A Bayesian Design for Agreement Studies for Medical Devices

Lorenzo Manchini, Ph. D. and Rajat Mukherjee, Ph. D.

August 11, 2022

1 Introduction

A common way to validate an investigational device measuring a certain physiological function such as the heart-rate (bpm) is to validate this device with respect to a gold-standard measurement or a reference measurement made, for example, using a predicate device. Validation is performed by assessing the agreement between the measurements made by the two devices.

A well-known traditional (frequentist) method to assess the statistical agreement between two devices is the application of the Bland-Altman plot. This plot, introduced by Bland and Altman (1986), is obtained by constructing a scatter plot of the difference of each pair of measurements from the two methods against their average.

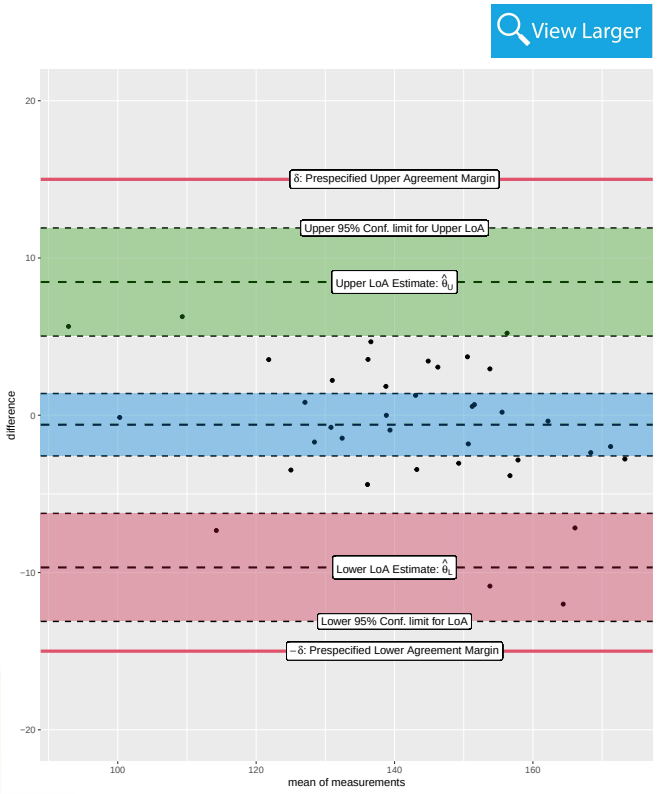

Figure 1: Bland-Altman plot example using simulated data

Limits of Agreement (LoA) are then added to the plot which are calculated using:

$$LoA = \mu_d \pm 2\sigma_d,$$

where, μ_d is the (unknown) population mean and σ_d^2 is the (unknown) population variance of differences between the two methods respectively. We denote these population quantities by $\theta_U = \mu_d + 2\sigma_d$ and $\theta_L = \mu_d - 2\sigma_d$. The upper and lower LoAs are estimated from the data as $\bar{d} \pm 2SD_d$, where $\bar{d}$ and SD_d are the sample mean and the sample standard deviation of the paired differences between the two measurements. The upper and lower LoA estimates for a simulated data set is shown in Figure 1. The blue region is the 95% confidence interval for the bias. The $100 \times (1-\alpha)\%$

confidence interval (CI) for the LoA are then obtained with the bounds given by

$$\text{lower} = \bar{d} - 2SD_d - t_{n-1,1-\alpha/2} SD_d \sqrt{\frac{1}{n} + \frac{z^2_{1-\alpha/2}}{2(n-1)}}; \text{ and}$$

$$\text{upper} = \bar{d} + 2SD_d + t_{n-1,1-\alpha/2} SD_d \sqrt{\frac{1}{n} + \frac{z^2_{1-\alpha/2}}{2(n-1)}}.$$

These confidence intervals are given by the region shaded in green (upper) and red (lower) in Figure 1. For normally distributed data and for $\alpha = 0.05$, 95% of the future differences are expected to lie between the above two bounds. In order for the agreement to be acceptable these upper and lower bounds should be within a tolerance margin ($\pm\delta$) which needs to be prespecified. The null hypothesis (H_0 : lower $< -\delta$ OR upper $> \delta$) is rejected at 5% alpha if the 95% upper is less than δ AND the 95% lower is greater than $-\delta$.

Powering for such agreement studies are needed to ensure that even when the bias and variability are acceptable, the confidence interval for the LoA is not too wide. Sample size calculation for normal endpoints can be carried out using an iterative process suggested by Lu et al. (2016). Careful planning of such studies is also required to ensure that they are also not over-powered. Prior data may be available from previous pilot studies which can be used to guide the selection of the maximum bias and variability values for which the trial is powered for. Power curves for fixed (no interim analysis) designs for different assumed values of μ_d and $\sigma_d = 5$ and a tolerance margin of $\delta = 15$ are given in Figure 2. We see that there is substantial variation in sample size required for say 80% power, depending on the true values of these parameters.

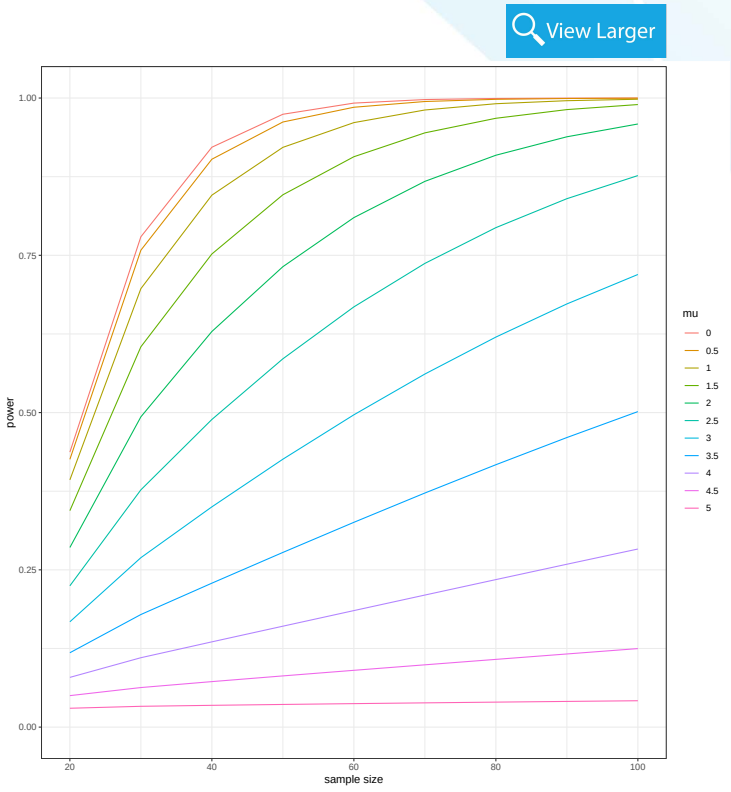

Figure 2: Power curves for different values of μ_d with $\sigma_d = 5$.

In absence of prior data, adaptive group sequential designs that allow for early stopping for efficacy or futility or adapt sample size based on interim data can help in mitigating the risk of an under-powered or over-powered trial. Below we propose a Bayesian adaptive two-stage design as an example of such a design.

2 Why a Bayesian Design?

Adaptive designs can be constructed under the traditional Frequentist as well as the Bayesian framework. In the frequentist framework, interim decision for adaptations such as sample size re-

estimation rely on the conditional power (CP). For agreement studies, there is no straightforward CP computation procedure available. Monte-Carlo estimates using simulations are possible, however, in our point of view a better alternative is the use of the Bayesian predictive power as defined by Spiegelhalter et al. (2004). In the Bayesian framework we assume non-informative or weakly-informative priors on μ_d and σ_d. For the example below, these priors are $Normal(0, 10^6)$ and $Gamma(0.5, 10^6)$ respectively. These priors densities are shown in Figure 3. The Normal prior is essentially a flat curve centered at 0 and thus allowing for all possible values a priori. The Gamma prior also allows for small as well as large variances.

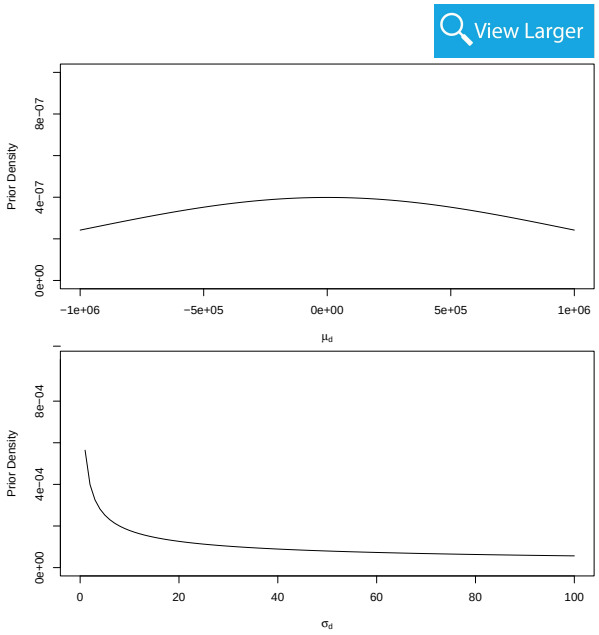

Figure 3: Weakly informative prior densities for μ_d with σ_d for the Bayesian analysis.

Once data from the target trial is available the above priors are updated to get the posterior densities of μ_d and σ_d from which we can get the posterior densities for the upper ($\theta_U = \mu_d + 2\sigma_d$) and lower ($\theta_L = \mu_d - 2\sigma_d$) LoAs (Figure 4). The use of weakly informative priors ensure that the posterior densities are almost fully informed by the data from the target trial.

For making interim decisions suppose $\mathcal{D}_{n_1}$ are the data available at the interim from n_1 observations. The posterior predictive density of future differences (d^*) in paired measurements ($p(d^*|\mathcal{D}_{n_1})$) can then be derived using the Bayes theorem from the posterior distribution $p(\mu_d, \sigma_d|\mathcal{D}_{n_1})$ obtained at the interim analysis. This posterior predictive distribution can then be used for calculating the predictive power for different sample sizes for stage-2 (n_2). We describe the proposed two-stage adaptive design with the help of an example with simulated data in the next section.

3 Bayesian Adaptive Two-Stage Design

The two-stage adaptive design described here has only one interim look but can be generalized into multistage design with several interim looks. The increase in number of looks generally requires a marginally larger sample size.

3.1 Bayesian Success Criteria

As described above, at the final analysis we want to test the null hypothesis that $\theta_U > \delta$ or $\theta_L < -\delta$, i.e., the upper (lower) LoA is greater (less) that δ ($-\delta$). In the Bayesian framework, this null hypothesis can be rejected if the posterior probability that the $\theta_U < \delta$ and $\theta_L > -\delta$, is greater than a pre-specified high threshold, usually greater than 0.975. This threshold (γ) is the success

criteria and needs to be fixed at the planning stage based on simulations, mainly to ensure that the type-I error is controlled at the nominal level of 2.5% (one-sided). For illustration purpose, for the example below we set γ at an arbitrarily high value of 0.99 applicable to both the interim (for possible early stopping for success) and the final analyses. In practice however, a Bayesian group sequential design can be considered in order to set different values of γ for the different interim and final analyses with the γ threshold being larger for earlier interim than the later ones and the final analysis. A nice example of such Bayesian group sequential design is the Pfizer COVID vaccine trial, see Polack and et al. (2020). Also see Mukherjee et al. Mukherjee et al. (2022) for a Bayesian sequential trial based on predictive power for COVID vaccine trials.

3.2 Bayesian Interim Analysis

At the interim analysis, the posterior probability that the 95% upper LoA is less than δ and the 95% lower LoA is greater than $-\delta$ is to be calculated. If this probability already exceeds the $\gamma = 0.99$ threshold then stop the trial for success, otherwise calculate the predictive power to decide on the final sample size. Another possibility is early stopping due to futility when the calculated predictive power at the maximum sample size (respecting time and budgetary constraints) is still small, say below 20%. The futility threshold is also typically pre-fixed using simulations under various scenarios and clinical and logistical considerations and in general is considered to be non-binding.

3.3 Example with Simulated Data

To illustrate the proposed Bayesian Adaptive design we use simulated data where the original measurements (mimicking heart rate, bpm) were generated using a normal distribution with mean 140 and SD 20. The mean bias of paired differences (μ_d) has been set equal to 2, and the standard deviation of differences (σ_d) equal to 5. Agreement margin and success threshold have been set to $\delta = 15$ and $\gamma = 0.99$ respectively. Figure 4 (left) shows the Bland-Altman plot at the first interim look at $n_1 = 40$ while the Bayesian posterior distribution for the upper and lower LoAs are shown in Figure 4 (right).

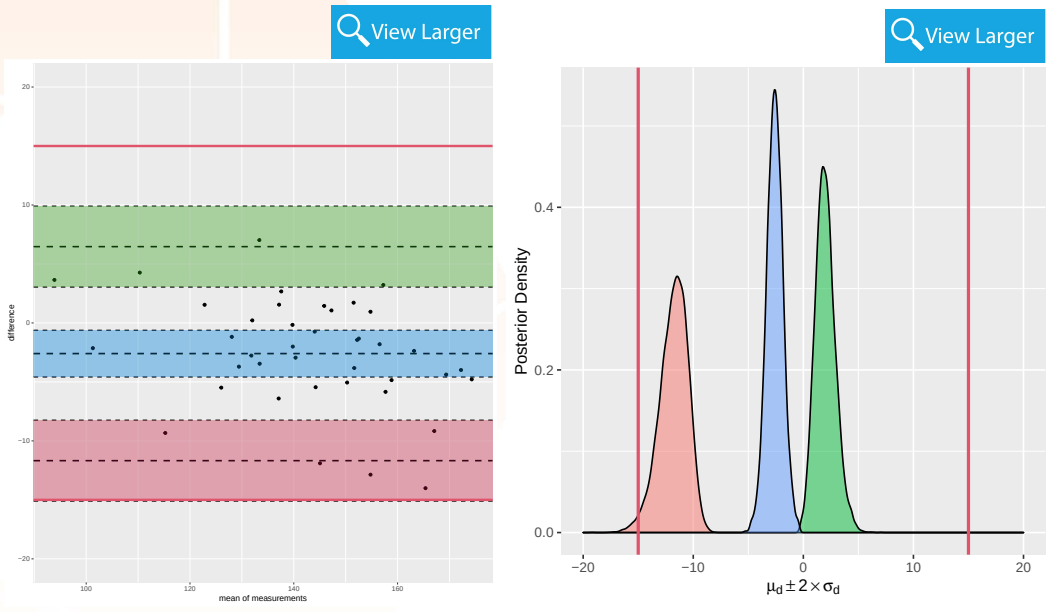

Figure 4: Bland Altman plot (left) and the posterior distributions (right) for the upper and lower LoAs at the interim with $n_1 = 40$ subjects. The red vertical lines (right) correspond to $\pm\delta$.

Starting from weakly-informative prior, the posterior probability of success for Bayesian analysis is 0.988, just below the success threshold. Also, lower LoA confidence interval in Bland-Altman plot includes the value $\delta = -15$. Thus, both the Bland-Altman analysis as well as the Bayesian posterior distributions in Figure 4 (right) suggest not stop the trial early (with n_1 subjects) for success and continue to the next interim.

Following the proposed Bayesian design, we calculate the predictive power at different final sample sizes. These are given in Table 1 below. We see that a predictive power of at least 80% is guaranteed at a final sample size of $N = n_1 + n_2 = 70$ where the predictive power given the interim data is 82.12%. We thus take a final analysis at this sample size. The final analysis Bland-Altman plot is given in Figure 6. The Bland-Altman plot with LoAs and the final Bayesian posterior distribution (right) for the upper and lower LoAs are shown in Figure 5 below. The posterior probability that the upper LoA is less than $\delta = 15$ and lower LoA is greater than -15 is now 0.9979 which is statistically significant.

Table 1: Predictive power computed with the interim data with various final sample sizes.

Final N	Predictive Power
50	0.7441
60	0.7918
70	**0.8212**
80	0.8494
90	0.8691

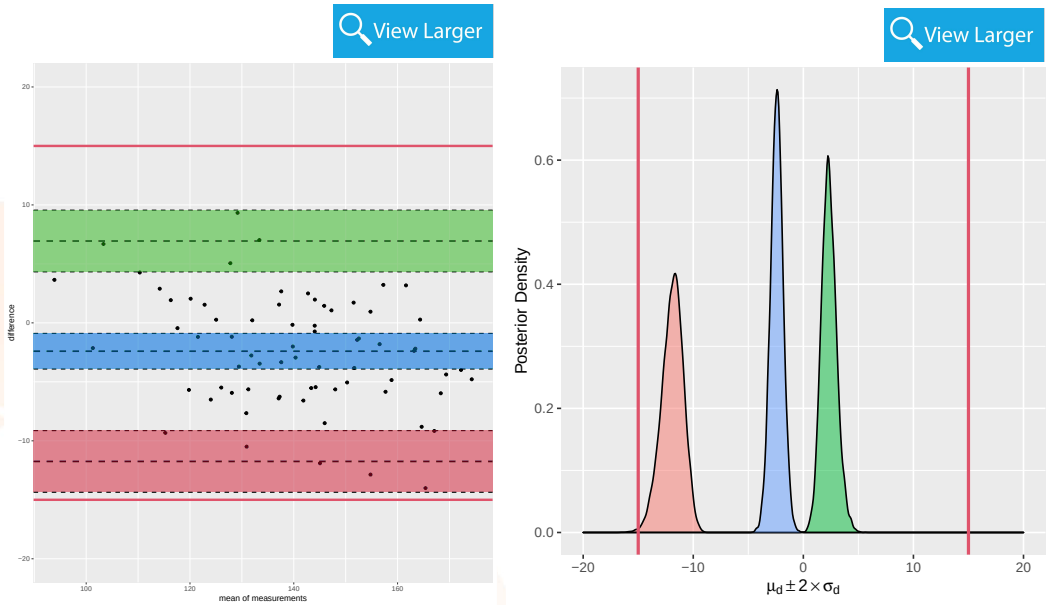

Figure 5: Bland Altman plot (left) and the posterior distributions (right) for the upper and lower LoAs at the final analysis with $n_1 + n_2 = 70$ subjects. Posterior probability of success is 0.9979.

4 Discussions

In this article, with an example using simulated data, we have presented a Bayesian adaptive design that allows for early stopping for success as well as for futility or re-estimation of the final sample size

based on the interim data and using the Bayesian predictive power. Note that we could have carried out the interim analysis in the Bayesian framework with the final analysis being the traditional Bland-Altman analysis. This would be considered a hybrid design. Bayesian designs in general offer more flexibility in terms of interim adaptations. We have only focused on correcting the sample size at the interim but adaptations like population enrichment, dose selection and hypothesis selection in an exploratory trial are also possible. Bayesian designs are also better suited for Master protocols such as Basket and Umbrella trial designs. Our example focused on using weakly informative priors for the parameters of interest, however, if reliable historical data, for example from pilot or historical trials are available, then informative priors can be used for Bayesian designs. Using informative priors can result in savings in sample size for the future trial. This is particularly interesting for trials in rare diseases and medical devices with predicate(s) in the market. Extensive simulations are required at the planning stage in order to establish frequentist operating characteristics of the design for regulatory approval of the study design. Communication with regulatory agencies (FDA, EMA) should start early for Bayesian designs especially when using informative priors. To sum up Bayesian designs may need more rigorous upfront planning but the promise to save at the end and/or mitigating risk through proper pre-specified adaptation is real.

References

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1(8476):307–310.

Lu, M. J., Zhong, W. H., Liu, Y. X., Miao, H. Z., Li, Y. C., and Ji, M. H. (2016). Sample Size for Assessing Agreement between Two Methods of Measurement by Bland-Altman Method. *Int J Biostat*, 12(2).

Mukherjee, R., Yajnik, P., Muhlemann, N., and Morgan-Bouniol, C. (2022). A sequential predictive power design for a covid vaccine trial. *Statistics in Biopharmaceutical Research*, 14(1):42–51.

Polack, F. P. and et al. (2020). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N Engl J Med*, 383(27):2603–2615.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, Chichester.