Proof of Concept and dose estimation with binary responses under model uncertainty

B. Klingenberg*,†

Dept. of Mathematics and Statistics, Williams College, Williamstown, MA 01267, U.S.A.

SUMMARY

This article suggests a unified framework for testing Proof of Concept and estimating a target dose for the benefit of a more comprehensive, robust and powerful analysis in phase II or similar clinical trials. From a pre-specified set of candidate models we choose the ones that best describe the observed dose-response. To decide which models, if any, significantly pick up a dose effect we construct the permutation distribution of the minimum P-value over the candidate set. This allows us to find critical values and multiplicity adjusted P-values that control the familywise error rate of declaring any spurious effect in the candidate set as significant. Model averaging is then used to estimate a target dose. Popular single or multiple contrast tests for Proof of Concept, such as the Cochran-Armitage, Dunnett or Williams tests are only optimal for specific dose-response shapes and do not provide target dose estimates with confidence limits. A thorough evaluation and comparison of our approach to these tests reveals that its power is as good or better in detecting a dose-response under various shapes, with many more additional benefits: It incorporates model uncertainty in Proof of Concept decisions and target dose estimation, yields confidence intervals for target dose estimates and extends to more complicated data structures. We illustrate our method with the analysis of a Phase II clinical trial.

KEY WORDS: dose finding; contrast tests; dose-response, minimum effective dose; model selection; trend test.

*Correspondence to: Dept. of Mathematics and Statistics, Williams College, Williamstown, MA 01267, U.S.A.
†email: bklingen@williams.edu

1. Introduction

Dose-response studies are important tools for investigating the existence, nature and extent of a dose effect in drug development, toxicology and related areas. The following four questions, usually in this order, are of prime interest [1]: i.) Is there any evidence of a dose effect (i.e., *Proof of Concept*), ii.) Which doses exhibit a response different from the control response, iii.) What is the nature of the dose-response relationship and iv.) What dose should be selected for further studies/marketing (i.e., *target dose estimation*)? Multiple comparison procedures in the form of single or multiple contrast tests are usually applied to address i.) and ii.), whereas statistical modeling additionally answers questions iii.) and iv.) at the expense of more elaborate assumptions. With the failure rate of current phase III trials reaching 50%, several of these failed trials are attributed to improper target dose estimation/selection in phase II and incorrect or incomplete knowledge of the dose-response (FDA meeting on "Good Dose Response", Oct. 2004). Furthermore, the FDA reports that 20% of the approved drugs between 1980 and 1989 had the initial dose changed by *more than 33%*, in most cases lowering it. Clearly, new or revised concepts are needed to make the decision and dose estimation process more efficient.

For independent and homoscedastic normal data, Bretz, Pinheiro and Branson [2] recently presented a unified framework to better address PoC and target dose estimation in phase II clinical trials. They construct optimal contrast tests from each of several candidate models for the underlying dose-response profile. Subject to establishing PoC with these tests, they pick the model that gives the maximal contrast for further model-based inference. For categorical responses, the construction of optimal contrasts tests is not straightforward due to their mean-variance relationship and the selection of contrast coefficients is often subjective. Further, they only establish the presence (and sometimes size) of a dose effect, but do not provide confidence intervals for a target dose estimate or allow the consideration of clinical or regulatory requirements. However, handling *all* of the questions i.) through iv.) in a *unified framework* is a desirable goal and explored in this article.

In Section 2 we start with a candidate set of several plausible dose-response models for a binary response to incorporate model uncertainty in the PoC decision and subsequent target dose estimation. Since the target dose heavily depends on the assumed shape, considering several shapes a priori makes the procedure more robust to model misspecification. To decide which of the candidate models, if any, significantly pick up the dose-response signal we compare each one to the no-effect model via a penalized deviance difference statistic. To adjust the significance levels of these multiple tests of

PoC (one under each candidate model), we use the step-down approach with the minimum P-value of Westfall and Young [3] that controls the familywise error rate (FWER) of declaring any spurious signal in the family of candidate models as significant. In Section 4 we move to target dose estimation such as the minimum effective dose (MED), using model averaging. Through an extensive simulation study in Sections 3 and 4, we evaluate and compare our proposed framework to many popular contrast tests, such as the Cochran-Armitage, Dunnett or Williams procedures. For a variety of possible dose-response shapes (and known MED's), we find the probability that our approach establishes PoC, and we find the median bias, precision and median squared error in the estimation of the MED. The simulation results show that our approach easily competes with the most powerful contrast tests for establishing PoC under all shapes considered, and additionally provides an estimate of the MED with error bounds. In Section 5 we investigate the sensitivity of the analysis to the candidate set, while Section 6 briefly discusses some extensions.

The need to adjust dose estimation (but not so much PoC) for model uncertainty has been recognized for some time now, addressed mostly in a Bayesian fashion via Bayesian model averaging (BMA, [4, 5]). Recently, Morales et al. [6] illustrated BMA for benchmark dose estimation in quantitative risk assessment. Here, however, we are interested in comparing substantially different, mostly non-nested models, whose parameters are on different scales and have different interpretations, which complicates prior specifications and assigning prior odds to candidate models. For further discussion of these and related topics see [7]. Burnham and Anderson [8] present the issue of selecting among and averaging across candidate models in a frequentist framework, using AIC values as selection criteria and model weights. We will adapt these ideas to our context of identifying dose-response models that show a significant dose effect. However, since we operate in a regulatory environment, our procedures will adhere to a strong error control, controlling the probability of erroneously declaring PoC with at least one candidate model when in fact the compound under investigation has no effect. As the PoC decision is an extremely important one with expensive consequences, this is a crucial feature. Controlling the risk of further pursuing an ineffective drug or a drug given at the wrong dose seems important both in terms of protecting humans from unnecessary/uncertain exposure and saving resources for other, more promising venues. Further, according to the ICH-E4 guideline [9] on dose-response information to support drug registration, "a well controlled dose-response study [...] can serve as primary evidence of effectiveness", in which case error control becomes essential. Our simulations will show that merely using the candidate model with the smallest AIC (or the one that gives the smallest P-value in a test for PoC) leads to an inflation of this error over its nominal level, declaring PoC too often. In fact, for typical sample sizes used in phase II trials, without multiplicity adjustments this error may more than double, leading to greater uncertainty about the dose-response profile and the appropriateness of the selected dose.

2. PoC under model uncertainty

Let Y_{ij} be the binary response of subject j at dose d_i, $i = 1$ (placebo) $,\dots,k$, $j = 1,\dots,n_i$. We assume a parallel group design, with responses Y_{ij} independent within and across the k treatment arms. The extension to other layouts is briefly mentioned in Section 6. Let $\pi(d_i) = P(Y_{ij} = 1)$ denote the probability of a successful efficacy (or safety) outcome for subjects at dose d_i. To illustrate with some actual data, let Y_{ij} be the binary indicator for relief of abdominal pain over a three week period for patients suffering from Irritable Bowl Syndrome (IBS). A phase II clinical trial was set up to investigate the efficacy of a compound against IBS in women at $k = 5$ dose levels $\boldsymbol{d} = (0, 1, 4, 12, 24)mg$. Preliminary studies with only two doses indicated a placebo effect of around 30% and a maximal efficacy of roughly 65%. However, prior to the trial, investigators were uncertain about the shape of the dose-response profile in between these two extremes. In particular, they could not rule out strongly concave or convex patterns, or even a downturn at higher doses.

2.1. Defining a candidate set and test statistic

To address these issues of model uncertainty, we start by considering a candidate set $\mathcal{M} = \{M_1,\dots,M_m\}$ of m models for describing the unknown dose-response relationship. These models can vary with respect to a link function used and/or the nature of the effect of the dose as expressed in a (linear) predictor. However, with k doses there are k sufficient statistics (the number of success at each dose level), so the number of parameters appearing in the model must be kept below k to avoid overfitting or simply reproducing the data. At the same time, the models must be flexible within the predetermined efficacy range to allow for various plausible dose-response shapes. To illustrate, Table I presents $m = 10$ candidate models M_1 through M_{10} (plotted in Figure 1) that use at most 3 parameters and show a variety of potential dose-response shapes for the efficacy profile of the IBS compound, with the predetermined placebo effect of 30% and the maximum efficacy at 65%. Most of these models are generalized linear models (GLMs) with flexible fractional polynomial (Royston and Altman [10]) linear predictor form, successfully employed in similar applications of estimating a benchmark dose in quantitative risk assessment (e.g., [11, 12]). The candidate set covers a broad dose-response space and should include the scenarios anticipated by the clinical team developing the drug. For plotting the models, initial parameter estimates are computed from educated guesses of the placebo and maximum possible effect, for instance from pilot or previous studies. For three-parameter models, it is necessary to include additional information, e.g. at which dose the maximum efficacy is expected. An R function, available at the author's website `www.williams.edu/∼bklingen` automates this procedure and plots many flexible candidate models for a set of given specifications. (The Appendix shows the R code used to generate Figure 1). This function can be used as an interactive tool by the clinical team to add, modify or delete models from the candidate set until a reasonable choice covering plausible shapes is found. It should be noted that initial parameter estimates are only needed for illustration of the candidate dose-response curves, but not for the following methodology to work.

After defining a candidate set, we are interested in identifying those models that pick up a potential dose-response signal from the data. To this end, we compare each model $M_s \in \mathcal{M}$ to the model $M_0 : \pi(d_i) = \beta_0$ of no dose effect via a signed and penalized likelihood ratio statistic

$$T_s = (-1)^{I(\hat{\pi}_s(d_{\max}) \leq \hat{\pi}_s(d_1))} \{-2 [\log L(\boldsymbol{y}, \boldsymbol{n}; M_0) - \log L(\boldsymbol{y}, \boldsymbol{n}; M_s)]\} - 2\mathrm{df}_s.$$

Here, $I(.)$ is the indicator function and $L(\boldsymbol{y}, \boldsymbol{n}; M_s)$ is the maximized binomial likelihood corresponding to model M_s, with $\boldsymbol{y} = (y_{1+}, \ldots, y_{k+})$ and $\boldsymbol{n} = (n_1, \ldots, n_k)$ the observed number of successes and sample sizes at the k dose levels, respectively. The part of T_s in curly brackets is the deviance difference between the two models and equals $2 \sum y_{i+} \log[\hat{\pi}_s(d_i)/\hat{\pi}_0(d_i)] + 2 \sum (n_i - y_{i+}) \log[(1 - \hat{\pi}_s(d_i))/(1 - \hat{\pi}_0(d_i))]$, where $\hat{\pi}_s(d)$ is the ML estimate of $\pi(d)$ under model M_s and $\hat{\pi}_0(d) = \sum y_i / \sum n_i$. Naturally, we are only interested in a positive, i.e., beneficial dose effect (without loss of generality, we assume Y_{ij} is coded such that an increase in $\pi(d)$ with dose is desirable). Although straightforward for monotone profiles, in general we define a dose effect as positive for model M_s if $\hat{\pi}_s(d_{\max}) > \hat{\pi}_s(d_1)$, where $d_{\max} = \mathrm{argmax}_d |\hat{\pi}_s(d) - \hat{\pi}_s(d_1)|$ is the dose at which the maximum absolute effect relative to placebo occurs. This rules out shapes (e.g., certain quadratic ones) where a drop relative to placebo is too drastic, but still allows some shapes (e.g., certain J-shapes) with an initially negative effect. We use this signed version of the difference in deviance to filter out (and move to the lower tail) those models with a negative estimated dose effect, so that we can reject the null hypothesis and conclude a beneficial effect for large values of T_s. In addition, we penalize fitting more complex models by subtracting two times the difference in the number of parameters between M_s and M_0. Note that the unsigned version of T_s is identical to the difference between the AIC values for the two models, a statistic discussed for model selection by Lindsey and Jones [13] and strongly favored by [8]. Next, we turn to establishing the significance of T_s under simultaneous inference with *all* candidate models.

2.2. A permutation test for PoC

We are interested in models that are "best" suited for picking up the dose-response signal, i.e. that distance themselves as much as possible from M_0 on the likelihood scale, as measured by T_s. Let p_s be the P-value corresponding to T_s. We seek out those models that have the smallest P-values in the candidate set $\mathcal{M}$. A dose-response effect is established (i.e., PoC) if the minimum P-value is small enough, that is, if $\min_s p_s \leq c$, where c is a suitable chosen critical value (i.e., an adjusted significance level) such that the type I error rate of erroneously declaring PoC under simultaneous inference is controlled at the desired level α. We prefer working on the P-value scale and not directly on the scale of the test statistics T_s (e.g., by using $\max_s T_s$), as these are not standardized. In fact, the asymptotic distribution of T_s is proportional to a Chi-square with df_s degrees of freedom, see below.

Under the null hypothesis M_0 of no dose effect, doses are interchangeable. By finding $(T_1, \dots, T_m)$ for a random sample of the $\left[\sum_{i=1}^{k} n_i\right]!/\prod_{i=1}^{k} n_i!$ permutations or assignments of subjects to different dose levels, we obtain the exact (to any desired degree of accuracy) P-values for the observed test statistics T_s^{obs}: $p_s^{\text{obs}} = \frac{1}{B}\sum_{b=1}^{B} I(T_s^{(b)} \geq T_s^{\text{obs}})$, where $T_s^{(b)}$ is the value of T_s under the b-th permutation, $b = 1, \dots, B$ and $I(.)$ is the indicator function. From the permutation distribution of $(T_1, \dots, T_m)$, we also get the permutation distribution of the minimum P-value, $\min_s p_s$, by finding the minimum P-value $\min_s p_s^{(b)}$ for each permutation b, where $p_s^{(b)} = \frac{1}{B}\sum_{l=1}^{B} I(T_s^{(l)} \geq T_s^{(b)})$ is the P-value for $T_s^{(b)}$ under the b-th permutation. Then, by choosing c equal to the α percentile of the distribution of $\min_s p_s$, the overall type I error rate of our PoC decision criterion to reject if the smallest observed P-value is less than c is controlled at level α.

With a random sample of $B = 50,000$ permutations of the IBS data, Table II displays the observed T_s and their corresponding permutation P-values p_s^{obs}, under the 10 candidate models of Table I. Viewed individually, PoC can be established with all candidate models except M_7 at a significance level of $\alpha = 2.5\%$, say. From the permutation distribution of the minimum P-value, we obtain $c = 0.0083$

when $\alpha = 2.5\%$. (A plot of the histogram of $\min_s p_s$ is available via the R code shown in the Appendix). Since the minimum observed P-value $p_5^{\text{obs}} = 0.00001$ (only one permutation resulted in a larger T_5) is smaller than c, PoC can be established at the 2.5% significance level, which is not surprising given the sample proportions plotted in Figure 3. Note that finding the asymptotic distribution for $\min_s p_s$ is not straightforward, as it is based on the correlated (to various degrees) statistics $\{T_s\}$. The Bonferroni adjusted critical value $c_{\text{Bonf}} = \alpha/10 = 0.0025$ leads to more conservative inference and hence less power to detect PoC. If we assume independence among test statistics T_s, the asymptotic distribution of $\min_s p_s$ is $\text{Beta}(1, m)$, with 2.5 percentile equal to $c_{\text{Beta}} = 0.0025$ when $m = 10$, again a conservative estimate of the critical value.

2.3. Multiplicity adjusted P-values for testing PoC

Rather than finding the appropriate critical value, we can adjust the observed P-values to reflect multiplicity when testing PoC with several models simultaneously. Following the definition in Westfall and Young ([3], chpt. 2), the multiplicity adjusted P-value for testing PoC with model M_s is the proportion of permutations for which the minimum P-value over all models is smaller than the one observed for model M_s. In their more powerful step-down procedure, the permutation distribution of $\min_s p_s$ is used to find the adjusted P-value for the model with the smallest P-value, e.g. M_5 for the IBS data. Subsequent steps delete all permutations for the model with the smallest P-value (e.g., all $T_5^{(b)}, b = 1, \ldots, B$ are deleted) and find the permutation distribution of the minimum P-value over the remaining models. The multiplicity adjusted P-value corresponding to the second smallest observed P-value (p_9^{obs} in our example) is then the proportion of permutations for which the minimum P-value over the $m-1$ remaining models is smaller than the observed one. Using these step-down adjustments based on the minimum P-value, Table II displays all adjusted P-values for the IBS data and shows that PoC can be established (e.g., at a 2.5% FWER) with almost all of the candidate models, even when adjusting for simultaneous inference. The use of these adjusted P-values guarantees that the familywise type I error rate for one or more incorrect PoC decisions when the drug is actually ineffective is controlled at the desired level (e.g., at 2.5%). This control is in the strong sense, i.e., no matter which or how many of the m PoC null hypothesis are actually true. As mentioned before, by taking advantage of the correlation among the T_s statistics, the adjusted P-values are considerably smaller than ones based on the universally applicable Bonferroni or the more powerful Bonferroni-

Holm ([14]) correction, which would multiply the smallest permutation (or asymptotic, see below) P-value by 10, the second smallest one by 9, etc. Finally, based on the Chi-square approximation for the difference in deviance, unadjusted asymptotic P-values displayed in Table II are given by $1/2 + 1/2\,\Pr\{\chi^2_{df_s} \leq -(T_s + 2df_s)\}$ if $T_s + 2df_s \leq 0$ (i.e., under an estimated negative dose effect) and by $1/2\,\Pr\{\chi^2_{df_s} \geq T_s + 2df_s\}$ if $T_s + 2df_s > 0$ (i.e., under an estimated positive dose effect), where χ^2_q is a Chi-square random variable with q degrees of freedom. For large sample sizes in each arm, as in our example, these asymptotic P-values are virtually identical to the unadjusted permutation P-values p_s^{obs}. A second R function available at the author's website provides all output displayed in Table II for a given candidate and data set, and the Table can be reproduced with the R-code in the Appendix.

We would like to stress that the above approach controls the familywise error rate in this multiple testing scenario and is in contrast to the common habit of "model fishing", whereby several models are fitted (often a posteriori, after looking at the data) to account for the uncertainty on which model the PoC decision should be based. Often, one model is selected based on the most significant difference in deviance, largest Pearson's goodness of fit statistic or minimum AIC, without any regard to multiplicity issues. It is clear that these practices ignore uncertainty in the decision process and inflate actual type I error rates of hypotheses tests (such as PoC, see our simulation results in Section 3) or fail to meet nominal coverage rates for confidence intervals (such as ones for target dose estimates) as they are carried out conditional on the selected model ([7, 15]). In this sense and in a regulatory environment, the candidate set $\mathcal{M}$ must be specified a priori, before the data gathering stage, in a study protocol, using information up to that point. To protect against model misspecification, models that yield different shapes can be included. Section 5 includes further discussion on the sensitivity of results to the choice of the candidate set.

3. A simulation study with comparisons to powerful trend and contrast tests

In order to evaluate type I error rates and the power of establishing PoC with the suggested framework, and to compare its performance against popular contrast tests, we ran several simulation studies. We used the same (unequally spaced) dose levels as in the IBS study but now with balanced sample sizes of $n_i = 25$ or 50 subjects per arm (any larger sample size gave similar results). Table III shows type I error rates using 5000 simulations from the no-effect model $M_0 : \pi(d_i) = 0.3$, when the goal is to control the FWER at $\alpha = 5\%$ or 2.5%, respectively, and with the candidate set consisting of the models in Table I. As expected with our exact procedure, the actual type I error rate of erroneously declaring PoC is well controlled at the prescribed level. On the contrary, type I error rates of nominal 5% or 2.5% PoC tests that ignore the multiplicity in selecting a model for the PoC decision are inflated to about 11% and 6%, respectively, leading to many more false positive PoC decisions than desired. We demonstrate this in our simulations in two ways, by evaluating the procedure that simply fits all candidate models (which is often done implicitly in practice, when no candidate set is specified) and then bases the decision of PoC on the model that yields the smallest asymptotic P-value. Alternatively, we evaluate the procedure where the PoC decision is based on the candidate model with the minimum AIC. In both cases, we use as test statistic the difference in deviance (i.e., likelihood ratio test) between the no-effect model M_0 and the selected model, given the model shows a positive dose effect (otherwise, we declare no PoC). Both procedures lead to very similar inflations of type I error rates (because the model with the smallest P-value for the difference of deviance is also often the one with the minimum AIC) and hence Table III shows results only for the latter one. This inflation is in line with current results for normal data obtained by Hu and Dong [15], who also explore the issue of drawing inference after dose-response model selection.

To evaluate power, we assumed a placebo effect of 30% and a maximal efficacy of 65%, consistent with the prior information in the IBS study. For models with a downturn, we assumed that the maximal efficacy occurs at doses of $8mg$ and $14mg$, respectively. Table III shows the power of establishing PoC under selected dose response shapes from Table I (i.e., when the true dose response shape is a member of the candidate set), based on 1000 simulations from each. As mentioned in the introduction, PoC is often established via trend tests or multiple comparison procedures. We compare type I error rates and power for our procedure to some of the most popular and powerful ones, such as the Cochran-Armitage (CA) trend test or Dunnett's multiple comparison procedure (see, e.g. [16]). We also include multiple contrast tests that take the maximum over a larger set of contrasts to cover a broader dose-response space. Well known examples are Williams' trend test [17, 18] which is particularly powerful to detect concave shapes and Hirotsu's [19] step contrasts (see also [20] and, for a related test, [21]) which are very sensitive for single-jump shapes. The maximum of Helmert, linear, and reverse Helmert contrast coefficients are designed for linear and quadratic shapes. Another powerful test is based on Marcus' [22] multiple contrast coefficients which include the Williams' contrasts, supplemented by their reflections and some additional contrast vectors. Table IV shows contrast coefficients for all these tests for a balanced 5 dose layout. Table V compares the performance of our procedure and the contrast tests under model misspecification, i.e., when none of the models in the candidate set corresponds to the true dose response profile from which the data were generated. Some of these shapes (plotted in Figure 2) can be approximated by models in the candidate set, while others severely deviate and represent extreme cases.

The CA test is one of the most powerful trend tests for monotone dose response shapes [23]. The power of our approach is close to the power of the CA test for all monotone dose response shapes included in the simulation (M_1, M_3, M_5, M_7 and all but the first three shapes in Figure 2), sometimes even outperforming it (e.g., under models M_5 and the Emax and sigmoid Emax model). This is no surprise, as the CA test is closely related to a score test (it equals the score test under a logit model) and score and likelihood ratio tests are asymptotically equivalent for the monotone shapes. As expected, for non-monotone shapes, the CA test performs poorly. Interestingly, Dunnett's test performs inferiorly for all monotone shapes, while the other contrast tests show power similar to our approach for some but not all shapes. Overall, across Tables III and V, our approach consistently shows high power *regardless* of the underlying dose-response shape and easily competes with every contrast test considered. Summing up, almost nothing is lost in terms of power when using our robust procedure instead of the optimal contrasts test that would correspond to the (unknown) shape.

4. Dose estimation

Once we decide on a best approximating model (or models) for the true dose-response shape, the focus shifts towards target dose estimation. Here, we consider estimating the minimum effective dose (MED), defined as the smallest dose that is both clinically relevant and statistically significant [1]. Following results in [2], we estimate the MED for model M_s by

$$\widehat{\mathrm{MED}}_s = \mathrm{argmin}_{d \in (d_1, d_k]} \{\hat{\pi}_s(d) > \hat{\pi}_s(d_1) + \Delta, \hat{\pi}_s^L(d) > \hat{\pi}_s(d_1)\},$$

where Δ is the clinically relevant effect and $\hat{\pi}_s^L(d)$ is the lower limit of a $100(1-\gamma/2)$ confidence interval for $\pi(d)$. Obtaining the MED estimate is straightforward (i.e., amounts to solving a polynomial) for models such as M_1 to M_4. For others, a basic line search starting at the lowest dose yields the MED. We say that the MED does not exist if it is outside the observed dose range. Other target doses, such as ones with clinical relevance specified in absolute terms and not relative to the placebo effect can be considered accordingly.

For the IBS data, model M_5 has the smallest multiplicity adjusted P-value and is chosen for target dose estimation. With a clinical relevant effect of $\Delta = 15\%$ specified in the planning stage of the trial between the sponsor and the regulatory agency (for guidelines regarding IBS, see [25]), $\gamma = 0.05$ and maximum likelihood estimates (standard errors in parentheses) equal to $\hat{\beta}_0 = 0.63$ (0.13), $\hat{\beta}_1 = -1.10$ (0.26), $\widehat{\text{MED}} = 1.3mg$. The left panel in Figure 3 illustrates the construction of the MED under this model. Various asymptotic formulae (e.g., [24]) exist for obtaining the standard error of a target dose estimate, however, they do not take model uncertainty into account and may not work well if the distribution of the target dose estimate is very skewed. A 95% bootstrap percentile confidence interval is straightforward to construct, either by fitting model M_5 repeatedly to bootstrap samples taken within each dose group (non-parametric version) or via repeatedly simulating from the fitted model (parametric version). For the IBS data, these intervals are similar at $[0.6mg, 6.9mg]$ and $[0.7mg, 6.4mg]$, respectively. Both intervals are constructed conditional on establishing PoC, by discarding simulated data sets for which the MED did not exist, either because clinical relevance or statistical significance could not be established. However, they do not represent uncertainty due to model selection by pretending model M_5 is the only one under consideration.

4.1. Combining dose estimates from models

An alternative estimate of the true MED can be obtained as a weighted average of the estimated MEDs (so they exist) from each model. By letting

$$\widehat{\text{wMED}} = \sum_{s:\widehat{\text{MED}}_s \le d_k} w_s \widehat{\text{MED}}_s \Big/ \sum_{s:\widehat{\text{MED}}_s \le d_k} w_s,$$

where $\widehat{\text{MED}}_s$ is the estimated MED for model M_s and w_s are suitable weights, model uncertainty is incorporated into the target dose estimate. Note that for two models M_s and $M_{s'}$ with equal number of parameters, $\exp(\frac{1}{2}[T_s - T_{s'}])$ is the likelihood ratio of the two models. For instance, the observed data are $\exp(\frac{1}{2}[16.35 - 14.25]) = 2.9$ times more likely under M_5 than under M_4 and 1.4 times more likely than under M_9. For the latter comparison, the likelihood ratio is penalized by $\exp(df_{s'} - df_s)$ for the excess parameters in one model over the other. These considerations motivate weights $w_s = \exp(T_s/2)$ for constructing $\widehat{\text{wMED}}$, which have also been suggested by [26] in a similar context. In this way, the ratio of weights w_s and $w_{s'}$ attached to the MED's from models M_s and $M_{s'}$ reflects their relative (penalized) likelihood distance. For a slightly different approach of finding weights, see [27].

Relative weights for the IBS data are displayed in Table II, which yield $\widehat{\text{wMED}} = 1.7mg$, slightly larger to the estimate solely based on M_5 as it incorporates information from other models with different slopes and curvature. The right panel in Figure 3 illustrates this point by showing the fit of models that received the largest weights. We obtain a 95% bootstrap confidence interval by repeating the entire procedure that led to the estimate of the wMED over many resamples. In this way, all sources of uncertainty associated with testing PoC and estimation of the wMED are incorporated. With 5,000 bootstrap resamples and discarding those for which PoC could not be established (which was 1% of bootstrap samples with the IBS data) the interval is conditional on establishing PoC and equals $[0.4mg, 7.9mg]$. This interval is somewhat wider than the one based on M_5 alone as it incorporates model uncertainty in the MED estimate.

4.2. Performance of MED estimators

To evaluate bias and variability of $\widehat{\text{MED}}$ and $\widehat{\text{wMED}}$, we included the dose estimation step in the simulation study from the previous section. For each simulated data set for which PoC could be established, we computed $\widehat{\text{MED}}$ from the fitted model with the smallest adjusted P-value (using $\Delta = 0.15$ and $\gamma = 0.05$) or combined the MED's from all models to get $\widehat{\text{wMED}}$. Boxplots in Figure 4 for selected dose response shapes show the results graphically. The true MED for a particular model is given by the gray vertical line, and the median of $\widehat{\text{MED}}$ and $\widehat{\text{wMED}}$ over 1000 simulations from that model by a full circle. For most shapes the median bias in both estimates gets relatively small with $n = 50$ observations per dose group. With regards to the precision of both estimators, it is not surprising that the Interquartile Range (IQR) is smallest for shapes that have steep slopes at the true MED, such as M_5, M_8, M_{10} or the Emax model. However, large variability remains in both estimators for moderately monotone increasing shapes (such as M_1, M_3 or the sigmoid Emax model), even under per-arm sample sizes of $n_i = 100$. We see that for these shapes target dose estimation is considerably harder than merely establishing PoC, and sample sizes that are sufficient for establishing PoC may not be adequate for precise target dose estimation.

As the distribution of the estimates can be very skewed, Table VI displays robust summary statistics to evaluate the performance of $\widehat{\text{MED}}$ and $\widehat{\text{wMED}}$ when $n_i = 100$ (i.e., when establishing PoC is virtually certain), such as the median bias, the IQR and the square root of the median squared error

($\sqrt{\text{MSE}}$). Although $\widehat{\text{wMED}}$ is slightly more biased, its $\sqrt{\text{MSE}}$ is comparable to the one of $\widehat{\text{MED}}$ for the best fitting model. Table VI also presents the percentage that the minimum adjusted P-value occurs for the candidate model from which the data were generated (if this shape is in the candidate set) and the average relative weight associated with that model. For example, the first line in the last two columns reads as follows: For 35% of the 1000 simulations under shape M_1, the minimum adjusted P-value actually occurred for candidate model M_1. Other than M_1, the highest percentage (23%) occurred under model M_6. (M_1 and M_6 propose a similar shape, hence this split). The average relative weight assigned to the estimated MED from model M_1 equals 20%, while the next largest relative weight, 19%, was assigned to the MED estimate from model M_6. Models that are well identified and accumulate the largest weights are again the ones with steep slope at the MED and a distinct curvature. For instance, M_8 is correctly identified 86% of the times, and its MED estimate received an average weight of 69%. For others, such as M_3, competing models (in this case M_2 and M_4) result in nearly identical fits, and hence the wMED weights are distributed evenly across these models.

5. Sensitivity of results to models included in candidate set

An important aspect of the proposed methodology is the selection of models to be included in the candidate set. For the IBS data, the clinical team had little prior knowledge of the dose-response based on biological considerations and the goal was simply to cover a variety of plausible shapes. Adding candidate models that are similar to ones already in the set $\mathcal{M}$ will have almost no effect on the critical value and hence on the power of establishing PoC. For example, suppose we add 8 GLMs with linear predictor form equal to $M_1 - M_5$ and $M_8 - M_{10}$ but using a complementary log-log link and another 5 logit models with fractional polynomial powers not included in Table I, such as $\beta_0 + \beta_1 d^{-3/2}$ or $\beta_0 + \beta_1 \log(d) + \beta_3 d^{-1/2}$, for a new total of 23 candidate models. These 13 new models add little to the shape space spanned by the original candidate set. The critical value $c = 0.0073$ for this much larger candidate set is just slightly lower than $c = 0.0082$ from the original analysis. (Note that $c_{\text{Bonf}} = 0.025/23 = 0.0011$ or $c_{\text{Beta}} = 0.0011$ are considerably smaller.) Hence, the cost for multiplicity adjustments due to adding these 13 models are almost negligible, due to their high correlation with models already in $\mathcal{M}$. In particular, the model $\beta_0 + \beta_1 d^{-3/2}$ with the logit link fits slightly better than M_5 with the logit or asymmetric cloglog link (although adjusted P-values are virtually identical for the three). The MEDs $0.8mg$, $1.3mg$, and $1.5mg$ under these three models receive the largest relative weights of 16%, 14% and 11% in the computation of the wMED, which changes slightly from $1.5mg$ to $1.4mg$. The corresponding bootstrap confidence interval that takes into consideration all sources of uncertainty is almost unchanged at $[0.5mg, 7.2mg]$. On the other hand, leaving out models that are not well represented by those already in $\mathcal{M}$ has an impact on the power of establishing PoC. For example, the critical value with the initial candidate set that does not include the shapes M_8 and M_{10} that allow for a considerable downturn increases to $c = 0.0099$, resulting in a more powerful procedure for some alternative shapes except of course those that are well described by these two models.

Many popular dose-response models are non-linear in the parameters of the linear predictor, such as the compartment (M_{11}), Emax (M_{12}), sigmoid Emax (M_{13}), 4-parameter logistic (M_{14}) and Weibull (M_{15}) models, see Table VII. While the first two shapes can be approximated by fractional polynomials, the latter three allow for modeling a monotone but flexible S-shaped dose-response curve over the efficacy range of the drug, at the cost of an increase in the number of parameters. In fact, with only 5 dose levels, these models might not be appropriate and parameter estimation can be poor and instable. As a consequence, these models are often difficult to fit even for the observed data, but especially to permuted data where the natural structure is broken. For the IBS data, standard errors for the β_3 parameter that controls the slope of the S-shaped curves are huge in all three 4-parameter models (s.e.$(\hat{\beta}_3)/\hat{\beta}_3 > 6$), and ML-estimates are unstable and sensitive to starting values, indicating that some parameters are not well identified for the given data and spacing of doses. Moreover, all four models did not converge for a significant percentage of the randomly generated 50,000 permutations, mostly because automatic and constrained maximization (with constraints $\beta_2, \beta_3 > 0$) can be tricky on the permuted data sets. Table VIII gives details of the permutation approach when the original candidate set of Table I is amended with $M_{11} - M_{13}$. If, for a given permutation, a model M_s does not converge, we set $T_s = -\infty$ as PoC cannot be established with that model. We observe that although key characteristics such as the critical value or the estimate of the wMED do not change, there might be considerable bias in the adjusted P-values based on the permutation approach. For other data sets, we observed an even higher percentage of non-converging fits, or fits for which a positive definite estimate of the variance-covariance matrix of the estimated parameters could not be obtained. In such situations, we recommend not to perform the permutation analysis but instead adjust the asymptotic P-values by some suitable method that does not require permutations. For instance, the column labeled "B-H P-value" in Table VIII applies Bonferroni-Holm multiplicity adjustments to the asymptotic P-values. However, in our simulations to construct Tables III and V, we did note some loss of power in the PoC test (e.g., an average of 10% over the models in Table III at $n = 50$) when using the B-H multiplicity adjusted P-values over the ones based on the permutation approach. This is because the step-down multiplicity adjustment of the minimum P-value incorporates the correlation information in $(T_1, \ldots, T_m)$, leading to a larger critical value c.

6. Conclusions and Extensions

In this article we presented a framework that combines formal hypothesis testing for PoC under a strong error control with flexible modeling of the dose response relationship. The simulation studies showed that for moderate sample sizes (e.g., 25 and more subjects per arm) the suggested framework is as powerful (even under candidate model misspecification) as methods based on popular trend or contrast tests that are only powerful for a particular subsets of dose-response shapes. By design, our framework allows to power the trial for shapes that the clinical team thinks are reasonable, instead of taking an "off the shelf" method such as the CA-test or Dunnett's procedure that are only powerful for a much narrower alternative. Unequally spaced doses and/or unbalanced sample sizes are naturally handled, and adjustments due to covariates such as baseline measures or subject-specific characteristics can be implemented at the modeling stage. Contrary to contrast tests, we can obtain and assess the variability in the target dose estimate, which, somehow surprisingly, is shown to be quite large under some monotone shapes, even for large sample sizes.

One restriction of the permutation approach is that it is likely to fail when non-linear (on the link scale) candidate models are desired. For such cases, an alternative is to adjust the asymptotic P-values of a statistic such as T_s with a suitable procedure (such as Bonferroni-Holm). However, this can lead to some loss of power in the PoC test. Convergence of non-linear models can be checked a priori via a simulation study similar to the one in Section 3.

The generality of our approach makes extension to continuous or count and multivariate or repeated responses (such as in cross-over studies) possible when the fitting process is not too complex. If non-likelihood based methods are used in fitting e.g. repeated categorical responses such as in a GEE approach, an attractive alternative to the signed penalized deviance difference statistic T_s could be a penalized generalized score statistic [28]. Research is currently under way to implement our methods for establishing PoC and target dose estimation with multivariate binary responses, such as occur with two primary endpoints or when considering an efficacy and safety endpoint jointly. In the bivariate case, one can consider candidate models that describe different shapes for the two margins and a model for their association (e.g., Sect. 6.5 of [29]). For instance, in the IBS study mentioned in Section 2 one often differentiates between two primary endpoints, relief of abdominal pain and relief of overall symptoms excluding abdominal pain. The candidate set then specifies combinations of the models in Table I for these two margins and a model for their association. Similar to T_s and for a given definition of a positive dose effect, the permutation distribution of

$$T(M_r, M_s) = (-1)^{I(\text{neg. dose effect})}\left\{-2\left[l(M_0, M_0) - l(M_r, M_s)\right]\right\} - \text{penalty},$$

where $l(M_r, M_s)$ is the maximized multinomial log-likelihood under marginal models M_r and M_s and a particular model for the association can be used to test for PoC, incorporating model uncertainty but still controlling the FWER. Target dose estimation can then proceed from either the fitted margins (which is appropriate if one margin describes safety, which needs to be controlled regardless of efficacy), the model for the joint success probability or some other utility function. Other extensions of the methodology are towards semi-parametric models that could yet allow more flexibility in modeling the dose-response, at the cost of interpretability in discussions with the clinical team that are crucial in our first step of building the candidate set. Putting aside a possible power advantage or the difficulty of selecting a smoothing parameter with a small number of doses (knots), a fully parametric approach allows for straightforward simulation to find the necessary sample size for a given power under various scenarios of an assumed placebo or maximum dose effect, as well as under candidate model misspecification (for normal responses, see Pinheiro et al. [30]). Finally, the simulation study in Section 3 also serves to illustrate how the methodology can be used in the design stage to explore the power of establishing PoC under various sample size allocations, sets of candidate models or number and spacing of doses.

APPENDIX

The following code and additional information on the R implementation of the proposed method are available at `www.williams.edu/∼bklingen/PoC/Rcode`. Simply copy and paste the R code to reproduce all results in this paper or edit the code to create new candidate models.

Defining and plotting candidate models:

```
source("http://lanfiles.williams.edu/∼bklingen/PoC/plotModels.R")
source("http://lanfiles.williams.edu/∼bklingen/PoC/binomial1.R") #allows for identity link
source("http://lanfiles.williams.edu/∼bklingen/PoC/makelink.R") #allows for loglog-link
dose <- c(0,1,4,12,24)
M1 <- list(family=binomial())
M2 <- list(model=pow(dose,0.5), family=binomial())
M3 <- list(model=pow(dose,0), family=binomial())
M4 <- list(model=pow(dose,-0.5), family=binomial())
M5 <- list(model=pow(dose,-1), family=binomial())
M6 <- list(family=binomial(link=log))
M7 <- list(model=expo(dose,2,scale=max(dose)), family=binomial(link="identity"))
M8 <- list(model=pow(dose,c(1,2),dmax=14), family=binomial())
M9 <- list(model=pow(dose,c(0,-1),dmax=8), family=binomial())
M10 <- list(model=pow(dose,c(0,1),dmax=8), family=binomial())
models <- list(M1,M2,M3,M4,M5,M6,M7,M8,M9,M10)
plotModels(dose, models, low=0.3, high=0.65) #plots candidate dose-resp. models
```

Inference with candidate models (adj. P-values, MED, ...):

```
source("http://lanfiles.williams.edu/∼bklingen/PoC/perm_minP_GLM.R")
source("http://lanfiles.williams.edu/∼bklingen/PoC/nonlin_dr.R") # only needed if fitting
non-linear models
y <- c(38,52,67,59,58) # successes
n <- c(100,102,98,99,94) # sample sizes
resp <- cbind(y,n-y)
dr <- permT(dose,resp,models,perms=5000,trace=500,clinRel=0.15,alpha=0.025)
summary(dr)
hist(dr) # plots the histogram of the minimum P-value
plot(dr) # plots the fitted model with the smallest adj. P-value, see Figure 4
plot(dr,which.models=c("M5","M9","M10","M4") # similar to Figure 4
```

ACKNOWLEDGEMENTS

I would like to thank Frank Bretz and José Pinheiro for helpful discussions, Cornelia Dunger-Baldauf for making available the data and background/regulatory information for IBS and two referees for a careful reading and stimulating comments that helped improve the quality of this paper.

REFERENCES

1. Ruberg SJ. Dose-response studies. I. Some design considerations. *Journal of Biopharmaceutical Statistics* 1995; **5**:15–42.
2. Bretz F, Pinheiro J, Branson M. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* 2005; **61**:738–748.
3. Westfall P, Young S. *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment.* John Wiley & Sons: New York, 1993.
4. Hoeting J, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. *Statistical Science* 1999; **14**:382–417.
5. Clyde MA, George EI. Model uncertainty. *Statistical Science* 2004; **19**:81–94.
6. Morales KH, Ibrahim JG, Chen CJ, Ryan L. Bayesian model averaging with applications to benchmark dose estimation for arsenic in drinking water. *Journal of the American Statistical Association* 2006; **101**:9–17.
7. Hjort NL, Claeskens G. Frequentist model average estimators. *Journal of the American Statistical Association* 2003; **98**:879–899.
8. Burnham K, Anderson DR. *Model Selection and Multimodel Inference: A practical Information-theoretic Approach* (2nd edn). Springer: New York, 2002.
9. Guidline for Industry. Dose-response information to support drug registration *Federal Register* 1994 **59**: 55972–55976
10. Roystone P, Altman DG. Regression using fractional polynomials of continues covariates: parsimonious parametric modelling. *Applied Statistics* 1994; **43**:429–467.
11. Faes C, Geys H, Aerts M, Molenberghs G. On the use of fractional polynomial predictors for quantitative risk assessment in developmental toxicity studies. *Statistical Modelling* 2003; **3**:109-126.
12. Faes C, Aerts M, Geys H, Molenberghs G. Model Averaging Using Fractional Polynomials to Estimate a Safe Level of Exposure. *Risk Analysis* 2007; **27**:111–123
13. Lindsey J, Jones B. Choosing among generalized linear models applied to medical data. *Statistics in Medicine* 1998; **17**:59–68.
14. Hochberg Y, Tamhane, AJ. *Multiple Comparison Procedures.* Wiley: New York, 1987.
15. Hu C, Dong Y. Estimating the predictive quality of dose-response after model selection. *Statistics in Medicine* 2007; **26**:3114–3139
16. Piegorsch W. Multiple comparisons for analyzing dichotomous responses. *Biometrics* 1991; **47**:45–52.
17. Williams D. Tests for differences between several small proportions. *Applied Statistics* 1988; **37**:421-434.
18. Bretz F. An extension of the Williams trend test to general unbalanced linear models. *Computational*

Statistics & Data Analysis 2006; **50**:1735-1748.

19. Hirotsu C. Isotonic inference with particular interest in application to clinical trials. In *Industrial Statistics*, Kitsos C P, Edler L (eds). Physica-Verlag: Heidelberg, 1997.
20. Bretz F, Hothorn, L. Detecting dose-response using contrasts: Asymptotic power and sample size determination for binomial data. *Statistics in Medicine* 2002; **21**:3325–3335.
21. Liu Q. An order-directed score test for trend in ordered $2 \times k$ tables. *Biometrics* 1998; **54**:1147-1154.
22. Marcus R. The power of some tests of the equality of normal means against an ordered alternative. *Biometrika* 1976; **63**:177–183.
23. Tarone R, Gart J. On the robustness of combined tests for trends in proportions. *Journal of the American Statistical Association* 1980; **75**:110–116.
24. Morgan BJT. *Analysis of Quantal Response Data*. Chapman & Hall: London, 1992.
25. Corazziari E, Bytzer P, Delvaux M, Holtmann G, Malagelada JR, Morris J, Muller-Lissner S, Spiller RC, Tack J, Whorwell PJ. Consensus report: clinical trial guidelines for pharmacological treatment of irritable bowel syndrome. *Aliment Pharmacol Ther* 2003; **18**:569-580.
26. Buckland S, Burnham K, Augustin N. Model selection: An integral part of inference. *Risk Analysis* 2005; **25**:1147–1159.
27. Moon H, Kim HJ, Chen J, Kodell R. Model averaging using the Kullback Information Criterion in estimating effective doses for microbial infection and illness. *Biometrics* 1997; **53**:603–618.
28. Boos D. On generalized score tests. *The American Statistician* 1992; **46**:327–333.
29. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1989.
30. Pinheiro J, Bornkamp B, Bretz F. Design and analysis of dose finding studies combining multiple comparisons and modeling procedures. *Journal of Biopharmaceutical Statistics* 2006; **16**:639–656.

Table I. Candidate dose-response models for the efficacy of a compound against IBS.

$\mathcal{M}$	Link	Predictor	# parms
M_1:	logit	$\beta_0 + \beta_1 d$	2
M_2:	logit	$\beta_0 + \beta_1 \sqrt{d}$	2
M_3:	logit	$\beta_0 + \beta_1 \log(d+1)$	2
M_4:	logit	$\beta_0 + \beta_1 \sqrt{d+1}$	2
M_5:	logit	$\beta_0 + \beta_1/(d+1)$	2
M_6:	log	$\beta_0 + \beta_1 d$	2
M_7:	identity	$\beta_0 + \beta_1 \exp(\exp(d/\max(d)))$	2
M_8:	logit	$\beta_0 + \beta_1 d + \beta_2 d^2$	3
M_9:	logit	$\beta_0 + \beta_1 \log(d+1) + \beta_2/(d+1)$	3
M_{10}:	logit	$\beta_0 + \beta_1 \log(d+1) + \beta_2 d$	3

Table II. AIC, T_s and corresponding asymptotic, raw and step-down multiplicity adjusted permutation P-values for the PoC test with each candidate model. Adjusted P-values adjust the raw P-values for the multiple tests of PoC, one for each of the 10 candidate models. Further, MED estimates and their associated relative weight in the computation of the wMED.

			asympt.	raw	adj.	$\widehat{\text{MED}}$	$w_s/\sum w_s$
$\mathcal{M}$	AIC	T_s	P-value	P-value	P-value	(mg)	(%)
M_1:	45.4	3.68	0.0086	0.0088	0.0118	NA	0
M_2:	40.3	8.76	0.0005	0.0005	0.0011	12.3	1
M_3:	38.5	10.53	0.0002	0.0002	0.0006	8.0	2
M_4:	34.8	14.25	<0.0001	0.0001	0.0003	2.8	14
M_5:	32.7	16.35	<0.0001	<0.0001	0.0001	1.3	40
M_6:	45.8	3.25	0.0110	0.0113	0.0145	NA	0
M_7:	48.1	0.90	0.0442	0.0454	0.0454	NA	0
M_8:	42.0	7.01	0.0020	0.0021	0.0041	6.8	0
M_9:	33.4	15.63	<0.0001	<0.0001	0.0001	0.7	28
M_{10}:	34.9	14.20	0.0001	0.0001	0.0002	1.7	14
			critic. value $c = 0.0083$			$\widehat{\text{wMED}} = 1.7mg$	

Table III. Power (in %) of establishing PoC at a 2.5% FWER under selected dose-response profiles from Figure 1 for our approach based on the adjusted minimum P-value ($\min P$) vs. the minimum AIC approach and popular contrast tests. Based on 1000 simulations from each profile (simulation margin of error at most 3%), using at most $B = 3000$ permutations for each. To evaluate the type I error rate when controlling the FWER at 5% (third column) or 2.5% (fourth column), we used 5000 simulations from M_0 (simulation margin of error at most 0.8%), and at most $B = 5000$ permutations for each. The last two columns refer to median and average power over these selected models.

		Dose-response profiles								Med.	Avg.
n	Method	M_0^*	M_0^{**}	M_1	M_3	M_5	M_7	M_8	M_{10}	Power	Power
25	min AIC***	12.5	7.0	87	90	88	83	80	71	85	83
	$\min P$	5.0	2.6	77	76	77	77	64	64	77	73
	CA	5.4	3.0	84	79	59	82	35	10	69	58
	Dunnett	5.5	2.6	55	65	73	54	57	66	61	62
	Williams	5.0	2.5	67	76	82	65	52	56	66	66
	Hirotsu	5.2	2.6	77	76	76	77	59	53	76	70
	Helmert	4.8	2.6	77	75	77	80	46	47	76	67
	Marcus	5.1	2.7	78	78	78	79	58	52	78	71
50	min AIC***	11.1	6.1	98	99	99	99	96	97	99	98
	$\min P$	4.8	2.3	97	98	97	99	89	92	97	95
	CA	5.1	2.7	98	98	86	99	57	12	92	75
	Dunnett	5.4	2.9	90	90	96	90	86	90	90	90
	Williams	5.3	2.5	93	97	98	94	79	87	94	91
	Hirotsu	5.2	2.7	97	98	97	99	87	85	97	94
	Helmert	5.5	2.6	97	99	96	99	77	78	97	91
	Marcus	5.0	2.7	97	98	96	99	87	84	97	94

*Controlling FWER at 5% (i.e., $\alpha = 0.05$)

**Controlling FWER at 2.5% (i.e., $\alpha = 0.025$)

***The candidate model with the minimum AIC is selected without regard to multiplicity adjustments

Table IV. Contrast vectors $\boldsymbol{c}^{(l)}$ for maximum contrast tests with 5 dose levels and test statistic $T = \max_l T^{(l)}$, where $T^{(l)} = \sum_i c_i^{(l)} p_i \Big/ \sqrt{p_0(1-p_0)\sum_i [c_i^{(l)}]^2/n_i}$ and p_0 the pooled sample proportion. The multiple contrast vectors called "Helmert" are actually the linear, Helmert and reverse Helmert contrasts.

Contrast	Contrast Vectors
CA:	$(0, 1, 4, 12, 24)$
Dunnett:	$(-1, 1, 0, 0, 0),\ (-1, 0, 1, 0, 0),\ (-1, 0, 0, 1, 0),\ (-1, 0, 0, 0, 1)$
Williams:	$(-1, 0, 0, 0, 1),\ (-1, 0, 0, 1/2, 1/2),\ (-1, 0, 1/3, 1/3, 1/3),\ (-1, 1/4, 1/4, 1/4, 1/4)$
Hirotsu:	$(-1/4, -1/4, -1/4, -1/4, 1),\ (-1/3, -1/3, -1/3, 1/2, 1/2),$ $(-1/2, -1/2, 1/3, 1/3, 1/3),\ (-1, 1/4, 1/4, 1/4, 1/4)$
Helmert:	$(-4, -2, 0, 2, 4),\ (-1, -1, -1, -1, 4),\ (-4, 1, 1, 1, 1)$
Marcus:	$(-1, 0, 0, 0, 1),\ (-1, 0, 0, 1/2, 1/2),\ (-1, 0, 1/3, 1/3, 1/3),\ (-1, 1/4, 1/4, 1/4, 1/4),$ $(-1/4, -1/4, -1/4, -1/4, 1),\ (-1/3, -1/3, -1/3, 1/2, 1/2),\ (-1/3, -1/3, -1/3, 0, 1),$ $(-1/2, -1/2, 1/3, 1/3, 1/3),\ (-1/2, -1/2, 0, 1/2, 1/2),\ (-1/2, -1/2, 0, 0, 1)$

Table V. Power (in %) of establishing PoC under dose-response model *misspecification*, controlling the FWER at 2.5% (except under the row labeled "min AIC").

		Dose-response profiles										
		Peak			Step					Sig.	Med.	Avg.
n	Method	1	2	Plateau	1	2	3	Emax	Logist.	Emax	Power	Power
25	min AIC*	74	70	95	85	96	96	88	93	95	93	88
	min P	55	50	87	81	92	90	76	84	92	84	79
	CA	10	18	1	84	95	77	58	89	82	77	57
	Dunnett	54	62	75	55	70	81	73	64	72	70	67
	Williams	22	50	44	65	81	84	81	75	81	75	65
	Hirotsu	37	57	60	84	93	95	73	85	89	84	75
	Helmert	13	41	23	86	89	88	77	84	85	84	65
	Marcus	36	56	58	83	93	94	77	86	89	83	75
50	min AIC*	97	91	99	97	100	100	99	100	100	99	98
	min P	81	87	99	99	99	100	98	99	99	99	96
	CA	16	31	1	99	100	97	86	99	97	97	70
	Dunnett	90	87	95	90	94	97	95	91	95	94	92
	Williams	43	82	75	92	97	98	98	96	98	96	86
	Hirotsu	65	89	91	99	100	100	97	99	100	99	93
	Helmert	26	72	48	99	99	99	98	99	99	99	82
	Marcus	64	88	90	99	100	100	97	99	100	99	93

* Does not control FWER at $\alpha = 2.5\%$

Table VI. Median bias, IQR and square root of the median squared error ($\sqrt{\text{MSE}}$) for $\widehat{\text{MED}}$ and $\widehat{\text{wMED}}$ of the model with the smallest adjusted P-value under selected shapes of Tables I and VIII, when $n_i = 100$. The last two columns refer to the percentage of correctly identifying the model in the candidate set and its average weight in estimating $\widehat{\text{wMED}}$. For an explanation of these values see Section 4.2.

Sim.	True	$\widehat{\text{MED}}$ (in mg)			$\widehat{\text{wMED}}$ (in mg)			$\max T_s$	Avg. w_s
Model	MED	bias	IQR	$\sqrt{\text{MSE}}$	bias	IQR	$\sqrt{\text{MSE}}$	(in %)	(in %)
M_1:	10.6	0.4	4.4	2.3	0.2	4.6	2.3	35, 23(M_6)	20, 19(M_6)
M_3:	3.1	0.4	2.5	1.3	0.8	3.6	1.4	31, 27(M_2)	20, 18(M_2)
M_5:	0.7	0.2	0.6	0.2	0.5	1.0	0.5	56, 21(M_4)	34, 20(M_9)
M_7:	17.6	-0.3	3.1	1.3	-2.0	3.5	2.2	61, 17(M_6)	34, 19(M_8)
M_8:	3.5	-0.1	1.1	0.6	-0.3	1.1	0.7	86, 10(M_{10})	69, 15(M_{10})
M_{10}:	1.0	<0.1	0.4	0.2	0.1	0.8	0.3	78, 11(M_9)	66, 15(M_9)
M_{12}:	0.7	0.2	0.6	0.3	0.6	1.1	0.6		
M_{13}:	5.1	-0.8	2.6	1.6	-0.4	3.0	1.6		

Table VII. Some non-linear dose-response models.

$\mathcal{M}$	Link	Predictor	# parms
M_{11}:	logit	$\beta_0 + \beta_1 d \exp(-d/\beta_2), \beta_2 > 0$	3
M_{12}:	logit	$\beta_0 + \beta_1 \log(d+1)/[\beta_2 + \log(d+1)], \beta_2 > 0$	3
M_{13}:	logit	$\beta_0 + \beta_1 [\log(d+1)]^{\beta_3}/(\beta_2^{\beta_3} + [\log(d+1)]^{\beta_3}), \beta_2, \beta_3 > 0$	4
M_{14}:	logit	$\beta_0 + \beta_1/(1 + \exp\{[\beta_2 - \log(d+1)]/\beta_3\}, \beta_2, \beta_3 > 0$	4
M_{15}:	logit	$\beta_0 + \beta_1 \exp\{-\exp[\beta_3(\log(d+1) - \beta_2)]\}, \beta_2, \beta_3 > 0$	4

Table VIII. Analysis based on extended candidate set including non-liner models $M_{11} - M_{13}$. Same columns as in Table II, with the addition of the percentage of convergent fits over the 50,000 permutations ("Convg.") and the multiplicity adjustment of the asymptotic P-values based on the Bonferroni-Holm procedure ("B-H P-value") that does not require any permutation.

			asympt.	raw	adj.	$\widehat{\text{MED}}$	$w_s/\sum w_s$	Convg.	B-H
$\mathcal{M}$	AIC	T_s	P-value	P-value	P-value	(mg)	(%)	(%)	P-value
M_1:	45.4	3.68	0.0086	0.0093	0.0126	NA	0	100	0.0258
M_2:	40.3	8.76	0.0005	0.0007	0.0014	12.3	1	100	0.0026
M_3:	38.5	10.53	0.0002	0.0004	0.0007	8.0	2	100	0.0012
M_4:	34.8	14.25	<0.0001	0.0001	0.0002	2.8	10	100	0.0003
M_5:	32.7	16.35	<0.0001	<0.0001	0.0001	1.3	30	100	0.0001
M_6:	45.8	3.25	0.0110	0.0118	0.0126	NA	0	100	0.0220
M_7:	48.1	0.90	0.0442	0.0456	0.0456	NA	0	100	0.0442
M_8:	42.0	7.01	0.0020	0.0019	0.0036	6.8	0	100	0.0081
M_9:	33.4	15.63	<0.0001	0.0001	0.0002	0.7	21	100	0.0003
M_{10}:	34.9	14.20	0.0001	0.0001	0.0003	1.7	10	100	0.0005
M_{11}:	37.0	12.07	0.0002	0.0002	0.0004	2.4	3	87	0.0011
M_{12}:	34.7	14.32	<0.0001	<0.0001	0.0001	0.8	11	76	0.0005
M_{13}:	34.7	14.33	0.0001	<0.0001	0.0001	1.1	11	55	0.0006

critic. value $c = 0.0083$ $\widehat{\text{wMED}} = 1.5mg$

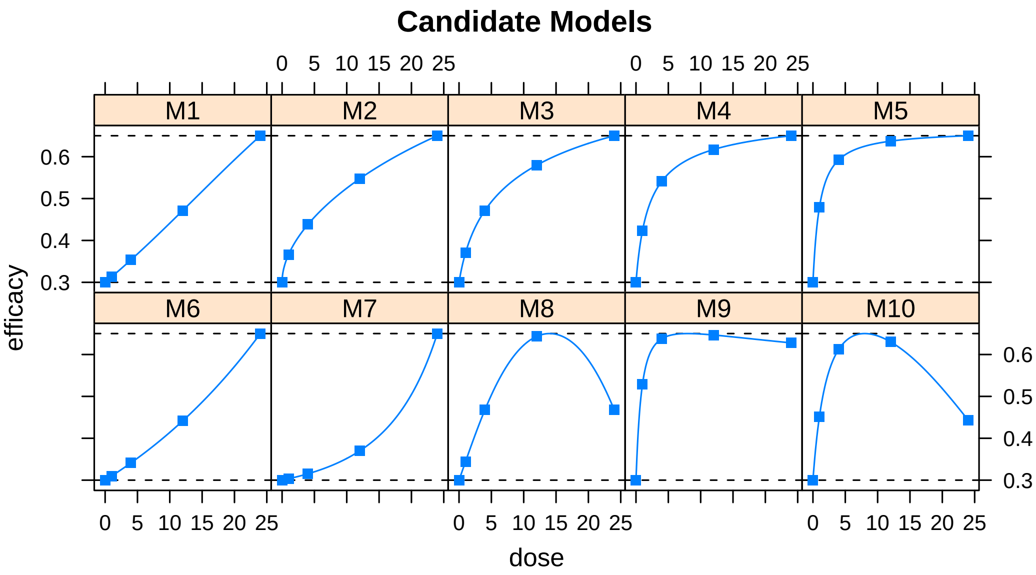

Figure 1. Dose-response profiles for models in Table I. Squares indicate success probabilities at dose levels $\boldsymbol{d} = (0, 1, 4, 12, 24)mg$. Initial guesses for parameters were chosen such that $\pi(0) = 0.30$ and $\pi(d_{\max}) = 0.65$, where $d_{\max}$ is the dose at which the maximum efficacy occurs. $d_{\max} = 24mg$ for all monotone shapes, but was selected equal to $14mg$ for $M8$ and $8mg$ under shapes M_9 and M_{10}.

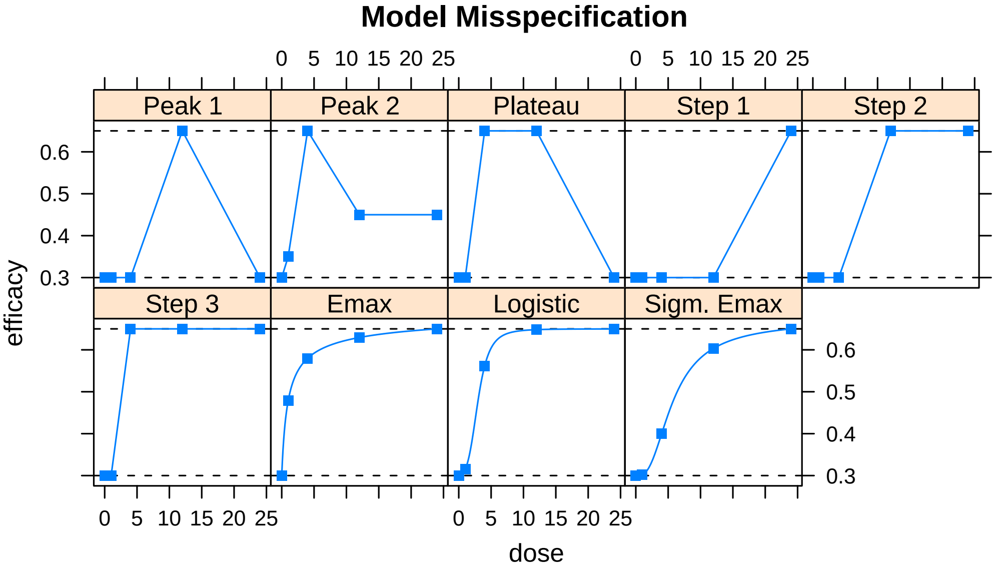

Figure 2. Dose-response profiles under model misspecification.

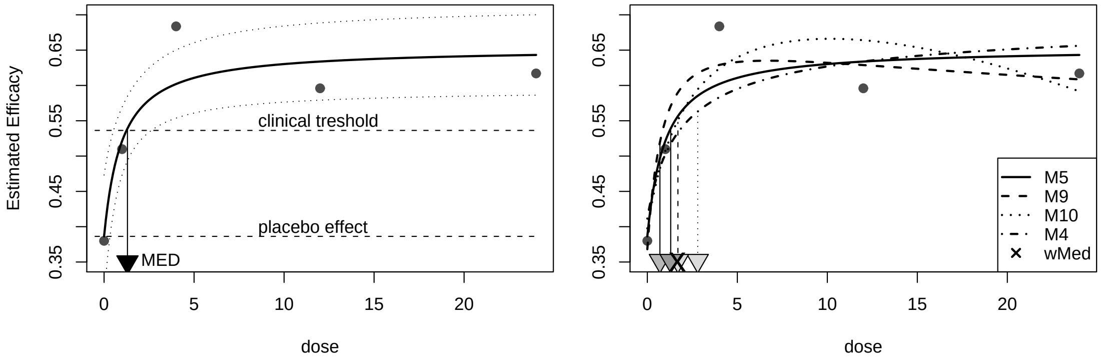

Figure 3. Left Panel: Fitted dose-response, pointwise 95% confidence intervals and MED for model M_5. Dashed lines indicate estimated placebo effect and clinical relevant effect (with Δ=15%). Full circles display observed proportions. Right Panel: Fitted dose-response for the four models receiving the largest weight and their MED's (triangles, using $\Delta = 15\%$). The intensity of gray of the triangles indicates the associated weight (the darker, the more weight, see Table II) in computing $\widehat{\text{wMED}}$, which is located at the "x" marker.

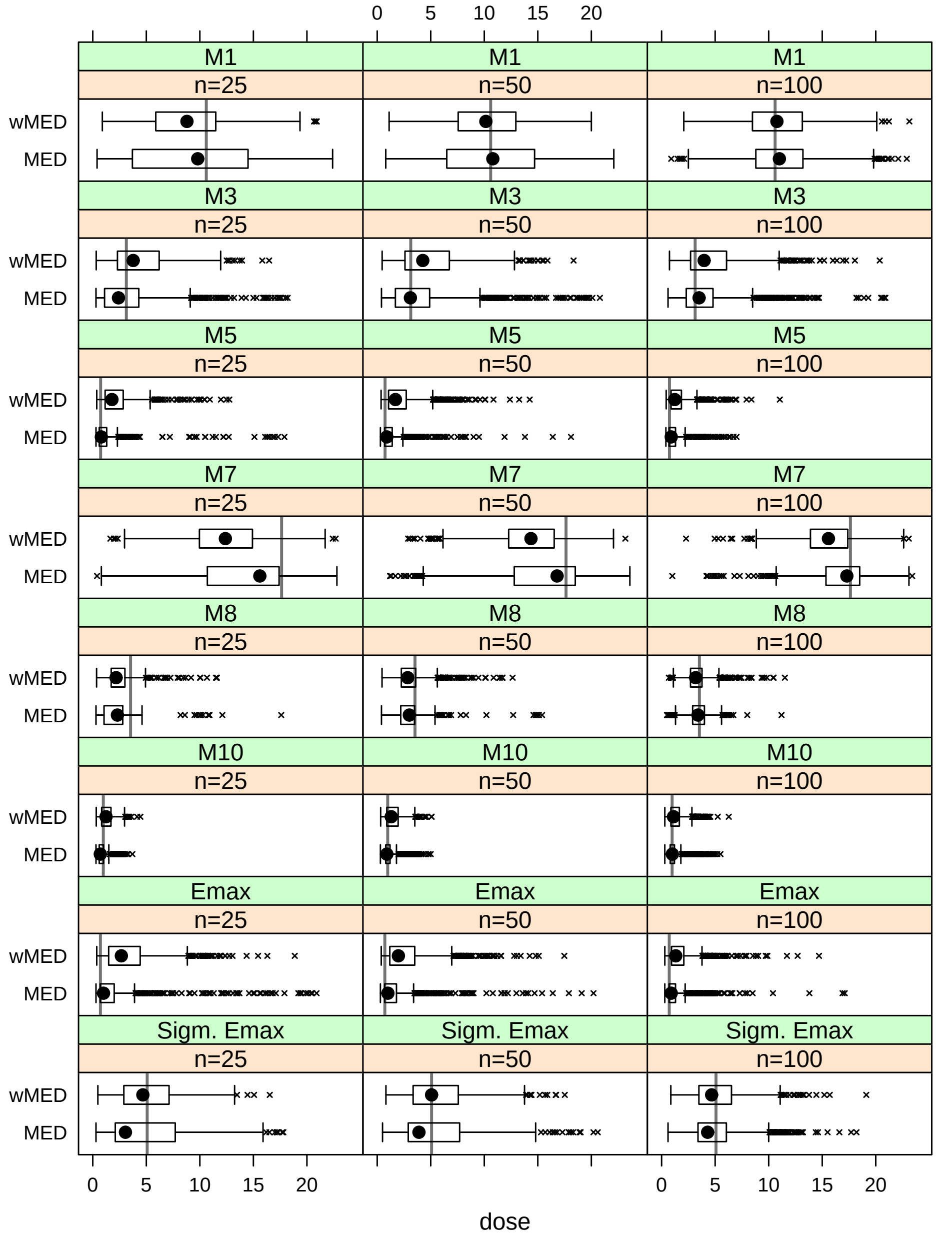

Figure 4. Dose estimation performance under selected dose-response shapes and balanced, per arm sample sizes of $n_i = 25, 50$ or 100. Vertical lines indicate the true MED for a particular shape with a clinical relevant effect of $\Delta = 0.15$.