Dose-response modeling with bivariate binary data under model uncertainty

Bernhard Klingenberg

1 Department of Mathematics and Statistics, Williams College, Williamstown, MA, 01267 and Institute of Statistics, Graz University of Technology, Steyrergasse 17/IV, 8010 Graz, Austria. E-mail: bklingen@williams.edu

Abstract: When modeling a dose-response for a drug based on bivariate binary data such as two co-primary efficacy endpoints in early stages of development, there is usually uncertainty about the form of the true underlying dose-response shape. Often, investigators fit several different models that are deemed plausible, but later fail to acknowledge this uncertainty in inference that is based on a single model selected via e.g. the minimum AIC criterion. This leads to an inflation in the error of the proof of activity decision and may also result in poor estimation of a target dose that is used in future trials. In this article we acknowledge model uncertainty by fitting several candidate models for a bivariate binary response and develop a principled approach to establish proof of activity and a target dose.

Keywords: Dose estimation; Multiplicity adjustment; Proof of Concept.

1 Introduction

Dose-response studies are important tools for investigating the existence, nature and extent of a dose effect on efficacy or safety outcomes in drug development, toxicology and related areas. The following four questions, usually in this order, are of prime interest: i.) Is there any evidence of a dose effect (i.e., Proof of Activity), 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)? Here we suggest to answer questions i) to iv) in a unified framework using statistical modelling. To meet the criticism that results based on modelling depend too much on assumptions on the underlying dose-response shape, we incorporate model uncertainty into our methods of establishing Proof of Activity (PoA) and estimating a target dose. We do this by considering several candidate models for the true dose-response which we simultaneously use for inference. Recently, Klingenberg (2008) has shown via simulation that for univariate binary dose-response data from simple parallel group designs, common methods of establishing PoA with a model based approach (i.e., basing the PoA decision on the P-value of a test using the model with the smallest AIC
among a set of candidate models) lead to inflated type I errors. This means
that the probability of carrying forward an ineffective drug into Phase III
is not well controlled. Motivated by the example below, in this article we
consider bivariate binary response data \((Y_{1j}, Y_{2j})\), where \(Y_{ij}\) is the binary
indicator of efficacy (or safety) for endpoint \(i\) \((i=1,2)\) for subjects ran-
domized to one out of \(k\) dose levels \(d_j\), \(j = 1\) (Placebo), \ldots \(k\). We develop
a principled approach to both, establishing PoA and estimating a target
dose (such as the minimum effective dose, MED) under model uncertainty,
controlling the type I error of an incorrect PoA decision.

2 Proof of Activity under model uncertainty

We assume independent multinomial distributions \(\text{Mult}(n_j; \pi_{00}(d_j), \pi_{01}(d_j),
\pi_{10}(d_j), \pi_{11}(d_j))\) for the counts in each of the \(j = 1, \ldots , k\) \(2 \times 2\) tables that
crossclassify the bivariate response at dose level \(d_j\). Here, \(n_j\) is the number
of subjects randomized to dose level \(d_j\) and \(\pi_{ab}(d_j) = P_{d_j}(Y_{1j} = a, Y_{2j} = b), a, b \in 0, 1\) is the probability of response \((a, b)\) at dose level \(d_j\). From now
on, we assume that \(Y_{1j}\) and \(Y_{2j}\) are both measuring efficacy, for instance
when there are two co-primary endpoints in a clinical trial (see example
below). One simple approach for establishing PoA and estimating the MED
would be to collapse the two efficacy endpoints into a single one, record-
ing an overall success if both endpoint show efficacy, and failure otherwise.
However, one drawback of this method is that it declares outcomes of type
\((0, 1)\) and \((1, 0)\) as failures, which might lead to a loss of power in es-
tablishing PoA and too high an estimate for the MED. In a multivariate
approach, we model separately the marginal success probabilities \(\pi_{1+}(d_j)\)
and \(\pi_{+1}(d_j)\) for each endpoint in terms of dose, taking into account the
dependence between them by modeling the log-odds (Palmgren, 1989).

Let \(M_s\) denote a specific constellation of two marginal models and one
model for the association in \((Y_{1j}, Y_{2j})\). To accommodate model uncertainty
in the PoA decision and subsequent target dose estimation, we start by
considering several plausible dose-response models \(M_s, s = 1, \ldots , m\) that
differ in the way they model the two margins and/or the association. Let
\(\mathcal{M} = \{M_1, \ldots , M_m\}\) be a candidate set spanned by \(m\) such models. Since
target doses such as the MED depend on the assumed shapes for each mar-
gin, 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 signed and penalized likelihood ratio statistic
\[T_s = \pm \left\{ -2 \left[ l_0 - l_s \right] \right\} - 2df_s,\]
where \(l_s\) is the maximized multinomial log-likelihood under candidate model
\(M_s\), and \(M_0\) and \(l_0\), respectively correspond to the no dose effect model
\[ \pi_{1+}(d_j) = \alpha_1, \pi_{+1}(d_j) = \alpha_2, OR_j = \frac{\pi_{0+}(d_j)\pi_{11}(d_j)}{\pi_{01}(d_j)\pi_{10}(d_j)} = \alpha_3 \] that assumes constant margins and odds ratios across all dose levels. Naturally, we are only interested in models that show a positive dose effect. Although straightforward for monotone marginal profiles, in general we define a dose effect as positive if \( \hat{\pi}(d_{\text{max}}) > \hat{\pi}(d_1) \), where \( d_{\text{max}} = \arg\max_d |\hat{\pi}(d) - \hat{\pi}(d_1)| \) is the dose at which the maximum absolute effect relative to placebo occurs. This condition must be met for both, the first (\( \hat{\pi} \equiv \hat{\pi}_{1+} \)) and second (\( \hat{\pi} \equiv \hat{\pi}_{+1} \)) margin, otherwise we declare the dose effect as negative. The purpose of the ± sign in \( T_s \) is then to give models with an estimated (but potentially significant) negative dose effect a small value of the test statistic, moving it to the lower tail. Finally, we penalize fitting more complex models by subtracting two times the difference in the number of parameters between \( M_s \) and \( M_0 \). Up to the ± sign, \( T_s \) is equivalent to the differences in the AICs of \( M_0 \) and \( M_s \), a statistic favored for model selection by Burnham and Anderson (2002).

Under the null hypothesis of no dose effect, bivariate responses \((Y_{1j}, Y_{2j})\) are exchangeable among the \( k \) dose levels. To evaluate the significance of \( T_s \) under simultaneous inference with all \( m \) candidate models, we build the permutation distribution of its maximum, \( \max_s T_s \), by fitting each \( M_s \) to a random sample of all possible assignments of the \((Y_{1j}, Y_{2j})\) responses to dose levels. (Although the asymptotic distribution of \( T_s \) is proportional to a Chi-square, the distribution of the maximum is not straightforward to derive due to correlation between the test statistics.) The permutation distribution yields raw and, using the full closed-testing methodology or the step-down approach in Westfall and Young (1993), multiplicity adjusted P-values for the PoA test with each candidate model. These adjusted P-values now appropriately account for the uncertainty in the dose-response shapes and control the familywise error rate of a wrong PoA decision under the family of candidate dose-response models. This control is in the strong sense, that is, under any combination of true and false null hypotheses, where the \( s \)-th null hypothesis concerns the testing of no dose effect under model \( M_s \).

### 3 MED estimation

After establishing PoA with at least one candidate model (smallest multiplicity adjusted P-value less than the chosen overall type I error rate), the procedure moves to estimating the MED. For a given model, the MED is the smallest dose that shows a clinical relevant and statistical significant improvement over placebo, i.e.,

\[ \tilde{\text{MED}} = \arg\min_{d \in (d_1, d_k)} \{ \hat{\pi}(d) > \hat{\pi}(d_1) + \Delta, \hat{\pi}^L(d) > \hat{\pi}(d_1) \}, \]

where \( \Delta \) is the clinically relevant effect (may differ by margins) and \( \hat{\pi}^L(d) \) is the lower limit of a 100\((1 - \gamma/2)\) confidence interval for \( \pi(d) \). We say
4 Dose-response for bivariate binary data under model uncertainty

that the MED does not exist if there is no \( d \in (d_1, d_k) \) for which the conditions are satisfied. There are several possibilities of defining the MED with bivariate data. Here, we estimate the MED from each of the two margins (i.e., with \( \hat{\pi} = \hat{\pi}_{1+} \) and \( \hat{\pi} = \hat{\pi}_{+1} \)) and take the maximum over the two MED estimates as the MED estimate for the model, so that clinical relevance and statistical significance is guaranteed for both margins. A different approach would derive the MED by setting \( \hat{\pi} = \hat{\pi}_{11} \). We can take the MED that corresponds to the model with the smallest adjusted P-value as the overall estimate for the target dose. However, a MED estimator that incorporates model uncertainty is based on a weighted average of the MEDs from each candidate model (so they exist), \( \hat{\text{wMED}} = \sum_s w_s \hat{\text{MED}}_s / \sum_s w_s \), with weights \( w_s = \exp(T_s / 2) \). In this way, the ratio of weights \( w_s \) and \( w_{s'} \) attached to the MED from two candidate models \( M_s \) and \( M_{s'} \) reflects their (penalized) likelihood distance.

4 Example: PoA and MED for a diarrhea compound

Guidelines from the European Agency for the Evaluation of Medicinal Products for proving efficacy for compounds against Irritable Bowel Syndrome (IBS) demand that two endpoints, relief of abdominal pain and relief from overall GI-tract symptoms be considered jointly. Figure 1 and Table 1 show several shapes for the marginal probability of each of these endpoints (plotted using information such as the expected placebo and maximum dose effect from prior studies) that are deemed plausible by the clinical team developing a compound against IBS at dose levels \( d = (0, 1, 4, 12, 24) \) mg. The different shapes of these models are generated with linear predictors of fractional polynomial form (Royston and Altman, 1994) that allow for a broad dose-response space. For instance, the clinical team was uncertain about the rate of increase in both margins at low doses and about the mono-
TABLE 1. Various shapes for the marginal efficacy of the IBS compound

<table>
<thead>
<tr>
<th>Shape</th>
<th>Linear Predictor</th>
<th>Shape</th>
<th>Linear Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L_0):</td>
<td>(\alpha)</td>
<td>(L_4):</td>
<td>(\alpha + \beta \exp(\exp(d_j/\max(d_j))))</td>
</tr>
<tr>
<td>(L_1):</td>
<td>(\alpha + \beta d_j)</td>
<td>(L_5):</td>
<td>(\alpha + \beta d_j + \gamma d_j^2)</td>
</tr>
<tr>
<td>(L_2):</td>
<td>(\alpha + \beta \log(d_j + 1))</td>
<td>(L_6):</td>
<td>(\alpha + \beta \log(d_j + 1) + \gamma/(d_j + 1))</td>
</tr>
<tr>
<td>(L_3):</td>
<td>(\alpha + \beta/(d_j + 1))</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE 2. Candidate dose-response models for the efficacy of the IBS compound.
The triplets in the first column refer to which linear predictor is used to model the two margins \(\logit[\pi_1(d_j)]\) and \(\logit[\pi_2(d_j)]\) and the log-odds ratio \(\log(OR_j)\).

<table>
<thead>
<tr>
<th>candidate model</th>
<th>AIC</th>
<th>(T)</th>
<th>perm. P-value</th>
<th>adj. P-value</th>
<th>MED (mg)</th>
<th>weight (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M_0): ((L_0, L_0, L_0))</td>
<td>900.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(M_1): ((L_1, L_2, L_0))</td>
<td>898.2</td>
<td>1.92</td>
<td>0.4196</td>
<td>0.5875</td>
<td>15.9</td>
<td>4.3</td>
</tr>
<tr>
<td>(M_2): ((L_2, L_2, L_0))</td>
<td>894.3</td>
<td>5.80</td>
<td>0.0031</td>
<td>0.0101</td>
<td>6.0</td>
<td>29.8</td>
</tr>
<tr>
<td>(M_3): ((L_3, L_2, L_0))</td>
<td>895.6</td>
<td>4.55</td>
<td>0.0060</td>
<td>0.0218</td>
<td>9.6</td>
<td>15.9</td>
</tr>
<tr>
<td>(M_4): ((L_4, L_2, L_0))</td>
<td>900.7</td>
<td>-0.55</td>
<td>0.6408</td>
<td>0.7593</td>
<td>NA</td>
<td>1.2</td>
</tr>
<tr>
<td>(M_5): ((L_1, L_6, L_0))</td>
<td>900.0</td>
<td>0.08</td>
<td>0.6212</td>
<td>0.7593</td>
<td>15.4</td>
<td>1.7</td>
</tr>
<tr>
<td>(M_6): ((L_2, L_6, L_0))</td>
<td>896.1</td>
<td>4.03</td>
<td>0.2682</td>
<td>0.4780</td>
<td>4.0</td>
<td>12.3</td>
</tr>
<tr>
<td>(M_7): ((L_3, L_6, L_0))</td>
<td>894.0</td>
<td>6.08</td>
<td>0.0030</td>
<td>0.0101</td>
<td>1.0</td>
<td>34.3</td>
</tr>
<tr>
<td>(M_8): ((L_4, L_6, L_0))</td>
<td>902.6</td>
<td>-2.45</td>
<td>0.7631</td>
<td>0.7631</td>
<td>NA</td>
<td>0.5</td>
</tr>
</tbody>
</table>

tonicity of the dose-response curve for the second margin. Hence, the candidate set should include models that incorporate various scenarios for the slope or non-monotonicity. A potential candidate set for the IBS compound is shown in Table 2. For instance, consider model \(M_6\) which postulates 
\[
\logit[\pi_1(d_j)] = \alpha_1 + \beta_1 \log(d_j + 1), \logit[\pi_2(d_j)] = \alpha_2 + \beta_2 \log(d_j + 1) + \gamma_2/(d_j + 1), \log(OR_j) = \log(\pi_{00}(d_j)\pi_{11}(d_j)/\pi_{01}(d_j)\pi_{10}(d_j)) = \alpha_3.
\] When fitting \(M_6\) to (slightly modified) clinical trials data on the efficacy of the IBS compound, a test of no dose effect compares this model to the no-effect model \(\beta_1 = \beta_2 = \gamma_2 = 0\). Its penalized likelihood ratio statistic \(T_6 = -2 \times \{-447.06 - (-442.04)\} - 2 \times 3 = 4.03\). However, the maximum value of 6.08 for the test statistic over all candidate models occurs under model \(M_7\) that allows for a steeper rate of increase in the marginal odd of abdominal pain. This model would also be considered for inference (testing PoA and estimating the MED) when the decision is based on the minimum AIC criterion. The likelihood ratio PoA test with this model has an asymptotic P-value of 0.0035, very similar to the raw permutation P-value displayed in Table 2 that shows the proportion of all 10,000 permutations that yielded a \(T_7\) value as large or larger than the observe one. However, this
statement of significance is conditional on the selected model and ignores the uncertainty the clinical team had at the start of the trial, as expressed in the candidate set. Under simultaneous inference with all candidate models, i.e., testing PoA under each model, the multiplicity adjusted P-value for the PoA test in model $M_7$ equals 0.0101, which is about three times larger but still significant at a conventional overall type I error rate of 2.5%, say. This multiplicity adjusted P-value is derived from the closed testing framework using the permutation distribution of $\max_s T_s$. The MED derived from $M_7$ (with a clinical relevant effect of $\Delta = 10\%$ for both margins and $\gamma = 5\%$) equals 1.0mg. Incorporating model uncertainty, the weighted estimate $w\text{MED} = 5.2\text{mg}$ uses the MED estimates from all candidate models with weights displayed in Table 2 and may be more appropriate.

5 Discussion

The use of multiplicity adjusted P-values guarantees that the type I error rate for one or more incorrect PoC decision (based on various candidate models) when the drug is actually ineffective is controlled at the desired level (e.g., 2.5%). This is in contrast to the common habit of “model fishing” or gambling on an a priori specified shape in the study protocol. A disadvantage of our method is that non-linear (on the link scale) dose response models are harder to incorporate because they would not provide convergent fits for many permutations. Here we treated the case where both variables describe efficacy. Equally interesting is the case where $Y_{1j}$ is a binary primary efficacy variable and $Y_{2j}$ a binary variable describing safety at dose level $j$.

References


