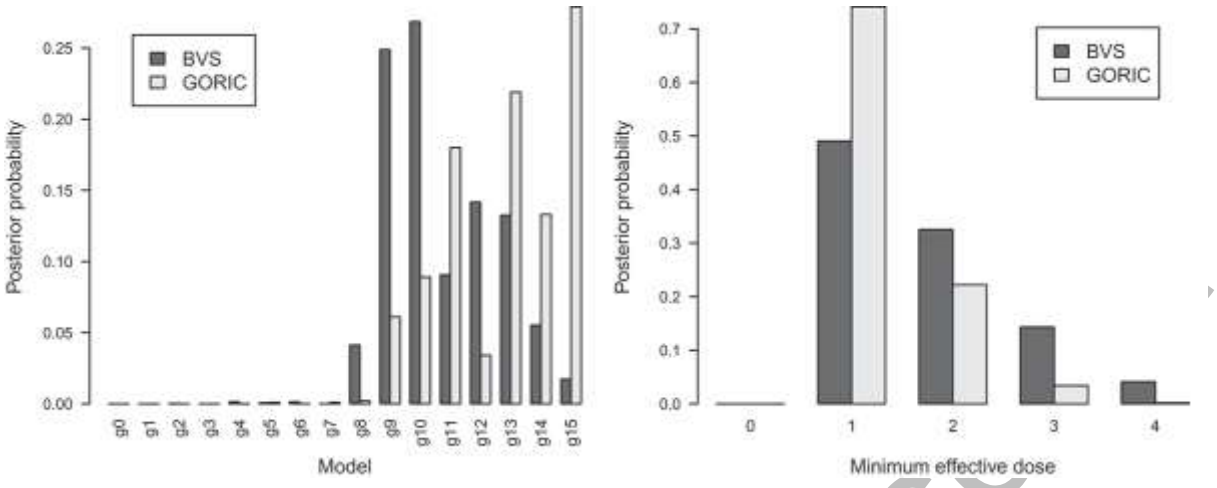


Figure 5: *The Toxicity data. The BVS results (black) and GORIC results (grey) comparison. Left panel: Posterior probability for* g_r, $r = 0,\mathrm{K},7$. *Right panel: Posterior probability of the MED.*

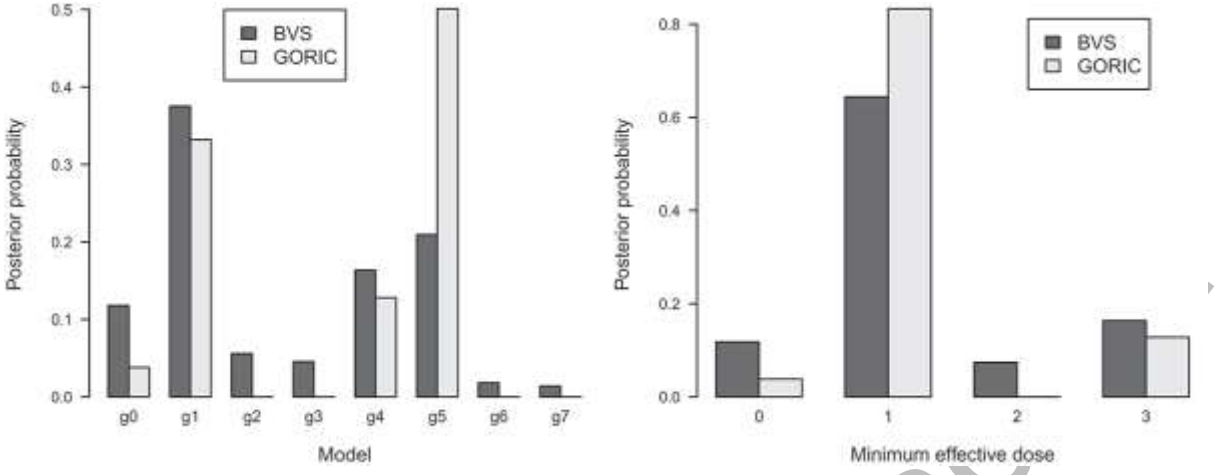


Table 1: The set of eight possible monotonic dose-response models for an experiment with four dose levels (including placebo). Denote μ_i the mean response of dose level. The model g_0 represents the null model of no dose effect.

Model	Up: Mean Structure	Down: Mean Structure
g_0	$\mu_0 = \mu_1 = \mu_2 = \mu_3$	$\mu_0 = \mu_1 = \mu_2 = \mu_3$
g_1	$\mu_0 < \mu_1 = \mu_2 = \mu_3$	$\mu_0 > \mu_1 = \mu_2 = \mu_3$
g_2	$\mu_0 = \mu_1 < \mu_2 = \mu_3$	$\mu_0 = \mu_1 > \mu_2 = \mu_3$
g_3	$\mu_0 < \mu_1 < \mu_2 = \mu_3$	$\mu_0 > \mu_1 > \mu_2 = \mu_3$
g_4	$\mu_0 = \mu_1 = \mu_2 < \mu_3$	$\mu_0 = \mu_1 = \mu_2 > \mu_3$
g_5	$\mu_0 < \mu_1 = \mu_2 < \mu_3$	$\mu_0 > \mu_1 = \mu_2 > \mu_3$
g_6	$\mu_0 = \mu_1 < \mu_2 < \mu_3$	$\mu_0 = \mu_1 > \mu_2 > \mu_3$
g_7	$\mu_0 < \mu_1 < \mu_2 < \mu_3$	$\mu_0 > \mu_1 > \mu_2 > \mu_3$

Table 2: Estimated posterior model probabilities for the Angina data for GORIC, AIC, BIC and BVS. First column: Order restricted log-likelihood.

Profile	ORLL	GORIC	AIC	BIC	BVS
g_0	–149.77	0.00	0.00	0.00	0.00
g_1	–144.55	0.00	0.00	0.00	0.00
g_2	–141.46	0.00	0.00	0.00	0.00
g_3	–140.80	0.00	0.00	0.00	0.00
g_4	–138.65	0.00	0.00	0.00	0.00
g_5	–136.92	0.00	0.00	0.00	0.00
g_6	–137.39	0.00	0.00	0.00	0.00
g_7	–136.61	0.00	0.00	0.00	0.00
g_8	–135.97	0.00	0.01	0.04	0.04
g_9	–132.31	0.06	0.13	0.21	0.25
g_{10}	–131.99	0.09	0.18	0.29	0.27
g_{11}	–131.01	0.18	0.17	0.11	0.09
g_{12}	–133.01	0.03	0.06	0.11	0.14
g_{13}	–130.82	0.22	0.21	0.13	0.13
g_{14}	–131.42	0.13	0.12	0.07	0.06
g_{15}	–130.43	0.28	0.11	0.03	0.02

Table 3: Estimated posterior model probabilities for the Toxicity data for GORIC, AIC, BIC and BVS. First column: Order restricted log-likelihood. Note that, as explained in Section 4, some of the models were not fitted for IC; due to the incorporated order restrictions they reduced to other models.

Profile	ORLL	GORIC	AIC	BIC	BVS
g_0	–82.98	0.04	0.08	0.16	0.12
g_1	–80.32	0.33	0.42	0.46	0.38
g_2	—	0	0	0	0.06
g_3	—	0	0	0	0.05
g_4	–81.28	0.13	0.16	0.18	0.16
g_5	–79.51	0.50	0.34	0.21	0.21
g_6	—	0	0	0	0.02
g_7	—	0	0	0	0.01

Table 4: Comparison of proportion of time the true model is selected based on 1000 simulated data sets for BVS, GORIC, AIC and BIC criterion for $K = 4$, $n = 3$.

λ	Profile	BVS	GORIC	AIC	BIC
	g_0	0.73	0.59	0.76	0.81
1	g_1	0.57	0.51	0.53	0.49
	g_2	0.46	0.42	0.47	0.46
	g_3	0.03	0.16	0.05	0.03
	g_4	0.55	0.48	0.51	0.48
	g_5	0.08	0.22	0.09	0.07
	g_6	0.02	0.16	0.04	0.02
	g_7	0.00	0.03	0.00	0.00
2	g_1	0.83	0.63	0.78	0.80
	g_2	0.78	0.54	0.73	0.77
	g_3	0.22	0.48	0.30	0.23
	g_4	0.82	0.61	0.78	0.79
	g_5	0.43	0.54	0.49	0.42
	g_6	0.23	0.46	0.29	0.24
	g_7	0.01	0.28	0.04	0.02
3	g_1	0.88	0.63	0.79	0.83
	g_2	0.84	0.55	0.76	0.81
	g_3	0.59	0.66	0.64	0.60
	g_4	0.86	0.62	0.80	0.83
	g_5	0.79	0.67	0.77	0.77
	g_6	0.57	0.65	0.63	0.59
	g_7	0.09	0.62	0.25	0.19

Table 5: Comparison of proportion of time the true MED is selected based on 1,000 simulated data sets for BVS, GORIC, AIC and BIC criterion for $K = 4$, $n = 3$.

λ	Profile	BVS	GORIC	AIC	BIC
	g_0	0.73	0.59	0.76	0.81
1	g_1	0.62	0.73	0.61	0.55
	g_2	0.47	0.51	0.49	0.47
	g_3	0.40	0.53	0.39	0.34
	g_4	0.55	0.48	0.51	0.48
	g_5	0.39	0.53	0.39	0.35
	g_6	0.32	0.40	0.36	0.34
	g_7	0.32	0.44	0.32	0.29
2	g_1	0.96	0.99	0.96	0.94
	g_2	0.83	0.72	0.82	0.83
	g_3	0.61	0.81	0.65	0.59
	g_4	0.82	0.61	0.78	0.79
	g_5	0.70	0.85	0.74	0.71
	g_6	0.57	0.60	0.59	0.59
	g_7	0.48	0.70	0.53	0.48
3	g_1	1.00	1.00	1.00	1.00
	g_2	0.91	0.72	0.86	0.90
	g_3	0.82	0.94	0.86	0.83
	g_4	0.86	0.62	0.80	0.83
	g_5	0.90	0.98	0.93	0.91
	g_6	0.75	0.69	0.76	0.76
	g_7	0.64	0.86	0.71	0.66