Fit for Purpose Integrated Review

Submission	Empirically Based Bayesian Emax Models for Dose Response Design and Analysis
Submitter	Pfizer
Submission Date	April 21, 2021
OCP Reviewer(s)	Junshan Qiu, Jingyu Yu
OCP Concurring Reviewers	Rajanikanth Madabushi, Hao Zhu
OB Reviewer(s)	Junghi Kim, Qianyu Dang, Donald Schuirmann,
Final Signatory	Issam Zineh, OCP Director Sylva Collins, OB Director

Contents

1. EXECUTIVE SUMMARY

Pfizer (the Applicant) submitted a proposal, under the Fit-For-Purpose (FFP) initiative, intended to support the use of the 'Empirically Based Bayesian Emax Models for Dose Response Design and Analysis' as a statistical methodology for dose finding clinical trials. The submission states that as a method to guide dose selection for drug development, the proposed Bayesian Emax model characterizes a relationship between drug efficacy and dosage level. Future submitters utilizing the proposed methodology in their regulatory submissions are referred to in this document as 'Sponsor(s)' or 'a Sponsor'.

The Applicant's submission document contains an overall procedure of the empirically-based Bayesian Emax models and refers to an R package called *ClinDR* by Thomas (2021) for more technical details. In the proposed Bayesian hierarchical model, the Applicant extends a standard Emax model using reparameterization. Meta-data from 199 compounds are utilized to assess the prior distributions of the Emax model parameters. When the dose-response relationship of a new compound is of interest, the Applicant proposes to apply the historical data-based priors to new data to update the posterior distributions for all parameters. Then, the Applicant states the model estimates a dose-response curve to be used for the dose selection, with the updated posterior distributions.

The review of this FFP submission, conducted by The Office of Clinical Pharmacology (OCP) and The Office of Biostatistics (OB), focused on (but was not limited to) the applicability of the proposed model for future applications, in terms of evaluations on identifiability of model parameters and evaluation of prior specifications, and assessing model fit and performance. Listed below are issues identified during the review, along with proposed Agency recommendations:

a. The Applicant assumes an identical treatment effect across multiple studies. In that sense, the proposed method could be applicable in situations where component studies are comparable in terms of study population, randomization allocation scheme, primary endpoint assessment timelines, etc. In future applications of the methodology to an investigational new drug, a Sponsor should provide justifications for inclusion of each individual component study prior to designing the meta-analysis and seek concurrence from the FDA review team in the relevant therapeutic area.

b. The Applicant uses predictive probability for non-monotonicity as a goodness-of-fit statistic (GOF). In general, this proposed test statistic appears reasonable. However, the GOF statistic is not universally useful under all possible scenarios. One example of an applicable scenario for the proposed GOF statistic may be dose-finding studies with a high-signal design. Different types of graphical summaries can be considered before the GOF testing. Other aspects of data, such as change in variabilities by doses, may also need to be checked to provide additional information on the model fitting, which can be useful when data from different sources are combined. When using the proposed methodology in the future, Sponsors should consider a systematic checking procedure to ensure completeness

of all aspects of model fitting in future applications, as many more abnormal curves could appear with expanded application of the tool.

c. The Applicant proposes potential decision criteria of an optimal dose based on two sets of simulation studies. The supporting evidence may not be sufficient to generalize the proposed decision criteria, because the decision criteria for thresholds of the posterior probabilities of the target efficacy and futility could be compound-specific and may be dependent upon the study design. Sponsors are recommended to consider study specific information to identify dose selection criteria (e.g., thresholds for the posterior probability of the target improvement and futility to be used for the dose selection).

d. When one study cannot fully inform the proposed Bayesian Emax model to ensure all model parameters are identifiable, additional studies – either historical or prospective – are needed to make best use of the R package. Otherwise, alternative model and/or methods should be considered and selected based on the data available.

The proposed Bayesian approach appears attractive for its ability to borrow information from historical data for the analysis of typically small dose response studies. We acknowledge that the proposed Bayesian method works well for the majority of listed compounds in the Applicant's database, based on results described in the submission document. While the proposed prior distributions are subjective, the proposed prior specification for the Emax model appears acceptable for dose-response predictions.

The proposed empirically-based Bayesian Emax model, including the GOF statistic, can be designated fit-for-purpose under the following conditions:

(1) component studies for a new compound are homogeneous
(2) the proposed GOF statistic is applicable
(3) the model is identifiable
(4) study-specific information is considered for dose selection

The Agency's determination is based on the Applicant's original submission, the Applicant's responses to Agency information requests during the review, and the relevant statistical literature, including Gelman et al. (2013) and FDA's Guidance for Industry: Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products (FDA, 2018). This recommendation does not preclude the availability and application of other methods for dose finding clinical trials. In practice, a Sponsor should carefully consider the specific study design characteristics when choosing candidate methods for a dose finding clinical trial; and when deciding on the trial design.

2. INTRODUCTION AND BACKGROUND

2.1 Background

In the field of clinical pharmacology, the sigmoidal Emax exposure-response models have been widely used to understand the safety and efficacy of drugs. The Emax model was initially developed to describe the kinetic interaction between drug and receptors leading to pharmacological responses (Wagner, 1968 and Gibaldi, 1968). Given reversible pharmacological responses, consistent with the laws of mass action, the model assumes that a drug acts at the level of the molecular target, forming a drug-receptor complex, and that the magnitude of drug response is proportional to the fractional number of occupied receptors. Since there exists a finite mass of receptors in the body, a maximum response can be attained when all receptors are occupied. The interaction between drug and receptor is sigmoidal in shape. The model was further modified via reparameterization to fit different purposes (Lalonde, 1992).

2.2 Overview of Submission

The Applicant's submission provides an overall procedure for empirically-based Bayesian Emax models for dose response design and analysis and refers to an R package called *ClinDR* by Thomas (2021) for more technical details. The Applicant describes the proposed Bayesian Emax model as a method to guide the dose selection for drug development by characterizing a relationship between drug efficacy and dosage level. Meta-data from 199 compounds are utilized in the Bayesian hierarchical model to assess the prior distributions of the Emax model parameters. This strategy to construct the prior distributions based on the historical data is called the meta-analytical predictive (MAP) prior (Schmidli et al. 2014). When the dose-response relationship of a new compound is of interest, the Applicant proposes to utilize the MAP prior in order to update the Bayesian posterior distributions for all parameters. Then with the use of updated posterior distributions, the Applicant states that the model guides the dose selection by computing the expected response over a grid of potential doses.

In the proposed Bayesian hierarchical model, the Applicant extended a standard Bayesian Emax model by introducing a new parameter, *difTarget,* and reparametrizing ED_{50} with a normalization scalar P_{50}. The Applicant states that "because *difTarget* describes an observable treatment difference, it is easier to understand than the theoretical effect at infinitely high doses (e.g., E_{max}), and its estimation is usually better than that of the E_{max} parameter". The Applicant estimated $\log(ED_{50}/P_{50})$ rather than estimating ED_{50} directly in the dose response studies by using a symmetric and diffuse prior (e.g. *t*-distribution). The Applicant indicates that the specification of such prior distribution requires a clinical team to supply the P_{50}, which is routinely assessed as part of phase II development activities. Also, the placebo response, E_0, is context-specific, and therefore the prior mean and scale parameters for E_0 must be pre-specified.

The submission document mentions examples of binary, continuous, and non-monotone responses where the proposed Bayesian Emax model has been used and presents an example demonstrating the implementation of the model for subsequent study planning. The submission presents simulation results evaluating several methods, including the proposed Bayesian Emax model, pairwise comparison of each dose to placebo, and a modified maximum likelihood Emax

model estimation. The performance of each estimation method is summarized by root mean square error (RMSE) and actual coverage (COV) of nominal 95% intervals.

Finally, the submission mentions software implementations of the proposed Bayesian Emax models for guiding dose selection and conducting simulations. The software R package *clinDR* implements the proposed methods. The dose response meta-data are distributed within the package and are publicly available for use by any future Sponsor.

2.3 Data Sources

Efficacy data from 199 compounds have been collected and summarized to assess dosing designs and common features of clinical dose response curves. The meta-data include FDA-approved compounds between 2009-2017, the Applicant's compounds that demonstrated efficacy in at least one phase II study between 1998-2017, and biological compounds demonstrating efficacy in one or more published phase II studies between 1985 and 2014. The primary exclusions were oncology compounds and vaccines, which have different dose-finding objectives, designs and analyses. There were 128 small-molecule compounds, 61 biologicals (monoclonal antibodies), and 10 classified as 'other' (e.g., therapeutic proteins).

3. CLINICAL PHARMACOLOGY EVALUATION

Review Question 1: How does the proposed Bayesian Emax model perform with non-informative priors comparing with the corresponding frequentist approach?

To address this concern, the OCP review team conveyed the following to the Applicant on November 12, 2021:

- We understand that the poor performance of the proposed Bayesian method under a setting of using diffusive priors for all model parameters seems unfavorable. However, we believe it is important to ensure comparable performance of the Bayesian method under noninformative priors with the frequentist approach since it is not feasible to evaluate whether the priors are beneficial enough as the magnitude of informativeness will depend on the data from the new drug.
- We speculate that the poor performance of the Bayesian method under noninformative priors is likely due to reparameterization including EC_{50} to $\log(EC_{50}/P_{50})$ and E_{max} to *difTarget*. The former used a historical P_{50} as a constant scaler which may not be efficient and the latter created a complicated correlation among the model parameters. We recommend you investigate the potential concerns and make efforts to further improve the model to facilitate the review.

The Applicant responded to our comments on December 15, 2021. The simulation studies described in Sections 3.3.1, 3.3.2 and Appendix H of the original submission were repeated using a diffuse prior distribution for the E_{max} model parameters. Independent uniform prior distributions were applied to each parameter (Table 1). The parameters were not normalized (e.g., the P_{50} was not used) or transformed except the uniform prior distribution was applied to $\log(ED_{50})$ rather than ED_{50}, as is commonly done. Additionally, the E_{max} parameter was used

instead of the *difTarget* parameter. The priors were chosen to be diffuse with respect to the dose-response designs and parameters in the simulation study, but the uniform bounds were not set to extremely large values to avoid numerical underflow and overflow problems.

Table 1. Summary of Diffuse Priors

Parameters	Distributions
E0	Uniform (-20,20)
Emax	Uniform (-20,20)
Log(ED50)	Uniform (-log(1000), log(1000))
λ	Uniform (0, 10)

Low and high information settings are considered.

Question 1 Conclusions

The performance of Bayesian Emax model with non-informative priors is worse or similar when compared to the frequentist approach (e.g., more model options can be considered to inform dose selection), depending on the applied information settings.

- Low information settings in the simulation studies were selected to represent the most common situations in drug development. "Low information" refers to the setting including fewer doses (e.g., n<3) and lower signal-to-noise (e.g., <50% variation explained). The MCMC methods perform worse in the realistic settings with fewer doses (e.g., n < 3) and when there is lower signal-to-noise where the Emax model is un-identifiable in many settings. When further combined with diffuse prior distributions, the resulting posterior distributions are far from the asymptotic normal ideal and difficult to numerically evaluate.
- High information settings include more well-targeted doses and higher signal-to-noise. The results for the diffuse Bayes and modified ML estimation are more similar, and neither uniformly dominates the other as measured by root mean squared error (MSE). The diffuse Bayes posterior intervals do have better repeated-sampling coverage probabilities than the modified ML methods in these settings.

Review Question 2: How does the Bayesian Emax model perform under a variety of scenarios that model parameters are not fully informed by data?

To address this concern, the OCP review team conveyed the following to the Applicant on March 7, 2022.

- We simulated data without informative sampling points for Emax model parameters such as ED_{50} and E_{max}. With simulated data, we found parameter identifiability issues became a hurdle for applying the proposed Bayesian Emax model even with the informative priors derived from the meta-data analyses. We also observed that some study data in the meta-data pool are sparse to inform the model parameters as shown in Appendix C of the initial submission package (e.g., on page 69 for ID1046 1160.2; on page 80 for ID4009 R668−AD−1224; and on page 81 for ID17 1008−009).

- To address our concerns, pull some typical data from the meta-data pool within the following categories and analyze these data separately either with informative or non-informative priors. Submit for further review the results with the data used in the analyses.

The Applicant responded to the comments on April 4, 2022. They evaluated the performance of the proposed Bayesian Emax modeling when applied to poor dosing designs in four extreme settings (See Table 2). In practice, it is common to encounter dose ranging studies with low signal-to-noise and only few doses covering a limited dosing range (e.g., high/low dose < 10).

Table 2. Summary of Examples.

Example ID	**Endpoint**	**Design Feature**
1021	Continuous	Appear linear, Including placebo
Details	Compound ID=1021 is a subcutaneous injection for the treatment of Homozygous Familial Hypercholesterolemia. There was a single dose finding study with endpoint LDL-C percent change from baseline, 5 active doses (dosing range < 10), and a placebo group.	
Performance	The data supply information about the lower bound for the ED50 and the upper bound for the Emax, but the upper bound for the ED50 and lower bound for the Emax are set largely by the prior distribution, and they are thus somewhat arbitrary.	
1046	Continuous	Without placebo
Details	Compound ID 1046 (requested) is a small molecule for cardio-vascular prevention with primary endpoint change from baseline in Activated Partial Thromboplastin Time (APPT). There are only 3 doses covering a 6-fold range, and as noted in Section I, there is no placebo group.	
Performance	The impact of the lack of placebo data is visible in the upper bound for the effect parameter *DifTarget*, and the lower bound for the placebo response parameter, E0. Comparing the results for the diffuse and weakly informative placebo priors shows that these bounds are largely determined by the placebo response prior.	
4009	Binary	Two dose groups at plateau with placebo
Details	It is an extreme case with a binary endpoint and only two dose groups, both of which appear to be on the plateau of the dose response curve. The endpoint is an	

	investigator assessed global improvement responder variable for Dermatitis.	
Performance	Two features to note are 1) the wide interval for the ED50, which is nonetheless bounded by the lowest studied dose, and 2) the much closer agreement between the *DifTarget* and Emax parameters due to the fact that the data supply much information about the plateau.	
1035	Continuous	doses from two studies on or very close to the plateau
Details	Compound ID=1035 is an inhaled small molecule for the treatment of COPD with change from baseline in FEV1 as the endpoint. There were two studies providing limited dosing information. This example differs from the others included in our response because it includes two studies, and there is some indication that the lowest tested dose is below the plateau although the doses included in the studies clearly do not characterize most of the dose response curve well.	
Performance	The high uncertainty in the estimation of the Emax and ED50 parameters is unsurprising given the lack of data on the steeper portion of the dose response curve. The upper tails of the ED50 and Emax quantitatively reveal something less apparent from simple visual inspection of the sample means, which is the possibility that additional efficacy might be possible if it feasible to test a higher dose.	

Question 2 Conclusions

- Including a placebo group is not only valuable for estimating the dose response curve, but also for informing the re-parameterized model parameter '*difTarget*', which is the difference in response between the placebo and the specified target dose. The absence of a placebo group can alter the responses in the active dose groups and create potential reproducibility issues.
- The analyses results based on the 4 examples indicate that the proposed Bayesian Emax model cannot be adequately identified. In practice, it is highly recommended to use historical experience from other compounds and additional dose response studies with expanded dosing range to predict the unobserved dose response curve to inform dose selection.

Summary

When one study cannot fully inform the proposed Bayesian Emax model to ensure all model parameters are identifiable, additional studies – either historical or prospective – are needed to

make best use of the R package. Otherwise, alternative model and/or methods should be considered and selected based on the data available.

4. STATISTICAL EVALUATION

4.1 Heterogeneity in Meta-analysis

The Applicant proposes a Bayesian Emax model to characterize the relationship between drug efficacy and dose based on multiple data sources to guide the dose selection for further drug development. In the proposed model, the Applicant assumes an identical treatment effect across multiple studies by sharing common parameters for *difTarget*, λ, ED_{50} and E_{max} across the studies, while allowing for heterogeneity in the placebo effect using different placebo parameters (e.g. E_0). The Agency has the following comments on this parameterization:

- One of the most common purposes for performing the meta-analysis is to provide an estimate of the overall treatment effect across the studies. In general, heterogeneity of the treatment effect among the component studies is expected, which can be caused by differences in study populations such as age of patients, dose level, patient follow-up schedules, and other factors. In the statistical literature (see, e.g., *E9 Statistical Principles for Clinical Trials*, International Council on Harmonisation (ICH) of Technical Requirements for Pharmaceuticals for Human Use), it is often preferred that the statistical analysis accounts for such potential variation. In a Bayesian meta-analysis framework, the differences are not expected a priori in general, but study-specific parameters are "borrowing strength" across studies and are "exchangeable" (Gelman et al. 2013). Accordingly, the overall average effect across all studies could be estimated by the median (or mean) of the distribution of study-specific parameters, which can be formulated in the Bayesian hierarchical model. The proposed hierarchical model does not consider the potential heterogeneity in the treatment effects, in that the parameters defining the treatment effect (e.g., *difTarget*, λ, ED_{50} and E_{max}) are assumed to be identical across multiple studies. In accordance with the FDA meta-analysis Guidance (U. S. Food and Drug Administration, 2018), it may be desirable to allow treatment effects to both vary by study and average across studies with appropriate methods to achieve the treatment effect of interest.

- For a new compound, the proposed method will be applied to integrate information from multiple randomized controlled studies. To align with the FDA Guidance (U.S. Food and Drug Administration, 2018) regarding the meta-analysis of randomized trials, the randomized comparisons of each study should be maintained. In other words, when estimating the treatment effect by comparing drug to placebo, subjects randomly assigned to drug in a single study should be compared to subjects assigned to placebo from the same study and not to subjects from other studies. Though the study-specific placebo effect is incorporated, the summary statistics (e.g., mean of the primary endpoint) at each dose level from the multiple studies are pooled to estimate the common treatment effect

(e.g., *difTarget*, λ, ED_{50} and E_{max}). This pooling strategy deviates from a one-to-one randomization scheme. When there are large sample size disparities among the studies with different randomization allocations, the results could be biased.

- The current submission package does not investigate the performance of the proposed method for multiple studies with heterogeneity. Per Agency's recommendation in an Information Request, the Applicant proposed additional simulations for Low Signal and High Signal simulation designs for continuous and binary data, in a total of 8 simulation settings. To create heterogeneity, the Applicant plans to consider different number of doses and placebo effects across studies with a fixed treatment effect and consistent sample size per dose group across studies. The Agency recommends that future Sponsors who intend to use the proposed method explore the performance of the proposed method in multiple studies.

Summary

In the proposed model, the Applicant assumes an identical treatment effect across multiple studies by sharing common parameters for *difTarget*, λ, ED_{50} and E_{max} across the studies, while allowing for heterogeneity in the placebo effect using different placebo parameters (e.g. E_0). The proposed method could be applicable in situations where individual component studies are comparable in terms of study population (e.g., exclusion/inclusion criteria), randomization allocation scheme, primary endpoint assessment timelines, etc.

4.2 Priors

4.2.1 Priors for the historical data

The Applicant used Bayesian hierarchical modeling methods to conduct the meta-analysis. In such modeling, a prior is specified for each study. However, the values of the parameters of the priors are not specified as they would be in a Bayesian analysis of the single study. Rather, the parameters for each study prior are assumed to represent random draws from a hyperprior distribution. The values for the parameters of the hyperprior are then specified.

In this case, the Applicant actually assumed that "The model includes compound-specific E_{maxj} , ED_{50j} and λ_j parameters, and study-specific E_{0jk} parameters." Thus, all of the studies in the "meta-data" for a specific compound *j*, are assumed to have the same prior, except for the placebo effect parameter, E_{0jk}, which is given its own prior for each study *k*. Because the placebo effect parameter is assumed to be study-specific, its parameters are not included in the hyperprior.

The Applicant used transformations of the original Emax model parameters in their analyses. Instead of E_{maxj} , the Applicant used the parameter $difTarget_j$, defined as

$$difTarget_j = E_{maxj}\left(\frac{dTarget^{\lambda_j}}{dTarget^{\lambda_j} + ED_{50j}^{\lambda_j}}\right)$$

which is the difference from the placebo response at dose *dTarget*, a specified large dose. Instead of ED_{50j}, the Applicant used $log\left(\frac{ED_{50j}}{P_{50j}}\right)$, where P_{50j} is the a priori prediction of ED_{50j} for compound j. Instead of λ_j, the Applicant used $log\lambda_j$.The Applicant's submission document does not specify whether "log" in these definitions means natural log, log to the base 10, or some other type of logarithm.

The Applicant's submission (in Section 2.2.1) seems to describe the individual prior distributions as multivariate-t distributions on 5 degrees-of-freedom, at least for the parameters $log\lambda_j$ and $log\left(\frac{ED_{50j}}{P_{50j}}\right)$ (note: the submission is ambiguous regarding the individual prior for *difTarget*.) The package goes into detail about the hyperpriors that were tried. These are summarized in Table 5 of Appendix A in the submission package. The Applicant considered hyperpriors where all of the hyperparameters were independent, and also considered hyperpriors where the hyperparameters $\mu_\lambda, \mu_{ED50}, \mu_{difTarget}$ (the prior means for $log\lambda_j, log\left(\frac{ED_{50j}}{P_{50j}}\right)$, and $difTarget_j$ respectively) may have correlations.

For each hyperprior, the meta-data were analyzed with MCMC methods resulting in a posterior distribution for the hyperparameters. The medians of the posterior hyperprior distributions are presented in Tables 6 and 7 of Appendix B in the submission document.

4.2.2 Priors for future dose response studies

The Applicant states that the posterior predictive distribution obtained from the meta-analysis (presumably based on the hyperprior of their choice) would serve as the prior distribution for future studies. The MCMC calculations would produce a distribution in numerical form (i.e. not in analytical form, such that the formula for the density could be written down.) However, the Applicant in teleconference with FDA dated on April, 4th, 2022, indicated that the resulting numerical distribution, at least for the location parameters, could be well-approximated by a multivariate Student's *t*-distribution (suitably scale and location translated) on 5 degrees-of-freedom (hereafter called "t(5)".)

The proposed prior distribution for $log\lambda_j$ and $log\left(\frac{ED_{50j}}{P_{50j}}\right)$ is a bivariate version of the multivariate t(5). $log\lambda_j$ is centered at 0.0 with a scale multiplier of 0.425. $log\left(\frac{ED_{50j}}{P_{50j}}\right)$is centered at 0.0 with a scale multiplier of 1.73. The correlation between these two variates is set at -0.45. By default, both E_0 and *difTarget* are assigned independent (i.e., independent of $log\lambda_j$ and

$log\left(\frac{ED_{50j}}{P_{50j}}\right)$ and of each other) diffuse t(5) distribution, with a location of 0.0 and a scale parameter of 2.0 or 4.0. The Applicant does note that in specific cases there may be prior information about E_0 and/or *difTarget* that would warrant specifying an informative prior different from these defaults.

For σ(a parameter of the likelihood), a uniform distribution is used, with limits chosen wide enough to include the plausible values. For β (the regression coefficient(s), within the likelihood, for any covariates included in the model/analysis), a diffuse multivariate normal is specified.

Summary

The proposed parameter transformations of λ and ED_{50}, where symmetric priors are specified for $log\ (\lambda)$ and $log\left(\frac{ED_{50j}}{P_{50j}}\right)$, appear reasonable given the nature of these parameters. The transformation of E_{max} to *difTarget* does not seem as intuitive, but the Agency accepts the Applicant's statement that the transformation improves model fit.

The proposed Bayesian approach appears attractive for its ability to borrow information from historical data for the analysis of typically small dose response studies. The Agency acknowledges that the proposed Bayesian method works well for the majority of listed compounds in the Applicant's database, based on results described in the submission document. While the proposed prior distributions are subjective, the proposed prior specification for the Emax model looks acceptable for dose-response predictions.

4.3 Assessing Model Fit

Statistical assessment of model fitting is meant to check adequacy of model prediction by comparing the prediction to the observation. In this submission, model diagnostics were used to identify deviations in the observed data from the Emax model prediction. The Applicant uses Bayesian GOF by drawing simulated samples from the joint posterior predictive distribution of replicated data and comparing the samples to the observed data. As in any GOF test, test quantities must be defined to represent aspects of the data we are checking. One important aspect to be checked in this application is non-monotone dose response. The Applicant proposes "the difference in response between the best of the lower doses and the highest dose" as the test quantity. The Agency acknowledges that such test quantity is important, but the test statistic may not be the best choice and not universally useful under all possible scenarios (see Section 3.3.3). Unlike the Bayesian Emax model itself, the GOF predictive value does involve hypothesis testing, so the power and type I error need to be considered. Using the GOF predictive value of 0.05 as a boundary makes a reasonable compromise between type I and II errors. In addition, graphic checks can be performed before any quantitative approach. In this submission document, all the compounds where the model failed can be relatively easily identified with the graphic check. Furthermore, there is a chance that the graphic check can still display "clear lack of fit by other measures and visual inspection" when GOF > 0.05. Thus, the P-value of the GOF should

be combined with other types of model checks and other test quantities in model assessment. Different graphic displays can be implemented for test statistics of interest before the GOF testing. Other aspects of data, such as change in variabilities by doses, may also need to be assessed in order to provide additional information on the model fitting, because this could be useful when data from different sources are combined. In general, adjusting multiple comparisons when using more than one test statistic in the GOF testing are not recommended.

The Applicant also assesses other posterior predictive quantities, such as the high fit statistics, where "the highest dose performed better than expected compared to the lower doses", but its interpretations and implication in drug development is not clear. A higher fit probability for one dose compared to another does not necessarily indicate inadequacy of the model fitting.

Summary

In this submission, an overwhelming majority of the compounds listed show good model fit. This may be because many compounds with improper dose-response curves had been ruled out in pre-clinical studies. There is still a chance that other types of unusual curves could be undetected with the current approach.

Overall, the proposed model fitting assessing tools are acceptable. However, a more systematic checking procedure may be needed in future submissions to ensure completeness of all aspects of model fit, as many more abnormal curves could appear with expanded applications.

4.4 Examples and Simulations

Heterogeneity in a meta-analysis example

In Section 3.2.1 of the submission, the Applicant analyzes dose response of tofacitinib for the treatment of rheumatoid arthritis (RA). By using the proposed Bayesian Emax model, the Applicant combines the drug responses (ACR20) from two dose response studies 'A3921019' and 'A3921035' whose primary visits are at week 6 and week 12, respectively. The Applicant states that data from the two studies are similar and the proposed method was used for meta-analysis with these two, assuming identical treatment effect at week 6 and week 12. It is known that a drug effect is a function of dose and time, but the proposed model includes parametric functions for the relationship between dose and response only; the dependence of time and response is not considered. Combining drug effects assessed at different timepoints could result in different clinical interpretations.

Criteria for dose selection

In Section 3.2.4 of the submission, the Applicant describes how to derive decision criteria for dose selection. The Applicant states that "*Exploration of the distribution of the posterior probabilities from the futile and optimal doses suggests a potential decision criteria for establishing doses for further evaluation could require a probability ≥ 0.75 to exceed futility and*

a probability ≥ 0.1 to meet or exceed the targeted efficacy … These rates are indicative of the precision in dose selection achievable with a binary endpoint and moderate effect size." The Agency acknowledges that the decision criteria for thresholds of the posterior probabilities of the target efficacy and futility were derived based on two sets of simulation studies with a pre-defined optimal dose (e.g., 50 mg or 450 mg), but supporting evidence is not sufficient to generalize the proposed decision criteria. For example, in Section 3.2.1, the Applicant assumes that the target improvement probability is ≥ 0.3 and a dose providing < 0.2 improvement would be unsuccessful. In Table 2, the Applicant presents the predictive posterior probabilities of the target improvement and futility at each dose level to demonstrate how the model predictions correspond to the results of four phase III studies. Based on the criteria proposed in Section 3.2.4, a candidate for the optimal dose is 6 mg with safety considerations, while the Applicant proposes the dose level of 10 mg, at which a probability to exceed futility is 1 and a probability to exceed the targeted efficacy is 0.8. It appears that there is still some ambiguity regarding how to determine decision criteria. The Agency recommends future Sponsors consider study-specific information to identify dose selection decision criteria (e.g., thresholds for the posterior probability of the target improvement and futility to be used for the dose selection).

GOF for Low-signal design

Unlike the Bayesian Emax model itself, the GOF predictive value does involve hypothesis testing, so the power and type I error need to be considered. In the simulation study of Section 3.3.3 of the submission, the power of the GOF test is extremely low for any Low-signal design; the proposed model may not be useful under such scenarios. Instead, pairwise comparisons or other diagnostic tools could be considered in such a scenario.

4.5 Others

Use of prior knowledge of P_{50} and E0

The Applicant notes that P_{50} and the placebo response (E_0) are compound-specific and not derived from the historical data. The Applicant states P_{50} can be predicted from pre-clinical and early-stage clinical data before initiating clinical dose finding studies, and the prior mean and scale parameters for the placebo effect E_0 must be specified by the study team. In future submissions for new compounds, a Sponsor should justify the choices of prior information regarding P_{50} and E_0. In addition, meetings with the Agency should be requested to discuss the prior information. If there are multiple predictions based on different data sources that differ substantially, Sponsors should perform a sensitivity analysis to check the robustness of their models to different choices of priors for P_{50} and E_0.

Prior update from historical data

In the Applicant's communication dated September 28, 2021 in response to Agency's information request, the Applicant describes that when data from a substantial number (e.g., approximately 100) of new compounds are available, the hierarchical modeling and the

predictive prior distributions will be updated. The Agency notes that any change in the prior specifications should be made at an early stage and prespecified prior to a future submission.

5. REFERENCES

U. S. Food and Drug Administration/Center for Drug Evaluation and Research/Center for Biologics Evaluation and Research. (2018). Guidance for Industry Meta-Analyses of Randomized, Controlled, Clinical Trials (RCTs) to Evaluate the Safety of Human Drugs or Biologic Products.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B. Vehtari, A. and Rubin, D.B. (2013). Bayesian Data Analysis (Third ed.). Boca Raton, Fl: CRC Press.

Thomas, N., Wu, J. and Smith, M.K. (2021). *clinDR*: Simulation and Analysis Tools for Clinical Dose Response Modeling. R package version 2.3.5.

Schmidli, H., S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics 70, 1023–1032.

Gibaldi, M. (1968). Measurement and interpretation of certain biopharmaceutic and pharmacodynamic parameters. Chemotherapy 13, 1-15.

Lalonde, R. (1992). Applied pharmacokinetics. In W. Evans, J. Schentag, and W. Jusko (Eds.), Principles of Therapeutic Drug Monitoring. Philadelphia, PA: Lippincott Williams Wilkins.

Wagner, J. (1968). Kinetics of pharmacologic response. i. proposed relationships between response and drug concentration in the intact animal and man. Journal of Theoretical Biology 20 (2), 173-201.

RESPONSE TO FDA REQUEST FOR INFORMATION

received 14 Sep 2021

FDA Qualification Opinion
"Empirically Based Bayesian Emax Models for Dose Response Design and Analysis"

BACKGROUND

21 April 2021, Pfizer submitted a Fit-for-Purpose (FFP) Request for an FDA Qualification Opinion on "Empirically Based Bayesian Emax Models for Dose Response Design and Analysis." On 14 September 2021, Pfizer received comments and requests for information on this submission from the FDA.

This document includes Pfizer's formal responses to these FDA information requests.

AGENCY REQUEST FOR INFORMATION AND SPONSOR RESPONSE

Clinical Pharmacology

Question 1

We recommend conducting a performance comparison between a "Population analyses under frequentist paradigm" and your proposed Bayesian approach under the non-informative prior setting. This comparison serves as a basic assessment of the submitted package without influence of priors given that the two approaches should yield similar results under the setting of using non-informative priors. Propose a detailed analysis plan based on simulation studies for review if you choose to proceed.

Response 1

Section 3.3, "Simulation results", contains an extensive evaluation of a "population analyses under frequentist paradigm" compared to our proposed Bayesian methods using non-informative prior distributions. Our prior distributions are constructed with diffuse prior distributions for the placebo-response and effect-size parameters. The ED_{50} and Hill (λ) parameters are assigned the prior distributions derived from the predictive distributions of these parameters in the large meta-analysis with approximately 200 compounds. Simulations of formal Bayesian methods (i.e., MCMC evaluation) with fully diffuse prior distributions on all parameters is unnecessary because past simulations have shown this method will perform badly. Indeed, the poor performance of the Bayesian methods with a fully diffuse prior distribution was one of the primary motivations for conducting the large meta-analyses to provide an empirical basis for stabilizing the Bayesian estimation of the most difficult parameters to estimate.

Several repeated-sampling simulations are included in the test code for the R package, ***clinDR***, which are executed with each new version to validate the code. These simulations are conducted under different conditions with known theoretical approximations that provide reliable references for checking the validity of our code. The prior distributions mentioned in the previous paragraph are utilized. Dose response studies with very large sample sizes are simulated because statistical theory assures that the Bayesian methods, if correctly implemented, will produce confidence intervals having the correct coverage in this setting. Several such simulations are included for different types of applications (e.g., continuous and binary data). We can supply more detail and perform more such simulations if this is helpful. We have also performed test simulations where lightly informative distributions are applied to the placebo and effect size parameters. In this setting, by drawing the model parameters

from their prior distribution and setting them to be the population parameters for each simulation replication, the resulting Bayesian posterior intervals should also have the correct frequentist coverage even with small simulated sample sizes (Cook and Gelman 2006). The simulation code to perform this Bayesian frequentist simulation is already part of the R package. We can supply such results if they will increase your confidence in the numerical methods.

Statistical Comments

Question 1

Clarify how the results of your meta-analysis are to be used for the design and analysis of future dose-response studies.

Response 1

The meta-analysis provides a strong empirical basis for specifying a simple parametric model (Emax) that describes the clinical dose response of a high proportion of compounds in drug development. This empirical evidence supplements theoretical support for this model from clinical pharmacology. Although the Emax model has a simple mathematical form, it is non-linear in its parameters and often difficult to estimate with data that can be practically collected. Maximum likelihood or non-linear least squares estimation of the Emax model does not perform well in clinical settings. The estimation can be improved with Bayesian prior distributions for the non-linear parameters. The hierarchical modeling of the meta-data produces predictive distributions for these parameters that can serve as empirically based prior distributions for the difficult-to-estimate parameters when combined with commonly available information specific to each future compound.

The resulting Bayesian Emax model serves as the foundation for both the planning and analysis of subsequent dose response studies. An example illustrating the modeling approach to select a dosing design, determine appropriate sample sizes to differentiate between active doses, and assess the operating characteristics of different dose selection criteria is in Section 3.2.4. Section 3.2.1 describes an application of the model that was used to plan the doses to include in a second dosing study, combine the results from the two studies which have differing doses, and to predict the performance of different doses in future Phase 3 studies, which was used to select two doses for future development.

Question 2

In Section 3.1.2, you proposed to analyze future dose response studies by specifying a prior distribution for the transformed Emax model parameters log (λ) and log (ED_{50}/P_{50}). Both of these transformed parameters will have a symmetric prior (a scaled t_5 distribution) centered at zero, implying that the prior median for λ is 1 and the prior median for ED_{50} is P_{50}. The prior for $\ln(\lambda)$ is scaled by 0.425, and the prior for $\ln(ED_{50}/P_{50})$ is scaled by 1.73. A correlation coefficient between $\ln(\lambda)$ and $\ln(ED_{50}/P_{50})$ is also specified as -0.45. However,

the document does not state explicitly how this prior distribution is to be implemented. We recommend providing the details of your proposed method with the following points:

Question 2a

Clarify a hierarchical structure and likelihood of the proposed model.

Response 2a

In a future application to a new compound, the prior distributions of the $\ln(\lambda)$ and $\ln(ED_{50}/P_{50})$ parameters are based on the predictive hierarchical distributions estimated in the meta-analysis described in Section 2.2. To simplify the estimation of a future dose response curve, analytical approximations (t-distributions) to these predictive prior distributions were developed, which is described in Section 2.3. The hierarchical distributions in the meta-analysis are not updated with each new compound, which greatly simplifies the analysis of a new compound, because it does not require hierarchical modeling. This is a common strategy in Bayesian analyses. It has been described as the meta-analytic-predictive prior approach (Schmidli, et al, 2014). Because we have numerous (approximately 200) compounds in the meta-analysis, data from a single new compound would yield very little change to the estimation of the hierarchical distributions. When data from a substantial number (e.g., approximately 100) of new compounds are available, the hierarchical modeling in Section 2.2 will be updated, and the predictive prior distributions will be updated. Over time this will allow the empirical basis for the Bayesian Emax model to improve while maintaining the simplicity of its application to each new compound.

Question 2b

Include the nature of each prior and justify its applications. For example, clarify why you set specific priors for each parameter and how the correlation between $\ln(\lambda)$ and $\ln(ED_{50}/P_{50})$ is to be implemented.

Response 2b

As noted in the response to comment 2a), all of the numerical values (e.g., correlation of -0.45 between $\ln(\lambda)$ and $\ln(ED_{50}/P_{50})$) are derived and explained in Section 2.3 based on the modeling of the meta-data described in Section 2.2. As referenced in Section 2.3, additional computational and numerical details for the prior predictive distributions are in Appendices A and B. The approach used to combine these predictive prior distributions with compound-specific information to complete the specification of a full prior distribution for the model parameters is in Section 3.1.2. Clinical teams have been able to understand and implement this approach.

The use of the prior distribution in the dose response package, ***clinDR***, does not require any new programming. The numerical values describing the prior distribution (e.g., expected placebo response) are input to a function called emaxPrior.control that creates an object specifying the prior distribution. The inputs to the prior derived from the meta-analysis are supplied by default to emaxPrior.control, so it is not necessary for users to input them (users

can over-ride the defaults, but this is not recommended except for some prior sensitivity checking). The prior object is then passed to the model fitting and simulation functions. The MCMC evaluation of the posterior distribution is performed by these functions using the STAN MCMC Bayesian computing package (STAN Development Team, 2015). The inputs describing the prior distribution can also be assigned in a simple GUI interface using the shiny app that is supplied with the ***clinDR*** package. The shiny app then creates the function calls for users.

Question 3

In Section 3.2.2, you state that the results of the study 'A7941005' are displayed in the left panel of Figure 8, and the results of combining two studies 'A7941005' and 'A7941006' are presented in Figure 8. However, we expect the results from the single study and the results from the combined studies would be different. We recommend updating the results accordingly.

Response 3

Thanks for highlighting our sentence on the second line of p. 31, which is confusing:

"The results of the first study are displayed in the left panel of Figure 8."

The sentence should have stated:

"The plot of the fitted Emax model from the final analysis, which included data from both studies, is in Figure 8. The final fit for the first study displayed in the left panel of Figure 8 is indistinguishable from the results when the model was initially applied to the first study alone (not shown)."

The plot of the model fit to the first study alone is displayed here in Figure 1 below.

Figure 1. Dose response plot for first study fitted only with data from the first study in the example in Section 3.2.2, corresponding to Figure 8 in the report

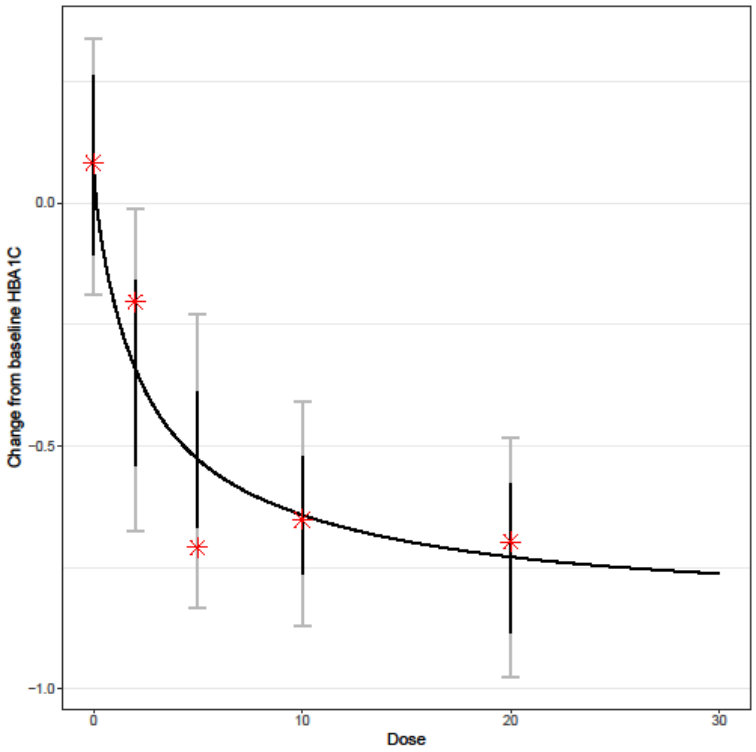
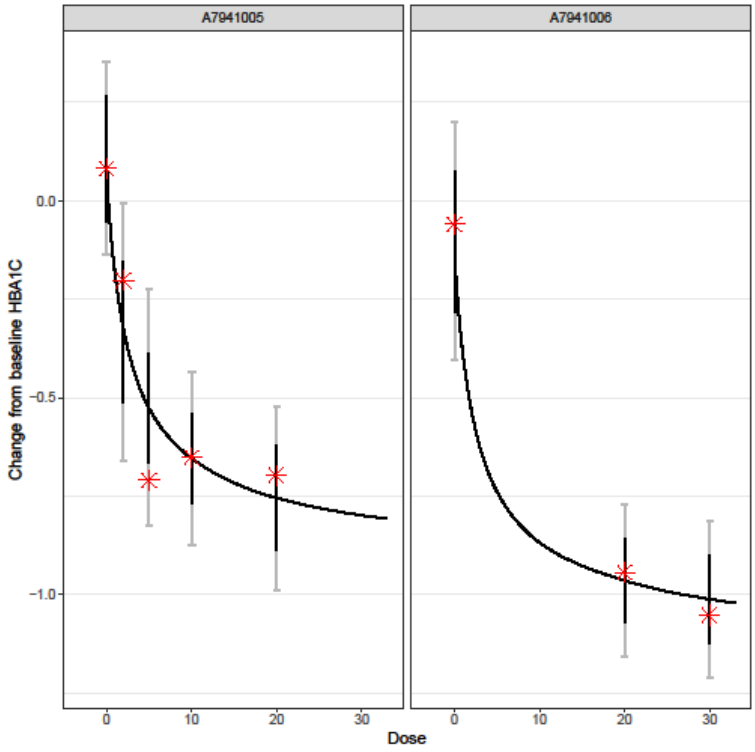


Figure 8 in the manuscript is included here as Figure 2 for convenient comparison

Figure 2. Figure 8 Reproduced from the Report



We propose to update the description of the plot in the manuscript as presented here if/when the manuscript is revised. If you prefer, we can add Figure 1 to the manuscript.

Question 4

In Section 3.3, simulations are used to describe the operating characteristics of the proposed Bayesian Emax model under a wide range of conditions. We recommend the following:

Question 4a

It appears that all simulation scenarios include only one study. Since the proposed model will be applied to meta-analyses, simulation studies considering certain characteristics of meta-analyses, such as the number of studies, their sample size and level of heterogeneity across studies, will be helpful to demonstrate the proposed method's performance. Propose a detailed analysis plan for the simulation studies.

Response 4a

We propose to use the simulation design from Section 3.3 as the foundation for the additional simulations. The population Emax models with λ =0.8 and ED_{50}/ P_{50}=4 will be used (see Figure 22 in the report for displays of these curves), because they produce data similar to many real dose response studies, and the population parameters differ from the prior distribution by amounts common in real applications. Simulations will be done for the Low Signal and High Signal simulation designs for continuous and binary data (4 combinations).

The first simulated study for each new simulated compound will match those in the report, but then a second 'study' will be simulated for each compound that matches the common situation in which only the higher doses and placebo are included in the second study. Specifically, for the Low Signal design, the lower dose will be dropped in the second study, and for the High Signal design, the lowest two doses will be dropped. The second studies will thus have lower total sample sizes and fewer doses. Finally, the data in the second study will be simulated with the same population placebo rate as in the first study, and then it will be repeated with the population placebo response set to 1/2 the difference between the highest dose and placebo in the first study (a large change in placebo response, for a total of 8 simulation settings). The simulation results will be summarized like those in Section 3.3, including the same alternative methods for benchmarks.

Question 4b

In Section 3.3.3, the proposed model is evaluated under a simulation where the true population dose-response relationship does not follow the Emax model. Table 4 displays the proportion of the simulated data sets with the GOF fit probability less than 0.05. In Table 4, the GOF test power was low for any Low-signal design. Explain whether a different model should be applied instead in such a Low-signal design.

Response 4b

A different model is not indicated in the Low-Signal settings. A detailed discussion of considerations in the Low-Signal setting is given on p. 56 of the report. The data from such studies will poorly determine the dose response curve, and it will yield large uncertainty for most decision metrics even when the model is correct. Note that the power to differentiate from placebo is only 80% when the Emax model is correctly assumed. The power to reject the null is <60% for the models with non-monotone dose response even when the optimal dose is known in advance. There is little opportunity to differentiate between active doses.

Further, if it is known in advance that we are in a Low Signal setting, either the design will require change, or the compound will be terminated for lack of sufficient efficacy. The fact that the design is in a Low-Signal setting is typically known only after the data are collected. Selecting alternative models at this stage is unlikely to be helpful, and it can be misleading.

Question 5

You extended a standard Bayesian Emax model by introducing a new parameter difTarget and reparametrizing ED50 with a normalization scalar P50. You state that because difTarget describes an observable treatment difference, it is easier to understand than the theoretical effect at infinitely high doses, and its estimation is usually better than that of the Emax parameter. In the submitted document, benefits of the proposed method are not clearly demonstrated compared to the standard Bayesian Emax model. Provide an example of comparison between the proposed method and the standard Bayesian Emax model.

Response 5

The two parameterizations yield equivalent models, although the subsequent prior specifications may be somewhat different. The primary benefit of using the difTarget

parameterization is interpretability. When data are available to support a more informative distribution, the data almost always pertain to one (usually the highest) of the doses under study. When more diffuse prior distributions are specified, perhaps imposing only biological plausibility, we found that teams considered the plausible response range for the highest dose even when the Emax parameter was intended, because issues such as safety limitations made consideration of extremely high doses nonsensical.

The difference in the resulting posterior inferences about dose response within the observed dosing range between the parameterizations is typically inconsequential. The posterior medians and 90% posterior intervals for the placebo adjusted response at each tested dose using the two different parameterizations for the example in Section 3.2.2 are in Table 1 below. The prior distributions for the placebo response, difTarget, and the Emax parameter were set to diffuse values. Not only are the parameterizations different for the two fits in Table 1, but an older prior distribution based on an earlier version of the meta-data was used for the Emax parameterization (still available in the ***clinDR*** package as function prior.control, but now deprecated). This prior did not include any prior correlation between parameters, and it used a different mathematical form for the Hill parameter, l (Thomas and Roy, 2017). The impact of the differing parameterizations and prior distributions on inferences for the key dose response measures is very limited. Bigger differences are sometimes observed for the individual model parameters, some of which may be ill-determined by the data, which makes them less reliable and more dependent on prior assumptions. We routinely de-emphasize the reporting of the model parameters for this reason.

A final point is the difTarget parameterization tends to yield better behaved parameter distributions that are closer to the optimal conditions for the MCMC algorithms generating them. Figure 3 has a scatterplot matrix of the MCMC generated parameters using the difTarget parameterization. Figure 4 has the corresponding plot for the Emax parameterization. The difTarget parameterization yields posterior distributions for the parameters closer to ideal independent normal distributions. This behavior has been typical in our experience. The one known disadvantage of the difTarget parameterization is that it does not directly constrain the Emax parameter, so when the Emax parameter is poorly determined, the resulting posterior distribution for the Emax parameter can include a few extreme values. This is managed by educating users and using posterior medians and posterior percentiles by default in the package reporting functions rather than posterior means and standard deviations.

Table 1. Posterior median and 90% posterior interval for model-based estimation of the placebo adjusted dose response at each tested dose computed using the fitDif and Emax parameterizations/prior distributions

	Placebo-adjusted dose response				
Parameterization	2	5	10	20	30
difTarget	-0.43 (-0.67,-0.22)	-0.63 (-0.84,-0.43)	-0.76 (-0.94,-0.58)	-0.86 (-1.04,-0.68)	-0.91 (-1.10,-0.71)
Emax	-0.42 (-0.64,-0.21)	-0.60 (-0.78,-0.41)	-0.74 (-0.91,-0.57)	-0.86 (-1.04,-0.69)	-0.93 (-1.13,-0.74)

Figure 3. Matrix scatter plot of the MCMC generated parameters using the fitTarget parameterization for the example in Section 3.2.2.

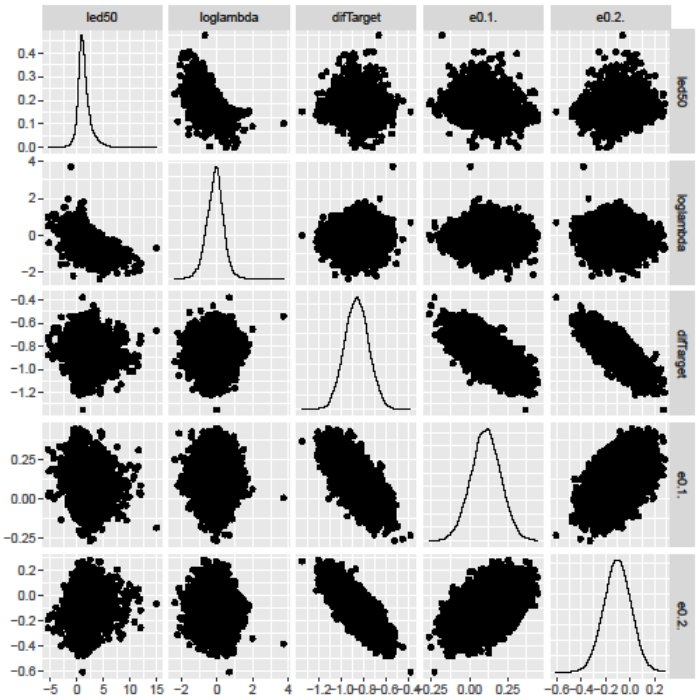
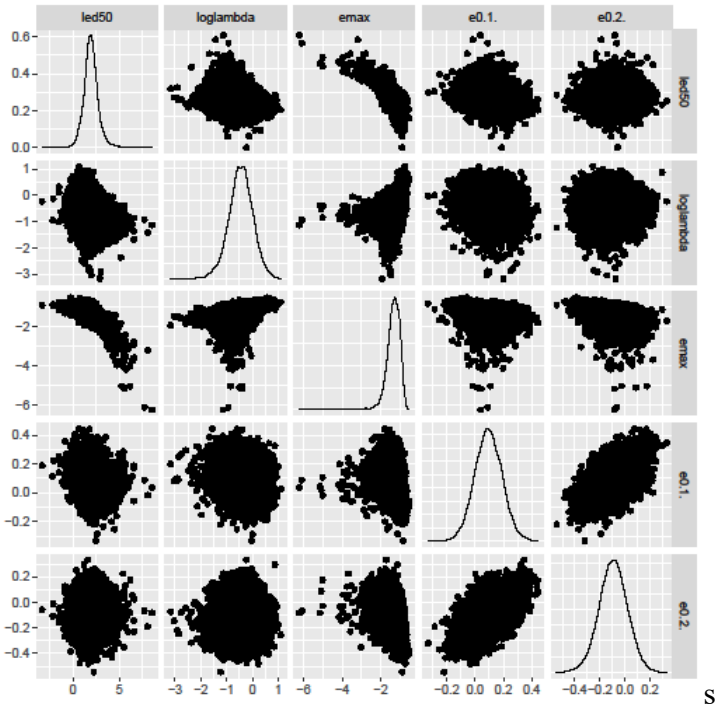


Figure 4. Matrix scatter plot of the MCMC generated parameters using the Emax parameterization for the example in Section 3.2.2



s

Question 6

In Section 2.2.2, you state that the binary endpoints were analyzed along with the continuous endpoints using approximate weighted normal likelihoods. From the information provided, it is not clear how you used the weighted normal likelihoods. Clarify.

Response 6

For continuous endpoints, the data for a compound consists of a sample mean for each dose group and its standard error. Fully efficient estimation (under the usual normal-distribution modeling assumptions) can be achieved from this aggregated data using weighted least squares (and its Bayesian analog) with the weight assigned to each dose group sample mean inversely proportional to its squared standard error. While the weighting produces efficient estimation of the parameters determining mean response, the naive estimation of the residual variance term is poor (the degrees of freedom is the number of dose groups less the number of model parameters). The sample means and their SE were thus supplemented by the pooled within dose group sample (squared) SD with chi square distributions in the modeling to create a complete set of sufficient statistics. This is straightforward to implement with flexible general-purpose Bayesian computation software such as STAN.

A similar approach was used for binary endpoints. The dose group sample means become sample proportions. To both improve the normal approximation to the distribution of the sample proportions and perform the dose response modeling on a more appropriate scale, the

dose group sample proportions were logit-transformed. The standard errors of the logit-transformed sample proportions were computed using the usual asymptotic formula. The logit-transformed proportions were then input with weights inversely proportional to their squared standard errors. The binary contributions are somewhat simpler because they do not require estimation of residual variance. The normal approximation to the logit-transformed proportions was utilized because preliminary checking showed it had little impact on the hierarchical model fitting, and it was much faster when fitting the data from 200 compounds simultaneously, which can require several hours.

When implementing the Bayesian Emax model for binary endpoints for a new compound, the ***clinDR*** package implements the full Bernoulli/Binomial likelihood.

Question 7

In Section 2.3, you provided 90% prediction intervals for the logλ and logED50/P50 from several different hierarchical prior distributions to evaluate the sensitivity of the resulting predictive distributions. Appendix A and Appendix B describe prior distributions and posterior estimates of each model respectively. We recommend providing corresponding posterior/predictive posterior estimates (and 90% intervals) for the difTarget so to clinically interpret the impact of different hierarchical prior distributions. Provide the results in a table for the median and 90% intervals.

Response 7

The hierarchical models provide succinct summaries of the meta-analysis results for the λ and ED_{50} parameters. They also form the basis for the proposed prior distributions for these parameters in future applications. We provided detailed summaries from the meta-analyses for these parameters for these reasons. A hierarchical model for difTarget was not developed for reasons given in Section 2. Instead, separate prior/posterior distributions were created for each of the 208 historical compounds included in the meta-analysis. We have not planned to tabulate the results for difTarget because this would yield 208*8=1664 rows (8 models were evaluated). Note that the posterior distributions of difTarget for these historical compounds do not have a direct role in future analyses. Our response to Comment 5) compares summaries of posterior distributions of placebo-adjusted dose response for the example in Section 3.2.2 for two different prior distributions, which differ in more ways than the 8 priors evaluated in the report. It may serve as a better indication of prior sensitivity.

Question 8

In Section 2.4, you proposed to use the difference in response between the best of the lower doses and the highest dose for goodness-of-fit statistics (GOF) to detect non-monotone response. We need further information on the following points:

Question 8a

Clarify whether you used any cut off value for the GOF posterior predictive in identifying non-monotone dose response. Also, please clarify whether you checked other aspects of the data other than non-monotone response for accessing the model fit.

Response 8a

In applications to new compounds, a GOF predictive value ≤0.05 will initiate a careful review of available data (of different types) to determine if primary reporting should be shifted to pairwise comparisons rather than remain model based. The 0.05 boundary is partially due to convention, but exploration of false positive and false negative findings of non-monotone trends using simulation (Thomas and Roy, 2017) support the use of the 0.05 boundary as a reasonable compromise between these two types of errors. Any boundary is of course somewhat arbitrary, so the GOF assessment is always supplemented by graphical displays, assessment of other data sources (e.g., drug concentrations), biological plausibility supported by mechanistic biomarker data, experience with related compounds, etc. Based on other data sources, a GOF predictive p-value somewhat >0.05 could still result in a tentative rejection of the model.

The two compounds with non-monotone response documented in the report have GOF predictive values <0.05. They were highlighted in the report because they are the only compounds with substantial data from other sources confirming the findings. They were not included in the hierarchical model fitting. There were four compounds with non-monotone GOF predictive values <0.05 included in the hierarchical modeling.

Two other GOF predictive tests were computed for each compound as part of the meta-analysis. As discussed in response to comment 8 c), the upper tail of the non-monotonicity check (or equivalently, one minus the predictive probability so low values serve as a warning) also serves as a check for unexpected high deviations from the dose response predicted by the fitted model. Another GOF test included in the meta-analyses checked for more increase in response at the lowest doses than predicted by the model. As noted on p. 20 in the report, the results from this GOF test were unremarkable and thus not reported. The latter test has not been implemented within the ***clinDR*** package to date, but if it or other tests prove useful, they can be rapidly implemented because they are simulation-based and do not require analytical approximations.

Question 8b

You calculated GOF statistics from the compound-indications in the meta-data. Gelman et al (2013) suggested not to adjust for multiple comparisons when using more than one test statistic. Explain why this should or should not be applied in this application.

Response 8b

All GOF tests in our report, and in applications to dose response modeling, the fit probabilities are not adjusted for multiplicity. Figure 5 in our report, and the text describing it, do place the GOF fit probabilities with the context of evaluating numerous compounds with the objective of determining how frequently we expect to encounter non-monotone dose response. The primary purpose of that display is not to assess non-monotonicity for a specific compound, although it does have value for that objective. The aggregate assessment in Figure 5 does not contradict the advice from the reference, which is intended for reporting of multiple distinct features of model fit within a single study. While assessment of multiplicity regarding the GOF tests is not part of our routine practice, understanding the

impact of sampling variability and multiplicity is important when interpreting the GOF results and the need for supplemental supporting information.

Question 8c

You identified TAID = 1001 with the high fit statistics where the highest dose performed better than expected compared to the lower doses. Clarify how the drug development is impacted when the model prediction is better for the highest dose.

Response 8c

This is an older compound from a company ended by a merger, so it difficult to know how the biomarker data may have guided dose selection. We have not had a similar internal example and thus suspect this may be a rare situation. Numerically, we expect the effect at the highest dose to be under-estimated, but the estimated curve will still display an increasing effect at higher doses due to the presence of a large effect at the highest dose. The plot for TAID=1001 illustrates this as the model under-estimates the effect for the highest dose while over-estimating the effect for the second highest dose, but it still displays a large increase in effect between these doses.

Most settings where the model-based estimate appreciably exceeds the observed sample mean/proportion at the highest dose are covered by our evaluation of non-monotone dose response, which is extensively discussed elsewhere. This occurs because the highest dose has leverage in the model fitting, so the model tends to fit the observed response at the highest dose well except when the monotonicity of the model makes this impossible. Based on visual inspection of the fitted dose response curves in Appendix C, the results for protocol A3071017, compound TAID=4, displays the largest discrepancy between the model fit and the sample mean without also displaying potential non-monotone behavior. The dose response for the compound was well studied and has high signal-to-noise (endpoint: HDL). There were 3 studies with QD dosing and 2 studies with BID dosing. The one potentially discrepant data point is from the QD dosing. Without accounting for multiplicity, this data point is of some concern because it lies slightly below the 95% posterior prediction interval, so a one-sided GOF test would yield a value below 0.025 (Our response to Comment 10 includes more information about using these intervals to form GOF tests where we have not pre-planned one.) In a similar situation in the future, we would likely evaluate exposure response to check if this discrepancy might be due to differences between the QD and BID regimens. However, all other data from the 5 studies are consistent with the model and they do not display a difference between QD and BID doses, so we think this is likely sampling variability expected from a search of over 200 examples with more than 4 sample means/proportions per compound. Note also that the lower sample mean in study A3011017 still exceeds the mean and model fit at the next lower dose, so the qualitative conclusion that dose response is increasing in the upper portion of the studied dosing range would not change if the less favorable sample mean value were used.

Question 9

In Appendix A, you stated that to simplify specification of prior distributions for the non-hierarchical parameters E0jk and Emaxj, the continuous endpoints were normalized to have an overall mean of 0 and SD of 1. Each binary endpoint was normalized so that the mean of the logit of the dose group proportions weighted by their sample sizes is 0. However, based on the clinDR package, it appears that there is no procedure for the data normalization. Please clarify this discrepancy. If the normalization is used, clarify how the data normalization affects the study results.

Response 9

Appendix A gives details of the prior distributions utilized in the meta-analysis of approximately 200 historical compounds. Hierarchical models are specified for the distributions of the $\ln(\lambda)$ and $\ln(ED_{50}/P_{50})$ across compounds. Hierarchical modeling of variation in placebo and drug response across compounds was not performed for reasons noted in Section 2. Instead, independent diffuse prior distributions were specified for these numerous parameters (at least 2 per compound, with additional placebo parameters when there are multiple studies for a compound). Specifying appropriately chosen distributions specific to each compound to achieve 'diffuse' prior distributions would require considerable programming and context review. An efficient approach to achieve this was to 'normalize' the data for each compound so the location (overall mean) and scale (SD) are the same. This allowed us to re-use one diffuse prior distribution for each placebo response parameter, and one diffuse prior distribution for each drug response. The final results were back-transformed to the original scales for reporting results of the meta-analysis. Note that this norming does not change the λ or ED_{50} parameters, which are invariant to these scale changes.

The hierarchical distributions for $\ln(\lambda)$ and $\ln(ED_{50}/P_{50})$ from the meta-analysis provide the empirical basis for the prior distribution of these parameters for future compounds. The distributions of the placebo and drug response from the meta-analysis are not used. When the prior distributions for the placebo and drug response parameters are specified for a new compound under study, the scale of the endpoint is known and very important. The prior distributions, whether diffuse or informative, are thus created specifically for the endpoint. There is no norming of the endpoint data, this is a computational convenience only utilized in the large-scale meta-analysis of the historical data.

Question 10

In Appendix C, you identified two compounds, TAID 33 or 2051, for potential non-monotone response based on the goodness of fit statistics. Yet simple visual inspections identified many cases, which raises questions about their model fittings (e.g., ID 2034, 2043, 2057, etc.), though they are not as extreme as TAID 33 or 2051. Clarify those compounds' model fits as well as model diagnostics.

Response 10

As noted in the response to comment 8a), compounds 33 and 2051 were highlighted in the report because we have additional evidence of non-monotone response for them, and they were excluded from the hierarchical modeling.

The compounds 2034 and 2057 have non-monotone GOF predictive fit probabilities of 0.091 and 0.078, respectively. These are external compounds, so we have little information available to further assess potential non-monotonicity for these compounds. For internal compounds, a reported fit probability, and plots like these would likely initiate a careful review. The power to detect non-monotonicity for any given compound is typically limited. However, if non-monotone trends were prevalent, we should observe an excess of lower fit probabilities (i.e., between 0.05 and 0.25) like those identified in comment 10. When viewed in aggregate, Figure 5 in the report shows that the number of compounds displaying such trends is what would be expected if the data from all compounds were generated from monotone Emax models. If there were fewer plots like those for compounds 2034 and 2057, it would provide evidence that the data had been sanitized and selectively reported. It is apparent that all doses studied in these examples were on the plateaus of the dose response curves. Under these conditions, it is more likely than not that the highest dose will not have the best sample mean. We suspect that these common poor dosing designs producing this sampling artifact have created the perception that non-monotone response is prevalent.

The outlier sample proportion for an interior dose of compound 2043 is again consistent with the deviations expected for sample means/proportions about the Emax fitted curves. As noted in response to comment 8a), we constructed specific GOF predictive probabilities to assess model deviations near the low and high doses, and to check for non-monotone response. A more general predictive check, which includes assessment of outliers of unspecified forms, was also supplied by plotting 95% posterior predictive intervals for each dose group sample mean/proportion. These are the grey intervals on the dose response plots. The outlier for compound 2043 is at the low end of the predictive interval so a corresponding unadjusted one-sided posterior predictive probability for this outlier is approximately 0.025. Placed into context, this is an unremarkable outlier because there are 208 compounds included in the hierarchical modeling with more than four dose groups per compound. Fewer such outliers would again be evidence of selective reporting. Compound 2043 is also an external compound so we have little additional information to assess the possibility that this outlier could be invalid data or be due to inadequacy of the Emax model. We do have considerable experience with immunology compounds for treating rheumatory arthritis, however, and compounds in this therapeutic area using the ACR20 endpoint have not displayed unusual dose response, so we think it likely this outlier is due to the expected sampling variability.

References

Cook, S., Gelman, A. and Rubin, D. (2006), Validation of Software for Bayesian Models Using Posterior Quantiles, Journal of Computational and Graphical Statistics, 15, 675-690.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., Neuenschwander, B. (2014), Robust Meta-Analytic-Predictive Priors in Clinical Trials with Historical Control Information, Biometrics, 70, 1023-1032.

STAN Development Team (2015), RStan: The R interface to Stan.

Thomas, N., and D. Roy (2017). Analysis of Clinical Dose-Response in Small-Molecule Drug Development: 2009-2014. Statistics in Biopharmaceutical Research, 9 (2), 137-146.

RESPONSE TO FDA REQUEST FOR INFORMATION

received 12 Nov 2021

FDA Qualification Opinion
“Empirically Based Bayesian Emax Models for Dose Response Design and Analysis”

BACKGROUND

21 April 2021, Pfizer submitted a Fit-for-Purpose (FFP) Request for an FDA Qualification Opinion on "Empirically Based Bayesian Emax Models for Dose Response Design and Analysis." On 14 September 2021, Pfizer received comments and requests for information on this submission from the FDA and submitted a response. On 12 November 2021, Pfizer received additional comments and requests for information based on the Pfizer response. The Pfizer responses to these follow-up comments are provided below. The question numbering of the 14 September 2021 FDA request is maintained; not all questions received follow-up comments from the FDA.

AGENCY REQUEST FOR INFORMATION AND SPONSOR RESPONSE

Clinical Pharmacology

Question 1 Follow-up

We acknowledge that you conducted a comparison between maximum likelihood or non-linear least squares estimation with the proposed Bayesian method in section 3.3.

- "Population analyses" refers to a typical practice of implementing a non-linear mixed effect model included in PK/PD packages such as NOMEM, Monolix, etc. Model parameters can be fixed or random. Relevant covariates can be used to explain partial random effect. This feature is helpful when multiple studies are included in the analyses.

- We understand that the poor performance of the proposed Bayesian method under a setting of using diffusive priors for all model parameters seems unfavorable. However, we believe it is important to ensure comparable performance of the Bayesian method under non-informative priors with the frequentist approach since it is not feasible to evaluate whether the priors are beneficial enough as the magnitude of informativeness will depend on the data from the new drug.

- We speculate that the poor performance of the Bayesian method under non-informative priors is likely due to reparameterization including EC50 to log(EC50/P50) and Emax to difTarget. The former used a historical P50 as a constant scaler which may not be efficient and the latter created a complicated correlation among the model parameters. We recommend you investigate the potential concerns and make efforts to further improve the model to facilitate the review.

Response 1

The Pfizer response to the FDA follow-up comment is provided in the attached file "diffuseTabs.pdf".

Statistics

Question 2 Follow-up

FDA Comment to response 2:

- In Sections 2 and 3, you described how you derived parameters for the prior distributions of the ln(λ) and ln(ED50/P50) based on the historical data. However, it is still not clear how you will implement these priors and correlation structures in a future application to a new compound. It appears that the source code in a *ClinDR* package includes this information. Provide in a written form how you will implement the priors for the ln(λ) and ln(ED50/P50) and their correlation structures in the future application.

Response 2

The Bayesian model is implemented in STAN using the rstan interface to R in the clinDR package. The STAN code implementing the model is in the file mrmodel.pdf, which is included in our response and distributed as part of the package source code. It can be compiled once [compileStanModels()] when the package is installed, which substantially speeds up subsequent use. The model code is dense because it implements several options: continuous and binary data, sigmoidal and hyperbolic Emax models, optional baseline covariates, 0, 1, or multiple E0 intercepts, and patient-level and grouped data. We describe the basic use case with a sigmoid model, patient-level data, and one intercept (protocol) with no baseline covariates. Applications of the prior distribution like those for future applications are in Section 3.2 of the original submission.

The prior means, scale parameters, and correlation defining the prior distribution are collected into a 'prior' object that is passed to the STAN model using a package function:

```
emaxPrior.control(epmu=NULL,epsca=NULL,
        difTargetmu=NULL,difTargetsca=NULL,
        dTarget=NULL,p50=NULL,
        sigmalow=NULL,sigmaup=NULL,
        effDF=parmDF,parmDF=5,
        loged50mu=0.0,loged50sca=1.73,
        loglammu=0.0,loglamsca=0.425,parmCor=-0.45)
```

The default values, which were derived from the analysis of the large collection of dose response studies, are strongly recommended. They are ordinarily changed only for testing and sensitivity analyses. The prior mean and scale parameters are given on the logit scale when the response is binary.

Prior distributions for E0 and difTarget:

These are described first because they are the least complex. They are also described in the clinDR documentation. Both parameters have independent t-distributions.

emaxPrior.control	STAN Role	Description
epmu	epmu	prior mean of intercept
epsca	epsca	t-distribution scale parameter for the intercept
effDF[1]	e0DF	Degrees of freedom for the prior t-distribution
dTarget	dTarget	Dose matching difTarget efficacy
difTargetmu	difTargetmu	Prior mean of difTarget at dTarget
difTargetsca	difTargetsca	t-distribution scale parameter for difTarget
effDF[2]	diftDF	Degrees of freedom for the prior t-distribution

The code implementing these priors is:

```
e0[1]~student_t(e0DF,epmu,epsca);

difTarget~student_t(diftDF,difTargetmu,difTargetsca);
```

Note that parameters are independent unless they are given a multivariate distribution or share a common stochastic parent in their definition.

Prior distributions for the ln(λ) and ln(ED50/P50):

emaxPrior.control	STAN Role	Description
loged50mu	loged50mu	Prior mean of log(ED50/P50) [typically 0]
loged50sca	loged50sca	t-distribution scale parameter for the log(ED50/P50)
loglammu	loglammu	Prior mean of log(λ) [typically 0]
loglamsca	loglamsca	t-distribution scale parameter for the log(λ)
parmCor	parmCor	Prior correlation of log(ED50/P50) and log(λ)
parmDF	parmDF	DF for the bivariate t-distribution

The code implementing the bivariate-t prior distribution first assigns the input values to bivariate vectors and a 2x2 variance-covariance matrix:

```
scavar[1,1]=square(loged50sca);

scavar[1,2]=loged50sca*loglamsca*parmCor;

scavar[2,1]=scavar[1,2];

scavar[2,2]=square(loglamsca);


cscavar=cholesky_decompose(scavar);

zvec[1]=0.0;

zvec[2]=0.0;

muvec[1]=loged50mu;

muvec[2]=loglammu;
```

The Cholesky decomposition is computed to accelerate STAN execution because it can be done once here, and not repeated for each MCMC simulation. zvec is just a bivariate zero vector for computational convenience. Recall that muvec will also be zero in most applications.

The one minor complication is that STAN can generate draws from a multivariate normal, but it does not have a built-in function for the multivariate t-distribution. This can easily be created by generating an independent chi-square variable, and then scaling the multivariate normal draw by the square root of the chi-square variable divided by its DF:

```
parmvec~multi_normal_cholesky(zvec,cscavar);

chi2var[1]~chi_square(parmDF);

parmvect=muvec+parmvec/sqrt(chi2var[1]/parmDF);

ed50=p50*exp(parmvect[1]);

loglambda=parmvect[2];

lambda=exp(parmvect[2]);
```

Finally, the Emax parameter can be derived from the other generated parameters:

emax=(difTarget)*(ed50^lambda+dTarget^lambda)/dTarget^lambda;

For continuous endpoints, a prior distribution for the residual SD is also required. The lower and upper bounds of a uniform distribution are also part of the emaxPrior.control object:

sigma[1]~uniform(sigmalow,sigmaup)

Likelihood

First, the mean (or logit probability) is computed for each patient. For continuous data, the SE for the mean is also computed. With patient-level data, nv[i] is 1.

```
for(i in 1:N){
  emx[i] = e0[1] + (emax *
          pow(dv[i],lambda))/(pow(ed50,lambda)
         +pow(dv[i],lambda));
  sex[i]=sigma[1]/sqrt(nv[i])
 }
```

The 'pow' function evaluates the first argument raised to the power given in the second argument. STAN then implements the likelihood in vectorized form

For continuous data:

yv ~ normal(emx,sex);

For binary data:

yvb ~ binomial(nvb,inv_logit(emx))

The mrmodel code file is attached as a pdf.

Question 6 Follow-up

FDA Comment to response 6:

- Clarify whether the weighted normal likelihoods were utilized only in a meta-analysis of historical data and will not be used in a future application to a new compound.

Response 6

The weighted normal likelihood and its use approximating logit binary response data was only for the meta-analysis of historical data. The response to the first question provides details on the implementation of the Bayesian method to future dose response data sets. The only exception to this statement occurs if the method is applied to external data sets where only summaries at the dose group level are available. The code in the clinDR package can fit such data, and for continuous endpoints, this is a weighted normal likelihood. For binary data, even when only summary data are available, the binomial distribution is utilized, not the weighted normal approximation to it.

Question 7 Follow-up

FDA Comment to response 7:

- In our email dated September 14, 2021, we requested that you provide posterior/predictive posterior estimates for difTarget. On September 28, 2021, you responded that you had not planned to tabulate the results for difTarget because this would yield 208*8=1664 rows (eight models were evaluated) and the posterior distributions of difTarget for the historical compounds did not have a direct role in future analyses. A prior sensitivity analysis of difTarget is needed to assess the uncertainty of the parameter. Provide the sensitivity analysis for difTarget for a specific compound using the eight models you considered.

Response 7

We supply summaries of the posterior distributions derived from the eight different prior models for two well-known compounds. The first is atorvastatin (TAID=6), a small molecule with continuous endpoint (percent change in LDL), high signal-to-noise dose response (effect size for highest dose >5), and a 32-fold dosing range studied. The second is adalimumab (TAID=2002), a biological with a binary endpoint (clinical remission) for the indication of Crohn's disease, a moderate signal-to-noise (observed responder proportions of 0.36 and 0.12 for the high dose and placebo), and only a 4-fold dosing range studied. These compound characteristics cover much of the range of the compounds in the meta-data. Percentiles from the posterior distribution corresponding to the prior distributions are summarized in Table 1 (atorvastatin) and Table 2 (adalimumab). The prior models are described in Appendix A of the submission document. The posterior distributions are in practical agreement, and do not display systematic differences across the two sampled compounds. Note when reviewing the results in Table 2, the difTarget parameter is the placebo-adjusted difference on the logit-transformed scale.

Table 1. Percentiles of the posterior distribution corresponding to the eight prior models: atorvastatin (TAID=6)

	mod1	mod2	mod3	mod4	mod51	mod52	mod101	mod102
5%	-70.9	-70.1	-71.0	-70.4	-71.7	-72.0	-69.0	-68.1
10%	-69.5	-68.7	-69.7	-69.2	-70.4	-70.5	-67.7	-66.9
25%	-67.2	-66.4	-67.5	-66.9	-68.0	-68.2	-65.6	-64.6
50%	-64.6	-63.9	-65.0	-64.4	-65.4	-65.6	-63.2	-62.2
75%	-62.2	-61.3	-62.5	-61.9	-62.8	-63.0	-60.9	-59.8
90%	-60.1	-58.9	-60.3	-59.6	-60.5	-60.7	-58.8	-57.5
95%	-58.8	-57.6	-58.9	-58.3	-59.1	-59.2	-57.7	-56.1

Table 2. Percentiles of the posterior distribution corresponding to the eight prior models: adalimumab (TAID=2002)

	mod1	mod2	mod3	mod4	mod51	mod52	mod101	mod102
5%	0.66	0.65	0.54	0.75	0.67	0.68	0.66	0.66
10%	0.79	0.79	0.71	0.88	0.81	0.82	0.79	0.81
25%	1.03	1.03	0.98	1.10	1.06	1.06	1.03	1.03
50%	1.30	1.29	1.27	1.34	1.33	1.34	1.30	1.31
75%	1.57	1.56	1.55	1.59	1.60	1.61	1.57	1.58
90%	1.84	1.81	1.81	1.82	1.85	1.86	1.83	1.83
95%	1.99	1.97	1.96	1.97	2.01	2.01	1.98	1.98

Additional FDA comment

For the proposed Fit For Purpose review, simulation results for meta-analyses may not be necessary since the package does include the meta-analyses with real data. However, simulation studies for meta-analyses are recommended for a complete manuscript. During the design phase for future drug development programs, we recommend that you complete simulation studies demonstrating your method is applicable for multiple studies with different sample size or different levels of heterogeneity.

Response to Additional Comment

Please see the attached file “combStudies.pdf”.

```
data {
    int<lower=0> N; // number of patients/groups
    int<lower=0> nprot; // number of protocols-strata
            int<lower=0> protv[N];
            int<lower=0,upper=1> cont;  //indicator of data type
            int<lower=0,upper=1> sigmoid;  //indicator of model type
            int<lower=0,upper=1> gp;
            int<lower=0,upper=1> intercept;
            int<lower=0> nbase;

            //continuous data
    real yv[N]; //  outcome variable
            real<lower=0> nv[N]; // number patients per group; real for
arithmetic
            //binary data
    int yvb[N]; //  outcome variable
            int<lower=0> nvb[N]; // number patients per group

    real<lower=0> dv[N];
    matrix[nbase ? N : 0,nbase ? nbase : 0] xbase;

            real<lower=0> df2;   // saturated df divided by 2
            real<lower=0> ssy;
    real epmu;
            real<lower=0> epsca;
    real difTargetmu;
            real<lower=0> difTargetsca;
            real<lower=0> dTarget;
            real<lower=0> sigmalow;
            real<lower=0> sigmaup;
            real<lower=0> p50;
    real<lower=0> loged50mu;
    real<lower=0> loged50sca;
    real<lower=0> e0DF;
    real<lower=0> diftDF;
    real<lower=0> parmDF;
    real<lower=0> loglammu;
    real<lower=0> loglamsca;
            real parmCor;
            vector[nbase ? nbase : 0] basemu;
            cov_matrix[nbase ? nbase : 1] basevar;
}
transformed data{
            cholesky_factor_cov[nbase ? nbase : 1] cbvar;
            vector[2] zvec;
            vector[2] muvec;
            cov_matrix[2] scavar;
            cholesky_factor_cov[2] cscavar;

            cbvar=cholesky_decompose(basevar);

            if(sigmoid){
                        scavar[1,1]=square(loged50sca);
                        scavar[1,2]=loged50sca*loglamsca*parmCor;
```

```
                        scavar[2,1]=scavar[1,2];
                        scavar[2,2]=square(loglamsca);
                        cscavar=cholesky_decompose(scavar);
                        zvec[1]=0.0;
                        zvec[2]=0.0;
                        muvec[1]=loged50mu;
                        muvec[2]=loglammu;
            // assign defaults to avoid compiler error
            }else{
                        scavar[1,1]=1.0;
                        scavar[1,2]=0.0;
                        scavar[2,1]=scavar[1,2];
                        scavar[2,2]=1.0;
                        cscavar=cholesky_decompose(scavar);
            }
}
parameters{
    vector[intercept ? nprot : 0] e0;
    real difTarget;
           vector[sigmoid ? 2 : 1]parmvec;
           vector<lower=sigmalow,upper=sigmaup>[cont ? 1 : 0] sigma;
           vector[nbase ? nbase : 0] bslope;
           vector<lower=0>[sigmoid ? 1 : 0] chi2var;
}
transformed parameters{
    real <lower=0> lambda;
           real loglambda;
    real<lower=0> ed50;
           real led50;
           real emax;
           real<lower=0> tau2[cont ? 1 : 0];
           vector[sigmoid ? 2 : 1]parmvect;

    if(sigmoid){
                parmvect=muvec+parmvec/sqrt(chi2var[1]/parmDF);
                ed50=p50*exp(parmvect[1]);
                loglambda=parmvect[2];
      lambda=exp(parmvect[2]);
    }else{
                parmvect=loged50mu+parmvec;
                ed50=p50*exp(parmvect[1]);
                loglambda=0.0;
                lambda=1.0;
           }
           led50=log(ed50);
           if(cont){
                tau2[1]=1/(2*sigma[1]*sigma[1]);
           }
           emax=(difTarget)*(ed50^lambda+dTarget^lambda)/dTarget^lambda;
}

model{
    vector[N] emx;
           vector[N] sex;
```

```
            if(intercept){
                  for(i in 1:nprot){
                        e0[i]~student_t(e0DF,epmu,epsca);
                  }
            }
      if(sigmoid){
                  parmvec~multi_normal_cholesky(zvec,cscavar);
                  chi2var[1]~chi_square(parmDF);
            }else parmvec[1]~student_t(parmDF,0.0,loged50sca);
      if(cont){
                  sigma[1]~uniform(sigmalow,sigmaup);
      }

      difTarget~student_t(diftDF,difTargetmu,difTargetsca);

      for(i in 1:N){
                  if(intercept){
                              emx[i] = e0[protv[i]] + (emax *
pow(dv[i],lambda))/(pow(ed50,lambda)

        +pow(dv[i],lambda));
                  }else{
                              emx[i] = (emax *
pow(dv[i],lambda))/(pow(ed50,lambda)

        +pow(dv[i],lambda));
                  }
                  if(cont) sex[i]=sigma[1]/sqrt(nv[i]);
      }

            if(nbase){
                  bslope~multi_normal_cholesky(basemu,cbvar);
                  emx = emx + xbase*bslope; //emx;
            }

      if(cont){
            yv ~ normal(emx,sex);
                  if(gp){
                        ssy ~ gamma(df2,tau2[1]);
                  }
            }else{
                  yvb ~ binomial(nvb,inv_logit(emx));
            }
}
```

1 Diffuse prior distribution

The simulation studies described in Sections 3.3.1, 3.3.2 and Appendix H of the original submission were repeated using a diffuse prior distribution for the E_{max} model parameters. Independent uniform prior distributions were applied to each parameter. The parameters were not normalized (e.g., the P_{50} was not used) or transformed except that the uniform prior distribution was applied to $\log(ED_{50})$ rather than ED_{50}, as is commonly done. The E_{max} parameter was used instead of the difTarget parameter. The priors were chosen to be diffuse with respect to the dose response designs and parameters in the simulation study, but the uniform bounds were not set to extremely large values to avoid numerical underflow and overflow problems:

$$\begin{aligned} E_0 &\sim \text{uniform}(-20, 20) \ , \\ E_{max} &\sim \text{uniform}(-20, 20) \ , \\ \log(ED_{50}) &\sim \text{uniform}(-log(10000), log(10000)) \ , \\ \lambda &\sim \text{uniform}(0, 10) \ . \end{aligned}$$

The results are graphically summarized in the appendix at the end of this document. The Figure numbers in the appendix match the corresponding Figure numbers in the submission document. The contents and format of the results are described in Section 3.3 of the submission document. The results for the pairwise comparisons and modified ML estimation are unchanged. The results in the appendix are for the 'Diffuse Bayes Emax' method. In addition to comparing the Diffuse Bayes Emax performance to the pairwise comparision and modified ML methods, the Diffuse Bayes results can be compared to the Bayes Emax results using the more informative prior distributions in the submission document with the same Figure number. Note however that the scales on the plot may differ substantially. Common scales or merged plots were not attempted because the scales are often so different that the results for most methods would compressed into a small portion of the display to include one ill-behaved method. Note also that the simulated population models are defined relative to the informative prior distribution (e.g., how much the ED_{50} differs from the P_{50}), but these values are not used in the construction of the diffuse prior distribution.

With the diffuse prior distribution, the MCMC numerical methods evaluating the posterior distributions perform poorly in many of the simulation settings. Diagnostic checking of the MCMC draws from the posterior distributions raise concerns for many of the simulated data sets. This is consistent with the failure the numerical optimization methods to converge to maximum likelihood estimates in a majority of the simulated data sets for a sigmoidal E_{max} model. Unlike the modified ML approach, however, no model substitution or other alterations (e.g., fixing a parameter value) were implemented for the diffuse Bayesian method. As noted below, the MCMC methods perform worse in the realistic settings with fewer doses and lower signal-to-noise. All of the results for the diffuse Bayesian approach should be interpreted cautiously considering their problems with numerical evaluation. Note that the numerical problems are NOT due to the choice of parameterization. They occur because the E_{max} model is not identified in many applications. When com-

bined with diffuse prior distributions, the resulting posterior distributions are far from the asymptotic normal ideal and very difficult to numerically evaluate.

2 Low information settings

The low information settings in the simulation studies were selected to represent the most common situations in practical applications. With few exceptions, the diffuse Bayesian Emax model approach performed very poorly in absolute terms, and when compared to the other methods (even pairwise comparisons). The modified ML approach also performs poorly, but it is typically better than the diffuse Bayesian approach that proceeds even when there is evidence that the numerical methods implementing it have failed. While it was not attempted here, it is likely that performance of the diffuse Bayes approach would be more similar to the modified ML approach if similar ad hoc methods to swap in easier-to-estimate models yielding nominally good fits to data were developed.

The Bayesian Emax modeling using the more informative empirically-based prior distribution performed well, and superior to the other methods over a very wide range of conditions, including most of the low-information settings. It does not require post hoc adjustments. As expected, it is most advantageous when the prior distribution is accurate. Simulated conditions that deviated from the informative prior distribution were intentionally included to quantitate when performance of the informative method would deteriorate. The most pronounced deterioration occurred with unusually steep dose response (high λ, for example, Figures 43, 51). The primary purpose of the extensive evaluation of past dose response studies was to conservatively ensure that the deterioration, such as that in Figures 43 and 51, will be rarely encountered in practice.

3 High information settings

High information settings include more well-targeted doses and higher signal-to-noise. There are studies in the dose response meta-data with similar conditions, but such conditions are the exceptions. The likelihood and posterior distributions are closer to the asymptotic normal ideal in these settings, so there are fewer and less severe numerical problems with the diffuse Bayesian methods. The results for the diffuse Bayes and modified ML estimation are more similar, and neither uniformly dominates the other as measured by root mean squared error (MSE). The diffuse Bayes posterior intervals do have better repeated-sampling coverage probabilities than the modified ML methods in these settings.

The performance of the diffuse and informative Bayesian methods are also more similar. As expected, the more informative prior distribution yielded better results in most settings, except those already noted (e.g., Figures 43, 51) when the population model deviated from the prior distribution.

A few additional simulations (not shown) were conducted with additional dose groups and very high signal-to-noise that yield the theoretical asymptotic normal behavior of the estimated dose response modeling. The three estimation methods perform similarly and

well in these settings. Simulations in these settings are useful for testing the computing code, but they are not indicative of performance in real applications.

Appendix A Plots of performance with diffuse prior distributions

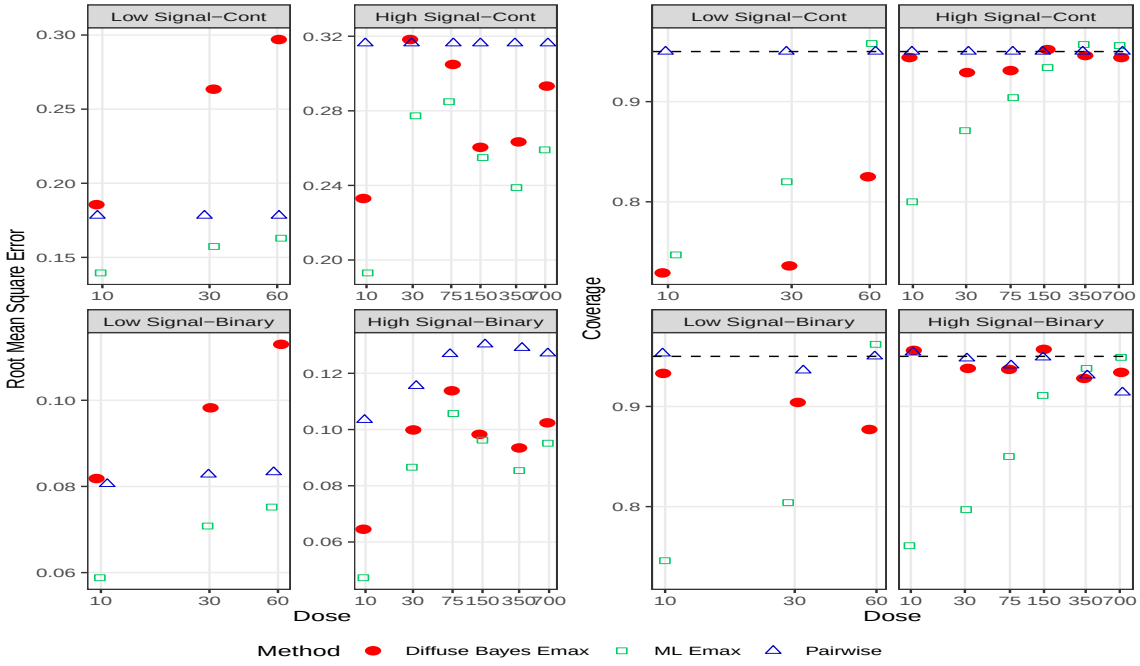


Figure 13: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 2P_{50}$, $\lambda = 1$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

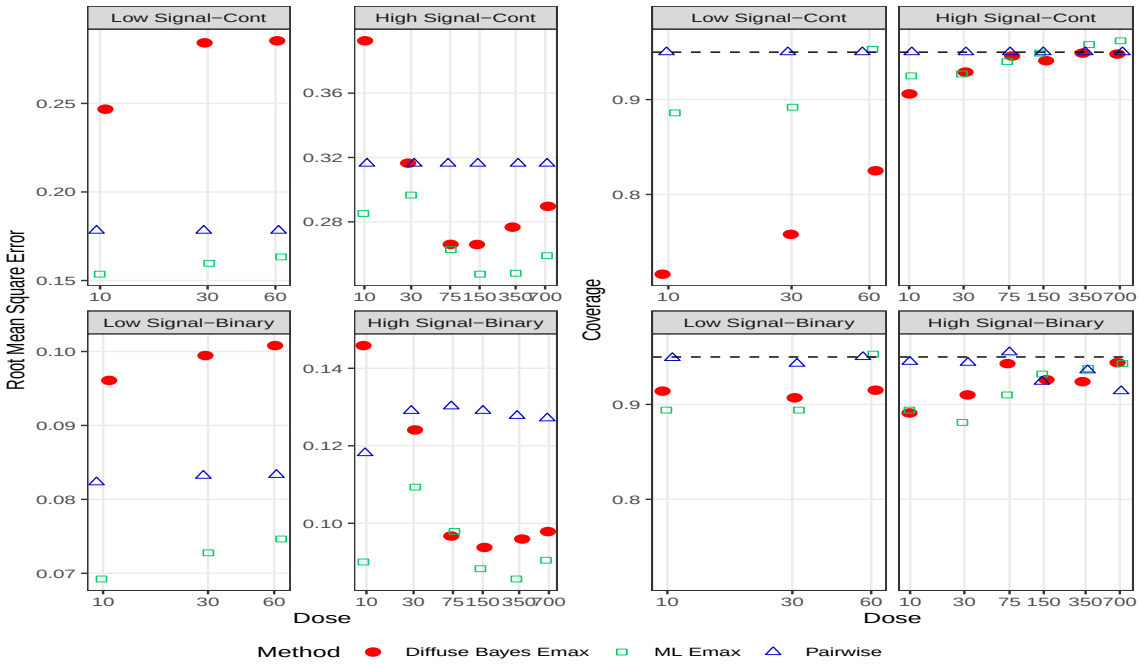


Figure 15: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 0.5P_{50}$, $\lambda = 1$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

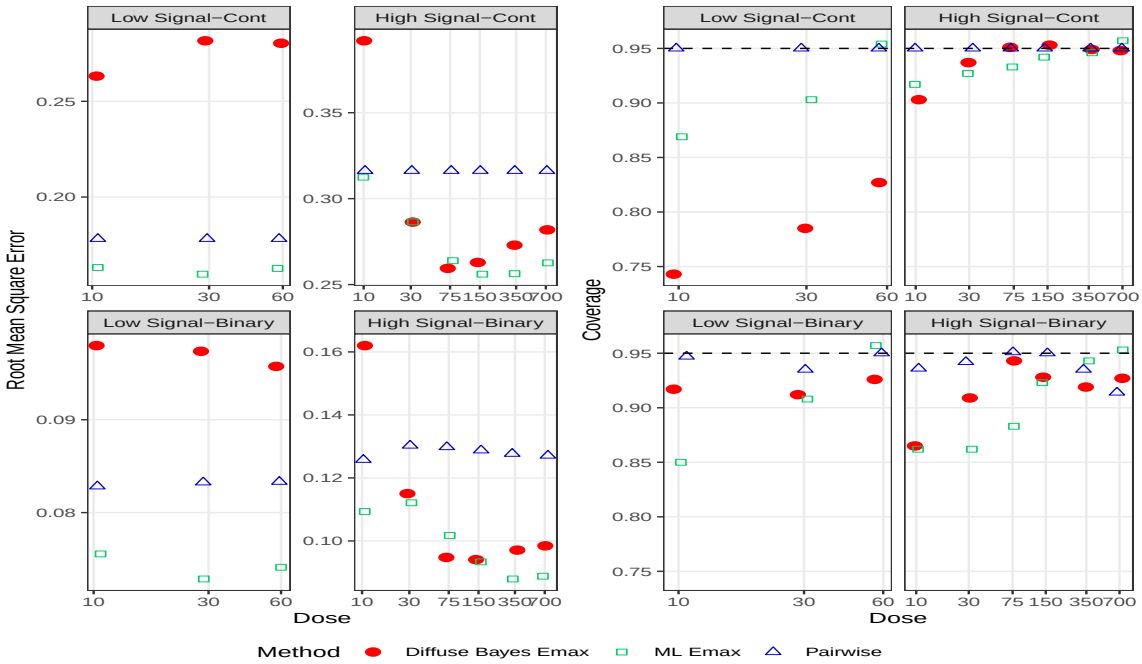


Figure 17: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = P_{50}/4$, $\lambda = 0.8$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

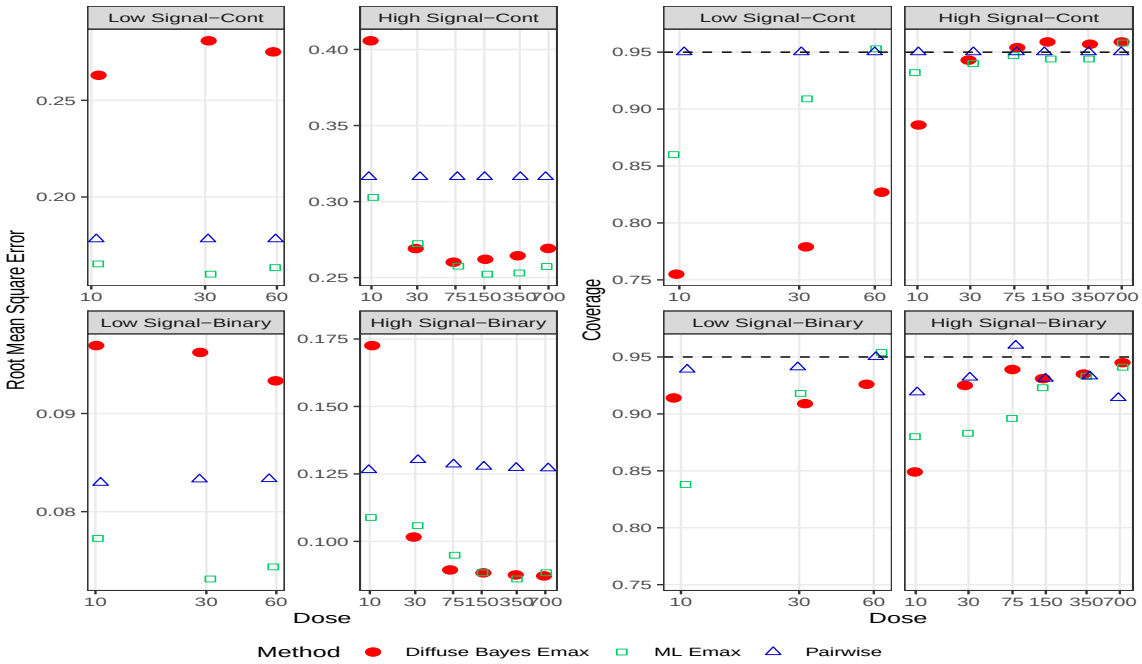


Figure 19: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = P_{50}/4$, $\lambda = 1.25$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

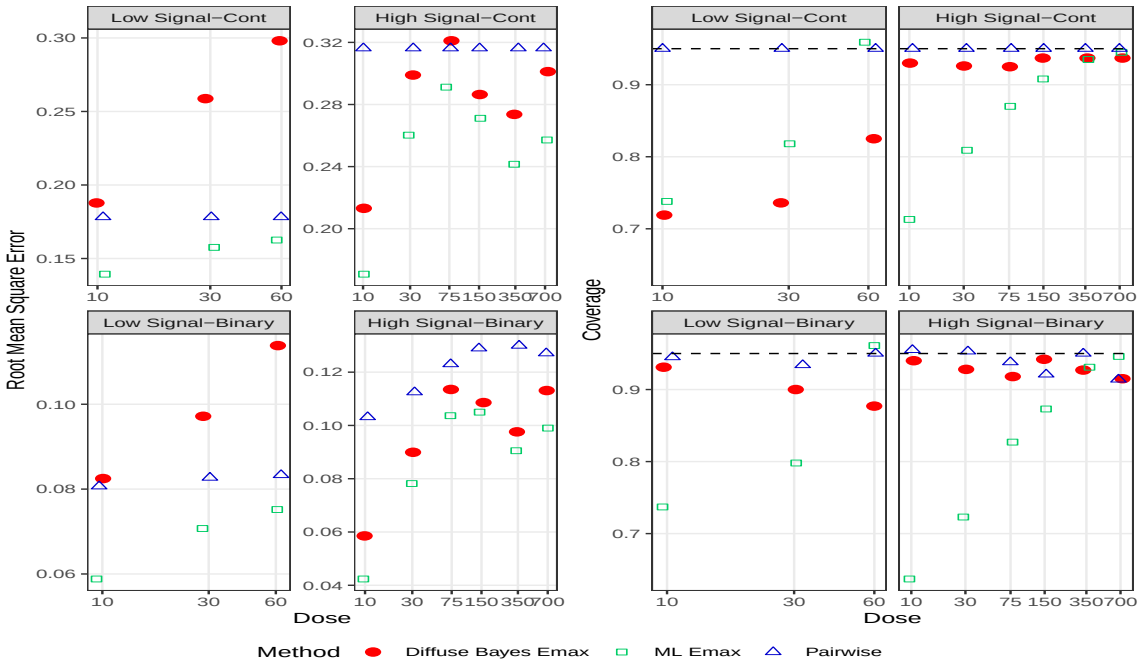


Figure 21: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 4P_{50}$, $\lambda = 0.8$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

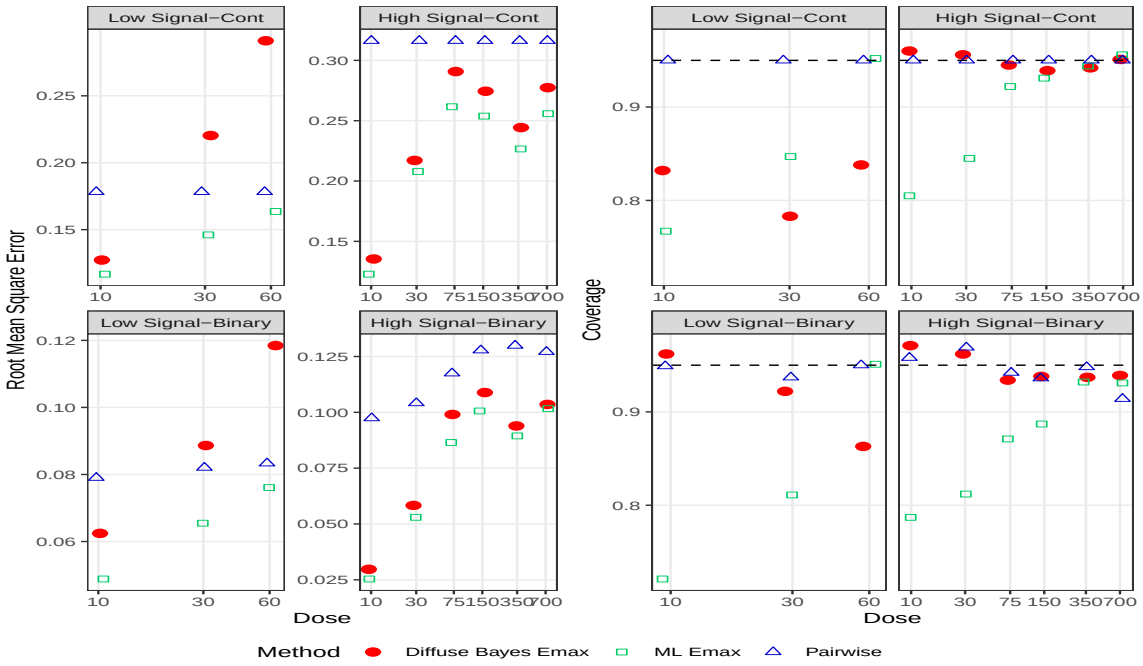


Figure 23: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 4P_{50}$, $\lambda = 1.25$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

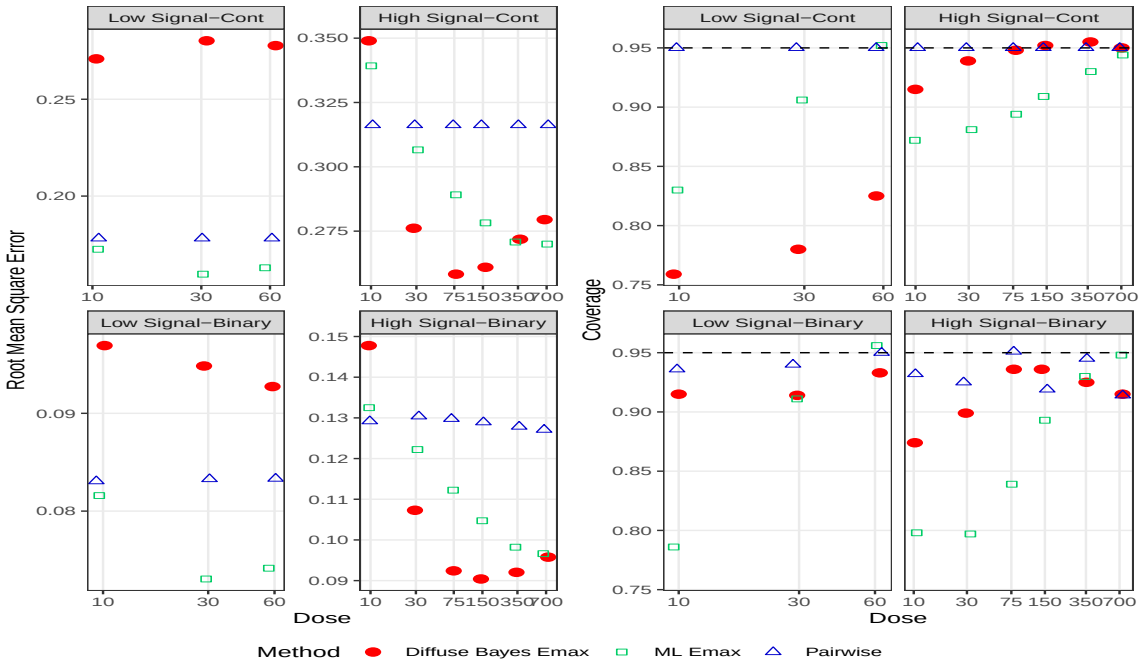


Figure 29: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = P_{50}/10$, $\lambda = 0.5$. For the continuous endpoint, $E_0 = 1$ and difTarget = 0.5, 1.5. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

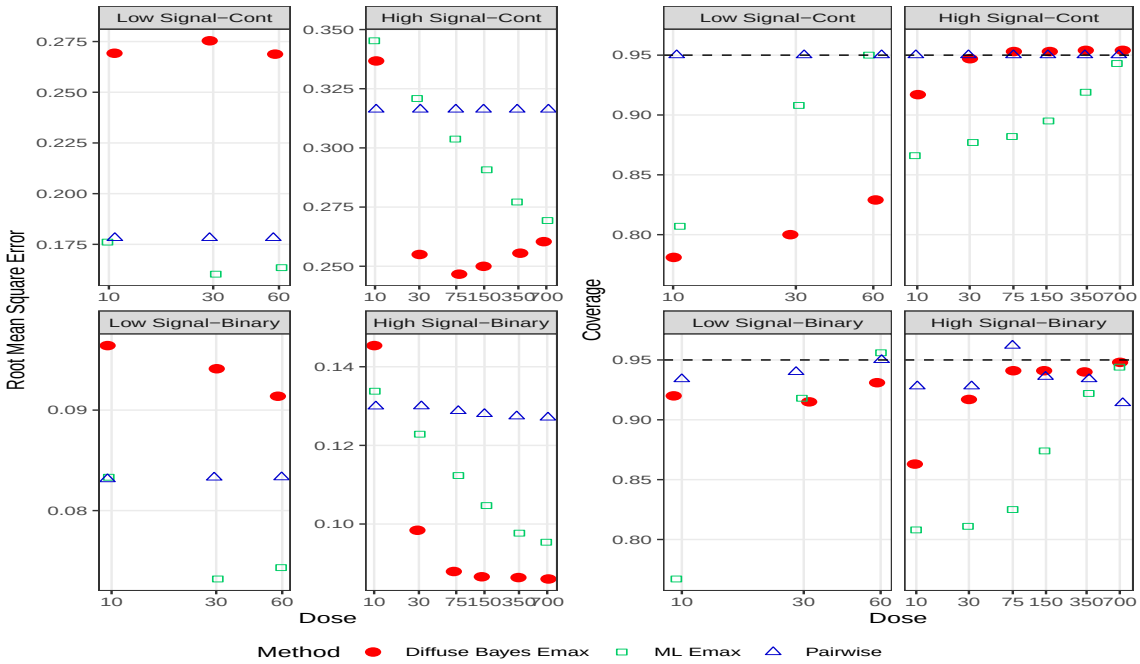


Figure 31: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = P_{50}/10$, $\lambda = 0.8$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

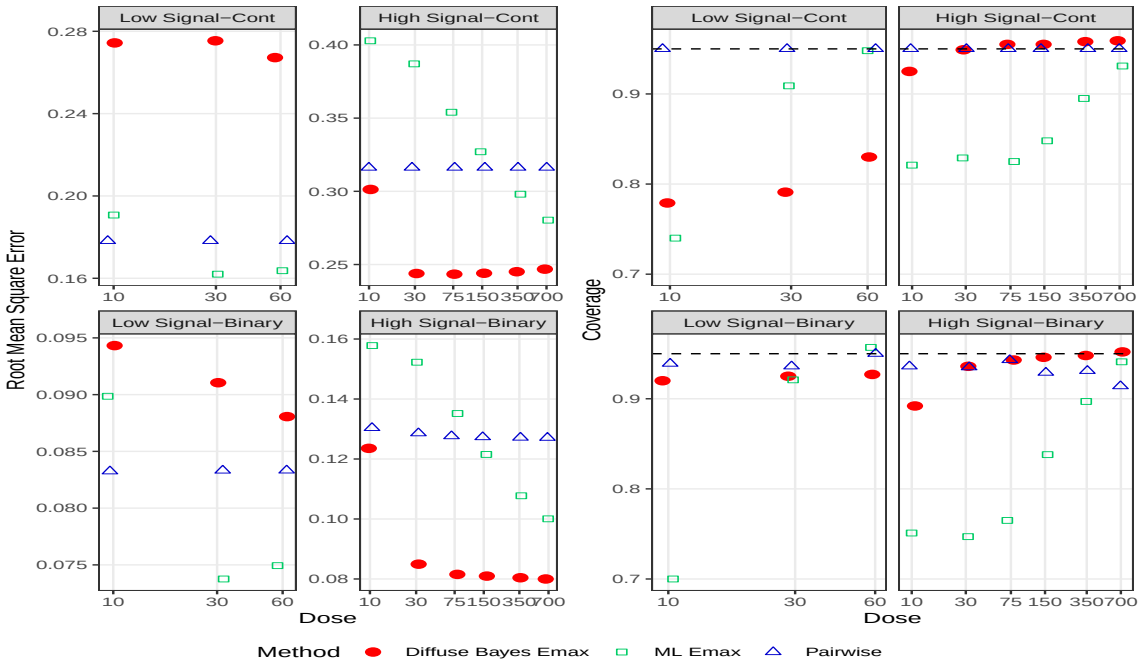


Figure 33: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = P_{50}/10$, $\lambda = 1.25$. For the continuous endpoint, $E_0 = 1$ and difTarget = 0.5, 1.5. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

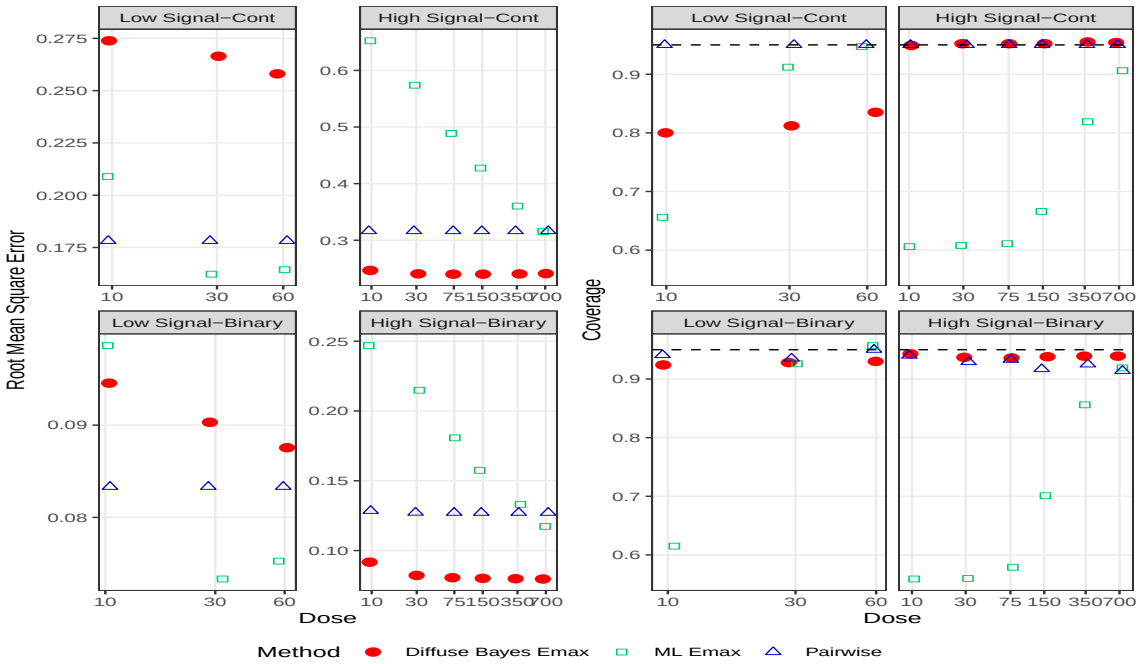


Figure 35: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = P_{50}/10$, $\lambda = 2.5$. For the continuous endpoint, $E_0 = 1$ and difTarget = 0.5, 1.5. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

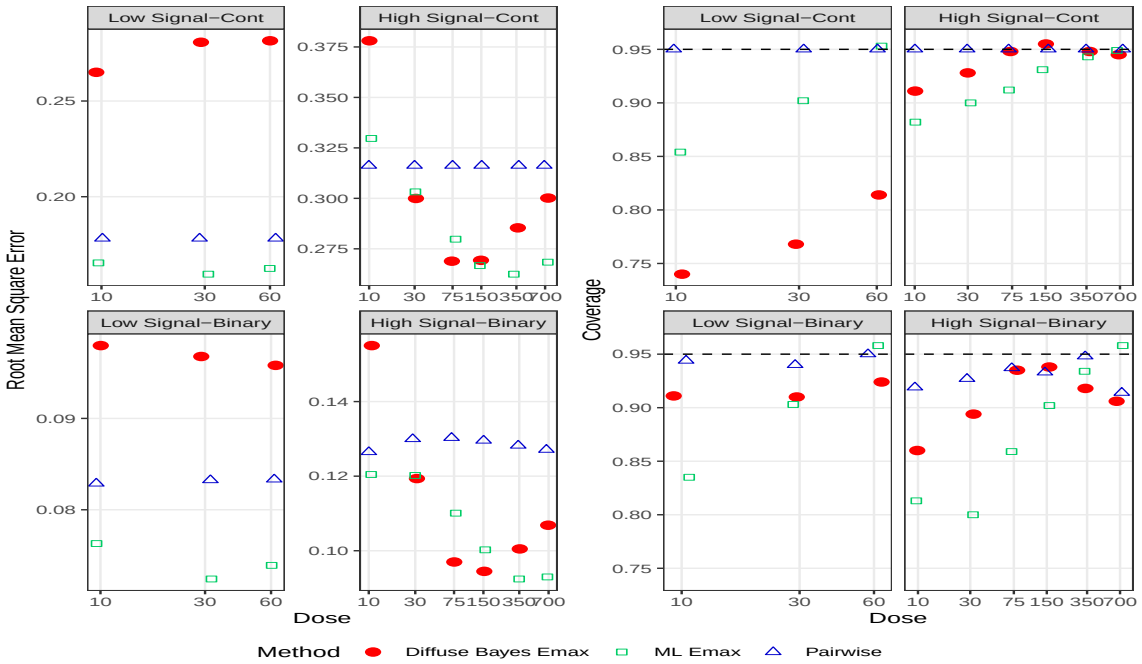


Figure 37: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = P_{50}/4$, $\lambda = 0.5$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

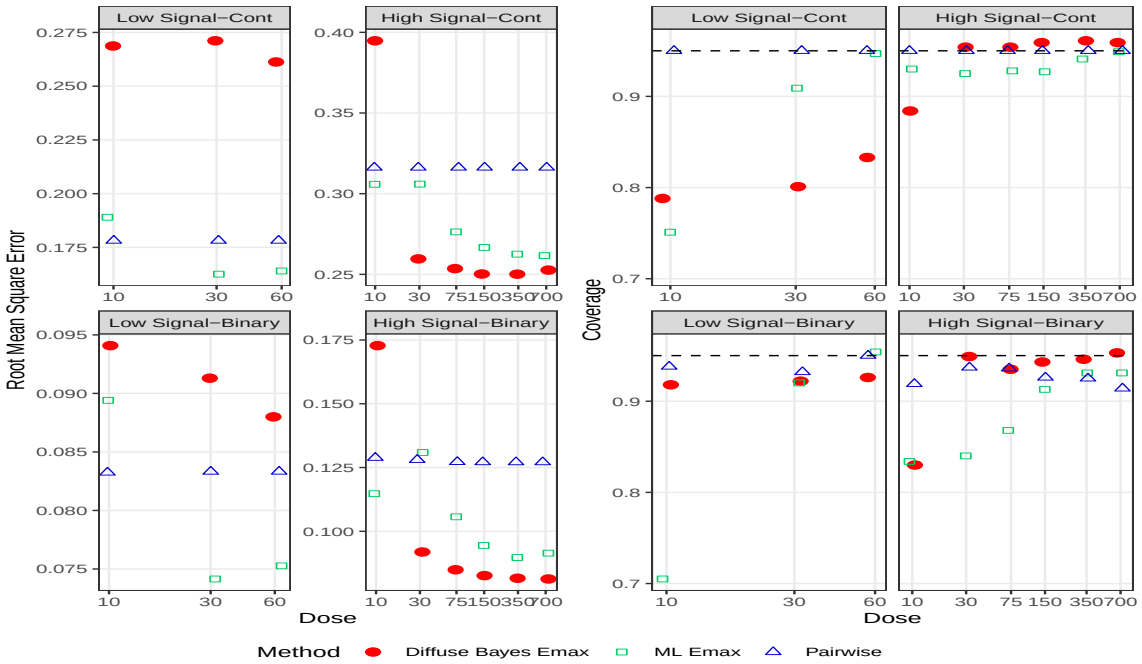


Figure 39: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = P_{50}/4$, $\lambda = 2.5$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

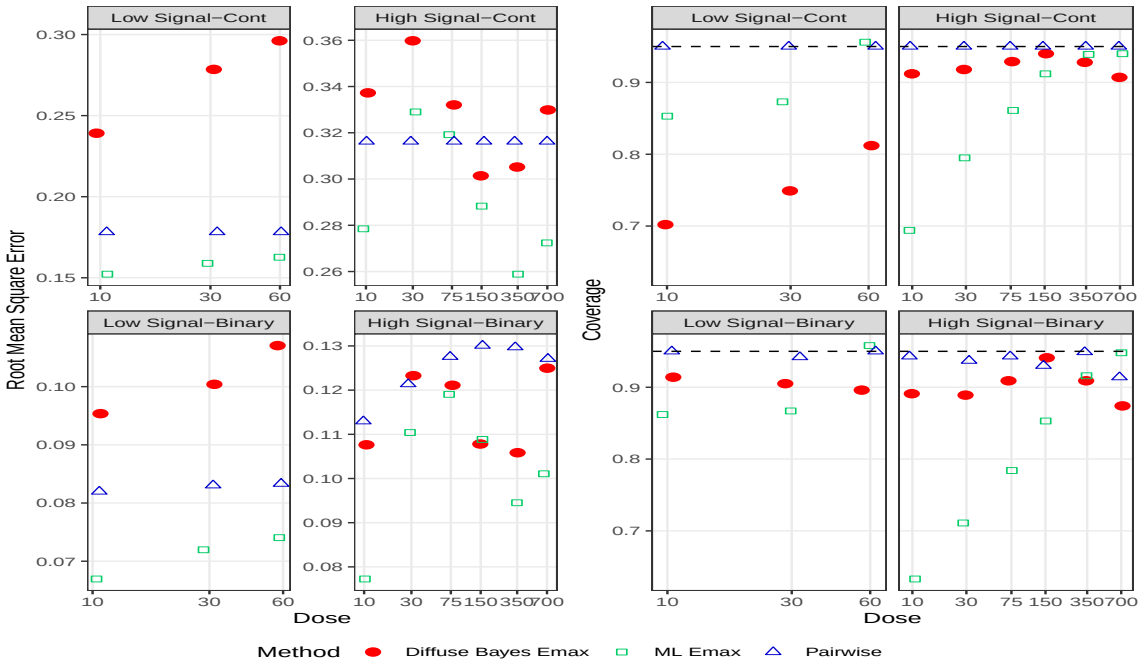


Figure 41: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 4P_{50}$, $\lambda = 0.5$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

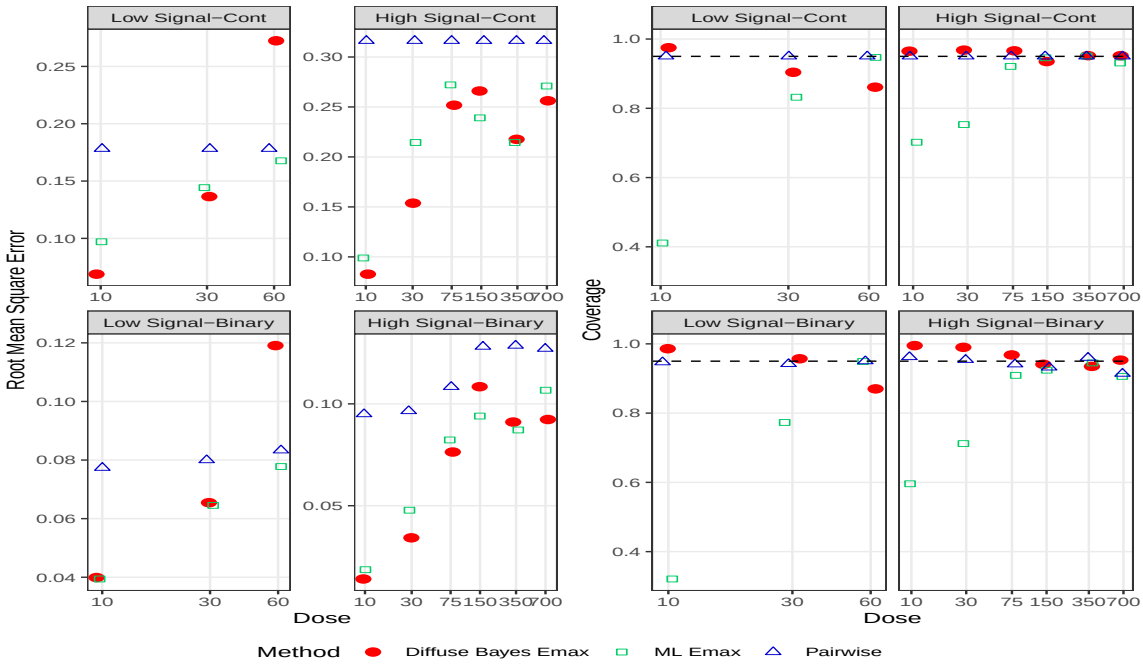


Figure 43: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 4P_{50}$, $\lambda = 2.5$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

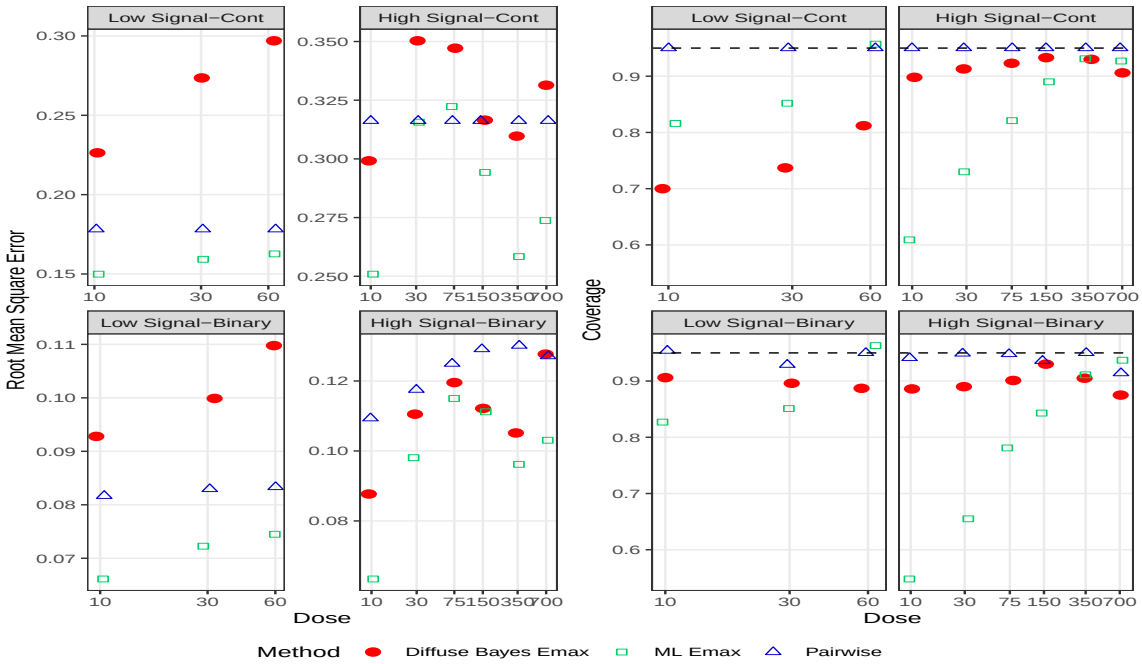


Figure 45: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 10P_{50}$, $\lambda = 0.5$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

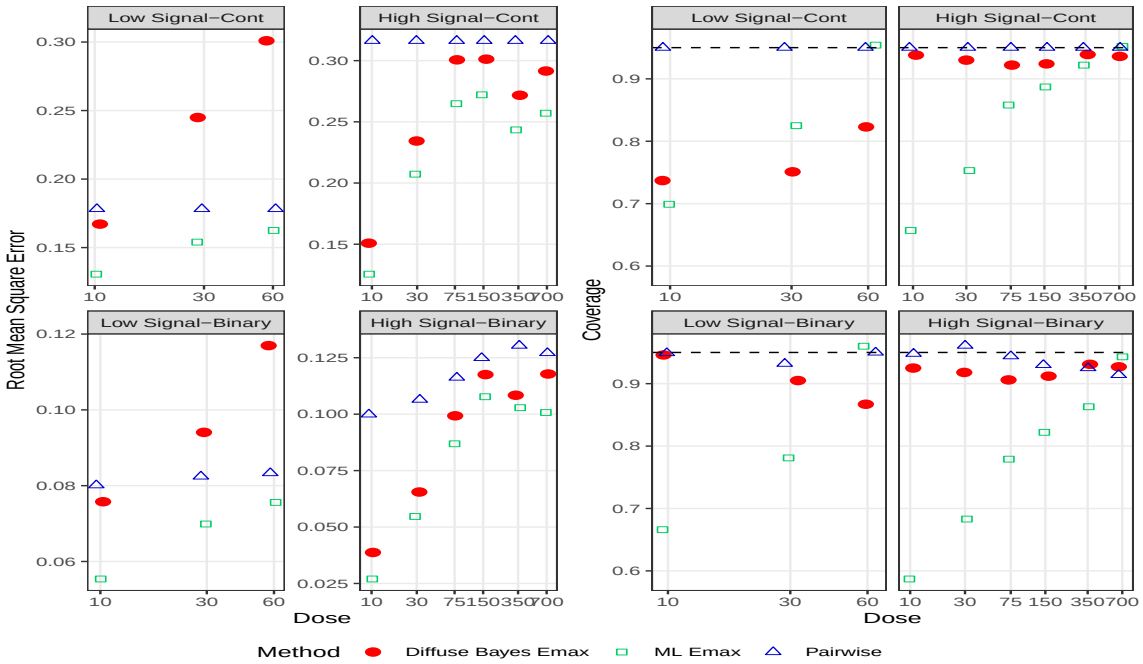


Figure 47: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 10P_{50}$, $\lambda = 0.8$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

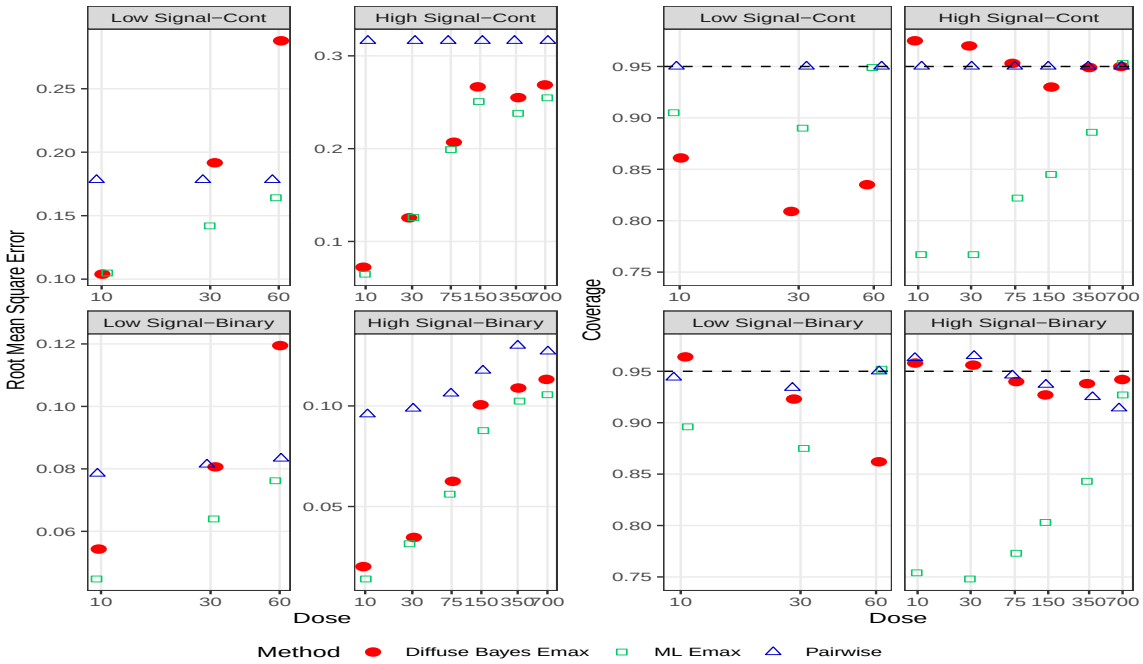


Figure 49: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 10P_{50}$, $\lambda = 1.25$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

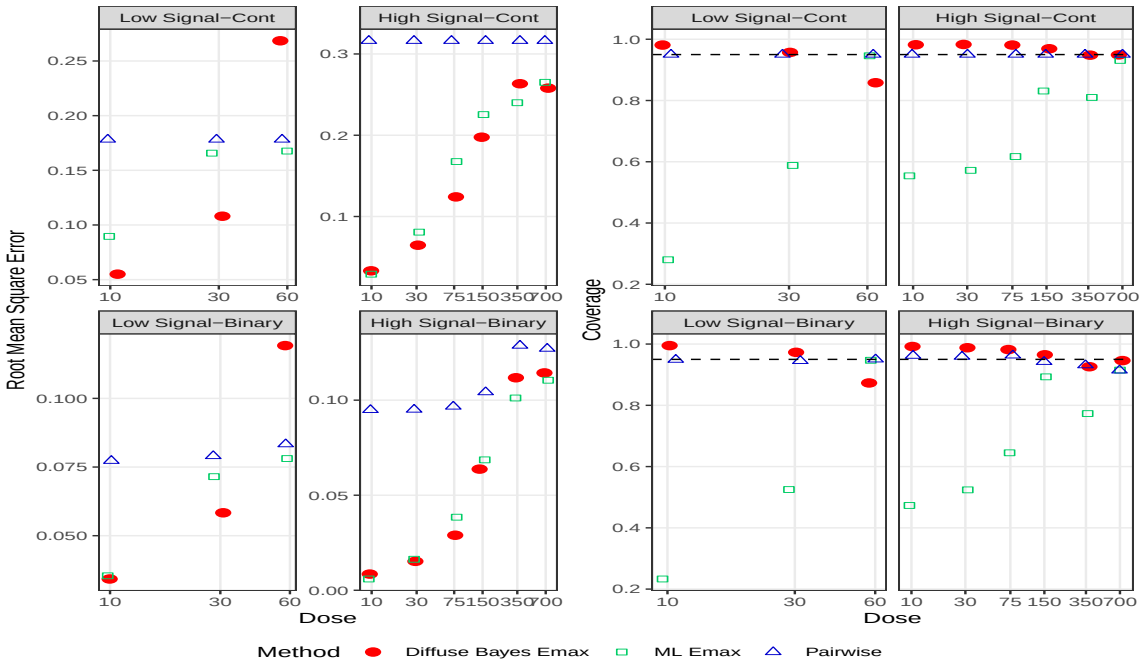


Figure 51: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 10P_{50}$, $\lambda = 2.5$. For the continuous endpoint, $E_0 = 1$ and difTarget $= 0.5,\ 1.5$. For the binary endpoint, the placebo response is 0.25 or 0.1, and the most effective dose has response of 0.49 or 0.63, respectively.

1 Evaluating dose response based on two studies

Data from more than one study are often combined to estimate dose response. Examples include 1) combining data from a single active dose and placebo in an initial proof of concept study with a subsequent dose finding study, and 2) combining data from two dose finding studies where the second study focuses on the lower or upper dosing range in the first study. The latter situation is simulated here.

The simulation design for the two studies will be based on simulation designs from the original submission that are graphically summarized in Figures 20-21 of the submission. The setting with $\lambda = 0.8$ and $ED_{50}/P_{50} = 4$ was selected as the base case because it produces data similar to many real dose response curves, and the differences between the populations parameters and their prior distributions is typical of real applications.

The Low and High Signal designs are simulated. The first study is unchanged from the original simulations. A second study is then simulated that focuses on the higher doses. For the Low Signal design, the lowest dose is dropped in the second study leaving two active doses. For the High Signal design, the two lowest doses are dropped leaving three active doses. The sample size in each group is the same in both studies, so the total sample size in the second study is smaller than in the first study. The simulations were repeated for a continuous and binary endpoint. The first set of four simulated conditions assume that the placebo response is the same across the two studies. The simulations were then repeated with a large shift in response between the two studies (1/2 the difference in response between placebo and the highest dose).

The Bayesian *Emax* model was fit with common parameters across the two studies except that a separate intercept (placebo response) was fit for each study, as recommended in the submission document. Pairwise comparisons were also assessed. For the pairwise comparisons, the data from common dose groups across the two studies were naively pooled, as is often done. The modified ML estimation was not implemented because complications arise when this approach is generalized to fit data from multiple studies.

2 Simulation results

The Bayesian E_{max} estimation performed well, and better than the pairwise comparisons in all of the simulations settings. The results are summarized in Figures 1 and 2. Figure 1 corresponds to Figure 21 in the submission. The addition of data from a second trial reduced the RMSE as expected. The upward (placebo) increase in responses did not noticeably impact the performance of the Bayesian E_{max} modeling in Figure 2.

The RMSE of the pairwise comparisons for the higher doses are also improved because their sample sizes are doubled by the addition of the second design. Note the change in scales between Figures 1 and 2. The RMSE of the pairwise comparisons for the lower doses are much larger in Figure 2 because the change in placebo creates bias in the comparisons for the lower doses that were not included in the second design. The bias is also reflected in the coverage probabilities for the lower doses. The performance of the pairwise comparisons could be improved by incorporating appropriate modeling assumptions. These improvements are built into the Bayesian E_{max} approach.

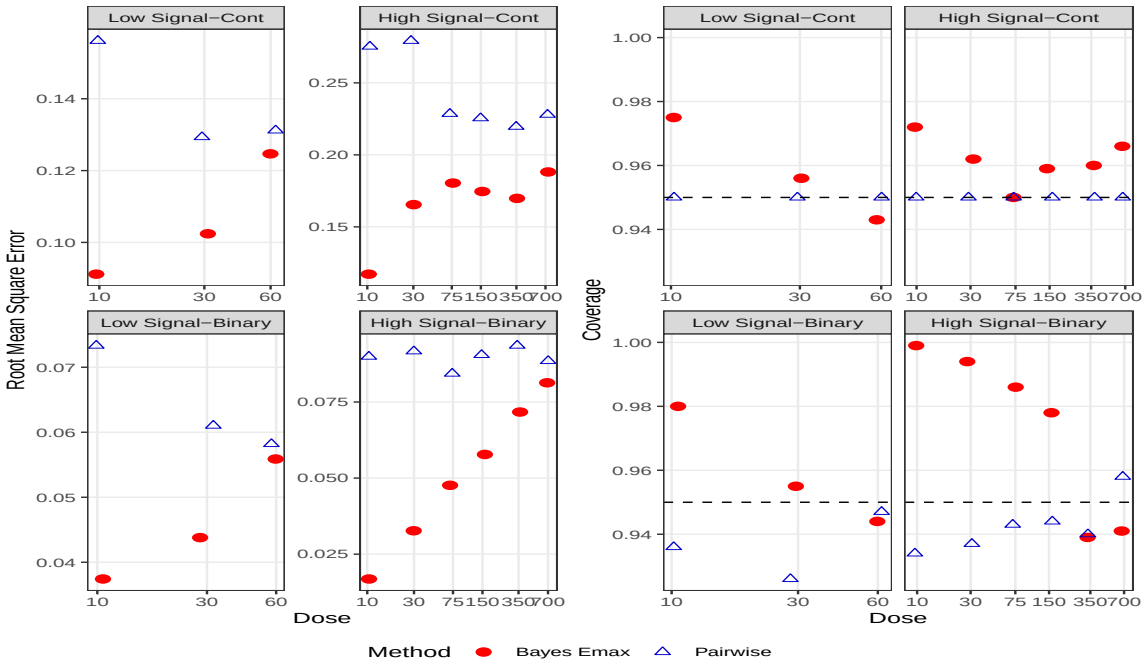


Figure 1: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 4P_{50}$, $\lambda = 0.8$. This figure corresponds to Figure 21 in the submission.

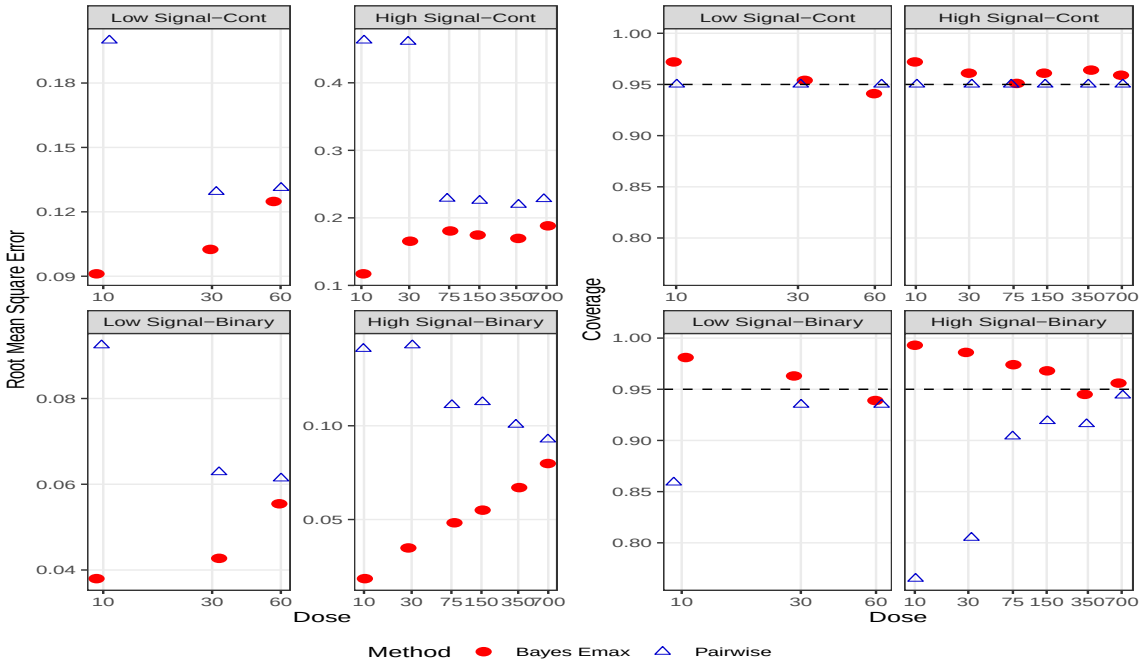


Figure 2: Comparison of RMSE (left) and Coverage (right) for each dose versus placebo. The $ED_{50} = 4P_{50}$, $\lambda = 0.8$. The placebo response increased by 1/2 the difference between placebo and the highest dose compared to the first set of simulations.

RESPONSE TO FDA REQUEST FOR INFORMATION

received 07 Mar 2022

FDA Qualification Opinion
“Empirically Based Bayesian Emax Models for Dose Response Design and Analysis”

BACKGROUND

21 April 2021, Pfizer submitted a Fit-for-Purpose (FFP) Request for an FDA Qualification Opinion on "Empirically Based Bayesian Emax Models for Dose Response Design and Analysis." On 14 September 2021 and 12 Nov 2021, Pfizer received comments and requests for information on this submission from the FDA and submitted responses. On 07 March 2022, Pfizer received additional comments and requests for information based on the Pfizer response. The Pfizer responses to these follow-up comments are provided below. The question numbering refers to 07 March 2022 comments and requests for information.

AGENCY REQUEST FOR INFORMATION AND SPONSOR RESPONSE

Clinical Pharmacology

Question

We simulated data without informative sampling points for Emax model parameters such as ED50 and Emax. With simulated data, we found parameter identifiability issues became a hurdle for applying the proposed Bayesian Emax model even with the informative priors derived from the meta-data analyses. We also observed that some study data in the meta-data pool are sparse to inform the model parameters as shown in Appendix C of the initial submission package (e.g., on page 69 for ID1046 1160.2; on page 80 for ID4009 R668−AD−1224; and on page 81 for ID17 1008−009).

To address our concerns, pull some typical data from the meta-data pool within the following categories and analyze these data separately either with informative or non-informative priors. Submit for further review the results with the data used in the analyses.

Potential scenarios

Response variables	Missing sampling points	Number of studies
Continuous	Missing to inform Emax	Each per therapeutic area if available
	Missing to inform ED50	Each per therapeutic area if available
Binary	Missing to inform Emax	Each per therapeutic area if available
	Missing to inform ED50	Each per therapeutic area if available

Response

The Pfizer response to the FDA follow-up comment is provided in the attached file "cpSectResp3.pdf".

Statistics

Question 1

Confirm that the proposed tool will be used to help select the doses in the next phase trial or find the optimal design for future studies. Confirm that no hypothesis testing (either Bayesian and/or Frequentist) is involved.

Response 1

The purpose of the analyses we propose is to select dose(s) for the next phase (typically phase 3). It is also used to plan phase 2 dose ranging studies and follow-up phase 2 studies when needed. We do not use hypothesis testing as an approach to achieve these objectives. As noted in Section 3.1.1 of the submission document, a preliminary alpha-controlled test to confirm an effect for the compound is often performed. If the test is not significant, along with a simple plot of the dose response data, it provides a rapid cautionary analysis showing that the efficacy of the compound or the design of the study was not as planned. These simple hypothesis tests are not the primary output of the study and do not contribute substantially to achieving the objectives of the trial.

Question 2

In your response submitted on December 15, 2021, you referred a *ClinDR* package to describe your model. Your original submission and subsequent responses to our information requests are not clear enough to allow FDA to thoroughly review the submitted information. Submit a detailed step-by-step algorithm including mathematical formulations of the methods you are proposing. The algorithm should be self-contained without any need to refer to the software codes. Clearly state all the assumptions and choices (e.g., the model, choice of the priors and hyperpriors), and what data will be used at each step. Justify any choices you make regarding the model, priors and hyperpriors.

For example, describe the whole procedure for the use case with a sigmoid model, patient-level data, and at least two intercepts (protocols) with no baseline covariates. Provide the following information for each step of your algorithm and any additional information that may be useful.

- Purpose of the step
- Data to be used
- Mathematical formulation (hierarchical structure including priors, hyperpriors, likelihood, etc.)
- Justification for the use of proposed prior
- Derived estimates from the procedure
- How the derived estimates will be used in the next step

If any test for goodness of fit is needed, explain in which step such tests should be implemented.

Additional clarity is needed. For example, it is stated or implied that the prior distribution for a hypothetical future dose response study - at least for ln(lambda) and ln(ED50/P50) - will be the posterior predictive distribution from the Bayesian Hierarchical meta-analysis of the “metadata”. However it is also stated that the proposed prior for such a hypothetical future dose response study is described as a multivariate t-distribution (location and scale translated). Clarify if the MCMC software used for the meta-analysis produce a posterior distribution that may be characterized as such a multivariate t-distribution. If so, thoroughly describe how this is done without referring to R-code.

Response

The Pfizer response to the FDA follow-up comment is provided in the attached file “statSectResp3.pdf”.

I Overview

In response to concerns about the performance of the proposed Bayesian E_{max} modeling when applied to poor dosing designs, we evaluate the procedure in four extreme settings. It is common in current practice to encounter dose ranging studies with low signal-to-noise and only few doses covering a limited dosing range (e.g., high/low dose < 10). This differs from some pre-clinical applications of similar modeling. The most common extreme settings occur when the doses appear to cover only the lower 'linear' portion of the dose response curve, or when all of the doses appear to be on or very near the plateau of the dose response. We evaluate four such examples under somewhat differing conditions, two with 'linear' response data, and two with data on the plateau only.

The first example in Section II is an evaluation of the dose ranging study for compound ID=1021. This compound was not included in the query, but we evaluated it because it has a continuous endpoint that appears to be nearly linear, and it includes a placebo group. The second compound (ID=1046) evaluated in Section III was identified in the query. Note that it does not include a placebo group. The methods we proposed can effectively evaluate studies without a placebo group, but the lack of a placebo group does alter their operating characteristics. While we applied the modeling here, we do not promote such applications because the placebo group is not only valuable for estimating the dose response curve, the absence of a placebo group can alter the responses in the active dose groups and create potential reproducibility issues. The third example in Section IV evaluates a study for compound ID=4009, also identified in the query. It is an extreme case with a binary endpoint and only two dose groups, both of which appear to be on the plateau of the dose response curve. The fourth example in Section V evaluates compound ID 1035, which has two studies with a continuous endpoint where the doses from both studies are on or very close to the plateau.

Some of the designs, for example the two-dose group design in Section IV, are very poor. No method of analysis can salvage such a poor design, as it is apparent that it does little other than characterize the plateau of the dose response curve and place an upper bound on the ED_{50}. For such limited designs, the Bayesian E_{max} model with the empirically based prior distribution for the λ and ED_{50} achieve the criteria that can be reasonably supported by the data:

1) it remains numerically stable,

2) it recovers the dose response curve within the observed dosing range,

3) it provides a reasonable prediction of the remainder of the dose response curve based on historical experience from other compounds, and

4) it provides an assessment of the uncertainty of the dose response curve estimated within and outside of the observed dosing range.

Because the dose response curve over its full range is often not well determined for the reasons noted, the parameters in the model that determine the full dose response are often unstably estimated. In addition, we have not found them to be very helpful when discussing dose selection. Thus we tend to minimize their role in reporting, viewing them as an intermediate output. Because the query seems focused on the model parameters, we include more focus on them in our response here than is our usual practice.

Two other clarifying points. The first is that we will act as if each compound is new, and we are analyzing its first dose response study, or in the case of compound ID 1035, its first two dosing studies. This is clearly not the case; it is done here to illustrate future applications of the modeling. The second is that some of the compounds evaluated here included other dose response studies, which considerably expanded their studied dosing ranges. This is one reason some of the dosing designs appear to be so poor.

II Example ID=1021: continuous endpoint with near linear observed trend

Compound ID=1021 is a subcutaneous injection for the treatment of Homozygous Familial Hypercholesterolemia. There was a single dose finding study with endpoint LDL-C percent change from baseline, 5 active doses (dosing range < 10), and a placebo group.

For our numerical illustration, the prior distribution was constructed using the default predictive prior distribution for the ED_{50} and λ. The P_{50}, inferred from the dosing design (this is an external study), is $P_{50} = 10.715$. The prior distributions for the placebo response and drug effect are centered at 0.0, with diffuse scale parameters set to 10 times the reported within-group SD for the endpoint. The prior construction here follows the recommendations in Section 3.1.2 of the submission document.

The sample means and fitted model are graphically summarized in Figure R.1. The fitted curve displays slight curvature within the observed dosing range, but it is not appreciably different from a linear fit. The prior distribution stabilized the resulting posterior distribution so the MCMC algorithm performed well even though the data provide little information about the E_{max} (and thus ED_{50}) parameter. The MCMC traceplot after burn-in is displayed in Figure R.2. The convergence and mixing of the chains is good, even though the $log(ED_{50})$ distribution is right skewed. The autocorrelation plot in Figure R.3 shows

that good quality near-independent chains were achieved. This assessment was further supported by the Gelman-Rubin convergence statistics.

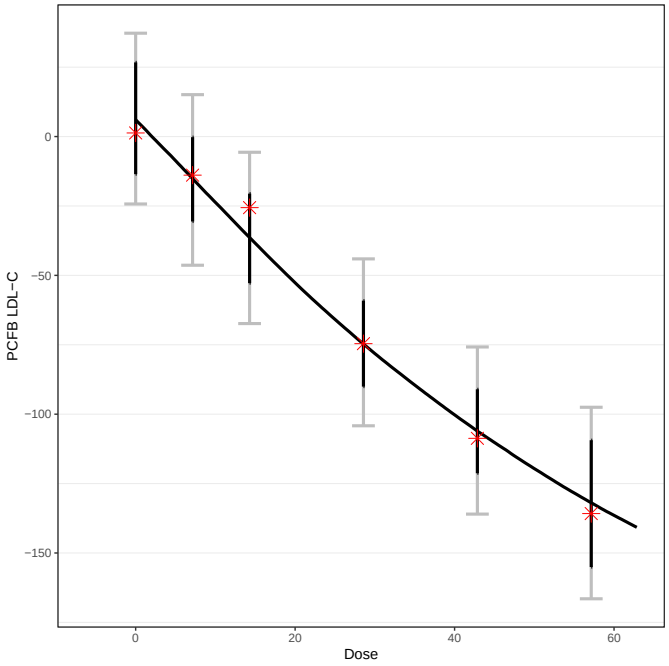


Figure R.1: Dose response curve for compound ID=1021. The red asterisk are observed dose group proportions. The solid curve is the posterior median estimator of response computed over a grid of doses. The solid bars are 90% posterior intervals for the population values, and the gray bars are 90% prediction intervals.

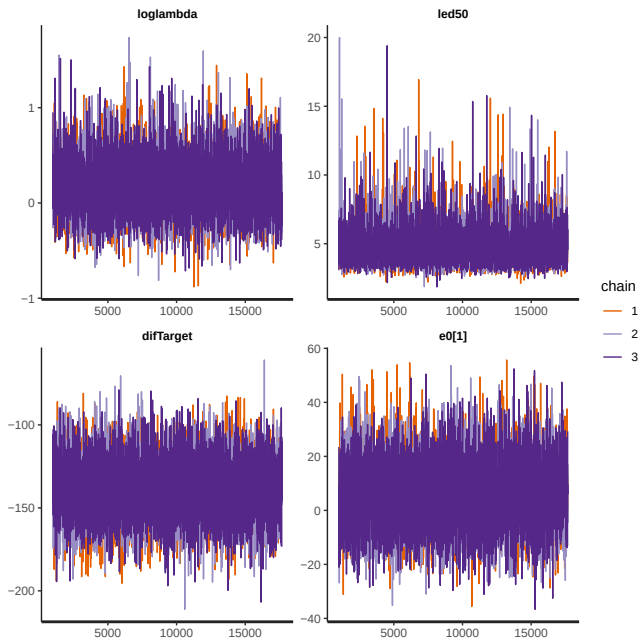


Figure R.2: Traceplot based on 3 chains from the MCMC fit for compound ID=1021. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

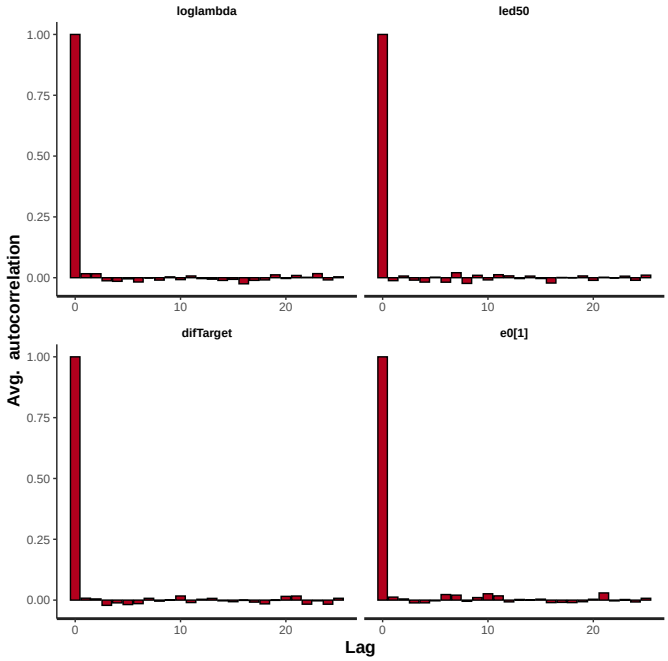


Figure R.3: Auto-correlations from the MCMC generated parameters for compound ID=1021. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

A summary of the posterior distribution for the model parameters is in Table R.1. The most notable features are the very high upper bound for the ED_{50}, and the corresponding lower bound for the E_{max} parameter. The E_{max} parameter is derived from the other model parameters, and note that the measured drug effect here is negative. The extreme skewness occurs because when λ is near one, a linear trend in the observed dosing range is achieved by increasing the ED_{50} and E_{max} while fixing the ratio E_{max}/ED_{50}, which is the slope of the line. The data supply information about the lower bound for the ED_{50} and the upper bound for the E_{max}, but the upper bound for the ED_{50} and lower bound for the E_{max} are set largely by the prior distribution, and they are thus somewhat arbitrary. Note that the difTarget parameter estimation is not appreciably impacted by this indeterminacy.

Parameter	Post Median	90% Post Interval
λ	1.16	(0.77,1.95)
ED_{50}	81.9	(24.8, 1182.7)
difTarget	-138.32	(-167.39,-108.99)
E_{max}	-349.02	(-2955.39,-158.61)
E_0	6.07	(-13.54,26.71)
σ	43.95	(37.44,53.15)

Table R.1: Parameter estimation for compound ID=1021.

The operating characteristics of the Bayesian E_{max} modeling in conditions like those for compound ID=1021 were evaluated in two simulation studies (1000 replications per setting) implemented using the *emaxsimB* function in R package *clinDR*. The same prior distribution and dosing design are utilized. The simulation population parameters were selected to produce simulated data similar to that observed in Figure R.1.

The simulation population parameters are given in Table R.2. For the first simulation, the ED_{50} was set to approximately 2 times the highest dose, and 10 times the P_{50}. The difTarget was set to match the observed effect at the highest dose, rather than no effect (with diffuse scale) as assumed in the prior distribution. The other parameters closely match those assumed in the prior distribution. The simulation results for the the model parameters are summarized in the left portion of Table R.2. The ED_{50} was shrunk back toward the P_{50} though it remained high, and the magnitude of the E_{max} parameter was correspondingly reduced. The central tendency of the other parameters remained focused on the popoulation/prior values. The repeated-sampling coverage of the 90% posterior intervals was $\geq 90\%$.

The second simulation setting was designed to be more extreme. The ED_{50} is 5 times the highest dose and approximately 25 times the initial projected P_{50}. The other sim-

ulation population parameters were also changed to differ from the corresponding prior expectations. Under these conditions the model fitting strongly shrinks the ED_{50} toward the P_{50}, and likewise for the offsetting E_{max} posterior distribution. The other parameters are not appreciably impacted. The posterior intervals continue to include the population values at the nominal 90% levels. If the ED_{50} is set to even higher values, the coverage of the posterior intervals for the ED_{50} and E_{max} parameters will drop below 90%. This is not practically damaging in this setting, however, because the fitted models will still predict large improvements if dosing at higher levels is feasible.

	Prior Quality Good			Prior Quality Poor		
Parameter	Pop	Mean (Post Median)	Cov (90% Int)	Pop	Mean (Post Median)	Cov (90% Int)
λ	1	1.05	0.99	0.9	1.05	0.98
$\log(ED_{50})$	4.74	4.06	1	5.65	4.1	0.98
difTarget	-137.1	-135	0.9	-137.1	-133.82	0.89
E_{max}	-411.3	-302.14	1	-720.69	-307.31	0.87
E_0	0	3.09	0.91	1	3.76	0.92
σ	44.36	44.99	0.9	44.36	45.01	0.9

Table R.2: Simulations for ID=1021. The columns are the means over 1000 simulation replications of the posterior median estimators, and the proportions of the 1000 90% posterior intervals that include the population values. The prior distribution was constructed with the $\log(P_{50}) = 2.37$, a placebo prior mean of 0.0, a treatment effect mean of 0.0, diffuse scale parameters equal to 10 times the pooled SD for compound ID=1021, and the meta-analytic default prior distribution for the remaining parameters.

The performance of the modeling for the primary task of estimating dose response within the observed dosing range was excellent under both simulation conditions. A subset of the routinely reported simulation results for the model-based effect estimates at each observed dose are in Figures R.4 and R.5. The repeated-sampling coverage of the Bayesian intervals achieves or exceeds the nominal levels. Note the large improvements in the estimation compared to paired comparisons. The very large improvement for the lowest dose is typical, and occurs because it is closest to placebo. Comparison of the two highest doses also has a similar small mean square error (not shown), which can be very useful when deciding whether to increase the recommended dose. In summary, the Bayesian E_{max} modeling achieved the four objectives for it given in Section I.

```
Coverage probabilities for nominal 0.9 intervals [Dose-PBO]:
Bayesian Dose response modeling posterior intervals:
 7.14 14.29 28.57 42.86 57.14
0.921 0.890 0.903 0.914 0.897

.
.
.

Square Root Mean Squared Error [Dose-PBO]:
Bayesian dose response modeling (EST=posterior median) :
  7.14  14.29  28.57  42.86  57.14
12.829 16.550 16.824 16.091 18.037

 Pairwise comparisons:
  7.14  14.29  28.57  42.86  57.14
20.660 21.149 21.037 21.149 21.427
```

Figure R.4: Summary of the simulation evaluation for compound ID=1021 in the setting with the prior distribution in closer agreement with the population value.

```
Coverage probabilities for nominal 0.9 intervals [Dose-PBO]:
Bayesian Dose response modeling posterior intervals:
 7.14 14.29 28.57 42.86 57.14
0.938 0.908 0.904 0.910 0.891

.
.
.

Square Root Mean Squared Error [Dose-PBO]:
Bayesian dose response modeling (EST=posterior median) :
  7.14  14.29  28.57  42.86  57.14
12.823 16.194 16.655 16.231 18.473

 Pairwise comparisons:
  7.14  14.29  28.57  42.86  57.14
20.660 21.149 21.037 21.149 21.427
```

Figure R.5: Summary of the simulation evaluation for compound ID=1021 in the setting with the prior distribution far from the population values.

III Example ID=1046: continuous data with near linear trend, no placebo group

Compound ID 1046 (requested) is a small molecule for cardio-vascular prevention with primary endpoint change from baseline in Activated Partial Thromboplastin Time (APPT). There are only 3 doses covering a 6 fold range, and as noted in Section I, there is no placebo group. The dose response data and the fitted curve are graphically summarized in Figure R.6. This is another example where the dose response appears near linear in the observed dosing range.

The prior distribution used for the curve fit in Figure R.6 uses the empirically constructed prior for the ED_{50} and λ with the recommended default values, and the $P_{50} = 200$. The placebo and effect parameter priors are centered at 0.0 with diffuse scale parameters set to 10 times the reported within-group SD. The diffuse prior distribution for the residual SD parameter is uniform on $(0.1, 100)$.

The traceplot of the MCMC output is in Figure R.7. The distributions of the E_0 and difTarget parameters are noticeably skewed and there is a trade off occurring between these parameters. Recall that difTarget is the difference between the high dose and placebo, so the indeterminacy in placebo has created a second issue distinct from the failure to observe the plateau of the response curve. Figure R.8 shows some auto-correlation even after thinning, which also suggests the posterior distribution is ill-behaved and the use of the MCMC for this distribution is borderline. Note that curve displayed in Figure R.6 is minimally impacted because it is predicting absolute response, not the difference with placebo. Likewise, differences between doses within the studied range are little impacted by the lack of placebo data.

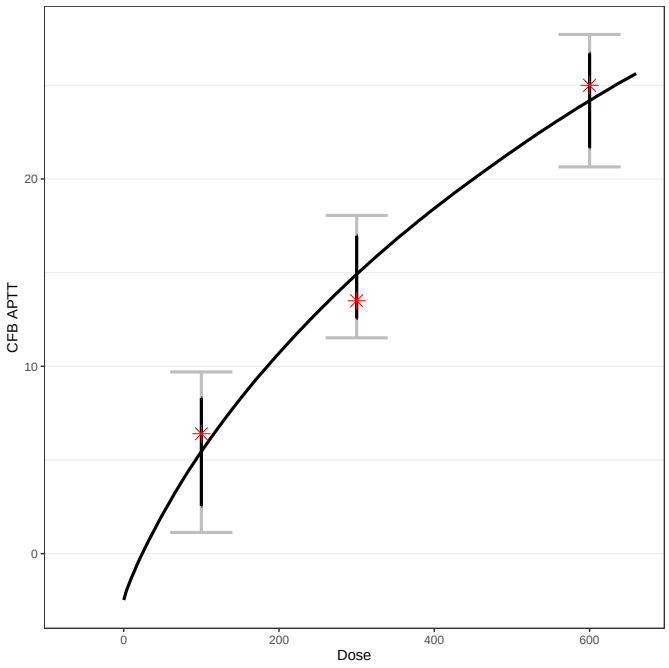


Figure R.6: Dose response curve for compound ID=1046, with diffuse prior distribution for the placebo response. The red asterisk are observed dose group proportions. The solid curve is the posterior median estimator of response computed over a grid of doses. The solid bars are 90% posterior intervals for the population values, and the gray bars are 90% prediction intervals.

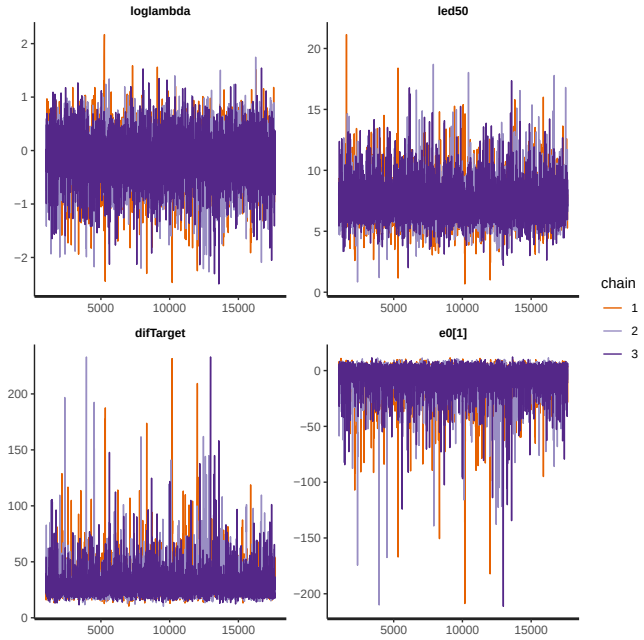


Figure R.7: Traceplot based on 3 chains from the MCMC fit of the E_{max} model with a diffuse prior distribution for the placebo response. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

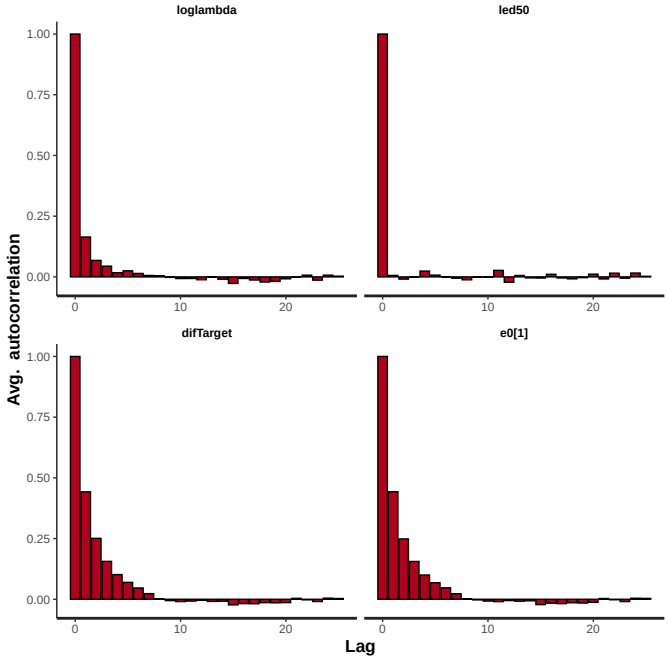


Figure R.8: Auto-correlations from the MCMC generated parameters of the E_{max} model with a diffuse prior distribution for the placebo response. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

If a trial is planned without placebo (not recommended), it is important to have some external information regarding placebo response. A second prior distribution was evaluated changing to a weakly informative prior distribution for the placebo response with a scale parameter of one SD. The resulting dose response plot is indistinguishable from Figure R.6, so it is not displayed. The resulting traceplot (Figure R.9) and auto-correlation plot (Figure R.10) demonstrate that even a weak prior distribution for the placebo response largely eliminates the indeterminacy and the borderline performance of the MCMC evaluation.

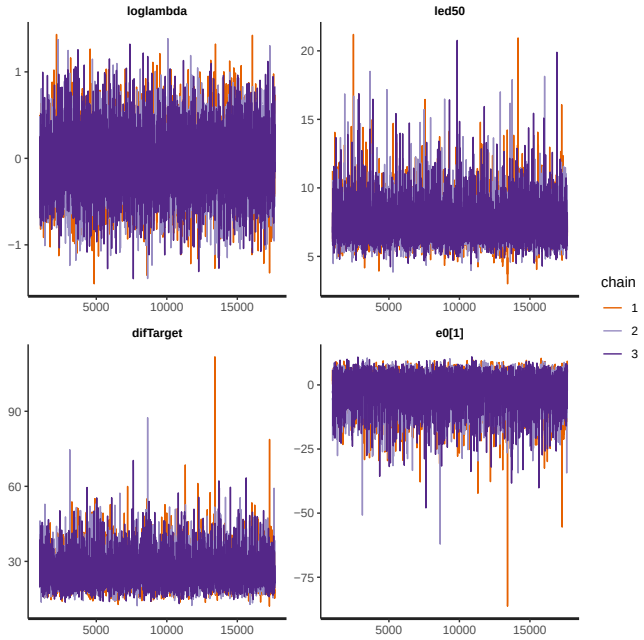


Figure R.9: Traceplot based on 3 chains from the MCMC fit of the E_{max} model with a weakly informative prior distribution for the placebo response. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

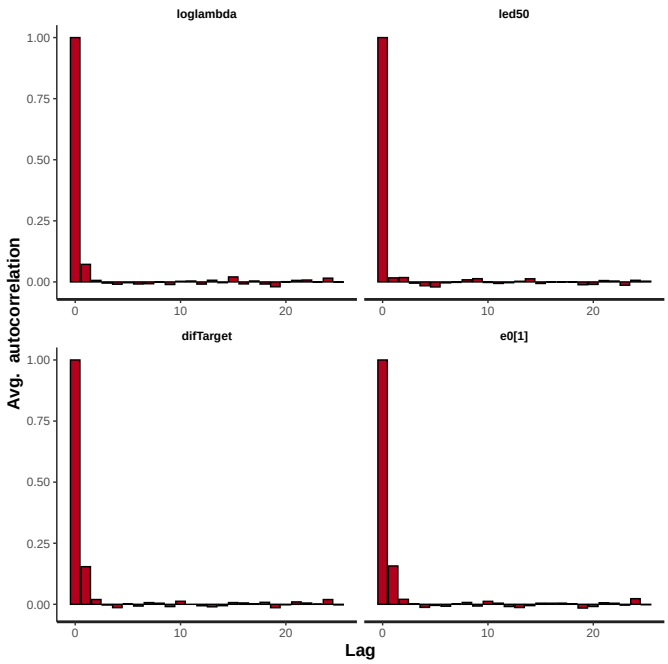


Figure R.10: Auto-correlations from the MCMC generated parameters of the E_{max} model with a weakly informative prior distribution for the placebo response. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

A summary of the posterior distribution of the model parameters is in Table R.3, with the results for the diffuse placebo prior on the left and the results for the weakly informative on right. The impact of the linear trend on the ED_{50} is similar to that for compound ID 1021 in Section II, so it is not discussed here. The impact of the lack of placebo data is now visible in the upper bound for the effect parameter difTarget, and the lower bound for the placebo response parameter, E_0. Comparing the results for the diffuse and weakly informative placebo priors shows that these bounds are largely determined by the placebo response prior. This is unsurprising. If you intend to estimate differences with placebo, you need to include placebo in the trial or be prepared to supply external information about placebo response. The problem with the latter approach is that the absence of placebo in the trial may alter the response to the active doses.

	Diffuse Prior for E_0		Weakly Informative Prior for E_0	
Parameter	Post Median	90% Post Interval	Post Median	90% Post Interval
λ	0.9	(0.39,1.74)	0.97	(0.53,1.77)
ED_{50}	1170	(212.4, 24321.9)	1297.9	(294.6, 25505.5)
difTarget	26.79	(18.21,56.02)	25.32	(17.96,38.85)
E_{max}	84.07	(36.32,503.53)	81.65	(35.35,659.96)
E_0	-2.48	(-32.4,6.07)	-0.92	(-14.81,6.18)
σ	14.08	(13.06,15.24)	14.07	(13.06,15.21)

Table R.3: Parameter estimation for compound ID=1046.

IV Example ID=4009: binary endpoint near plateau only

The compound ID=4009 was studied in four qualifying protocols. The protocol referenced in the query, R688-AD-1224, is not appropriate as an example of a standalone protocol because it has only two doses and no placebo group. This is not a reasonable application for any type of dose response modeling, and such applications were excluded from the proposed use cases in Section 1 of the submission document. We utilize protocol R668-AD-1334 instead, which is itself an extreme case with a binary endpoint and only two doses and placebo, Another similar situation is the suggested use of study 1008-009 from compound ID 1017. This study has a single dose and placebo, which is not a valid use case when viewed as a stand-alone study. Dosing for compound ID 1017 was evaluated in 3 protocols that provide a moderate amount of dose response information, so it is not included in our response as an example with a poorly determined dose response.

The dose groups for study R668-AD-1334 of compound ID=4009 are 0, 150 and 300 mg with sample sizes of 224, 224 and 223. The two active doses appears to be on or near the plateau of the dose response, as displayed in Figure R.11. The endpoint is an investigator assessed global improvement responder variable for Dermatitis. The data and fitted model are graphically summarized in Figure R.11.

The prior distribution for the curve fit in Figure R.11 uses the empirically constructed prior for the ED_{50} and λ with the recommended default values, and the $P_{50} = 25$. The prior distribution for the placebo response is centered at logit(0.1), and effect parameter prior is centered at 0.0; both have diffuse scale parameters set to 4 on the logit scale.

The MCMC diagnostics in Figures R.12 and R.13 support the use of the method in this setting.

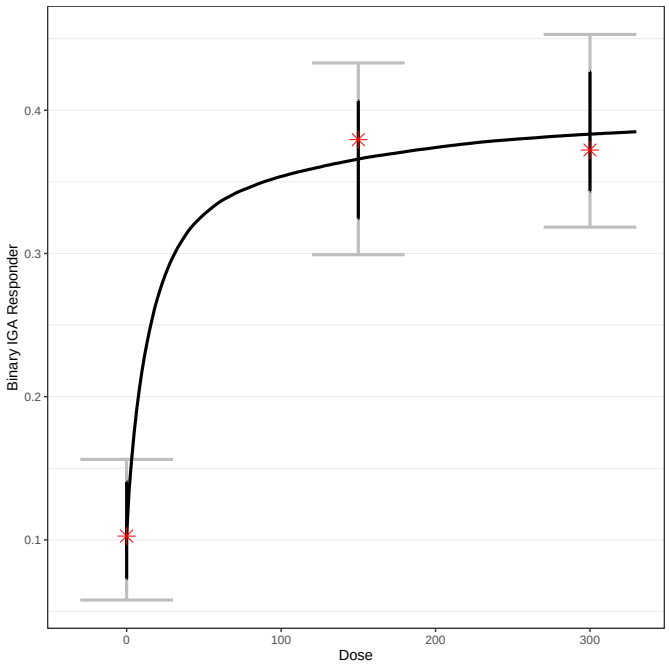


Figure R.11: Dose response curve for compound ID=4009. The red asterisk are observed dose group proportions. The solid curve is the posterior median estimator of response computed over a grid of doses. The solid bars are 90% posterior intervals for the population values, and the gray bars are 90% prediction intervals.

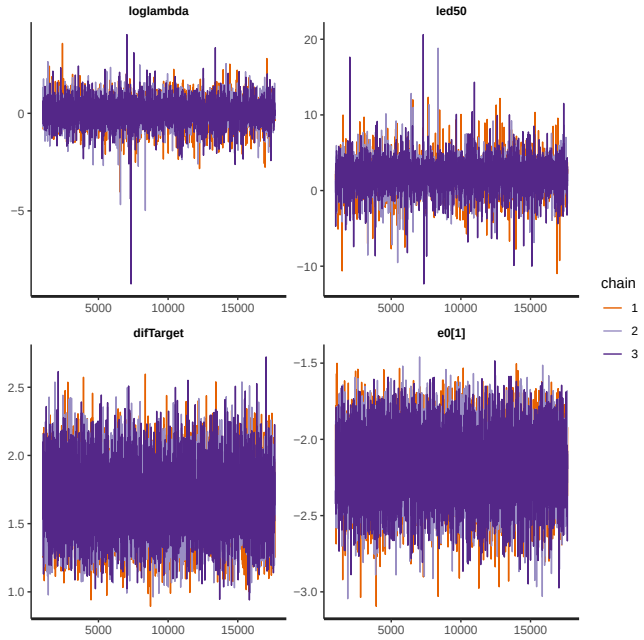


Figure R.12: Traceplot based on 3 chains from the MCMC fit for compound ID=4009. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

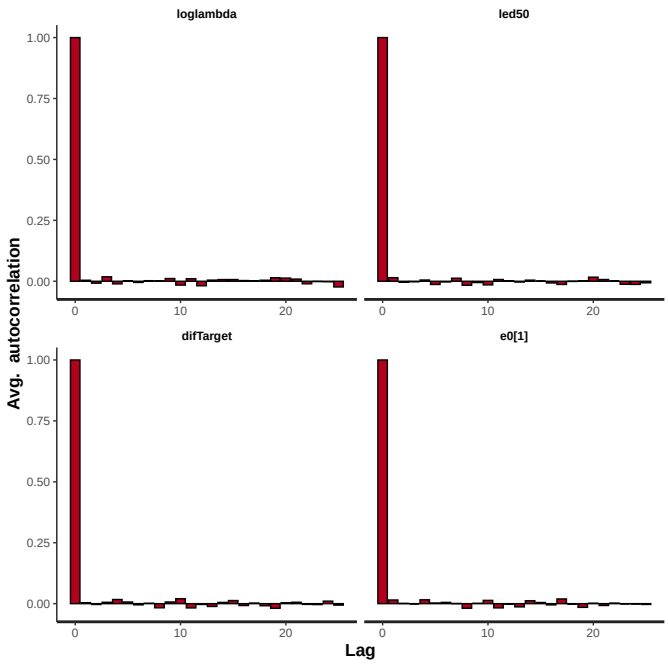


Figure R.13: Auto-correlations from the MCMC generated parameters for compound ID=4009. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

The posterior distribution for the model parameters are summarized in Table R.4. Two features to note are 1) the wide interval for the ED_{50}, which is nonetheless bounded by the lowest studied dose, and 2) the much closer agreement between the difTarget and E_{max} parameters due to the fact that the data supply much information about the plateau.

Parameter	Post Median	90% Post Interval
λ	1.11	(0.45,2.64)
ED_{50}	11.4	(0.4, 99.7)
difTarget	1.69	(1.29,2.11)
E_{max}	1.8	(1.34,2.75)
E_0	-2.16	(-2.54,-1.81)

Table R.4: Parameter estimation for compound ID=4009.

Simulation studies similar to those in Section II were conducted to evaluate the operating characteristics of the modeling when there are only two dose groups, with both apparently on the plateau of the dose response curve. The λ and ED_{50} simulation population parameters in the first simulation match the central tendencies of the prior distribution. The E_0 and difTarget were selected to match the sample rates from the real data. The results are summarized in the left side of Table R.5. The operating characteristics are

very good with repeated-sampling coverage of the posterior intervals for the ED_{50} and λ parameters exceeding the nominal 90% level because the population and prior values were in close agreement and the data have little information about these parameters.

The second simulation created a steeper dose response with a much lower ED_{50} (the population values are in Table R.5). The model now substantially over-estimates the ED_{50} because the data have very little information about the parameter to update the prior distribution for it. The posterior intervals for the ED_{50} and λ still have high coverage and appropriately warn of very large uncertainty about the lowest effective doses.

The performance of the method within the observed dosing range was good in both simulation settings, as summarized in Figures R.14 and R.15. In summary, the method achieved the four objectives given in Section I.

	Prior Quality Good			Prior Quality Poor		
Parameter	Pop	Mean (Post Median)	Cov (90% Int)	Pop	Mean (Post Median)	Cov (90% Int)
λ	1	1.01	1	1.5	1.07	1
$\log(ED_{50})$	3.22	3.03	1	0.92	2.68	0.98
difTarget	1.67	1.68	0.91	1.67	1.74	0.9
E_{max}	1.8	1.92	0.96	1.67	1.91	0.89
E_0	-2.2	-2.22	0.9	-2.75	-2.78	0.91

Table R.5: Simulations for ID=4009. The columns are the means over 1000 simulation replications of the posterior median estimators, and the proportions of the 1000 90% posterior intervals that include the population values. The prior distribution was constructed with the $P_{50} = 25$, a placebo prior mean of logit(0.10), a treatment effect mean of 0.0, diffuse scale parameters on the *logit* scale (4.0), and the meta-analytic default prior distribution for the remaining parameters.

```
Coverage probabilities for nominal 0.9 intervals [Dose-PBO]:
Bayesian Dose response modeling posterior intervals:
  150   300
0.930 0.918

.
.
.

Square Root Mean Squared Error [Dose-PBO]:
Bayesian dose response modeling (EST=posterior median) :
  150   300
0.030 0.032

 Pairwise comparisons:
  150   300
0.037 0.037
```

Figure R.14: Summary of the simulation evaluation for compound ID=4009 in the setting with the prior distribution in closer agreement with the population value.

```
Coverage probabilities for nominal 0.9 intervals [Dose-PBO]:
Bayesian Dose response modeling posterior intervals:
  150   300
0.882 0.904

.
.
.

Square Root Mean Squared Error [Dose-PBO]:
Bayesian dose response modeling (EST=posterior median) :
  150   300
0.028 0.028

 Pairwise comparisons:
  150   300
0.033 0.033
```

Figure R.15: Summary of the simulation evaluation for compound ID=4009 in the setting with the prior distribution far from the population values.

V Example ID=1035: continuous endpoint near plateau in two studies

Compound ID=1035 is an inhaled small molecule for the treatment of COPD with change from baseline in FEV1 as the endpoint. There were two studies providing limited dosing information. This example differs from the others included in our response because it includes two studies, and there is some indication that the lowest tested dose is below the plateau although the doses included in the studies clearly do not characterize most of the dose response curve well.

For our numerical illustration, the prior distribution was constructed using the default predictive prior distribution for the ED_{50} and λ. The P_{50}, inferred from the dosing design (this is an external study), is $P_{50} = 225$. The prior distribution for the placebo response is centered at 1.2, and the prior distribution for the drug effect is centered at 0.0, with diffuse scale parameters set to 10 times the reported within-group SD for the endpoint. The prior construction here follows the recommendations in Section 3.1.2 of the submission document. A separate independent placebo response parameter, E_0, was fit for each study.

The sample means and fitted model are graphically summarized in Figure R.16. Visual examination and the fitted curve suggests a an increase in response for doses between 150 and 300 mg, but this is not certain. The placebo response in the second study appears to have increased, which is not uncommon, but the same curve after placebo adjustment describes the data well.

The usual diagnostics for the MCMC evaluation of the posterior distribution, such as the traceplot (Figure R.17) and auto-correlation plot (Figure R.18) support its use. The traceplot does reveal the very heavy tail of the posterior distributions for the ED_{50} and λ parameters.

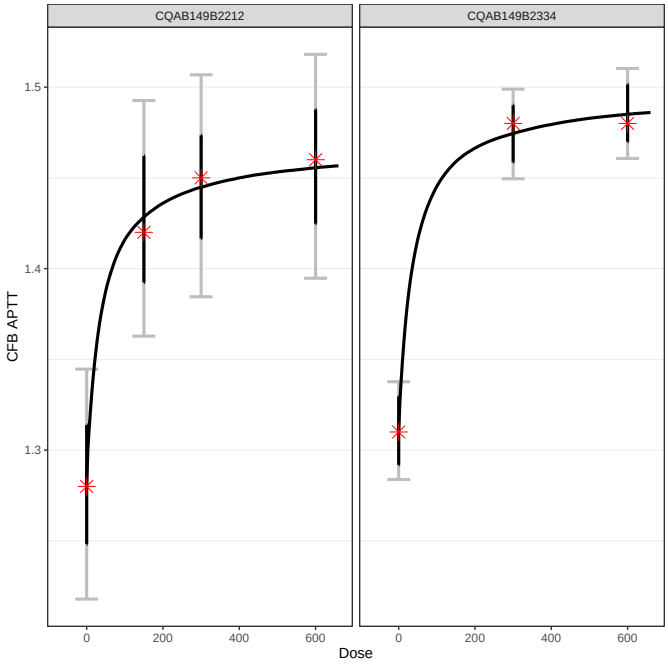


Figure R.16: Dose response curve for compound ID=1035. The red asterisk are observed dose group proportions. The solid curve is the posterior median estimator of response computed over a grid of doses. The solid bars are 90% posterior intervals for the population values, and the gray bars are 90% prediction intervals.

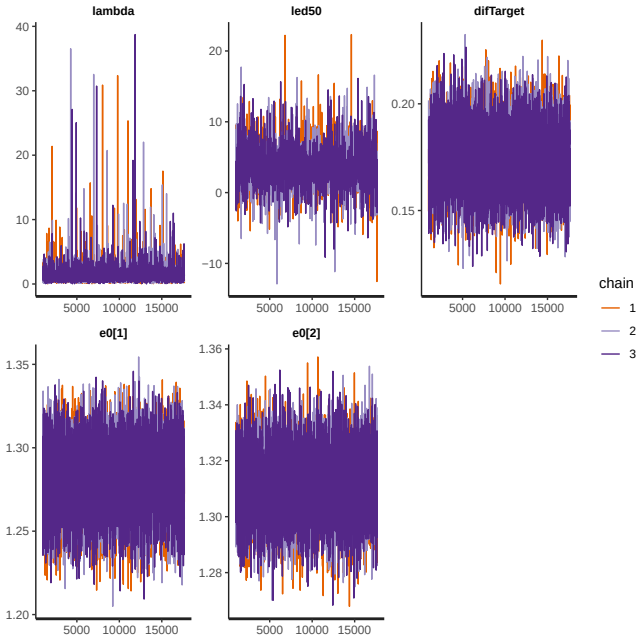


Figure R.17: Traceplot based on 3 chains from the MCMC fit for compound ID=1035. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

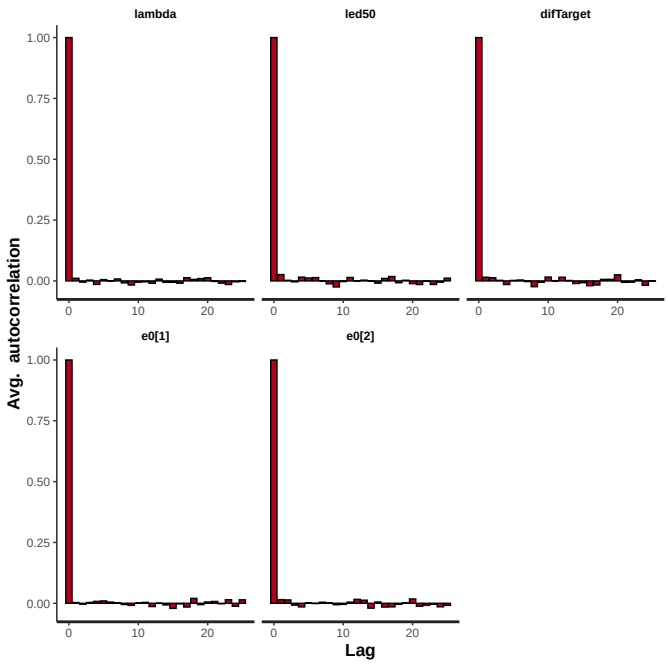


Figure R.18: Auto-correlations from the MCMC generated parameters for compound ID=1035. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

A summary of the posterior distribution for the model parameters is in Table R.6. The high uncertainty in the estimation of the λ and ED_{50} parameters is unsurprising given the lack of data on the steeper portion of the dose response curve. The upper tails of the ED_{50} and E_{max} quantitatively reveal something less apparent from simple visual inspection of the sample means, which is the possibility that additional efficacy might be possible if it feasible to test a higher dose. The modeling also makes the consistency of the results from the two studies more apparent than visual inspection of the sample means alone. As in the other examples, the modeling utilizes the limited information available effectively, but it is not a substitute for better dose response designs.

Parameter	Post Median	90% Post Interval
λ	1.17	(0.35,2.96)
ED_{50}	46.4	(2.3, 692.4)
difTarget	0.17	(0.15,0.2)
E_{max}	0.19	(0.15,0.36)
E_{01}	1.28	(1.25,1.31)
E_{02}	1.31	(1.29,1.33)
σ	0.23	(0.22,0.24)

Table R.6: Parameter estimation for compound ID=1035.

I Overview

We begin by responding to the final paragraph of Question 2 because once the construction of the prior distribution is understood, the remainder of the procedure and requested example apply routine Bayesian statistical methods. References to sections from the submission document are noted by including (SDOC) in the reference. The strategy for developing the prior distribution for the E_{max} model is described in Section II.1. The details of this strategy are then reviewed in Sections II.2-II.6. A detailed example involving two protocols is reviewed in Section III.

II Constructing an empirically-based prior distribution

II.1 Meta-analytical predictive (MAP) prior

The strategy we use to construct the prior distribution has been called the meta-analytical predictive (MAP) prior by Schmidli et al. (2014). Using notation similar to theirs, we have historical dose response data, $\mathcal{Y}_h$, $h = 1, \ldots, H$ for H compounds assumed to follow E_{max} models. The variation in the model parameters from different compounds are described by a hierarchical distribution. The dose response parameters display central tendencies (after appropriate transformation and normalization) summarized in a distribution called the posterior predictive distribution for a future compound. The posterior predictive distribution is derived from the hierarchical distribution, but it is not the same as the hierarchical distribution as will be explained. When data for the new compound, denoted by $\mathcal{Y}_*$, are available, it can be added to the historical data and the hierarchical model fit can be updated to yield a Bayesian posterior distribution for all of the historical and new parameters. This approach is called meta-analytical combined (MAC). It is a cumbersome and computationally intensive approach. When interest in focused on the parameters for the new compound, the posterior distribution for the new parameters can be computed using the predictive distribution based on the historical data as the 'prior' distribution, and it is combined with the usual likelihood function for the new data only. This approach is called meta-analytical predictive (MAP). Due to the properties of Bayesian updating, the MAC and MAP approaches yield the same correct posterior distribution for the new parameters conditional on both the historical and new data.

There are two distinctive features of the MAP approach we apply. The first is that the predictive distribution is computed for only two of the E_{max} model parameters: the ED_{50}

and λ (with appropriate transformation and norming). Hierarchical models and resulting predictive distributions were not applied to the placebo response and E_{max} parameters. The reasons for this difference in the handling of parameters is reviewed in Section II.6. The prior distributions for the placebo and effect parameters are usually assigned diffuse distributions when applied to a new compound. This completes the full specification of the prior distribution for a new dose response study. The creation of the prior is reviewed in Section II.6. We have noted that our software can accommodate an informative distribution for the placebo parameter(s) that might be derived from historical placebo data. There are many issues with the use of historical placebo data, but they are not specific to the dose response setting, so we have not included this potential use in our submission.

The second distinctive feature of our MAP implementation is that we approximated the posterior predictive distribution by a bivariate $t-distribution$. The predictive distribution does not have a closed analytical form, but we can generate a nearly unlimited sample from it using the MCMC (Markov Chain Monte Carlo) output from the modeling of the historical data. This approximation is reviewed in Section II.5. With this approximation, the E_{max} modeling of dose response for a new compound becomes a routine Bayesian application. Note in particular, using the MAP approach, there are NO 'hyper' parameters involved in the Bayesian modeling of dose response for a new compound.

The hierarchical distribution for the parameters, which is conditional on unknown hyper-parameters, also has an assumed bivariate $t-distribution$. Its relationship to the predictive distribution is reviewed in Section II.5. Finally, the diffuse prior distributions for the placebo and effect parameters are also specified in the form of $t-$distributions. The three distinctly different uses of the $t-distribution$ may be a source of confusion.

II.2 E_{max} model and its parameterization

The E_{max} models for continuous and binary data are introduced in equations (1.1) and (1.2) of the introductory section 1.3.1(SDOC). The description is standard, but two transformations specific to our application are introduced. The first is the use of a 'known' projected (predicted) ED_{50} value denoted by the P_{50}. In most subsequent computations, the ED_{50} is transformed by $\log(ED_{50}/P_{50})$. The interpretation of the ED_{50}/P_{50} as a measure of the precision of our projected ED_{50} at the time of planning the first phase 2 study is introduced. The second transformation replaces the E_{max} parameter representing the maximal compound effect at an infinite dose with difTarget, the effect of the compound at a pre-specified dose, dTarget. The dTarget is typically set to the highest observed dose. We found clinical teams understood difTarget at the highest dose better than E_{max}, and MCMC

algorithms tend to converge better using it because it is directly estimable using data from the highest dose. Note that difTarget and dTarget generalize the E_{max} parameter; setting dTarget to a high dose, e.g., 20 times the highest observed dose, makes difTarget and E_{max} equivalent for any users with a strong preference for the E_{max} parameterization.

II.3 Hierarchical E_{max} model for historical data

Section 2.1(SDOC) describes the historical dose response data collection and organization. As there have not been any queries regarding this section, we will not review it here.

The E_{max} models for the numerous historical compounds re-use the notation reviewed in Section II.2. The only generalization is the addition of subscripts denoting the different compounds. For the placebo response parameter (E_0), there is a second subscript indicating the study for compounds that have data from more than one study. Sections 1.3.1(SDOC) and 2.2.1(SDOC) discuss the need for potentially different placebo response even when studies have similar designs. The equality of the other parameters across protocols for the same compound is a topic of the meta-analysis and best viewed in the plots of the model fits in Appendix C(SDOC).

The full model involves the specification of the likelihood for the data, specification of a hierarchical model, specification of a prior distribution for the 'hyper' parameters of the hierarchical model, and specification of prior distributions for the placebo and effect parameters whose between-compound variation was not modeled. These specifications are reviewed in the next four sub-sections.

II.3.1 Likelihood function for the dose response data

The proposed methods to analyze dose response data from a new compound are ordinarily applied to patient-level data, although they can accommodate aggregated data, which is necessary in some circumstances. It was not possible to collect patient-level data from the historical dose response studies, so the meta-analysis of approximately 200 past compounds to create a prior distribution for future compounds was conducted on aggregated data. As will be noted, this is not a major limitation for our primary objective. The meta-data has the form $\bar{y}_{ijk}$, SE_{ijk}, d_{ijk}, n_{ijk}, SSY_{jk}, and df_{jk}, where i, j, k index dose, compound, and study (details are in Section 2.1(SDOC)). The $\bar{y}_{ijk}$ are sample means or proportions (binary outcomes), the SE_{ijk} are the usual (i.e., within dose group) standard errors of the $\bar{y}_{ijk}$, the d_{ijk} are the doses, the n_{ijk} are the dose/compound/study sample sizes, and for continuous endpoints, SSY_j and df_j are the pooled within-group sum of squares and degrees of freedom summed across all dose groups from all studies for compound j. For the E_{max} models of

continuous and binary outcomes in Section 2.2.1(SDOC), the aggregated data contain the sufficient statistics needed to compute the full likelihood.

For continuous outcomes, the $\bar{y}_{ijk}$ enter the likelihood as:

$$P\left(\bar{y}_{ijk} \mid d_{ijk}, \lambda_j, \mathrm{ED}_{50j}, \mathrm{difTarget}_j, \mathrm{E}_{0jk}\right) \sim N\left(\bar{y}_{ijk}; \mathrm{E}_{ijk}, \mathrm{SE}_{ijk}\right), \tag{R.1}$$

where E_{ijk} is

$$\mathrm{E}_{0jk} + \frac{\mathrm{E}_{\max_j} d_{ijk}^{\lambda_j}}{\mathrm{ED}_{50j}^{\lambda_j} + d_{ijk}^{\lambda_j}}\ , \tag{R.2}$$

and the $\mathrm{E}_{\max_j}$ are derived from $\mathrm{difTarget}_j$ and the other parameters. The SSY_j for compound j enters the likelihood as the usual scaled chi square represented by a gamma distribution:

$$P(\mathrm{SSY}_j \mid \sigma_j) \sim \Gamma(\mathrm{SSY}_j; \mathrm{shape} = \mathrm{df}_j/2, \mathrm{rate} = 1/(2\sigma_j^2))\ . \tag{R.3}$$

For binary data, the proportions are transformed to $\mathrm{logit}(\bar{y}_{ijk})$, and the SE_{ijk} are adjusted accordingly. The normal approximation is then applied to the transformed binomial variate, so the contribution to the likelihood has the same form as (R.1), with the logit transformation applied to E_{ijk} in (R.2). Note that while the normal approximation to the transformed rates is used for meta-analysis of the historical data, the binomial likelihood is used when evaluating new compounds with binary endpoints.

The $\bar{y}_{ijk}$ and SSY_j from all of the dose/compound/study groups are independent conditional on the parameters, so the terms of the form (R.1) and (R.3) combine multiplicatively in the likelihood. Using boldface notation to represent the large collections of terms, the contributions from all of the $\bar{y}_{ijk}$ and SSY_j are denoted:

$$P\left(\bar{\boldsymbol{y}} \mid \boldsymbol{d}, \boldsymbol{\lambda}, \mathbf{ED_{50}}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma}\right)\ , \tag{R.4}$$

$$P\left(\boldsymbol{SSY} \mid \boldsymbol{\sigma}\right)\ . \tag{R.5}$$

II.3.2 Hierarchical model

The hierarchical model representing the variation in $(\log(\lambda_j), \log(\mathrm{ED}_{50j}/\mathrm{P}_{50j}))$ is a bivariate distribution that is specified conditional on the P_{50j}, and some hyper-parameters, which must be estimated. A heavy-tailed t_5 distribution is specified, and the hyper-parameters are its means, scales, and potential correlation given by (μ_λ, $\mu_{\mathrm{ED}_{50}}$, σ_λ, $\sigma_{\mathrm{ED}_{50}}$, ρ_{12}). The model is fully specified (p. 13, Section 2.1.1(SDOC)):

$$\begin{aligned} P\Big(\log(\lambda_j), \log(\mathrm{ED}_{50j}/\mathrm{P}_{50j}) \mid \mathrm{P}_{50j}, \mu_\lambda, \mu_{\mathrm{ED}_{50}}, \sigma_\lambda, \sigma_{\mathrm{ED}_{50}}, \rho_{12}\Big) \sim \\ t_5\Big(\log(\lambda_j), \log(\mathrm{ED}_{50j}/\mathrm{P}_{50j});\ \mu_\lambda, \mu_{\mathrm{ED}_{50}}, \sigma_\lambda, \sigma_{\mathrm{ED}_{50}}, \rho_{12}\Big)\ . \end{aligned} \tag{R.6}$$

An equivalent form of (R.6) that more clearly represents the conditioning on the P_{50j} is

$$P\left(\log(\lambda_j), \log(ED_{50j}) \mid P_{50j}, \mu_\lambda, \mu_{ED_{50}}, \sigma_\lambda, \sigma_{ED_{50}}, \rho_{12}\right) \sim$$
$$t_5\left(\log(\lambda_j), \log(ED_{50j});\ \mu_\lambda, \mu_{ED_{50}} + \log(P_{50j}), \sigma_\lambda, \sigma_{ED_{50}}, \rho_{12}\right). \qquad \text{(R.7)}$$

The posterior distribution includes an independent multiplicative term of the form (R.7) conditional on the hyper-parameters and the P_{50j} for each compound. This product is denoted symbolically by:

$$P\left(\log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}) \mid \mathbf{P_{50}}, \mu_\lambda, \mu_{ED_{50}}, \sigma_\lambda, \sigma_{ED_{50}}, \rho_{12}\right)\ . \qquad \text{(R.8)}$$

II.3.3 Prior distribution for the hyper-parameters

The prior distributions for the means, scales, and correlation follow common practice for these parameters. A detailed description fully specifying these prior distributions is in Appendix A(SDOC). As there have not been any queries about the prior distributions of the hyper-parameters from this section, we will not review it here. One topic of note is that we performed sensitivity analyses to this prior specification. Results for 8 different combinations of the hyper-priors are reported. The results are reviewed in Section II.5 along with their role in constructing the predictive distribution used as the prior distribution for future compounds. To shorten the representation of the full posterior distribution of all of the parameters, the set of hyper-parameters are denoted by $\boldsymbol{\eta} = (\mu_\lambda, \mu_{ED_{50}}, \sigma_\lambda, \sigma_{ED_{50}}, \rho_{12})$, and their joint distribution by $P(\boldsymbol{\eta})$.

II.3.4 Prior distribution for the placebo and effect parameters

This topic is covered in the bottom half of p. 63, Appendix A(SDOC). There have been multiple queries related to it, so we will describe it with additional explanation here.

We begin with prior distributions for placebo response. The intent is to specify diffuse but numerically reasonable prior distributions for the placebo response. The scales of the endpoint for different compounds vary widely, so different diffuse scaling is required for each compound. With two hundred compounds and even more protocols, this is very cumbersome to input and validate in standard Bayesian analysis software. A simple device from mathematical statistics can easily avoid this problem. We apply linear re-scaling to the data to create standardized scales across compounds. Note that re-scaling the data does not change the ED_{50} or λ parameters, and the placebo and effect parameters are just linearly transformed and the MCMC-simulated values for these parameters can be simply back-transformed in a fully invariant manner. The likelihoods in equations (R.1) and (R.3) are applied with the transformed $\bar{y}_{ijk}$ and SSY_j.

Two standardizations are applied. First, for binary endpoints, following Chinn (2000), the logit response rates are divided by $\pi/\sqrt{3}$, yielding differences that can be interpreted as effects sizes for a continuous endpoint. Comparison of the placebo-adjusted effects at the highest doses for compounds with binary and continuous endpoints confirmed the success of the recommended re-scaling. Second, the mean of the endpoint for each compound was subtracted, and the centered endpoint was then divided by its SD. For binary endpoints, the scaling by the SD is omitted. The mean and SD were computed without regard to dose group or protocol, resulting in endpoints with mean differences between treatment groups somewhat smaller than the standard effect sizes often reported. The same independent identically distributed diffuse normal distribution is then applied to the mean of each placebo response:

$$P\left(\mathrm{E}_{0jk}\right) \sim N\left(\mathrm{E}_{0jk}; 0, SD = 3\right). \tag{R.9}$$

The independent product of each term in (R.9) is represented by $P\left(\mathbf{E_0}\right)$.

The prior distributions for the $\mathrm{difTarget}_j$ are constructed similarly, but the sampling of compounds with demonstrated efficacy in the meta-data requires an adjustment to commonly-applied prior distributions. First, the linear transformations applied to the data were also constructed to reverse the effect of compounds that had a negative effects, so all compounds have a positive effect enabling a single prior distribution. While a diffuse distribution for the standardized effects is used, there can still be a prior/data conflict when the prior distribution is centered at 0.0 because all of the compounds in the meta-data have demonstrated effect. After standardization, the placebo-adjusted sample mean at the highest dose from each compound has an observed mean of 0.85 and sample SD of 0.63. The prior distribution for the mean treatment effect was assigned a log t_5 distribution with its median set to 0.85, and a scale parameter (1.0) yielding a prior SD for the effect > 10 times the observed sample SD. The log transform was used to ensure an effect and reflect the skewed distribution of observed effects at the highest dose across compounds yielding independent identically distributed prior distributions:

$$P\left(log(\mathrm{difTarget}_j)\right) \sim t\left(log(\mathrm{difTarget}_j); log(0.85), \mathrm{scale} = 1.0\right). \tag{R.10}$$

The independent multiplicative product of terms in (R.10) from each compound are denoted by $P\left(log(\mathbf{difTarget})\right)$. The prior sensitivity assessments noted in Section II.3 included a lower and higher prior scale for the effect parameters in (R.10).

Finally, for continuous endpoints, the within-dose group SDs (standard deviation after normalization) were assigned a diffuse uniform distribution:

$$\sigma_j \sim U(\sigma_j; 0.01, 10)\ , \tag{R.11}$$

and the independent multiplicative product of the contributions in (R.11) from compounds with continuous endpoints is denoted by $P(\boldsymbol{\sigma})$.

NOTE: The data standardization and resulting prior specifications for placebo response and compound effects applied in the meta-data modeling are NOT applied to modeling of future compounds. The large scale (approximately 200 compounds) and special nature of the meta-data collection necessitated these adjustments. They are unneeded and inappropriate for the modeling of a future compound. Note also that the resulting posterior distributions for the placebo and effect size parameters included in the meta-data model play no direct role in subsequent applications.

II.4 Fitting the model for the meta-data

The posterior distribution is evaluated using MCMC methods. The essential output of the model fit are the $10,000$ draws from the posterior distribution of the hyper-parameters $\boldsymbol{\eta}$=(μ_λ, $\mu_{\mathrm{ED}_{50}}$, σ_λ, $\sigma_{\mathrm{ED}_{50}}$, ρ_{12}). The MCMC-generated hyper-parameters are the only output of the modeling that is subsequently used when computing the predictive distribution for future compounds. The remaining parameters are specific to the historical compounds, so they are of limited interest. They are used primarily for assessing the adequacy of the modeling of the meta-data. For example, the MCMC output of most of the parameters were used to compute the dose response curves fit to the historical data displayed in Appendix C(SDOC).

The full posterior distribution is constructed multiplicatively following the usual sequence of conditional distributions. The sample sizes are assumed ancillary and all of the probabilities condition on them. The degrees of freedom are a function of sample size, so their conditioning is also omitted in the derivation. From Bayes theorem, the posterior distribution of the parameters $P(\log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}), \boldsymbol{\eta}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma} \mid \bar{\boldsymbol{y}}, \boldsymbol{SSY}, \boldsymbol{d}, \mathbf{P_{50}})$ is proportional to

$$\begin{aligned} &P(\bar{\boldsymbol{y}}, \boldsymbol{SSY}, \boldsymbol{d}, \mathbf{P_{50}} \mid \log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}), \boldsymbol{\eta}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma}) \times \\ &\quad P(\log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}), \boldsymbol{\eta}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma}) \ . \end{aligned} \tag{R.12}$$

Conditioning on the ancillary data, $\boldsymbol{d}$ and $\mathbf{P_{50}}$, note that $\bar{\boldsymbol{y}}$ and $\boldsymbol{SSY}$ depend on a subset of the parameters (equations (R.4, R.5)). They are also independent of each other and the $\mathbf{P_{50}}$ when the $\mathrm{E_{max}}$ model parameters are known, so (R.12) is equal to

$$\begin{aligned} &P(\bar{\boldsymbol{y}} \mid \boldsymbol{d}, \log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}), \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma})\, P(\boldsymbol{SSY} \mid \boldsymbol{\sigma}) \times \\ &\quad P(\boldsymbol{d}, \mathbf{P_{50}} \mid \log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}), \boldsymbol{\eta}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma}) \times \\ &\quad P(\log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}), \boldsymbol{\eta}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma}) \ . \end{aligned} \tag{R.13}$$

Applying Bayes theorem, the middle probability in equation (R.13) becomes

$$\frac{P\left(\log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}) \mid \boldsymbol{d}, \mathbf{P_{50}}, \boldsymbol{\eta}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma}\right) P\left(\boldsymbol{d}, \mathbf{P_{50}} \mid \boldsymbol{\eta}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma}\right)}{P\left(\log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}) \mid \boldsymbol{\eta}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma}\right)} \ . \tag{R.14}$$

The hierarchical model specifies that $\log(\boldsymbol{\lambda})$ and $\log(\mathbf{ED_{50}})$ are independent of all of the conditioning terms in the left probability in the numerator of equation (R.14) except for $\mathbf{P_{50}}$ and $\boldsymbol{\eta}$, so this terms reduces to the hierarchical model (R.8) in section II.3.2: $P\big(\log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}) \mid \mathbf{P_{50}}, \boldsymbol{\eta}\big)$. The right probability in the numerator can be discarded because the usual ancillary assumption regarding the $\boldsymbol{d}$ and $\mathbf{P_{50}}$ ensures this probability does not depend on the parameters (recall that $\boldsymbol{\eta}$ are the parameters of the conditional distribution of $(\log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}})$ given $\boldsymbol{d}$ and $\mathbf{P_{50}}$). The final probability in (R.13) factors as

$$P\left(\log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}) \mid \boldsymbol{\eta}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma}\right) P\left(\boldsymbol{\eta}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma}\right) \ . \tag{R.15}$$

The denominator of (R.14) cancels the first term in (R.15). Noting that $(\boldsymbol{\eta}, \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma})$ are a priori independent yields the standard form for a hierarchical model:

$$\begin{aligned} &P\left(\bar{\boldsymbol{y}} \mid \boldsymbol{d}, \log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}), \mathbf{difTarget}, \mathbf{E_0}, \boldsymbol{\sigma}\right) P\left(\boldsymbol{SSY} \mid \boldsymbol{\sigma}\right) \times \\ &\quad P\left(\log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}) \mid \mathbf{P_{50}}, \boldsymbol{\eta}\right) \times \\ &\quad P\left(\boldsymbol{\eta}\right) P\left(\mathbf{difTarget}\right) P\left(\mathbf{E_0}\right) P\left(\boldsymbol{\sigma}\right) \ . \end{aligned} \tag{R.16}$$

Each probability in (R.16) is supplied in Sections II.3.1-II.3.4. Some of the hierarchical models included correlations between $(\log(\boldsymbol{\lambda}), \log(\mathbf{ED_{50}}))$ and **difTarget**. The same derivation applies for this expanded model. The change required is to include the three terms together in the hierarchical model, and remove the $P\left(\mathbf{difTarget}\right)$ from the non-hierarchical priors in the last line of (R.16). While the additional correlations are estimated, the prior means and scale parameters for **difTarget** are still set to known diffuse values.

II.5 Computing the predictive distribution (MAP) for a future compound

The values of $\log(\lambda)$ and $\log(\mathrm{ED}_{50}/\mathrm{P}_{50})$ for a future compound are predicted from their bivariate hierarchical t_5 distributions in (R.6). This model is conditional on the P_{50} of the future compound, the parameters $(\mu_\lambda, \sigma_\lambda, \mu_{\mathrm{ED}_{50}}, \sigma_{\mathrm{ED}_{50}})$, and for some models, ρ_{12}. The MCMC fit produces $10,000$ random draws from the posterior distribution of $(\mu_\lambda, \sigma_\lambda, \mu_{\mathrm{ED}_{50}}, \sigma_{\mathrm{ED}_{50}})$, and ρ_{12}. Random samples from the posterior predictive distribution for a future compound can thus be generated by simulating a single prediction from the conditional t_5 hierarchical distribution for each set of randomly generated $(\mu_\lambda, \sigma_\lambda, \mu_{\mathrm{ED}_{50}}, \sigma_{\mathrm{ED}_{50}})$,

and ρ_{12}. The randomly generated values of λ and ED_{50} for a future compound are then computed by back-transformation. The simulation approach described here is equivalent to approximating the predictive distribution with $P_{50} = 1$, and then adding the $\log(P_{50})$ from a future compound to this approximation (see equation (R.7)).

With $10,000$ draws, the predictive distribution can be precisely approximated. This is common Bayesian practice (e.g., Schmidli et al. (2014)) because it then allows standard Bayesian methods to be applied to future compounds. The predictive distribution is a mixing of t_5 distributions with different means and scales. As the $t - distribution$ is itself a mixing of normal distributions, it is not surprising that another $t - distribution$ can yield an adequate approximation to the predictive distribution. Q-Q plots of the $10,000$ draws of each parameter from the final selected predictive distribution (discussed below) versus the quantiles of the t_5 distribution are displayed in Figures R.1 and R.2. Without even adjusting the degrees of freedom, a $t_5 - distribution$ can provide a fit-for-purpose approximation. Regression of the generated $\log(\lambda)$ on $\log(ED_{50}/P_{50})$ also confirmed the linear relationship between the variables required by the bivariate $t - distribution$. The mean and scale parameters of the approximating t_5 distribution were computing by applying maximum likelihood estimation to the $10,000$ simulated values. The final selected model fixes the means at zero (to be discussed) so the means were not approximated for this model. The correlation between the $\log(\lambda)$ and $\log(ED_{50}/P_{50})$ was approximated simply by the sample correlation. As reported in Appendix A(SDOC), the scale values and correlation are 0.425, 1.73, and -0.45. The scale parameters approximated by simple method of moments (scale=$\sqrt{(3/5)}$SD) change inconsequentially to 0.434 and 1.74.

Figures 1 (displaying $\log(\lambda)$) and 2 (displaying $\log(ED_{50}/P_{50})$) (Section 2.3(SDOC)) summarize the predictive distributions yielded by the eight different sets of hyper-parameter distributions. The different hyper-parameter distributions are described in detail in Appendix A (SDOC). The $\log(ED_{50}/P_{50})$ distributions display little sensitivity to the distribution of the hyper-parameters. The P_{50} is above the ED_{50} about as often as it is below the ED_{50}. The $\log(\lambda)$ predictive distributions vary more with different distributions for the hyper-parameters. There were three distinct differences: 1) priors that fix the expected value of $\log(\lambda)$ at zero versus those that estimate the mean, 2) priors that allow for correlation between $\log(\lambda)$ and $\log(ED_{50}/P_{50})$, and 3) priors that set the lower bound for λ at 0.5. Over limited dosing ranges, the dose response curves from models with $\lambda = 0.75$ are nearly indistinguishable from appropriately-matched hyperbolic models with $\lambda = 1$, so this difference is not important for the intended applications of the resulting prior distributions. Because users sometimes force $\lambda = 1$ to use the well-known hyperbolic model, we prefer this centering. The second source of differences is that prior distributions (mod51,

mod52) that include potential correlation between the parameters yield more dispersed distributions. Seeking to be conservative in the construction of the predictive model, this suggested the use of the model with λ centered at the hyperbolic model that includes correlation (mod52). Finally, the third source of differences were prior distributions that truncated the lower bound of λ at 0.5 (mod101, mod102). They were included to check for the possibility that they would favor larger values of λ than the other prior distributions. This is not the case; the truncated distributions instead resulted in tighter concentration around the central tendencies of the other models.

From Figures 1 and 2 (SDOC), the selected t_5 model (mod52, third from top) includes the high density region and support produced by the other hyper-priors. The variation between these predictive distributions is larger than the approximation errors in the $t-$ *distribution*. This is another reason why more complex approximating distributions are not warranted. Also important, subsequent exploration of the sensitivity of estimation of dose response curves utilizing the different predictive distributions for $\log(\lambda)$ and $\log(\mathrm{ED}_{50}/\mathrm{P}_{50})$ display low sensitivity. The sources of sensitivity considered in this section are small when compared to the much larger deviations explored in Section 3.3 (SDOC).

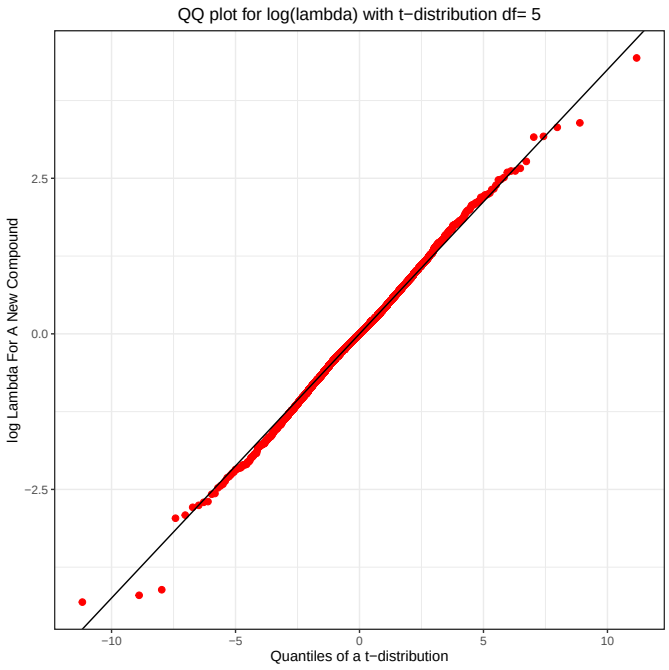


Figure R.1: QQ plot of the 10000 draws from the predictive distribution for $\log(\lambda)$ versus a t_5- *distribution*.

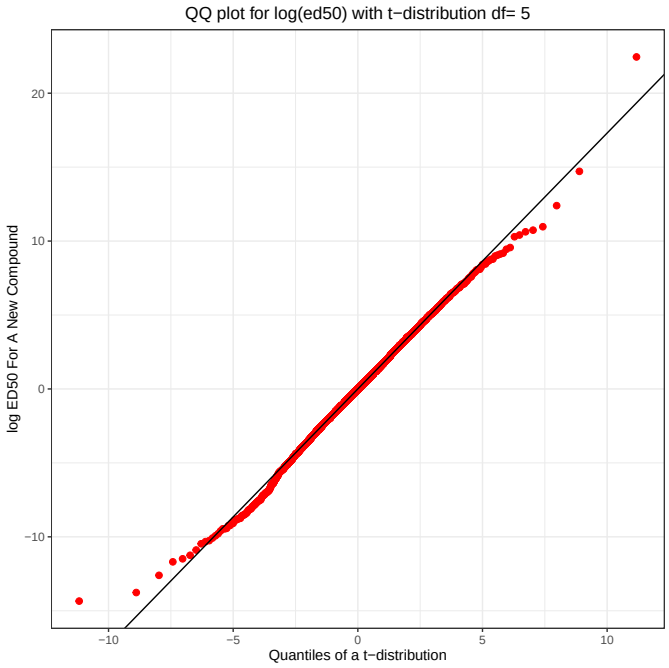


Figure R.2: QQ plot of the 10000 draws from the predictive distribution for $\log(\mathrm{ED}_{50}/\mathrm{P}_{50})$ versus a $t_5 - distribution$.

II.6 Combining compound-specific information with the predictive prior distribution (MAP) to complete the prior distribution for a new compound

The process of combining compound-specific information with the meta-analytic prior distribution for the λ and ED_{50} is described in detail in Sections 3.1.1 and 3.1.2 (SDOC), so it is not repeated here. Instead, our focus is on two topics related to the construction of the meta-analytic prior distribution.

The first is just to repeat that the predictive prior distribution for the λ and ED_{50} for a new compound under development is approximated by a bivariate t-distribution:

$$t_5\left(\log(\lambda), \log(\mathrm{ED}_{50}/\mathrm{P}_{50});\ \text{mean} = (0,0), \text{scale} = (0.425, 1.73), \rho = -0.45\right)\ . \qquad \text{(R.17)}$$

The specification of the P_{50} by the study team is discussed in Section 3.1.1 (SDOC). The prior distribution for the remaining parameters are specified independently of the λ and ED_{50} and usually assigned diffuse distributions.

The second consideration is the sampling of compounds that was feasible for the meta-data collection. Because of the large number of failed compounds and the lower quality of documentation for past studies with disappointing results, we were only able to collect data from compounds that demonstrated effect in phase 2 development, indeed, many of

the compounds were sampled from lists of approved compounds. Most, but not all future compounds beginning a dose ranging study, have a positive proof of concept (POC) study. The POC studies are typically small (single high dose and placebo) and compounds move forward because of the positive result, so it has been long observed that the results in the next development stage tend to suffer a 'regression' effect. The compounds in the meta-data, which were selected based on successful end of phase 2 data, tend to have large effect sizes than the compounds beginning dose ranging. Because of this and other technical challenges, the effect sizes from the meta-data are not incorporated into the formal Bayesian analysis of future compounds. Instead, the standard practice is to specify a diffuse prior distribution for the effect size of the new compound conservatively centered at no effect. An example of the prior construction is in Section III.1.

III Example with two protocols

The example in Section 3.2.1 (SDOC) is based on internal development of a compound. Patient-level data are available and it has two protocols, so it used here. The compound is older, but it is treated as if it is a new compound to illustrate the methods. The real analysis utilized the Bayesian E_{max} model approach, but its development pre-dated the collection of the meta-data, so the prior distribution implemented here differs somewhat from the original analysis. The two dose response protocols were executed sequentially. Because two protocols were requested, we describe the second analysis when data from both protocols are available. The first dose response analysis applied to the first protocol alone was used to guide design of the second protocol. It is described on the first two pages of Section 3.2.1 (SDOC), and not repeated here.

The two dose response studies were conducted as part of the development of tofacitinib for the treatment of rheumatoid arthritis (RA). The primary endpoint is the binary responder $ACR20$. Missing endpoints were treated as failure as was the standard for the $ACR20$ endpoint at the time. The twice-daily (BID) dosing regimen was studied throughout.

The first study had total daily doses of 0, 10, 30, and 60 with sample sizes of 62, 58, 68, and 63. A plot with the sample response rates for the $ACR20$ is in the left panel of Figure R.3. The red asterisks are the sample responder proportions. The second study had doses of 0, 2, 6, 10, 20, and 30 with samples sizes of $(46, 44, 46, 45, 56, 53)$. Additional details are given in the submission document. The sample response rates for the second study are in the right panel of Figure R.3.

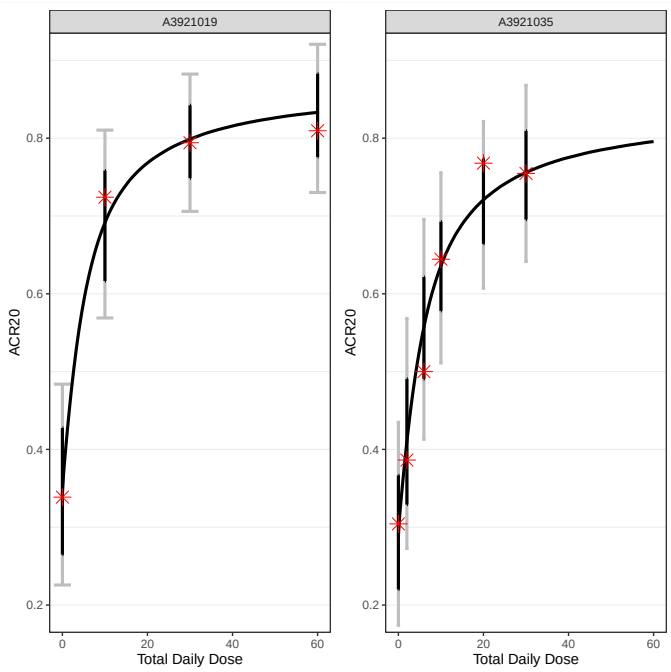


Figure R.3: Two dose response studies of a treatment for RA with $ACR20$ responder endpoint (binary). The red asterisk are observed dose group proportions. The solid curve is the posterior median estimator of response computed over a grid of doses. The solid bars are 90% posterior intervals for the population values, and the gray bars are 90% prediction intervals. The placebo response is allowed to differ in the fitted curves, but the other parameters in the model are the same across the two studies. The model output is back-transformed to the response rate scale.

III.1 Specification of the prior distribution

This section follows the description of prior specification in Section 3.2.1(SDOC). The projected $P_{50} = 10$ mg total daily dose, which was derived from pre-clinical data using clinical pharmacology methods. The P_{50} was also supported by limited clinical data in two other indications. This is the only compound-specific input required to specify the prior bivariate $t - distribution$ for the λ and ED_{50} using the predictive MAP distribution (Sections II.5, 2.3 (SDOC)):

$$t_5\left(\log(\lambda), \log(ED_{50});\ \text{mean} = (0, \log(10)), \text{scale} = (0.425, 1.73), \rho = -0.45\right)\ . \qquad \text{(R.18)}$$

A placebo response rate of 0.15 was predicted from historical data, but there was high between-study heterogeneity in the historical data, and there was evidence that placebo response was increasing over time. Diffuse t_5 prior distributions for E_{01} and E_{02} from the two studies on the logit scale were thus specified with a prior scale parameter of 4 and

mean of logit(0.15) for the logit-transformed rate:

$$\begin{aligned} P(\mathrm{E}_{01}, \mathrm{E}_{02}) =& t_5\left(\mathrm{E}_{01};\ \text{mean} = \text{logit}(0.15), \text{scale} = 4\right) \times \\ & t_5\left(\mathrm{E}_{02};\ \text{mean} = \text{logit}(0.15), \text{scale} = 4\right)\ . \end{aligned} \tag{R.19}$$

Following the broader guidance in Section 3.1.2 (SDOC), separate independent placebo responses were estimated from the two studies without any pooling of the placebo response across them. The placebo responses from the studies are not compared once data are available to test for possible pooling.

Similarly, the response difference between the highest dose (dTarget = 60 mg) and placebo was independently specified and centered at no effect on the logit scale with wide uncertainty in a t_5 distribution:

$$P(\text{difTarget}) = t_5\left(\text{difTarget};\ \text{mean} = 0, \text{scale} = 4\right)\ . \tag{R.20}$$

The full prior distribution, denoted $P(\lambda, \mathrm{ED}_{50}, \text{difTarget}, \mathrm{E}_{01}, \mathrm{E}_{02})$, is the product of the independent distributions in equations (R.18), (R.19), and (R.20).

III.2 Model fitting

The notation used in Section II is modified for patient-level data from two trials. The y_{ijk} are binary responder endpoints with corresponding doses d_{ijk}, where k is now the index for study ($k = 1, 2$), j is the index for dose group within study, $j = 1, \ldots, n_{dk}$, where n_{dk} is the number of dose groups in study k, and i indexes the n_{jk} patients in dose group j in study k. The $\mathrm{E}_{\max}$ model is applied on the logit scale as before:

$$P(Y_{ijk} = 1 \mid d_{ijk}, \lambda, \mathrm{ED}_{50}, \text{difTarget}, \mathrm{E}_{01}, \mathrm{E}_{02}) = \text{logit}^{-1}\left(\mathrm{E}_{0k} + \frac{\mathrm{E}_{\max} d_{ijk}^{\lambda}}{\mathrm{ED}_{50}^{\lambda} + d_{ijk}^{\lambda}}\right)\ , \tag{R.21}$$

where

$$\mathrm{E}_{\max} = \text{difTarget}\left(\frac{\text{dTarget}^{\lambda} + \mathrm{ED}_{50}^{\lambda}}{\text{dTarget}^{\lambda}}\right)\ ,$$

with dTarget = 60 mg. The Y_{ijk} are independent conditional on the model parameters and their doses. Denoting the collection of all responses and their doses by $\mathcal{Y}$ and $\mathcal{D}$, the likelihood can be written compactly as

$$\begin{aligned} & lik\big(\mathcal{Y}, \mathcal{D};\ \lambda, \mathrm{ED}_{50}, \text{difTarget}, \mathrm{E}_{01}, \mathrm{E}_{02}\big) \\ & \quad = P\left(\mathcal{Y} \mid \mathcal{D}, \lambda, \mathrm{ED}_{50}, \text{difTarget}, \mathrm{E}_{01}, \mathrm{E}_{02}\right) \\ & \quad = \prod_{k=1}^{2} \prod_{j=1}^{n_{dk}} \prod_{i=1}^{n_{jk}} P\left(Y_{ijk} = 1 \mid d_{ijk}\right)^{Y_{ijk}} \left(1 - P(Y_{ijk} = 1 \mid d_{ijk})\right)^{1 - Y_{ijk}}\ . \end{aligned} \tag{R.22}$$

Applying Bayes theorem to $(\mathcal{Y}, \mathcal{D})$, the posterior distribution, $P(\lambda, \mathrm{ED}_{50}, \mathrm{difTarget}, \mathrm{E}_{01}, \mathrm{E}_{02} \mid \mathcal{Y}, \mathcal{D}, \mathrm{P}_{50})$, is proportion to

$$P\left(\mathcal{Y}, \mathcal{D} \mid \lambda, \mathrm{ED}_{50}, \mathrm{difTarget}, \mathrm{E}_{01}, \mathrm{E}_{02}, \mathrm{P}_{50}\right) P\left(\lambda, \mathrm{ED}_{50}, \mathrm{difTarget}, \mathrm{E}_{01}, \mathrm{E}_{02} \mid \mathrm{P}_{50}\right) \ . \quad \text{(R.23)}$$

The left term in equation (R.23) factors into

$$P\left(\mathcal{Y} \mid \mathcal{D}, \lambda, \mathrm{ED}_{50}, \mathrm{difTarget}, \mathrm{E}_{01}, \mathrm{E}_{02}, \mathrm{P}_{50}\right) P\left(\mathcal{D} \mid \lambda, \mathrm{ED}_{50}, \mathrm{difTarget}, \mathrm{E}_{01}, \mathrm{E}_{02}, \mathrm{P}_{50}\right) \ . \quad \text{(R.24)}$$

The P_{50} is not predictive given the $\mathrm{E}_{\max}$ model parameters. The usual ancillary condition for the doses reduces the right term in equation (R.24) to $P\left(\mathcal{D} \mid \mathrm{P}_{50}\right)$, so the posterior distribution is proportional to the standard form:

$$lik\big(\mathcal{Y}, \mathcal{D};\ \lambda, \mathrm{ED}_{50}, \mathrm{difTarget}, \mathrm{E}_{01}, \mathrm{E}_{02}\big) P(\lambda, \mathrm{ED}_{50}, \mathrm{difTarget}, \mathrm{E}_{01}, \mathrm{E}_{02}) \ .$$

The posterior distribution is evaluated using MCMC stochastic simulation methods. The MCMC yields a large number of nearly independent draws from the posterior distribution that can be subsequently used to evaluate complex estimands because they do not require the derivation of asymptotic approximations for context-specific estimands.

III.2.1 Model output and checking

Standard approaches are applied to check the numerical validity of the MCMC fitting. Three chains were fit to check the mixing of the MCMC chains. Figure R.4 displays the sequence of parameters generated (trace plot) after an initial burn-in of 1000 draws. The chains are well-mixed, appear stationary, and have low auto-correlation. The low auto-correlation is confirmed in a plot of auto-correlations for each parameter in Figure R.5. Gelman-Rubin diagnostics (not shown) support the stationarity and convergence.

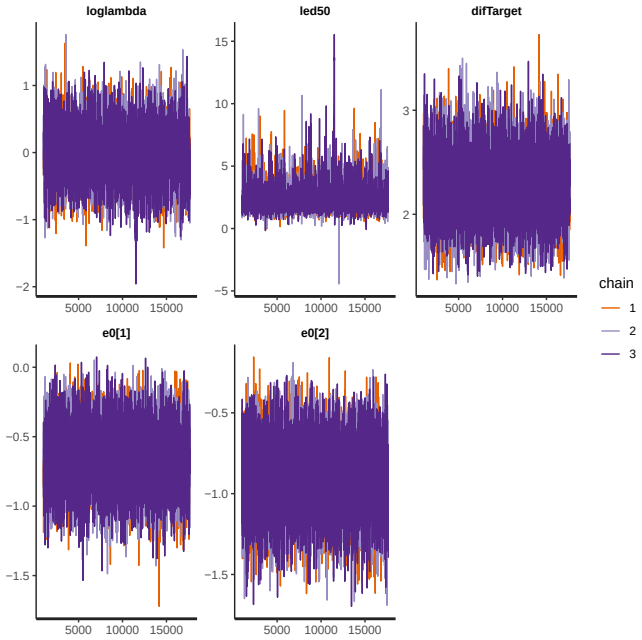


Figure R.4: Traceplot based on 3 chains from the MCMC fit of the E_{max} model. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

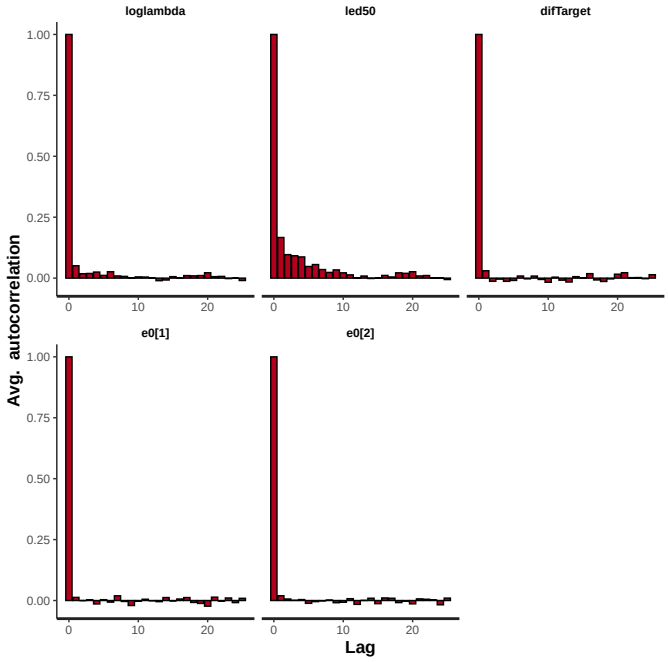


Figure R.5: Auto-correlations from the MCMC generated parameters of the E_{max} model. The burn-in was 1000 iterations, and a thinning gap of 5 was used to ensure low auto-correlation.

The output of the MCMC fit is $10,000$ draws of the parameters from the posterior distribution. Point estimates and intervals for the parameters can be computed from the draws, but we emphasize that these are not a primary objective of the analysis, and we encourage the practice of demoting them to appendices because they are unreliably estimated with some data sets, and they can be difficult to interpret. They are reported in Table R.1, but note that the results are on the logit scale, which is not helpful for clinicians. Further, the estimates in Table R.1 play no role in the subsequent analyses. All subsequent analyses are computed from the full $10,000$ posterior draws of the parameters

Figure R.3 is the first analysis of the model fitting that is routinely reported. It provides an interpretable graphical summary of the model results, and it is also useful for assessing model fit. The solid model curves are computed from the $10,000$ MCMC draws by creating a narrowly-spaced grid of doses. For each dose on the grid, the probability of response (after back-transformation) is computed for each of the $10,000$ sets of model parameters. The $10,000$ probabilities, which are conditional on the simulated model parameters, are denoted by $p_m, m = 1, \ldots, 10,000$. They are computed using the formula in equation (R.21) with the dose in that formula replaced by each dose on the grid. The medians of the $10,000$ p_m are plotted. The dark error bars are the 5^{th} and 95^{th} percentiles of the $10,000$ thousand p_m, which approximate the corresponding percentiles of the posterior distribution of the response rate at each tested dose.

The grey intervals at each tested dose are the posterior predictive intervals for the sample proportions in a new study matching the current study. They are computed by simulating a sample proportion, denoted by $\bar{y}_m$, corresponding to each of the $10,000$ p_m with the dose group sample size matching the size in the study. For example, 68 patients were assigned to the 60 mg dose in the first study. For each of the p_m computed for the 60 mg dose, a sample proportion was generated by drawing from a binomial distribution: Binomial$(n = 68, p_m)/68$. This process yields $10,000$ $\bar{y}_m$ drawn from the predictive distribution of the sample proportion that would be observed in a future study with 68 patients receiving the 60 mg dose conditional on the data actually observed in our two studies. Prediction intervals, etc, are computed from simple summaries of the $10,000$ simulated $\bar{y}_m$. Note that the predictive distribution is discrete (e.g., there are only 69 possible values), so a 90% intervals has at least 90% probability because exact achievement of 90% may not be possible.

The posterior predictive intervals are useful for judging outliers and other patterns of poor fit. The predictive intervals supplement the routinely computed posterior predictive check for non-monotonicity. The value of this predictive check in the current example applied to the second study is 0.323. Details of its calculation are given in Section 2.4(SDOC)

and not repeated here. Note that the predictive intervals and the non-monotonicity predictive check probability refer to the posterior predictive distribution derived from the posterior distribution for the new compound after observing its dose response study(ies). It does not refer back to the predictive distribution computed from the meta-data.

Parameter	Posterior Median	90% Posterior Interval
λ	1.04	(0.53,1.88)
ED_{50}	7.8	(3.5,49.4)
difTarget	2.26	(1.78,2.77)
E_{01}	-0.65	(-1.02,-0.29)
E_{02}	-0.9	(-1.27,-0.55)

Table R.1: Parameter estimation based on data from two studies.

III.3 Using the model output to guide dose selection

As can be observed from Figure R.3, the compound is highly active so there was very high confidence it would achieve statistical significance in phase 3 studies. Clinical assessment was that a dose yielding < 0.2 difference from placebo would be a failure, and a fully successful dose would need to yield ≥ 0.3 improvement versus placebo. Table 2(SDOC) was the primary model output used to select doses to satisfy these objectives; it is reproduced here as Table R.2 for ease of reference, and the label 'Mean' is replaced by 'Prop', which is a better descriptor for a responder endpoint.

	Population Responder Rates			Phase 3 Sample Responder Rates		
Total	Prop	Probability		Prop	Probability	
Dose	Diff	Diff> 0.2	Diff> 0.3	Diff	Diff> 0.2	Diff> 0.3
2 mg	0.12	0.1	0	0.11	0.1	0.01
4 mg	0.21	0.58	0.07	0.18	0.39	0.06
6 mg	0.27	0.89	0.31	0.23	0.65	0.17
8 mg	0.31	0.98	0.6	0.27	0.81	0.33
10 mg	0.35	1	0.8	0.3	0.9	0.48
12 mg	0.37	1	0.9	0.32	0.94	0.58
14 mg	0.39	1	0.95	0.33	0.96	0.66
16 mg	0.4	1	0.97	0.34	0.97	0.72
18 mg	0.41	1	0.98	0.35	0.98	0.76
20 mg	0.42	1	0.99	0.36	0.99	0.8
30 mg	0.46	1	1	0.38	0.99	0.89

Table R.2: The estimated difference from placebo and the probability the difference exceeds 0.20 or 0.30. The left columns are results for the population values. The right columns are based on the posterior predictive probabilities for observed responder rates from a phase 3 trial. The reduction in the differences for sample rates is due to dropouts.

The left side of Table R.2 provides Bayesian estimates and inference for the population values. For the dose in each row, the difference in the response rate between the dose and placebo is estimated by the median of the $10,000$ differences in responder rates computed from the MCMC model parameter output. The computation follows the same approach as for the dose response curve in Figure R.3 except now the placebo response must also be computed. It is denoted by p_{0m} and the difference by $d_m = p_m - p_{0m}, m = 1, \ldots, 10,000$. The MCMC-simulated E_{01} from the first study were used when computing p_m and p_{0m}. The placebo-adjusted difference is not sensitive to the which of the placebo responses is used because the E_0 nearly cancels in the difference, but unlike continuous data, the logit binary model does not produce exact cancellation. The posterior probability that the response rate from a dose exceeds 0.2 or 0.3 is precisely approximated by the proportion of the $10,000$ of d_m that exceed these levels.

The population difference in responder rates is a hypothetical construct. In practice, the phase 3 success criteria will be applied to the observed phase 3 estimates, which are simple observed sample proportions with patients who drop out regarded as failures (the handling of missing data has changed since the time these protocols were analyzed). The predictive distribution for the sample placebo adjusted responder rate follows closely the construction of the predictive intervals in Figure R.3. There are two differences: 1) the sample size for the active dose in the planned phase 3 trial is approximately 200, and it is 100 for placebo, and 2) 15% of the patients in the the phase 3 trial dropout and their response status is set to failure regardless of the responder status generated for them before consideration of dropout status. Dropout was assumed to be MCAR, and was implemented by assigning an independent 15% dropout chance to each simulated patient.

The population placebo-adjusted responder rate for the 10 mg dose has an 80% chance of exceeding 0.3, but after accounting for sampling variability and attenuation in observed effect due to dropout, this probability reduces to 0.48. The 20 mg dose has 80% chance of producing an observed 0.3 improvement criteria. Both doses are likely to achieve the minimally acceptable criteria of 0.2 improvement. Due to safety assessments (not evaluated here), both the 10 and 20 mg doses were evaluated in phase 3. The efficacy model predictions were accurate. The 10 mg dose was approved, but the 20 mg dose was not approved even though it was superior on efficacy by the amount predicted in Table R.2, validating the decision to carry both doses into phase 3 development.

The development context is different for each compound, but the example presented here illustrates how the MCMC output of the Bayesian E_{max} model can be flexibly adapted to decision criteria specific to each compound under development without requiring the derivation of new approximation formulas.

References

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine 19*, 3127–3131.

Schmidli, H., S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics 70*, 1023–1032.