Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document, contact (CDER) Scott Goldie at Scott.Goldie@fda.hhs.gov, or (CBER) Office of Communication, Outreach and Development, 800-835-4709 or 240-402-8010.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

January 2026
Biostatistics

Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products

Guidance for Industry

Additional copies are available from:

Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration
Phone: 855-543-3784 or 301-796-3400
Email: druginfo@fda.hhs.gov
https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugs

and/or

Office of Communication, Outreach, and Development
Center for Biologics Evaluation and Research
Food and Drug Administration
Phone: 800-835-4709 or 240-402-8010; Email: industry.biologics@fda.hhs.gov
https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics/biologics-guidances

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

January 2026
Biostatistics

TABLE OF CONTENTS

Use of Bayesian Methodology in Clinical Trials of Drugs and Biologics Guidance for Industry[1]

This draft guidance, when finalized, will represent the current thinking of the Food and Drug Administration (FDA or Agency) on this topic. It does not establish any rights for any person and is not binding on FDA or the public. You can use an alternative approach if it satisfies the requirements of the applicable statutes and regulations. To discuss an alternative approach, contact the FDA staff responsible for this guidance as listed on the title page.

I. INTRODUCTION

This document provides guidance to sponsors and applicants submitting investigational new drug applications (INDs), new drug applications (NDAs), biologics licensing applications (BLAs), or supplemental applications on the appropriate use of Bayesian methods in clinical trials. Bayesian methods can be used in various ways in clinical trials. For example, Bayesian calculations can be used to govern the timing and adaptation rules for an interim analysis in an adaptive design, to inform design elements (e.g., dose selection) for subsequent clinical trials, or to support primary inference in a trial. The primary focus of this guidance is on the use of Bayesian methods to support primary inference in clinical trials intended to support the effectiveness and safety of drugs.[2]

In general, FDA's guidance documents do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidances means that something is suggested or recommended, but not required.

II. BACKGROUND

A. Definition

Bayesian statistics is an approach to estimation or hypothesis testing to draw inference based on the use of Bayes' theorem. In a Bayesian analysis, data collected in a study are combined with a prior distribution that captures the pre-study information about a parameter of interest to form a

[1] This guidance has been prepared by the Office of Biostatistics in the Center for Drug Evaluation and Research and the Division of Biostatistics in the Center for Biologics Evaluation and Research at the Food and Drug Administration.

[2] For the purposes of this guidance, all references to drugs include both human drugs and biological products unless otherwise specified.

posterior distribution that expresses the updated, post-study information about the parameter of interest (e.g., the primary estimand[3]). The prior distribution often represents a summary of information and uncertainty available before the study begins. The posterior distribution can be used for inference and to draw conclusions about efficacy or safety.

B. Important Concepts

The following are definitions of important concepts used in this guidance:

- The *prior distribution* or *prior* is the pre-study probability distribution for model parameters.

- The *likelihood function* or *likelihood* describes the quantitative relationship between the parameters of interest and the study data. The mathematical form of the likelihood is determined by the model being used (for example, linear regression, logistic regression, ordinal regression).

- The *posterior distribution* or *posterior* is the post-study probability distribution for the parameter of interest. It is obtained by combining the prior distribution and the likelihood using Bayes theorem. It quantitatively summarizes what is known about the parameter of interest following collection of study data and can be used to draw inferences on the study hypotheses. Inference is often based on summary measures of this distribution. For example, evaluation of a treatment effect may be informed by the posterior mean to estimate the effect, a credible interval to quantify uncertainty around the estimated effect, and relevant posterior probabilities (e.g., the posterior probability that the effect is greater than zero). Credible intervals are intervals of possible values for the unobserved parameter that will contain the parameter value with a specified probability under the posterior distribution (e.g., 95% posterior probability).

- *Bayes theorem* is the mathematical rule for combining the prior distribution and likelihood together to form the posterior distribution.

III. SITUATIONS WHERE BAYESIAN METHODS HAVE BEEN USED

This section discusses settings and specific examples from development programs where Bayesian methods have been used in submissions to the Agency. Most of these case examples focus on the use of *borrowing* or *leveraging* of previously available trials or information across populations within a trial. Bayesian methods can also be considered in other settings.

[3] For discussion of estimands, see the ICH guidance for industry *E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials* (May 2021). For the most recent version of a guidance, check the FDA guidance web page at https://www.fda.gov/regulatory-information/search-fda-guidance-documents.

A. Borrowing from Previous Clinical Trials

Under certain circumstances, an informative prior for a clinical trial analysis can be formed based on results from previous clinical trial(s) of the same drug. For example, borrowing from a previous trial was used in a phase 3 study to evaluate REBYOTA, a fecal microbiota transplant product, for the prevention of recurrence of *Clostridioides difficile* infection (CDI) in individuals with recurrent CDI. *C. difficile* is a common cause of antibiotic-associated diarrhea and colitis and is a major public health burden. The primary analysis of the randomized, double-blind, placebo-controlled phase 3 study to evaluate the effectiveness of REBYOTA used a Bayesian model to formally incorporate data from a previous phase 2 placebo-controlled study of REBYOTA. This analysis supported the effectiveness of REBYOTA, which was approved in 2022. [4]

B. Augmenting a Randomized Concurrent Control Using an External Control or Nonconcurrent Control Data

In some cases, it can be challenging to conduct an adequately powered randomized trial due to limited population and/or ethical considerations. Borrowing data from an external or nonconcurrent control to augment the randomized concurrent control may be appealing in these situations, and Bayesian methods have been proposed to implement such approaches. For example, Bayesian methods have been proposed to augment the randomized concurrent control and leverage nonconcurrent control data in the oncology platform trials GBM AGILE[5] and Precision Promise,[6] which evaluate marker-targeted treatments for patients with glioblastoma and pancreatic cancer, respectively. The analyses use a Bayesian model to try to account for temporal shifts in efficacy outcomes such as tumor response (Saville et al. 2022).[7] As another example, a non-inferiority study of pediatric patients with multiple sclerosis[8] was proposed through the Complex Innovative Trial Design (CID) program[9] with a prespecified Bayesian analysis to leverage information from historical studies of the active comparator.

C. Pediatric Extrapolation

Pediatric extrapolation is defined in the ICH E11(R1)[10] guideline as "an approach to providing evidence in support of effective and safe use of drugs in the pediatric population when it can be assumed that the course of the disease and the expected response to a medicinal product would be sufficiently similar in the pediatric [target] and reference (adult or other pediatric)

[4] See prescribing information for Rebyota (fecal microbiota, live – jslm) suspension (https://www.fda.gov/media/163587/download?attachment).
[5] For additional details on GBM AGILE, see https://www.clinicaltrials.gov/study/NCT03970447.
[6] For additional details on Precision Promise, see https://www.clinicaltrials.gov/study/NCT04229004.
[7] See the FDA draft guidance for industry *Master Protocols for Drug and Biological Product Development* (December 2023) for additional discussion on the use of nonconcurrent control data in platform trials and potential for temporal shifts to lead to bias. When final, this guidance will represent the FDA's current thinking on this topic.
[8] See CID Case Study: A Study in Pediatric Patients with Multiple Sclerosis. (https://www.fda.gov/media/172313/download).
[9] Complex Innovative Trial Design Meeting Program (https://www.fda.gov/drugs/development-resources/complex-innovative-trial-design-meeting-program).
[10] See ICH guidance for industry *E11(R1) Addendum: Clinical Investigation of Medicinal Products in the Pediatric Population* (April 2018).

population." Pediatric extrapolation can extend what is known about the characteristics of interest in the reference population (e.g., efficacy, safety, and/or dosing) to those of the target population based on an assessment of the relevant similarities of disease, drug pharmacology, and response to therapy of the two populations. When some degree of pediatric extrapolation is justified, Bayesian methods can be considered to borrow data from adults in the analysis of a pediatric trial by using an informative prior distribution constructed based on results from previous adult trials. A discussion of safety considerations incorporated into pediatric extrapolation approaches is discussed in other guidance.[11] An example of a Bayesian approach to facilitate borrowing can be seen in supportive analyses in recent supplements for empagliflozin[12] and linagliptin[13] for the treatment of pediatric patients with type 2 diabetes mellitus (T2D). It is critical to consider the relevance[14] of the information from adults when considering borrowing. In these particular cases, the review team concluded that although there are differences in disease progression between pediatric and adult T2D populations, the pathophysiology of pediatric T2D is similar to that in adults and so the information was relevant, and borrowing was justified.

D. Borrowing Information Across Similar Diseases or Disease Subtypes

In some cases, distinct diseases or disease subtypes may have similar underlying causes and a history of similar responses to drugs. For example, there are groups of different types of cancer that share a specific molecular alteration and may be expected to respond to drugs targeting that alteration. In such cases, Bayesian methods might be considered to borrow information across the similar diseases or disease subtypes in evaluating an individual drug. For example, Bayesian analyses have been proposed for leveraging information about drug effects across related populations in basket trials that evaluate a drug for multiple diseases or disease subtypes under a common master protocol. As another example, a randomized, double-blind, placebo-controlled study in patients with epilepsy with myoclonic-atonic seizures[15] (EMAS) was proposed through the CID program that used a Bayesian primary analysis to borrow information from previous trials evaluating the effect of the drug in related conditions. The proposed approach leveraged data from previously conducted trials for different types of epilepsy using a Bayesian hierarchical model (BHM).

E. Borrowing Information Between Subgroups of a Patient Population (i.e., Subgroup Analysis)

It is important to try to understand drug effects in different subgroups of patients. There are statistical approaches that make use of results from every subgroup when estimating the treatment effect in each subgroup. One common approach is shrinkage estimation through a BHM. For a one-way BHM, the estimated treatment effect in one subgroup is a weighted

[11] For further discussion on pediatric extrapolation, see ICH guidance for industry *E11(A): Pediatric Extrapolation* (December 2024).

[12] See FDA Clinical Review (https://www.fda.gov/media/172973/download); FDA Statistical Review: (https://www.fda.gov/media/172972/download).

[13] See FDA Clinical Review (https://www.fda.gov/media/172628/download); FDA Statistical Review: https://www.fda.gov/media/172630/download).

[14] See section V.D.2 for a discussion of factors that impact the relevance of the data.

[15] See CID Case Study: A Study in Patients with Epilepsy with Myoclonic-Atonic Seizures (https://www.fda.gov/media/172312/download).

average of its raw estimated treatment effect (using only the data in that subgroup) and the overall estimated treatment effect. Shrinkage estimation can increase the precision of the subgroup estimates. As an example of this approach, a BHM was used to estimate treatment effects across regions in the Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome Results trial, which compared liraglutide to placebo in patients with T2D at high risk for cardiovascular disease. The analyses helped clarify potential differences in the drug effects across regions (i.e., Asia, Europe, North America, and the rest of the world).[16] BHMs have also been used for subgroup analyses that appear in some of FDA's Drug Trials Snapshots (Wang et al. 2024). One example is the Rinvoq Drug Trial Snapshot.[17]

F. Dose-Finding Trials in Oncology

Dose-finding trials for oncology drugs[18] have historically utilized non-randomized dose-escalation trials that seek to identify the maximum tolerated dose (MTD). Dose-escalation designs using Bayesian methods have been proposed with goals such as improving efficiency (e.g., reaching the MTD earlier), optimizing dose selection (i.e., minimizing toxicity and/or improving efficacy), and adding flexibility in terms of cohort sizes and timing of assessments. Designs which aim to identify the MTD include model-based designs (e.g., continual reassessment model [CRM], Bayesian logistic regression model [BLRM]) and model-assisted designs (e.g., Bayesian Optimal Interval Design [BOIN], modified toxicity probability interval [mTPI], mTPI2) (Ji et al. 2010; Quigley and Conway, 2010; Neuenschwander et al., 2008; Yuan et al., 2016;[19] Tighiouart and Rogatko, 2010). Although identifying the MTD has been the traditional paradigm for oncology drug development, for modern targeted therapies, such as kinase inhibitors and antibodies, identifying optimized dosage(s) based on alternative approaches may be more appropriate than selecting the MTD for further development.[20] Bayesian designs with the aim of identifying such dosages have been proposed for early-phase trials in oncology (Thall and Cook, 2004; Lin et al. 2020).

IV. SUCCESS CRITERIA AND OPERATING CHARACTERISTICS

A. Success Criteria: Definition and Role in Regulatory Decision-making

Clinical trial design includes pre-specification of criteria for determining whether the primary objectives of the trial have been met. In clinical trials intended to support the effectiveness and

[16] Additional discussion on the use of shrinkage estimation and BHMs for subgroup analysis, and additional details on the BHM model and results in the liraglutide trial, can be found in an FDA impact story: (https://www.fda.gov/drugs/regulatory-science-action/impact-story-using-innovative-statistical-approaches-provide-most-reliable-treatment-outcomes).

[17] Drug Trials Snapshots: RINVOQ Accessed August 29, 2024 (https://www.fda.gov/drugs/drug-approvals-and-databases/drug-trials-snapshots-rinvoq).

[18] See FDA guidance for industry *Optimizing the Dosage of Human Prescription Drugs and Biological Products for the Treatment of Oncologic Diseases* (August 2024).

[19] The BOIN design has received the fit-for-purpose designation for phase 1 dose-finding cancer trials: (https://www.fda.gov/media/155363/download).

[20] See footnote 17.

safety of drugs, success criteria of this type are useful as a goal to shape other design characteristics, such as sample size and power. Such criteria serve as a point of discussion between FDA and a sponsor on whether trial results could contribute to substantial evidence of effectiveness. Carefully chosen success criteria are important to trial interpretability and efficiency.

In clinical trials intended to support effectiveness and safety that are conducted with an overall frequentist statistical analysis plan, the efficacy success criteria are almost always chosen in such a way that the familywise Type I error rate (FWER) across all primary and secondary estimands is no greater than 0.025, one-sided. In the case of a trial with a single primary estimand, this means performing a statistical hypothesis test at a one-sided significance level of 0.025. For trials with multiple primary and/or secondary endpoints, the criteria can become more complex.[21]

In clinical trials with Bayesian inference for the primary estimand, this default success criterion may not be applicable or appropriate, such as in the case where there is borrowing of information (see Sections V.D and V.E), so careful specification of alternative success criteria is often critical when using Bayesian analyses. For these Bayesian approaches, specification of a success criterion is most often based on the posterior probability that the true treatment effect size exceeds some threshold. In mathematical notation, such a criterion might take the form $\Pr(\mathbf{d} > a) > c$, where **d** is a population-level summary of the size of the treatment effect, *a* is a minimum threshold for the treatment effect to be considered beneficial, and *c* is a minimum probability level that would support a conclusion of effectiveness. (In some cases, the criterion may be $\Pr(\mathbf{d} < a) > c$ instead, if lower values of **d** reflect greater benefit.) Choice of a success criterion of this kind thus means choice of specific values for *a* and for *c*. There are a variety of approaches to specifying these thresholds for Bayesian analyses. The choice of which approach to use depends on the trial objectives and specific Bayesian methods used.

1. Calibration to Type I Error Rate

For some Bayesian approaches, *a* and *c* can be chosen such that the overall FWER is controlled at a given level, typically 0.025 one-sided. This is referred to as calibrating the success criteria to Type I error rate. Such an approach may be appropriate for designs where Bayesian approaches are used not to synthesize multiple information sources, but instead to facilitate complex adaptive designs. Calibration to Type I error rate also may be useful in designs with noninformative prior distributions that express a lack of prior information relevant to the analysis.

When calibrating Bayesian success criteria to Type I error rate, *a* is chosen to be 0 for superiority designs or is based on the non-inferiority margin, *m,* for non-inferiority designs. Determining the appropriate choice of *c* to control FWER then becomes a computational problem only. For the simplest Bayesian approaches, including some designs with noninformative and/or conjugate prior distributions, *c* can be derived algebraically. For more complicated designs, including complex adaptive designs, *c* is instead approximated using clinical trial simulations.[22]

[21] See FDA guidance for industry *Multiple Endpoints in Clinical Trials* (October 2022).

[22] See FDA guidance for industry *Adaptive Designs for Clinical Trials of Drugs and Biologics* (December 2019).

2. Direct Interpretation of Posterior Probability

In Bayesian approaches where the prior distribution has been chosen to provide an accurate summary of the state of belief based on existing information before the trial begins, decision-making can be based on direct interpretation of the posterior probability distribution itself. With a prior chosen in this way, if the posterior probability $\Pr(\mathbf{d} > a) = c$ then the probability that the treatment effect is less than a is less than $1 - c$. For example, if the posterior probability of effectiveness is 0.98, the posterior probability that the treatment is ineffective is 0.02. The choice of success criteria can then be based on a determination of whether a $1 - c$ chance of ineffectiveness is sufficiently small in the specific context.

This approach can be appropriate in Bayesian analyses that explicitly leverage external data sources to support decision-making, though it is not limited to that setting. In such scenarios, it is critical that the prior be specified in a way that accurately and comprehensively reflects the external data to ensure credible conclusions.

3. Success Criteria Based on Benefit-Risk Assessment or Decision-Theoretic Approaches

Another general approach specifies success thresholds in a broader context incorporating product risk or additional considerations. The simplest form of such an approach would adopt a choice for the threshold a that ensures sufficient benefit to outweigh known or potential product risks.

More complex approaches might include consideration of external factors such as seriousness of a disease or availability of other approved therapies. A decision-theoretic approach might include assessment of the potential negative consequences of approving an ineffective drug or of not approving an effective drug. The statistical quantification of such negative consequences is sometimes referred to as "loss." One approach to incorporating such information is to form success criteria that minimize the expected loss. The loss function can include safety as well as effectiveness considerations to incorporate benefit-risk assessment into the formal decision-making process.

4. Additional Considerations

In trials that include interim decision-making, such as group sequential designs, success criteria should be specified for each decision point. In cases where success criteria are calibrated to Type I error rate, interim success criteria can be specified to ensure overall control of FWER across all decision-points for effectiveness.

Success criteria should be specified for all primary and key secondary endpoints that use Bayesian approaches. In those trials with success criteria calibrated to Type I error rate, clinical trial simulations may be required to ensure control of FWER across all endpoints. Sample size can be calculated to achieve the desired statistical power under the controlled FWER.

B. Operating Characteristics

In the design of a clinical trial, it is important to understand how the trial is likely to perform in terms of supporting correct conclusions and reliable estimation of treatment effects. In trials with frequentist inference, the most important operating characteristics related to hypothesis tests (long-run expectations of trial conclusions under assumptions about true parameter values) are the FWER and the power for the primary and other key endpoints. These quantities are fixed by design before the trial (conventionally, at 0.025 one-sided FWER and 80% or 90% power). This is possible because frequentist inference is based on the conditional probability of observing certain data given fixed assumptions about parameter values, and these probabilities can be calculated pre-test. The most important operating characteristics related to estimation in the frequentist paradigm are bias and mean squared error (MSE) of point estimates, and coverage probability and width of confidence intervals.

The situation is somewhat different in trials with Bayesian inference because Bayesian inference is based on the posterior distribution. Operating characteristics of the design and analysis therefore depend on both the prior distribution and the observed data. It is also important to note that the concept of a false positive conclusion in a Bayesian framework is conditional on a positive conclusion and not, as in the frequentist framework, on a true null hypothesis.

Some trials, including certain complex adaptive design trials, employ Bayesian analysis in an overall frequentist inferential framework (i.e., with calibration to Type I error rate control). In these cases, frequentist operating characteristics are of interest. Other trials, including most trials in which external information is incorporated into an informative prior distribution, use a Bayesian inferential framework that calls for a different approach to quantifying design characteristics. These two cases are discussed in the next two subsections.

1. Trials Calibrated to Type I Error Rate

For trial designs that calibrate Bayesian results to Type I error rate, the primary operating characteristics are the same as those described above for trials with frequentist inference (that is, Type I error rate and power for testing; bias and MSE of point estimates, and coverage probability and width of intervals, for estimation). Clinical trial simulations are generally used to estimate or demonstrate control of operating characteristics. Briefly, a large number of simulated trials, conditional on a chosen prior distribution and sample size, are generated under the assumption that the null hypothesis is true or that an alternative hypothesis is true. The proportions of simulated trials in which the null hypothesis is rejected is then used to estimate Type I error rate and power, respectively. An iterative simulation process is often used in which various sample sizes, prior distributions, success criteria, and other design elements (e.g., interim analysis boundaries) are adjusted to obtain desired operating characteristics. Simulations should comprehensively cover the plausible range of assumptions. This includes assumptions about statistical parameters such as the variance or background rate of the endpoint or operational parameters such as the accrual rate.[23]

[23] For additional considerations on these and other aspects of simulations, see Section VI.A of the FDA guidance for industry *Adaptive Designs for Clinical Trials of Drugs and Biologics* (December 2019).

2. Trials Not Calibrated to Type I Error Rate

In cases where a sponsor and FDA agree that a study design does not need to be calibrated to the Type I error rate, the accuracy of conclusions depends strongly on the choice of prior distribution, as well as many of the same features required in frequentist inference: an adequately fitting data model, an appropriate experimental design, and accurate measurements. Accordingly, design characteristics are calculated in reference to the prior distribution. For example, Bayesian power is the probability of meeting the success criterion, averaged over a prior distribution (Spiegelhalter et al., 2004). The sample size of a trial is chosen to achieve a desired Bayesian power conditional on a chosen prior distribution and other study design features. Another relevant operating characteristic is the probability of a correct decision (akin to calculating the positive predictive value for a diagnostic test) corresponding to a chosen prior. For example, simulations can be used to calculate the proportion of trials where a positive treatment effect was present, from among those trials that concluded effectiveness. For estimation of treatment effects, relevant operating characteristics include the expected bias and expected MSE of point estimates averaged across a prior. Similarly, the expected coverage probability or width of the corresponding credible interval can be assessed (Adcock, 1988; Joseph and Bélisle, 1997).

It is always critical for Bayesian analyses to have a prespecified prior, which is typically called the analysis prior. In a hypothetical situation where the choice of a prior distribution was unambiguous and clear, Bayesian power and other quantities could be calculated in reference to the same prior distribution that will be used in the final study analysis. In practice, however, there will usually be a range of plausible design priors (sometimes referred to as sampling priors), separate from the analysis prior, that are used as the basis for calculating study design characteristics. In simulation studies, the design prior serves as the prior from which parameter values are drawn, whereas the analysis prior is the prior that is used in the subsequent analysis of the data generated. An example of a design prior is a prior distribution on the treatment effect centered around the minimum clinically important difference (MCID) to evaluate Bayesian power. When the design prior is intended to explore scenarios corresponding only to an efficacious treatment, the design prior might be limited to an interval or a point mass indicating that the treatment is effective. Design priors that explore pessimistic assumptions about treatment effect should also be considered.

Operating characteristics can be calculated under various plausible design priors. Differences in design characteristics corresponding to different design priors can be used to quantify the sensitivity of the design to the choice of prior or to demonstrate that the probability of making an incorrect decision is very low even when the design and analysis priors do not match. In general, trial characteristics will be more sensitive to the analysis prior when the sample size is small or when there is an early interim analysis that makes the sample size effectively small.

3. Additional Considerations

The operating characteristics discussed above are typically provided for key objectives of a clinical trial. For any trial and within any development program, there are often multiple other objectives for which data collection is essential. For example, it is essential to generate evidence

regarding the safety and tolerability of a product. These other objectives should be considered in the overall trial design and drug development program.

V. PRIOR DISTRIBUTIONS

A. Overview and General Principles

Use of a prior distribution is the main feature distinguishing Bayesian from frequentist approaches. Priors allow the analysis to reflect the available information in the particular situation, whether positive, negative, or neutral. With any Bayesian analysis, the prior construction process should be designed, implemented, and documented in a systematic and transparent manner. Sponsors should pre-specify and justify the full details of the proposed prior distribution in the protocol. This justification should address the appropriateness of the prior distribution's influence and the operating characteristics of the design, given the proposed prior. An important distinction is between informative prior distributions that borrow external information into the analysis of the current trial, and noninformative or minimally informative prior distributions that express a stance of general uncertainty. The complexity of the prior construction process will depend on whether an informative prior is used, and on the sources of external information being considered. Informative priors will generally need a greater amount of justification (see Section V.D for details). The following sections provide guidance for these different types of prior distributions and for other aspects of prior distribution evaluation and selection.

B. Noninformative and Minimally Informative Priors

Noninformative and minimally informative priors are typically specified to reflect a stance of general uncertainty regarding the parameters to be estimated. A noninformative prior is designed to use no external information. There are a variety of approaches that aim to reflect such lack of information, and no single approach is universally preferred to others. A minimally informative prior distribution could be based on general information about a trial and the possible outcomes, such as the plausible magnitudes of change in the endpoints of interest and available information about the variability in previous studies. Noninformative and minimally informative priors are often used in cases where there is no relevant prior information. In many situations, priors that fall into these classes will tend to be overwhelmed by the observed data and therefore have minimal effect on the results such that the final conclusions are dominated by the observed data. In other situations, the available data may not provide much information for a particular parameter and so using a noninformative or minimally informative prior for the parameter may place weight on parameter values we know to be unlikely and may have a large effect on the estimates (for example, see Gelman, 2006). The properties of a particular prior can typically be determined using appropriate simulations during study design.

As with any prior distribution, a noninformative or minimally informative prior may not be invariant to parameter transformation, such that the induced distribution on such a transformation reflects an unintended understanding about the transformed parameter of interest. Care should be taken to understand such induced prior distributions during study design.

Noninformative and minimally informative priors can also play important roles in assessing the influence of informative prior distributions on the analysis. Such priors are often used in the calculation of effective sample size (see Section V.E) as a reference scale against which to measure the influence of another prior. In addition, when an informative prior is specified in the primary analysis, an analysis using a noninformative or minimally informative prior is often helpful in understanding the sensitivity of the results to the choice of prior.

C. Skeptical Priors

Skeptical priors express skepticism about the presence of very large treatment effects. One scenario where they may be appropriate is when there is prevailing information indicating the need to be more cautious than usual on drawing conclusions in favor of benefit. For example, if there have been a number of failed trials or drug development programs for closely related drugs in a therapeutic area, it would be natural to be skeptical about the potential for benefit with another similar trial or drug. Another possible scenario is in the evaluation of a new drug that is not likely to offer more than incremental improvement over an existing therapy. In this setting, it would be natural to be skeptical of a dramatic improvement over the existing therapy, and a skeptical prior could be used in the analysis. Such approaches have not been standard practice in drug development but could be considered in relevant circumstances such as those described above.

Skeptical priors may also be considered in trials with an adaptive design. In designs where there is a desire to calibrate to Type I error rate, skeptical priors can be used as an alternative to modification of the decision rule to maintain the desired error rate as the skepticism will counterbalance early random highs or lows and hence early stopping requires even greater early evidence of benefit. Enthusiastic priors (priors containing some degree of positive belief) can be used in a similar way to control early stopping for futility.

D. Informative Priors to Borrow External Information

1. General Recommendations

When proposing to use an informative prior that borrows external information for inference on the primary estimand, sponsors should provide strong justification that considers feasibility (e.g., of alternative approaches that do not involve borrowing) and the relevance of the available information. Areas where informative priors have been most often proposed include pediatrics and rare diseases. Additional areas can be considered on a case-by-case basis, and FDA advises early discussion of such proposals with the Agency.[24] The specific sources of information to be used in the informative prior and the degree of reliance on the information should be justified. The justification should include discussion of the influence of the prior and how it relates to the relevance of the borrowed data for the current trial, the sufficiency of the amount of prospective trial data that will be collected (e.g., for evaluating safety and benefit-risk), and the appropriateness of trial operating characteristics. In some cases, it can be advantageous to

[24] See FDA Guidance for Industry *Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products* (December 2020).

determine the details of the borrowing method while still blinded to the results of the trials that will be borrowed. The time needed for FDA and the sponsor to align on an appropriate prior should be considered in the development of the intended trial.

It is important to consider the possibility of prior-data conflict, defined as the scenario in which the data observed are notably inconsistent with the prior distribution. More formally, prior-data conflict occurs when the prior distribution places its mass "primarily on distributions in the sampling model for which the observed data is surprising (Evans and Moshonov 2006, p 894)." The potential impact of prior-data conflict on the interpretability of trial results should be explored in simulations at the study design stage by varying the size of the assumed treatment effect over a sufficiently broad range of scenarios that cover plausible degrees of conflict, including no effect in the target population. The outcome of these simulations should be considered carefully and discussed in the justification of the prior, and appropriate sensitivity analyses (see Section V.F) should be planned.

Prior construction, including making decisions on how much to borrow, is a multidisciplinary process that requires quantifying the degree of uncertainty in the relevance of the external information to the question of interest. Understanding relevance requires domain knowledge, while quantifying uncertainty and selecting appropriate statistical methodology requires thorough statistical evaluation. Close collaboration between disciplines throughout the process is essential.

2. Identification and Review of Available External Information

In general, the process of determining a prior should begin with the identification and review of all the available relevant external information. Possible sources of information may include relevant pharmacokinetic, pharmacodynamic, and clinical data (e.g., from previous trials or systematic reviews of trials evaluating the drug in the same or other related conditions), as well as nonclinical data, real-world data, and professional or expert guidelines or consensus opinions. Several factors should be considered in evaluating whether and how much to leverage external information to ensure that conclusions relying on such information are valid, reliable, and interpretable, such as:

- Data quality and reliability: The quality and reliability of information used to construct the prior should be adequate for the type of regulatory decision informed by the analysis. Data from clinical trials designed to support regulatory decisions will typically meet this bar, but other sources may require additional effort to ensure adequate quality and reliability. For real-world data, FDA has issued guidance[25,26,27] that discusses the processes and procedures that help ensure quality of the data. Other sources may be necessary in some circumstances, and in these cases comparable steps should be taken to ensure the quality and reliability of the information.

[25] See FDA guidance for industry *Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products* (December 2023).
[26] See FDA guidance for industry *Use of Electronic Health Record Data in Clinical Investigations* (July 2018).
[27] See FDA guidance for industry *Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products* (July 2024).

- Pre-specification: Statistical methods should be prespecified to minimize bias. Similarly, sponsors should pre-specify the construction process for the prior, including for the selection of sources for the same reason. As with any evidence synthesis approach, the standards for inclusion and the intended scope of use of the various sources for the prior should be predefined before starting a systematic search to identify specific sources.

- Relevance: The information being leveraged should be relevant to the applicable regulatory question. When multiple information sources are used, not all information may be equally relevant. It is important to consider and discuss any planned approaches to reflect the different degrees of relevance. Factors that may influence relevance include:

  - Similarity in estimand attributes such as the population (e.g., inclusion or exclusion criteria), endpoint, treatment conditions, or handling of intercurrent events

  - Any differences in measurement or assessment (e.g., of the endpoint)

  - Recency of data

  - Any potentially important changes (e.g., in aspects of standard of care such as concomitant medications) over time

- The design of studies providing the information: Borrowing of information based on randomized controlled comparisons typically relies on fewer and more plausible assumptions[28] than borrowing of information based on non-randomized comparisons or on a single treatment condition (e.g., historical control data).

- The availability of patient-level data: Typically, information should come from patient-level data as this allows for a thorough evaluation of the relevance of the external data and the potential to adjust for relevant covariates in the analysis, which is particularly important when inclusion and exclusion criteria are not fully aligned.

3. Prior Construction

After identification and review of the available external information sources, sponsors should decide how to use the relevant information sources. A thorough evaluation of all relevant sources and evidence for informing the prior should be completed, including evidence that may suggest skepticism of the existence or magnitude of a treatment effect. This evaluation should include a prespecified list of criteria and discussion of steps taken to ensure information was not selectively obtained or used to ensure that preferential selection of favorable studies (or the most favorable studies) did not occur. If multiple information sources will be leveraged, an evidence synthesis process should be pre-specified and employed to obtain the most plausible estimate of

[28] When borrowing information based on randomized controlled comparisons, it is still important to evaluate the potential impact of differences in effect modifiers between information sources.

the responses. Typical best practices for evidence synthesis in systematic reviews and meta-analyses, such as pre-specification of data collection processes, study inclusion or exclusion criteria, and synthesis methods, are relevant here.[29]

The modeling approach for constructing the prior will depend on several factors, including the similarity between the various data sources, the relative amount of information each source contains, and the relationships between the data sources. As the data sources increase in variety and the relationships becomes more varied, the necessary modeling approach becomes more complex. For example, suppose a sponsor is designing a pediatric study that will use an FDA-approved drug as an active control. For the investigational drug, there is available data from completed adult studies, whereas for the drug used as the active control there are pediatric as well as adult data. In this situation, not all data sources would be equally relevant since the adult data would be less relevant than the pediatric data and so a modeling approach that reflects this belief would be preferable. Modeling approaches can also be used to increase the relevance of the prior data (compared to taking the raw results from the previous studies) by adjusting estimates based on covariates or by selecting a subset of the data that more closely aligns with the question of interest. For example, the multiple sclerosis trial discussed in Section III.B modeled the relationship between age and the annualized relapse rate to improve the relevance of the data from the adult trials.

Typically, multiple sources of information and many assumptions[30] underlie a particular prior, and it is crucial to ensure that these are documented in the protocol with a discussion of the supporting evidence for each assumption to facilitate FDA's review. For example, in some cases it may be reasonable to assume that all data sources are equally relevant. A typical example of an assumption of equal relevance would be in a primary analysis in which all patients who are enrolled in a single clinical trial are analyzed simultaneously without any regard for possible subgroups. An example of where this assumption should be considered more closely is when pediatric patients (most commonly adolescents) are included in a single trial with adults. The appropriateness of an assumption of equally relevant data from adults and adolescents would be informed by the specific extrapolation concept and plan,[31] and the most appropriate type of relationship should be justified and discussed with the Agency.

In other cases, models that include exchangeability may be reasonable. The outcomes of a group of participants are considered exchangeable in a setting where if the values of the outcomes are revealed, but their labels are not, then the values are not helpful in predicting their labels. When the outcomes of a group of participants are exchangeable then all possible ordering of the outcomes are equally likely prior to observing the values for the outcomes. Statistical models routinely assume that the outcomes of participants within a treatment group are exchangeable or assume that the residuals are exchangeable. Outcomes modeled as independently drawn from the same distribution are exchangeable. Outcomes modeled as random draws without replacement

[29] See FDA draft guidance for industry *Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products* (November 2018). When final, this guidance will represent the FDA's current thinking on this topic.

[30] In this discussion, we use the taxonomy of the types of assumed relationships between data sources from Spiegelhalter, David J., Keith R. Abrams, and Jonathan P. Myles. Bayesian approaches to clinical trials and health-care evaluation. Vol. 13. John Wiley & Sons, 2004.

[31] See the ICH guidance for industry *E11A Pediatric Extrapolation* (December 2024).

from a set of possible values are also exchangeable. There are also other conditions that imply exchangeable outcomes (Yuan and Chai, 2008).

One setting where exchangeability may be considered is in the estimation of treatment effects within subgroups of a population (i.e., subgroup analysis). An assumption of exchangeability may be reasonable if there are no a priori expectations regarding effect modification by the subgroup factor. However, exchangeability of subgroup treatment effects is not always an appropriate choice. For example, if a drug is expected to be more effective in a subgroup of cancer patients who exhibit the molecular target than the complementary subgroup of patients without that target, the two possible orderings of treatment effects with the subgroups would not be equally likely. In some cases where exchangeability of subgroup treatment effects is not appropriate, all known effect modifiers can be included in the model as effect modifiers (e.g., as interaction effects with treatment when modeling individual observations). It then may be appropriate to regard the residual subgroup treatment effects as exchangeable.

There are settings where other potential assumptions may be reasonable. For example, there may be cases where it is reasonable to model the relationship between the source and prospective trial data using an assumed bias parameter in the model. This approach, which would allow borrowing of information on the precision while allowing for differences in the treatment effect, might be considered if there is reliable evidence of the magnitude of the bias or suspicion of a bias.

In other cases, it may be reasonable to use a model with a functional dependence assumption that predicts the distribution of responses using some function based on systematic predictable processes. For example, suppose two distinct populations were studied in two different trials. If a third trial was conducted that included both populations, it may be reasonable to construct the prior expectation for the effect in this trial using a combination of the results from the two previous trials while adjusting this expectation based on the relative sizes of the populations in the prospective trial. This type of assumption can be seen in some pediatric extrapolation settings where attempts are made to adjust for expected differences between adult and pediatric patients. For example, in the development programs for empagliflozin[32] and linagliptin,[33] the sponsor constructed pharmacometric models based on the previously conducted studies combined with the relevant observed baseline values of the pediatric patients to predict response.

Finally, there is often a need to implement a discounting approach to reflect any residual uncertainty regarding the question of interest. For example, when using Bayesian methods to implement pediatric extrapolation, if the adult trial results were used to derive the prior without discounting, the evidence from the adults alone would already meet typically used success criteria and if a trial were conducted, the adult data would often overwhelm the pediatric data regardless of the results in the pediatric study. This would not be consistent with the uncertainty on the question of benefit in pediatric patients. Therefore, it will often be more reasonable to use a prior centered on a degree of benefit similar to what is observed in adults, but with a greater

[32] See FDA Clinical Review (https://www.fda.gov/media/172973/download); FDA Statistical Review (https://www.fda.gov/media/172972/download).

[33] See FDA Clinical Review (https://www.fda.gov/media/172628/download); FDA Statistical Review: https://www.fda.gov/media/172630/download).

degree of variability and hence uncertainty. Discounting is discussed in more detail in the next section.

4. Discounting

Many prior discounting approaches fall into one of two categories – either static discounting where the method borrows the same information regardless of the observed trial data or dynamic discounting (sometimes called adaptive discounting) where the amount that is borrowed is determined by some measure of the similarity of the data, such as the difference between the observed and prior means. Static discounting methods tend to be easy to design and implement, as there are fewer decisions to be made, while dynamic discounting methods can provide protection against prior-data conflict because they borrow less when the data are less similar and more when the data are more similar. Dynamic approaches are used in many cases due to the more advantageous operating characteristics, such as with respect to bias and MSE, resulting from the lesser borrowing in cases of prior-data conflict. The challenge with dynamic discounting methods is that they introduce more parameters that need to be specified, for example, the similarity measure to be used and the rate at which borrowing declines based on this similarity measure.

There are many specific methods proposed in the scientific literature for implementing discounting. With any approach, the goal should be to identify parameter values that correspond to a reasonable degree of borrowing. What is reasonable will depend on the situation and the relevance of the borrowed data. A prior that is too informative risks overwhelming the data collected in the target population, regardless of what is observed. Leveraging too little information means not taking advantage of the available data and more information will be required in the target population, making the trial less efficient. Hence, finding the appropriate balance in the informativeness of the prior is crucial. Model parameter values that determine the degree of discounting and their interpretation depend on the method being used and may not always be easily compared across models. Hence, it is often useful to use more interpretable metrics (such as those discussed in Section V.D).

In some cases, it may be helpful to conduct a formal expert elicitation exercise (O' Hagan et al., 2006) with subject matter experts to incorporate the degree of consensus in the level of borrowing. It is important to consult with the Agency on the scope and utility of such an exercise before conducting it, and to prespecify the approach for elicitation.

The following paragraphs discuss considerations for the use of a few different methods for discounting but are not intended to provide an exhaustive list of methods. Applicability of a discounting method will be determined on a case-by-case basis and should be discussed with the Agency.

One approach uses power priors that are constructed by raising a prior distribution based on external data to a fixed power less than one, sometimes called a discounting factor. The discounting factor adjusts the relative informativeness of a single external subject's data compared to a single current subject's data. A discounting factor of 1 corresponds to equal weighting of external and prospective trial subjects, a discounting factor of 0 corresponds to no

borrowing of external information, and discounting factors in between correspond to borrowing with each external subject contributing less information than each trial subject. Once the discounting factor is chosen, the degree of borrowing is fixed, making this a static borrowing approach. Advantages of the power prior approach include simple implementation and a seemingly intuitive approach to discounting. A disadvantage is that it is static and therefore may have worse operating characteristics than dynamic approaches.

Although the basic power prior approach is a static borrowing approach, there are extensions proposed in the literature such as commensurate priors and supervised power priors which make the degree of borrowing dependent on different measures of the similarity between the observed and prior data. As the difference between the observed and prior means grows larger, less information will be borrowed. These approaches can allow control of how fast the discounting occurs as the difference grows, and determining an appropriate rate is one of the main challenges in using this approach.

Using a mixture prior is another simple way to implement a dynamic borrowing approach. Two or more individual prior components are combined with prespecified weights, and the result is a prior distribution that will adaptively reflect the similarity between the individual components and the observed data. The chosen weights must sum to one and can be interpreted as the probability of applicability of the particular data source. For example, a mixture prior might be constructed by combining an informative component based on estimates obtained from previous trial data with a noninformative component. The degree to which the resulting posterior will borrow the previous trial data will depend on the degree of similarity between the observed and previous data. The main difficulty is determining the level of weighting for informative components and hence the strength of borrowing. The advantages of this approach are simple programming implementation and well-developed methods for estimating the degree of borrowing.

Other discounting approaches include Bayesian hierarchical models and elastic priors. Bayesian hierarchical models are the main method used to implement exchangeable models. They induce borrowing by assuming that the parameters for various groups are at some level drawn from a common distribution. For example, in a basket trial evaluating a drug in multiple related diseases, a possible starting point for the modeling process might be that the treatment effects for the diseases are a random sample drawn from a single common distribution. Elastic priors (Jiang et al., 2023) are implemented in two steps. First, a measure of the degree of similarity called a congruence measure is used to quantify the strength of similarity between the external and current data. Then a function called an elastic function is used to map the congruence measure to a degree of borrowing. Like the power prior approach, the degree of borrowing takes values between 0 (no borrowing) and 1 (full borrowing). Elastic priors are flexible methods as they can use any of an extremely broad range of different functions to implement discounting that adapts as fast or as slow as desired based on the level of observed conflict between the prior and current data. However, this flexibility may make it harder to justify a particular choice of function.

E. Quantifying the influence of the prior distribution

The influence of a prior distribution should be discussed and documented. The influence can be measured in multiple ways using different metrics. Metrics that have been used in the regulatory setting include:

- The estimated treatment effect/difference or parameter(s) of interest based solely on the prior distribution. This is usually the mean of the prior distribution.

- Effective sample size (ESS): a measure of the information in a probability distribution in terms of the equivalent number of patients in the target population. For example, when applied to the prior alone we would call it the prior ESS. Note that this number may be larger than the sample size from the source population if there is a larger amount of variability in the target population than the reference population. Multiple methods that rely on different summary measures have been proposed for the ESS (Malec, 2001; Morita et al., 2008; Neuenschwander et al, 2020). In general, it is important to quantify and summarize the ESS for the entire plausible range of outcomes while also presenting relevant summary statistics such as the maximum and mean ESS.

These metrics are helpful to compare candidate priors. When an informative prior is used, values of these metrics should be compared to values obtained when using noninformative or minimally informative priors (or other relevant priors) to understand the effect of the informative prior. These metrics may be more easily understood than prior parameter values that depend on the specific model being used.

It can also be helpful to consider extensions of these metrics. For example, one might determine the ratio of the prior effective sample size to the intended prospective trial size or to the total amount of information (prior effective sample size plus prospective trial size) to measure the relative amount of information borrowed in an analysis. If a dynamic prior is used, then the effective sample size contribution of the prior will change as information is accrued and so it is important to reassess this measure following completion of the trial. After data are collected, additional measures regarding the prior's impact may be useful. Examples include measures of the conflict between the prior and the observed data and updated measures quantifying the strength of borrowing.

In some cases, the type I error rate has been proposed as a way to measure the degree of information in the informative prior. Borrowing information in an analysis will typically lead to an inflation in the Type I error rate compared to the nominal value (most often, a one-sided level of 0.025) used in the absence of borrowing. This degree of inflation has been proposed as one way of measuring the influence of a prior (Pennello and Thompson, 2007). However, there are multiple issues that make this method a poor way of measuring the degree of borrowing, and therefore it is not recommended for this purpose. First, by borrowing information, one is assuming that the borrowed information is relevant. Evaluating the degree of borrowing based on the expected outcome when there is no effect is philosophically inconsistent given a prior which assumes a non-zero effect. Second, dynamic methods lessen the impact of borrowing when effects that are smaller and close to the null value are observed. Consequently, borrowing will be

low when the observed data are incompatible with the prior, mitigating any potential increase in Type I error rate. Hence, the Type I error rate alone will not provide a complete assessment of the influence of the prior.

F. Sensitivity analyses

A necessary part of statistical inference in clinical trials is evaluating the sensitivity of the results and conclusions to plausible deviations from important analysis assumptions. In a Bayesian analysis, the choice of prior distribution is a critical assumption that can affect the results and conclusions, particularly when the prior distribution is informative and is associated with a large ESS. Therefore, sensitivity analyses should be planned that utilize a range of alternative reasonable choices for the prior distributions. For example, the amount or strength of borrowing could be varied (see Section IV.B on operating characteristics for additional examples). Comparing the results and conclusions (based on the respective posterior distributions) of such analyses can help one understand the sensitivity of the primary results and conclusions. Some approaches can build uncertainty about specific assumptions into the prior itself.

VI. ESTIMANDS AND MISSING DATA IN A BAYESIAN SETTING

The general considerations related to estimands and missing data described in the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9 and E9(R1) guidances are relevant for a trial that uses Bayesian methods as with any other trial. These considerations include the importance of constructing a relevant primary estimand of interest that can be estimated with plausible assumptions, implementing design and conduct approaches to prevent missing data, pre-specifying an appropriate primary analysis approach (i.e., estimator), documenting missing data assumptions, and planning sensitivity analyses to evaluate robustness to violations in those assumptions. Bayesian methods can be used to address missing data in the primary analysis or in sensitivity analyses.

There are additional important considerations in trials that use Bayesian methods to borrow external information. In particular, it is critical to consider whether there are any differences in the estimands and estimators between the external information source (e.g., previous trial) and the prospective trial. Ideally, the same primary estimands and estimators should be used in analyzing both data sources. The properties of the parameter estimates depend on the approaches used, and lack of alignment in the approaches between the data sources can make the external data less relevant and affect considerations about the degree of borrowing. However, FDA recognizes that it may not be feasible or advisable to use the same approach in all situations.

In general, where there is a lack of alignment between the estimands or estimators, sponsors can consider reanalyzing the external data using the approach that is planned for the prospective trial. This will produce the most compatible estimates but may not be possible if patient-level data are inaccessible or if a particular estimand or estimator that will be used in the prospective trial relies on data that were not collected in the external data source (for example, patient outcome data after treatment discontinuation may be necessary for some approaches and are not always collected). Reanalyzing completed studies risks overfitting to observed random highs, thereby

introducing bias. Sponsors should have early discussions with the Agency about the planned estimands, estimators, and approaches for handling missing data in the analyses of external data that will be borrowed, and any differences relative to the approaches planned for the prospective trial data.

VII. SOFTWARE AND COMPUTATION

When performing Bayesian inference, it is often necessary to rely on various approximate sampling algorithms to perform statistical inference. These sampling algorithms are implemented in a variety of general and specialized statistical software packages. These algorithms may also be implemented in specific situations using a variety of general programming languages. As stated in the Statistical Software Clarifying Statement,[34] "FDA does not require use of any specific software for statistical analyses." As noted in the statement and in ICH guidance *E9 Statistical Principles for Clinical Trials,*[35] "computer software used for data management and statistical analysis should be reliable, and documentation of appropriate software testing procedures should be available."

When performing Bayesian analyses, it is critical to ensure that the sampling algorithm being employed is reliable for the specific model (Gelman et al., 2020). Bayesian methods are dependent on algorithms to approximate the posterior distribution and approximations can fail for a variety of reasons. For example, a Markov Chain Monte Carlo (MCMC) algorithm may not adequately sample from a particular part of the posterior distribution, or an approximation used in an algorithm may not be accurate. Hence it can be necessary, especially for more complex models, to evaluate the reliability and accuracy of Bayesian computation using simulations before final selection of the model. In cases where sampling or convergence issues are encountered, more detailed documentation of the reliability of the final selected model should be provided, e.g., in a simulation report, including a discussion of any steps taken to ensure better sampling or convergence. The documentation should include the range of scenarios evaluated and the code used to implement the model to allow verification of the sampling properties.

VIII. DOCUMENTING AND REPORTING BAYESIAN ANALYSES

Clear documentation is necessary for FDA to review Bayesian proposals. The study design, estimands, and analyses should be pre-specified and justified in a protocol and consider applicable guidances.[36] A clinical study report should describe the design, analysis, and results. As with non-Bayesian approaches, an assessment of parametric assumptions associated with the likelihood should be provided in the clinical study report using appropriate diagnostics. Details specific to Bayesian approaches are discussed below.

---

[34] See FDA Statistical Software Clarifying Statement (https://www.fda.gov/media/161196/download).

[35] See ICH guidance for industry *E9 Statistical Principles for Clinical Trials* (September 1998).

[36] For example, see the ICH guidances for industry *E9 Statistical Principles for Clinical Trials* (September 1998) and *E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials* (May 2021), and the FDA guidance for industry *Adaptive Designs for Clinical Trials of Drugs and Biologics* (December 2019).

A. Documenting Plans for Bayesian Analyses

The protocol should describe and justify the design and the planned statistical analysis methods. For Bayesian methods, this includes detailed information to support the proposed prior distribution and any external information borrowing, likelihood function, success criteria, and trial operating characteristics. Additional details or supporting information may be provided in additional documents such as a statistical analysis plan or simulation report. All relevant information should be submitted to FDA during the design stage and as early as possible to ensure sufficient time for FDA feedback prior to initiation of the trial.

The following information should be provided to support the proposed prior distribution (see Section V for further discussion):

- A detailed description of the proposed prior distribution, including explicit specification of prior parameterization and underlying assumptions. If a function of parameter(s) is of interest, this should include a description of any induced priors.

- A rationale for the plausibility of the assumptions.

- Any data or other information that informed the prior distribution.

If the trial will use an informative prior to borrow external information for inference on the primary estimand the following additional information should also be provided:

- A strong justification for borrowing that considers feasibility (e.g., of alternative approaches that do not involve borrowing) and the relevance of the available information.

- A thorough evaluation of all the available relevant evidence for informing the prior, including evidence that may suggest skepticism. This should include a discussion of steps taken to ensure information was not selectively obtained or used.

- A detailed description of each external information source used to inform the prior, as well as any source that was excluded and why, including a discussion of data quality and reliability, relevance, type (e.g., patient-level vs. summary-level) and completeness of data available, and a rationale for the degree of borrowing incorporating such factors.

- If multiple information sources are leveraged, a description of how the data will be synthesized and the modeling approach for their synthesis.

- If a discounting method is incorporated, a description and justification of the selected approach, including a rationale for any parameters (e.g., that impact the degree of borrowing).

- A discussion of how the proposed analysis will handle prior-data conflict (e.g., supported by simulation results).

- Quantification and discussion of the degree of influence of the prior distribution (e.g., based on the ESS and other metrics).

- Planned sensitivity analyses that utilize a range of alternative reasonable choices for the prior distribution (e.g., that vary the amount or strength of borrowing from the prior distribution).

The following information should be provided to support the proposed success criteria (see Section IV.A for further discussion):

- Specification of the success criteria for all primary and key secondary estimands. For success criteria based on the posterior probability, both the treatment effect threshold and the minimum probability of effectiveness should be discussed and justified.

- For trials that include interim analyses with the potential for early stopping for efficacy, specification of the success criteria for each decision point.

- For trials with success criteria calibrated to Type I error rate, justification that the choice of success criteria leads to control of the FWER (e.g., based on a detailed report of simulations used to estimate the Type I error rate).

The following information should be provided to describe and support the appropriateness of the operating characteristics of the trial (see Section IV.B for further discussion):

- A discussion of how operating characteristics are estimated, with sufficient detail to facilitate FDA's verification of the values and evaluation of the proposed sample size and other design elements.

- For study designs calibrated to Type I error rate, an evaluation of the Type I error rate and power under a range of plausible effect sizes, and estimation properties including bias and MSE of point estimates and coverage probability and width of intervals.

- For study designs not calibrated to Type I error rate, an evaluation of Bayesian power, expected bias, expected interval width, expected MSE, and other quantities to assess the accuracy of conclusions and treatment effect estimates under relevant priors.

- When simulations are required to estimate operating characteristics, a comprehensive simulation report. This report should include simulation code and a detailed description of the simulation design, implementation, and results. The scenarios and assumptions used in the simulation should be pre-specified, comprehensive, and plausible.

For Bayesian analyses performed using approximate sampling algorithms (see Section VII), sponsors should provide a summary of the simulation setup. For example, for MCMC approaches, sponsors should provide a general summary of the length of warmup or burn-in, number of iterations, number of chains, convergence diagnostics, and any other important algorithm-specific settings (e.g., proposal distribution for the Metropolis algorithm, target

proposal acceptance probability when relevant for software implementation of Hamiltonian Monte Carlo). If computation requires specialized software, the software should be cited, and details of planned implementation provided. If such computation is required, there should be a description of how implementation issues (e.g., lack of convergence) will be addressed in the analysis.

For complex design proposals such as those with borrowing of external information based on informative priors, comparisons should be made between the operating characteristics of the proposed design and analysis plan and a variety of alternatives, including simpler designs. These comparisons should be documented as they help understand the potential advantages and limitations of the proposed design features. FDA also recommends that sponsors considering complex designs request a meeting with FDA prior to initiation of the trial. Including a detailed discussion of the proposed design and comparisons against alternative designs in meeting packages can help FDA reviewers understand and assess the appropriateness of trial operating characteristics.

B. Reporting Results from Bayesian Analyses

For completed clinical trials, sponsors should submit a clinical study report describing the design, analysis plan, and results from the trial. In addition to the typical content, the report of a trial with Bayesian analyses should include the following:

- The principal aspects of the design and analysis plan, as described in Section VIII.A.

- The results of the planned primary and secondary statistical analyses, including the treatment effect estimates, uncertainty in the estimates (e.g., with a credible interval), and whether pre-specified success criteria were met. The marginal posterior distributions of the estimands of interest should be described, including measures of location and variation. The results of sensitivity analyses, including quantifying and summarizing the sensitivity of results and conclusions to alternative reasonable choices for the prior distribution (see Section V.F). In the case of borrowing of external information with an informative prior in the primary analysis, the posterior results should be shown for other prior choices with different degrees of borrowing.

- The results from model checking, including an assessment of prior-data conflict and comparisons of the model predictions to the observed data.

- A report of sampling convergence diagnostics. For Bayesian analyses performed using non-direct sampling algorithms, this should include a description of the simulation settings as implemented. Any deviations from the planned implementation and a rationale for such deviations should be discussed.

- The software used to conduct analyses, including version details. Documented code should be provided for all primary and key secondary analyses and for any sensitivity analyses described in the study report. For Bayesian analyses performed using MCMC,

sufficient detail should be provided such that the results are reproducible. This requires reporting of the seed number(s) used during chain initiation.

- A discussion of the overall conclusions about the evidence related to the key trial objectives and analyses. This should briefly comment on the model chosen and the sensitivity of results and conclusions to alternative plausible assumptions and prior specifications.

In some cases, it may be reasonable to include highly technical information (e.g., reports of sampling convergence diagnostics) in an Appendix to the study report or in a separate dedicated document.

IX. GENERAL REFERENCES

Adcock, C. J. "A Bayesian approach to calculating sample sizes." Journal of the Royal Statistical Society: Series D (The Statistician) 37, no. 4-5 (1988): 433-439.

Evans, Michael, and Hadas Moshonov. "Checking for prior-data conflict." (2006): 893-914.

Gelman, Andrew. "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." (2006): 515-534.

Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. "Bayesian workflow." arXiv preprint arXiv:2011.01808 (2020).

Ji, Yuan, Ping Liu, Yisheng Li, and B. Nebiyou Bekele. "A modified toxicity probability interval method for dose-finding trials." Clinical trials 7, no. 6 (2010): 653-663.

Jiang, Liyun, Lei Nie, and Ying Yuan. "Elastic priors to dynamically borrow information from historical data in clinical trials." Biometrics 79, no. 1 (2023): 49-60.

Joseph, Lawrence, and Patrick Bélisle. "Bayesian sample size determination for normal means and differences between normal means." Journal of the Royal Statistical Society Series D: The Statistician 46, no. 2 (1997): 209-226.

Lin, Ruitao, Yanhong Zhou, Fangrong Yan, Daniel Li, and Ying Yuan. "BOIN12: Bayesian optimal interval phase I/II trial design for utility-based dose finding in immunotherapy and targeted therapies." JCO precision oncology 4 (2020): 1393-1402.

Malec, Donald. "A closer look at combining data among a small number of binomial experiments." Statistics in medicine 20, no. 12 (2001): 1811-1824.

Morita, Satoshi, Peter F. Thall, and Peter Müller. "Determining the effective sample size of a parametric prior." Biometrics 64, no. 2 (2008): 595-602.

Neuenschwander, Beat, Michael Branson, and Thomas Gsponer. "Critical aspects of the Bayesian approach to phase I cancer trials." Statistics in medicine 27, no. 13 (2008): 2420-2439.

Neuenschwander, Beat, Sebastian Weber, Heinz Schmidli, and Anthony O'Hagan. "Predictively consistent prior effective sample sizes." Biometrics 76, no. 2 (2020): 578-587.

O'Hagan, Anthony, Caitlin E. Buck, Alireza Daneshkhah, J. Richard Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley, and Tim Rakow. "Uncertain judgements: eliciting experts' probabilities." (2006).

O'Quigley, John, and Mark Conaway. "Continual reassessment and related dose-finding designs." Statistical science: a review journal of the Institute of Mathematical Statistics 25, no. 2 (2010): 202.

Pennello, Gene, and Laura Thompson. "Experience with reviewing Bayesian medical device trials." Journal of Biopharmaceutical Statistics 18, no. 1 (2007): 81-115.

Saville, Benjamin R., Donald A. Berry, Nicholas S. Berry, Kert Viele, and Scott M. Berry. "The Bayesian time machine: accounting for temporal drift in multi-arm platform trials." Clinical Trials 19, no. 5 (2022): 490-501.

Spiegelhalter, David J., Keith R. Abrams, and Jonathan P. Myles. Bayesian approaches to clinical trials and health-care evaluation. Vol. 13. John Wiley & Sons, 2004.

Thall, Peter F., and John D. Cook. "Dose-finding based on efficacy–toxicity trade-offs." Biometrics 60, no. 3 (2004): 684-693.

Tighiouart, Mourad, and André Rogatko, "Dose Finding with Escalation with Overdose Control (EWOC) in Cancer Clinical Trials". Statistical Science, 2010, 25: 217–226.

Wang, Yun, Wenda Tu, William Koh, James Travis, Robert Abugov, Kiya Hamilton, Mengjie Zheng, Roberto Crackel, Pablo Bonangelino, and Mark Rothmann. "Bayesian Hierarchical Models for Subgroup Analysis." Pharmaceutical Statistics 23, no. 6 (2024): 1065-1083.

Yuan, Ao, and Gen Xiang Chai. "Optimal adaptive generalized Pólya urn design for multi-arm clinical trials." Journal of Multivariate Analysis 99, no. 1 (2008): 1-24.

Yuan, Ying, Kenneth R. Hess, Susan G. Hilsenbeck, and Mark R. Gilbert. "Bayesian optimal interval design: a simple and well-performing design for phase I oncology trials." Clinical Cancer Research 22, no. 17 (2016): 4291-4301.