CID Case Study: A Study in Patients with Systemic Lupus Erythematosus

Study Design:

The proposed study is a randomized, double-blind, Phase 2 study in patients with systemic lupus erythematosus (SLE), a rare disease with a high unmet need. Patients are to be randomized to one of four treatment groups: three doses of investigational product (IP) or placebo. The primary endpoint of the study is Systemic Lupus Erythematosus Responder Index 4 (SRI-4) response at 52 weeks, a dichotomous outcome where response indicates success. This composite endpoint incorporates a hybrid Safety of Estrogens in Systemic Lupus Erythematosus National Assessment Systemic Lupus Erythematosus Disease Activity Index (SELENA-SLEDAI) score, British Isles Lupus Assessment Group index (BILAG) 2004 domain scores, and the Physician's Global Assessment (PGA). This endpoint will be evaluated using a Bayesian Hierarchical Model (BHM) with non-informative priors. Interim analyses will occur at 8 prespecified time points. The initial randomization ratio will be 1:1:1:1.

At each interim analysis except for the last one, a response adaptive randomization (RAR) procedure will take place where the randomization allocation probabilities for each of the three IP arms may be modified moving forward. The randomization allocation probability for the placebo arm will remain fixed at 25% throughout the study.

At each interim analysis except for the first one, a futility determination will be made, based on the performance of the three IP arms compared to placebo.

Innovative Characteristics:

FDA considers the following initially proposed study design features to be innovative, making it appropriate to review the design under the Complex Innovative Trial Design (CID) pilot meeting program:

- Use of an adaptive rule to allow for the possibility of changing the primary endpoint at Week 52 to Lupus Low Disease Activity State (LLDAS) or BILAG-Based Composite Lupus Assessment (BICLA), based on interim analysis results
- Use of an adaptive rule to allow for the possibility of pooling data from different dose levels in the comparison to placebo for the primary analysis
- Use of Bayesian methods and RAR to allow for modification of the randomization allocation probabilities for the IP arms in a study meant to reliably evaluate the IP in comparison to placebo

Potential Benefits of Design:

- It allows for the study to both serve as a dose-ranging study and reliably evaluate the IP in comparison to placebo, potentially saving time and resources relative to conducting separate studies for these two purposes.
- It tends to allocate more patients to most effective doses.
- It increases the power of comparisons of the most effective doses to placebo.

Considerations for the Proposed Design:

- What impact does RAR have on the comparability of treatment groups with respect to baseline measurements and characteristics?
- Does this study design have adequate operating characteristics (e.g., power, type I error rate, reliability of point estimates, probability of selecting the best dose, etc.) across different true dose-response relationships and across the multidimensional range of plausible values for the nuisance parameters?
- How does the use of RAR perform against arm-dropping approaches?
- For the primary analysis, how does the BHM perform compared to competitive conventional methods?
- What firewalls and other procedures will be put in place to ensure that interim analysis results will remain confidential and study integrity will be maintained?

Simulations:

The Sponsor conducted simulations to assess the operating characteristics of the proposed model under different true dose-response relationships and under a multidimensional range of plausible values for the nuisance parameters. Nuisance parameters included the true underlying placebo response rate, the patient enrollment rate, and the within-patient correlation of response status at adjacent visits. Simulations evaluated important operating characteristics such as type I error rate, power, reliability of point estimates, futility determination probabilities, and probabilities of selecting the most effective dose. The set of combinations of values for the nuisance parameters (under which simulations were to be performed) was expanded several times as a result of iterative feedback.

Discussion:

The study using this CID will have two purposes: (1) to compare several doses of the IP to each other; and (2) to reliably evaluate the IP in comparison to placebo. The former purpose reflects that the data will be used for dose selection. In contrast, the latter purpose reflects that the data from this study will be evaluated to determine whether there is evidence that the IP at the selected dose level is safe and effective.

When taking such an approach, a key consideration is whether the operating characteristics are appropriate across the multi-dimensional space of plausible ranges for nuisance parameters, which initially included the true underlying placebo response rate, the patient enrollment rate, the within-patient correlation of response status at adjacent visits, the within-patient correlation of response status across the different endpoints to potentially be used in the primary analysis, and the lag time between data cutoff and the time of adaptation at the relevant interim analyses. A potential challenge for earlier versions of this design was that the high dimensionality of the space would potentially make a comprehensive evaluation of the space infeasible. Ultimately, the dimensionality of this space was reduced by making changes to the CID such as removing the possibility of adapting the primary endpoint and fixing the lag time between data cutoff and the time of adaptation.

In addition, there were challenges in determining plausible ranges for at least one of the nuisance parameters. Based on FDA feedback, the sponsor conducted several sets of simulations to ensure that the true underlying placebo rates for SRI-4 response considered in the simulations covered the range of plausible values.

Regardless of whether operating characteristics are to be evaluated using analytical methods or via simulation, it is often beneficial to demonstrate the utility of a given CID by comparing it to more "conventional" study designs. When doing so, the extent to which these comparisons are informative depends on whether the selected comparator designs serve as adequate representations of the most competitive conventional study designs. The Sponsor conducted several comparisons, including a comparison of RAR to competitive study designs such as those where the randomization ratio remains fixed but poorly performing arms might be dropped partway through the study, as well as a comparison of use of the BHM to the use of multiplicity control procedures that leverage information known/believed prior to study initiation (e.g., hierarchical testing procedures in decreasing dose level if *a priori* the higher dose levels are thought to be most effective, Dunnett's test).

In addition to properly addressing statistical issues as illustrated above, it is important to ensure that clinical/scientific issues are also given proper consideration when determining whether a CID is adequately designed. The initial version of the CID used an adaptive rule to allow for the possibility of changing the primary endpoint based on interim results and used an adaptive rule to allow for the possibility of pooling data from different dose levels in the comparison to placebo for the primary analysis. FDA was concerned about the implication of these adaptations on the estimand[1] being targeted by the primary analysis. When pooling different dose levels in the comparison to placebo, it was not clear whether the primary analysis was targeting an estimand that was sufficiently clinically relevant. These two adaptive rules were ultimately removed from the CID.

The use of RAR can also impact the estimand being targeted in the final analysis. Even when the inclusion/exclusion criteria remain unchanged for the duration of the enrollment period of a study, patients entering the study toward the beginning of the enrollment period may have more severe forms of the disease in question, compared to patients entering the study toward the end of the enrollment period. If this were to occur in the study, any adaptation of the randomization ratios of the three different dose levels of the IP would potentially lead to a systematic imbalance across all four treatment arms with respect to potentially prognostic baseline characteristics. Results would therefore be difficult to interpret since they would be biased with respect to the estimand of interest. Given these concerns, stratification will be used, and the number of patients with a given baseline disease severity will be capped throughout the accrual periods to ensure stable proportions of high vs. low severity patients enrolled between formal analyses. These procedures are helpful in minimizing the probability of the mentioned imbalances. While the risk of such imbalances occurring due to RAR remains, FDA

[1] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9r1-statistical-principles-clinical-trials-addendum-estimands-and-sensitivity-analysis-clinical

determined that in this rare disease setting with a high unmet need the risks were acceptable in this design.

Compared to more conventional study designs, it appears that this CID will tend to allocate more patients to most effective doses and will increase the power of comparisons of the most effective doses to placebo. Careful review of the final analysis results will be necessary to ensure that the correct conclusions are drawn and that correct decisions are made after properly accounting for the study's complex and innovative elements.