A combined proof of concept and dose finding study with multiple endpoints: A Bayesian adaptive design in chronic prostatitis/chronic pelvic pain syndrome

Running title: Bayesian adaptive allocation procedure in CP/CPPS

Reynaldo Martina[1,2], Jos Houbiers[2], Joost Melis[2], and Olivier van Till[2]

[1] Biostatistics Department, University of Liverpool, 1-5 Brownlow Street, Liverpool, L69 3GL, UK
[2] Development, Astellas, Sylviusweg 62, 233BE, Leiden, The Netherlands

Received zzz, revised zzz, accepted zzz

There is a need for identifying effective drugs or terminating ineffective drugs as early as possible to optimize efficient and cost effective drug development. The aim of the proposed trial was to simultaneously establish Proof of Concept (PoC) and dose finding (DF) for a new drug with a novel mode of action in a new indication. We simulated and executed an adaptive allocation design to investigate the effects of a drug on male patients suffering from chronic prostatitis/chronic pelvic pain syndrome (CP/CPPS). This manuscript describes the clinical trial simulations and primary analysis results. A Bayesian adaptive allocation procedure was employed to allocate patients to treatment using a normal dynamic linear model. The study was to stop early for efficacy if the probability of a clinically significant difference between an experimental arm ~~drug~~ and placebo was at least 90%. The study was to stop for futility if the probability that the maximum effective dose was better than placebo by at least the futility difference was less than 20%. During the execution phase the study was stopped early, i.e. 32% less than planned maximum sample size, due to futility. The final results confirmed that the predefined stopping rules were met. In conclusion, the simulations showed that, if the drug was effective, this adaptive design could accomplish both the goals of PoC and DF. However, the study stopped early for futility in line with the simulation predictions for stopping. This resulted in the early stopping of a trial recruiting patients on ineffective treatment.

Key words: Bayesian adaptive allocation; Dose finding; Normal dynamic linear model; Proof of concept; Stopping rules

Corresponding author
Reynaldo Martina
Email: Reynaldo.Martina@liverpool.ac.uk
Phone: +447508711427

1 Introduction

The classical way of drug development includes various studies and small steps (e.g. Phase I (single ascending dose (SAD) and multiple ascending dose (MAD), Phase IIA (Proof of concept (PoC), Phase IIB (dose-finding (DF)), and ultimately the large Phase III studies). This way the early phase of development results in exposure of the new drug to many subjects, long timelines, related high costs, and late company decision points, while many drugs appear to fail at PoC or DF. Combining development phases is an option to bring decision points forward.

In this new project, we explored options to shorten drug development timelines, expose fewer subjects in early phase and reduce costs. In order to achieve these objectives we employed several of the following development design aspects at the early stages of development

1. Combine SAD and MAD into one study protocol

2. Partly overlap SAD and MAD

3. Include biomarker/surrogate endpoints in Phase I

4. Include patients in Phase I in MAD part

The Phase II study described in this manuscript was a combined PoC and DF study to investigate efficacy and the effective dose range of a drug with a novel model of action in a patient population that is difficult to recruit, using a Bayesian adaptive allocation procedure (Wagenlehner *et al*. 2017). Adaptive designs are being applied more often than not in the exploratory phase of drug development. The publication by Berry et al. (2002) on the practical applications of adaptive designs is arguably one of the most influential publications in the field of the practical application of novel designs. The ASTIN trial (Grieve and Krams, 2005) is a well-known case of the application of an adaptive trial, namely a Bayesian dose response adaptive trial. There are various reasons when an adaptive trial may provide substantial advantage compared to a conventional design, some of which were appropriate for this study:

1. Uncertainty in the choice of the primary variable
2. Uncertainty in the potential of the biomarker to predict clinical response
3. Patients recruitment is uncertain and patients therefore difficult to recruit

This was the first clinical study to test the effectiveness of the drug in patients. Due to the uncertainties associated with this drug an adaptive design was proposed to combine PoC and dose finding and include as many doses as possible to investigate the dose response relationship. The latter is also recommended by Grieve and Krams (2005). One of the key differences in the design as published by Grieve and Krams (2005) and the design we employed is that in our design it is insured that the sample size in the placebo group and the best dose group is similar, while in the published design there was a minimum allocation probability to placebo. Ensuring a similar number of patients in the placebo arm and best dose arm allows for a robust evaluation of PoC in a conventional 1:1 ratio of experimental drug vs placebo. Moreover, in our design at least 50 patients needed to be recruited in the placebo arm (and best dose arm) before the stopping rules could be activated, ensuring increased power to assess PoC. This requirement was not made in the design described by Grieve and Krams (2005).
Other combined PoC and DF studies, include the two-stage approach in migraine clinical trials described by Sagkriotis and Scholpp (2008) and a Bayesian dose-finding trial with adaptive dose expansion described by Berry et al. (2010). Another dose finding study (with possibility to include an evaluation of PoC) is described in Ivanova et al. (2008). An adaptive dose finding study based on a t-statistic was described. Whitehead and Zhou (2001) applied Bayesian methods in dose-escalation studies in healthy volunteers. Xie and Tremmel (2012) describe a Bayesian adaptive design for

randomised placebo-controlled combined phase I/II clinical trial, where the design included a dose escalation.

The doses used in the PoC and dose finding studies should also be carefully considered. As much doses as possible should be considered in a dose finding study (Grieve and Krams (2005)). If this is possible, adding a PoC part to the design by evaluating the best dose versus placebo is relatively easy and efficient.

Using the accruing data from the trial to estimate the dose response is in line with a Bayesian approach. A prerequisite is that the time spent on the interim data handling and delivery of results is kept to a minimum to ensure that only a minimum number of patients is randomized during that period to optimize the efficiency of the trial. This manuscript describes the employed Bayesian adaptive methods, simulations that were undertaken to evaluate the operating characteristics of the proposed Bayesian response adaptive design and final Bayesian results of the trial.

2 Study design

This was a randomised double blind response adaptive trial to investigate PoC, i.e. does the drug work in these patients, and to determine the dose response relationship of the drug (Wagenlehner *et al.* 2017). Five doses were investigated: placebo, 25mg twice daily (bid), 75mg (bid), 150mg once daily (qd), 150mg (bid) and 300mg (bid). In the dose response model these doses are represented by the notation d_0, d_1, d_2, d_3, d_4 and d_5.
The primary variable was the National Institute of Health Chronic Prostatitis Symptom Index (NIH-CPSI) total score at week 12. The NIH-CPSI combines aspects of 3 important symptom domains of Chronic Prostatitis/Chronic Pelvic Pain Syndrome (CP/CPPS). In particular, pain, micturition, and quality of life (QoL)), with a scale ranging from 0 to 45. Higher scores indicate higher disease severity (Litwin *et al.* (1999)).
The secondary variable was the change from baseline in the Pain sub score of the NIH-CPSI at week 12. The primary variable was used to re-estimate the allocation probability and evaluate the stopping rules for Go/No-Go (see Methods section). The secondary variable was used to re-estimate the allocation probabilities only.

A burn-in period was established comprising of 60 randomized patients in a 1:1:1:1:1:1 ratio (10 patients per arm). After the initial 60 patients the randomization allocation would be updated every four weeks following an interim analysis, which was to be performed by an independent data monitoring committee (IDMC). A separate and independent Data Safety Monitoring Board (DSMB) was responsible for periodic review of safety data and could stop the trial at any time for safety reasons. Stopping rules for efficacy were finalized before the final protocol and were defined as follows:

- If there are at least half of the patients enrolled in the study, at least 50 subjects randomized to the most likely maximum effective dose arm and the placebo, and the following condition is met, then the study would stop for success
  - The posterior probability that the *most likely maximum effective dose* is *clinically significantly better than placebo,* defined as a mean difference of at least 4 points on the change in NIH-CPSI total score, is at least 90%
- If there are at least half of the patients enrolled in the study and the following condition holds then the study would stop for futility:
  - If the posterior probability that the *most likely maximum effective dose* is *better than placebo*, defined as at least a mean difference of 2 points on the change in NIH-CPSI total score, is less than 20%.
- If the maximum enrollment is reached enrollment is stopped
- Study is successful if at the FINAL ANALYSIS the probability that the maximum effective dose arm is significantly better than placebo is at least 95%

The final analysis was performed on the Full Analysis Set (FAS) defined as:
The Bayesian interim and final analyses were performed on the Full Analysis Set (FAS) population, defined as patients who took at least one dose of double-blind study medication after randomization, had an NIH-CPSI total score at baseline and at least one NIH-CPSI total score post-baseline during the double-blind treatment period, or had an NIH-CPSI total score at baseline and dropped out due to an adverse event. The FAS was defined and finalized prior to un-blinding

The stopping rules included some novel ideas such as the inclusion of both a clinical significant difference and a futility difference for stopping for success or lack of effect. The inclusion of a once daily dosing arm in a twice daily dosing regimen is also unusual. This allowed for the investigation whether a single dose regimen could be considered for future development.

Sample size

A total of 350 patients was deemed sufficient from a clinical perspective. The adequacy of the total sample size for this study was explored through simulations. The simulated individual subjects' data used in the simulations of each of the scenarios were obtained by drawing a sample from a normal distribution with mean μ_d and variance σ^2, where μ_d is the mean change from baseline at week 12 for dose group d.
For the primary variable, a difference of at least 4 points in change from baseline in the NIH-CPSI total score compared to placebo was considered clinically relevant. Trial simulations with different numbers of subjects have shown that a total of 350 subjects will be sufficient for this trial to demonstrate that, assuming a difference of at least 4 points in change from baseline total score between the best dose and placebo and a standard deviation (SD) of 7, the probability that the best dose is superior to placebo is at least 95%, if the drug is effective. The simulations also showed that if the true dose response is similar to placebo, the trial would claim success in less than 5% of the simulated cases.

3 Methods

3.1 The dose response model

Let the change from baseline score for subject i, at week 4, week 8, and week 12 is $Y_{i,4}$, $Y_{i,8}$, and $Y_{i,12}$, respectively depicting a treatment duration of respectively 4, 8 and 12 weeks. Let μ_i be the mean of the treatment group to which ~~for~~ subject i pertains. The week 12 assessments are assumed to be normally distributed with mean μ_i and variance σ^2:

$$Y_{i,12} = N(\mu_i, \sigma^2), i = 1,2,..., N$$

Where N is the total number of patients with week 12 data. In this study N was capped at 350.
Let θ_i be the mean of treatment group i. A two dimensional first order normal dynamic linear model (NDLM) is constructed to model the dose response relationship (Berry et al., 2011), assuming non-informative priors:

$$\theta_0 \sim N(0,10^2)$$
$$\theta_1 \sim N(0,10^2)$$

Where θ_0 and θ_1 are the treatment effects of the placebo group and the lowest dose group, respectively. Two prior distributions were defined for the placebo group and the lowest dose group to ensure the treatment effects of the physiologically active dose groups were not diluted by the placebo effect. As a result, the placebo group is modelled separately. This avoids that the effects of the active doses are not imbedded in the placebo effect and vice versa. The prior distributions were considered non-informative

for this compound. However, the variance component was obtained from published studies that recruited similar type of patients as our target population. It should be noted that this variance component is large in relation to the mean, which allows the sampled values to vary considerably, allowing for (extremely) high or (extremely) low placebo response to be accounted for within the modelling framework, see also Section 4.1.
The NDLM structure for the twice daily doses was constructed as follows:

$$\theta_2 \sim N(\theta_1, \tau_1^2)$$
$$\theta_4 \sim N(\theta_2, \tau_1^2)$$
$$\theta_5 \sim N(\theta_4, \tau_1^2)$$

The once daily dose (150mg qd) was also modelled separately but was linked to the 150mg bid dose as follows, with a separate variance component:

$$\theta_3 \sim N(\theta_4, \tau_2^2)$$

The 150mg qd dose was modelled this way because it was expected that this dose would provide greater benefit than the 75mg bid dose, but would be inferior to the 300mg bid dose.
It should be noted that the modelling of a once daily dose in conjunction with a twice daily dose regimen was a novel application of this NDLM.
The variance components of the model (τ_1 and τ_2) were modelled assuming a non-informative inverse gamma (IG) prior distribution:

$$\tau_1 \sim IG(2,1)$$
$$\tau_2 \sim IG(2,1)$$

Due to the uncertainties already described for this drug, such as uncertainties regarding the effects in this patient population, the novel indication and unavailability of previously published, non-informative priors were chosen. The prior distributions are conjugate.

The key secondary variable was modelled in the same way as the primary variable.

It should be noted that the modelling of a once daily dose in conjunction with a twice daily dose regimen was a novel application of this NDLM.

3.2 The longitudinal model

At each interim analysis there will be subjects who could have complete or incomplete data, i.e. they are still participating in the study and not all their assessments have been carried out. Some subjects will have their week 12 observation, $Y_{i,12}$. These subjects may also have their interim values of the primary variable observed, $Y_{i,4}$ and $Y_{i,8}$. There may be subjects with interim observations, but no week 12 value. There may be subjects with no observations.
The data from subjects for whom their week 12 assessment has not yet become available will be used to estimate the week 12 value. A linear regression model is created for the correlation between the week 4 and week 8 values and the week 12 values. Let i=1,…,N (total number of patients in the trial). The following model structure was defined:

$$[Y_{i,12} \mid Y_{i,4}] \sim N(\alpha_{i,4} + \beta_{i,4} Y_{i,4}, \lambda_{i,4}^2)$$
$$\alpha_{i,4} \sim N(0,10^2)$$
$$\beta_{i,4} \sim N(0.80,0.25^2)$$
$$\lambda_{i,4}^2 \sim IG(0.01,1)$$

As before, the chosen prior distributions were non-informative. However, the underlying assumption for the mean of $\beta_{i,4}$ is that most of the effect of the dose, here set at 80%, is achieved at week 4, so this choice is not entirely uninformative.

A similar model was constructed for the distribution of $Y_{i,12}$ given $Y_{i,8}$. The prior distributions for week 8 were held similar to those for week 4. This approach allows for a conservative imputed week 12 value from week 8 data. Moreover, there is no reason to assume that if 80% of the response is achieved at week 4, that the effect will continue to increase up to week 12. The joint posterior distributions of $\alpha_{i,4}$, $\alpha_{i,8}$, $\beta_{i,4}$, $\beta_{i,8}$, $\lambda_{i,4}$, and $\lambda_{i,8}$ is updated based on all subjects with observed values of Y_{12}. This model is used (using Bayesian imputation within Markov chain Monte Carlo (Berry et. al. (2011) and Gelman et. al. (2004)) to update the dose-response model.

Week 12 values may also be missing during the final analysis after unblinding due to early discontinuation of subjects from the study. The same Bayesian model will be utilized to impute a missing Week 12 value. Note that when a subject has completed the week 12 assessment, this value will be used in the NDLM modelling and Bayesian probabilities (see next section) and his week 12 assessment will no longer be imputed.

Due to the novel nature of the compound and to ensure that a minimal number of patients will be exposed to a new drug with a still unknown safety profile, withdrawals due to an adverse events were considered treatment failures, i.e., change from baseline at week 12 was set to zero for both endpoints. This penalty (imputing a zero value to the change from baseline at week 12 for patients who withdraw due to an adverse event) ensures that the doses that are both efficacious (if drug is effective) and safe would be selected. A theoretical consequence is that the study could stop for futility because of the many drop outs due to adverse events leading to withdrawal on doses that are potentially effective. However, an effective dose with an unfavorable risk benefit profile is unlikely to be take forward for further development.

3.3 Bayesian probabilities and randomization vector

Let j in {1, 2} denote the parametrization of the primary and secondary variable, respectively. Let $\theta_{0j},\ldots,\theta_{5j}$ be the posterior mean of the effect of the 6 doses for each variable. The following Bayesian probabilities will be calculated at each 4-week update and for the final analysis for the primary and key secondary parameter from the joint posterior distribution:

P_{ij}^{MAX} : the probability that ~~each~~ dose i is the maximally effective dose for variable j. The dose with the largest P_{ij}^{MAX} for variable j is labelled the most likely maximum effective dose (best dose) for that variable.

$\Pr(\theta_{ij} < \theta_{0j} \mid \text{data})$: the posterior probability that dose i is superior to placebo for variable j.

$\Pr(\theta_{i1} < \theta_{01} - 4 \mid \text{data})$: the posterior probability that dose i is clinically significantly better than placebo, i.e., the NIH-CPSI total score has a difference to placebo of at least 4 points. Note that the clinically significant difference was defined for the primary endpoint only, i.e. for j=1.

$V(\theta_{ij})$: the posterior variance for the mean change from baseline in the primary (j=1) and secondary parameter for (j=2) dose i.

Furthermore, V_{dij} is defined such that:

$$V_{ij} = P_{ij}^{MAX} \sqrt{V(\theta_{ij})/(n_i + 1)}$$ (Berry *et al.* (2011), Connor *et al.* (2013))

Where n_i is the number of patients in dose i.

Note that at the time of the design of this study a different variant (earlier version) of V_{ij} was used compared to the published variant in Connor. In the variant above, the probability of a dose is the maximally effective dose (P_{ij}^{MAX}) has a greater impact on the allocation probability. This was done to ensure that patients will be allocated to doses with the largest effect with increased probability in order to increase the number of patients having the best possible doses.

The randomization probabilities (R_i) of the 6 treatment groups are calculated as the average of the two randomization probabilities associated with the primary and secondary variable:

$$R_i = \frac{1}{2}(\sum_{j=1}^{2}(V_{ij} / \sum_{i=0}^{5} V_{ij}))$$

4 Results

4.1 Simulation results

Extensive simulation studies were performed for each of 19 different scenarios including a scenario under which all doses are equal to placebo, several scenarios under which selected doses have a moderate, but no clinically relevant effect as well as scenarios under which selected doses had a clinically relevant effect, including different shapes of the dose-response curves. Some scenarios also consider opposing shapes of the dose-response relationship of the primary variable (NIH-CPSI total score) and the mean daily pain score. The response scenarios used for simulations were developed in collaboration with the internal medical team who provided the expertise for the treatment estimates. Literature reviews were performed to obtain estimates of variability and placebo response in this patient population [Nickel *et al.* (2004), Nickel *et al.* (2008), Wagenlehner *et al.* (2009)]. The margin for the clinical relevant effect to investigate was determined in collaboration with the medical team.

Table 1 provides an overview of the scenarios used in the simulations. For each treatment arm, the mean change from baseline in total score at 12 weeks compared to placebo is presented. The top value of each scenario is the mean change from baseline in the NIH-CPSI total score and the bottom value is the mean change from baseline in the mean daily pain score. Differences from placebo are presented to allow the reader to visualise which scenarios assumed superiority to placebo. The simulation scenarios could have included scenarios investigating the actual effect in each group. However, since the primary interest is the superiority compared to placebo the scenarios presented in Table 1 are more informative to the reader to view which scenarios assumed superiority to placebo and which do not. Moreover, based on the large variance (low precision) assumed, the simulated effects are allowed to vary considerably. As a result, the sampled values for each treatment (including placebo) could be larger than the values presented in this table, allowing for a reliable estimation of the mean change from baseline compared to placebo, which was the main focus of the simulation.

[Placeholder Table 1]

Under Scenario 1 in which all doses are equal to placebo, a result indicating final success (defined as the most likely maximum dose having a probability of superiority against placebo of at least 95% has a probability of 0.049 (see Table 2, row 1). This corresponds to a type I error rate of less than 5% in a conventional (frequentist) study. The study would be stopped prematurely for futility with a probability of 0.91 (see Table 2, row 1). Under Scenario 1a in which all doses are equal to placebo with respect to the NIH-CPSI total score and where there is a weak effect of the key secondary pain parameter, a result indicating final success has a probability of 0.047 (see Table 2, row 2). In both Scenarios (1 and 1a) the total number of subjects needed would be approximately 218, which is (on average) about 100 or 200 subjects less than the 330 or 432 subjects needed in a conventional study, see also comparative sample size section.

It was also investigated whether the design was able to detect the clinically relevant difference if the drug is effective. Under Scenario 5 where there is an increasing effect up to the clinically significant difference of -4, a result indicating final success has a probability of 0.961 (see Table 2, row 6). This corresponds to a type II error rate of about 4% in a conventional study.

Table 2 provides statistics of the 10000 simulations under each scenario. Mean and standard deviation (SD) of the number of subjects as well as the probability of final success (the most likely maximum effective dose has a probability of superiority against placebo of at least 0.95) and the probabilities of stopping for success, futility, or reaching the maximum number of subjects of 350 (MaxN) are provided.

[Placeholder Table 2]

Under Scenarios 2, 3 and 4, where moderate, but not clinically significant effects were simulated, a result indicating final success had a probability of 0.25, 0.58, and 0.85, respectively. Since this probability is lower than 95%, a trial with this response would be a failure. In scenarios under which at least one dose has a clinically significant effect the probability of a final success result is at least 0.94. This corresponds to a power of more than 90% in a conventional analysis. The probability of stopping for futility is at most 0.050.

It should be noted (see also Table 1) that based on the assumed prior for the variance, the sampled values for the treatment effect for each treatment arm are allowed to vary considerably. Moreover, the simulation scenarios were very flexible, allowing for monotic increasing, decreasing as well as fluctuating scenarios. Therefore, the simulation results were considered reliable.

Comparative sample size

With these assumptions (difference of 4 and SD of 7) a traditional parallel group trial having 80% power would require a sample size of 49 subjects per group. Assuming a commonly assumed dropout rate of 10%, the total number of subjects would be approximately 330. For a mere 20 extra subjects in the proposed adaptive design, the accruing data would be used more efficiently and the results will be more informative since the allocation of patients would be increased in the doses that are effective. As a result, the power in the effective dose arms is increased. Moreover, it is very common to use a power of 90% to maintain a low type II error rate and minimise company risk. This would result in a total of 432 patients equally divided among the dose groups.

4.2 Study results

The study stopped early for futility after 239 of the 350 patients (68%) were randomised. Two hundred and twenty six (226) patients had measurable data for the primary variable.

The probability that the most likely effective dose had a difference from placebo of at least the futility margin (FD) was 2%, which was significantly lower than the predetermined threshold (20%). It should be noted that, based on the pre-specified decision criteria, the study stopped soon after the rules for stopping were activated. In Table 3 the main results of the Bayesian analyses are presented, including the posterior probability of a dose being superior to placebo and the posterior probability that a dose is better than placebo by at least the futility margin (FD).

[Placeholder Table 3]

Table 4 presents the results of the Bayesian analyses for the NIH-CPSI pain sub score, which was used to update the allocation probabilities only. These full-analyses set (FAS) results confirm that the posterior probability of any of the doses were superior to placebo was low. This probability was at most 0.175.

[Placeholder Table 4].

A summary of treatment emergent adverse events are presented in Table 5. The treatment emergent adverse events were equally divided among the treatment group. There were no deaths and the number (%) of patients with an adverse event leading to discontinuation ranged from 0 (0%) in the placebo group to 2 (7.1%) in the 150mg twice daily group.

[Placeholder Table 5]

A sensitivity analysis using a per protocol set (PPS), which included all patients in the FAS who did not have any deviations to protocol that affected the NIH-CPSI total score, confirmed the results based on the FAS.

5 Conclusion

Extensive pre-planning was necessary for this combined PoC/DF study, from a medical, statistical and clinical/logistical perspective. The simulations showed that this Bayesian response adaptive design could accomplish both the goals of PoC and DF. The result of the simulations also showed that the probability of falsely claiming efficacy was less than 5%, in line with the magnitude of the conventional false positive rate. The simulations further indicated a strong probability of demonstrating efficacy and finding the most effective experimental dose if the drug was to be effective.

The study stopped early for futility in line with the simulation predictions for stopping. More than 32% of patients were not unnecessarily exposed to treatment that was not effective for their condition.

5 Discussion

The combined PoC and DF design was found to be adequate and sufficient to assess whether or not the drug is effective, and allows for the investigation of the most likely effective dose(s). In line with recommendations (Fisch et al. 2015) clear Go/No-Go criteria were defined up front. We used non-informative priors for the treatment effect and variance components. The choice for non-informative priors was reasonable because this was a novel compound in a new indication with hardly any publicly available data. An inverse gamma distribution was assumed for the prior variance of the treatment effect and the longitudinal model. Gelman (2006) argued that an inverse gamma prior could produce an improper posterior density if the parameters of the gamma distribution are close to 0 and proposed alternatives, such as uniform prior. However, there are also issues associated with the uniform prior and care should be exercised when choosing a prior distribution. In this study, the Bayesian posterior mean and variance were similar to those from the Frequentist analyses (Wagenlehner et al., 2017), indicating the chosen prior distributions were reasonable.

Riesenberg et al. (2012) published a combined PoC and DF study in depression. Sagkriotis and Scholp (2008) published clinical trial simulations in combined PoC and DF in migraine, combining Bayesian and frequentist methods in an adaptive setting. The authors showed that substantial savings in terms of patient exposure could be achieved. In a study described by Conner et al. (2013), a Bayesian adaptive trial design using adaptive randomization in a multi-arm trial was employed. It was found that when one treatment is superior, the trial design provides higher power and a lower expected sample size.

The study presented in this manuscript was unique in the combination of various design elements that enabled exposure of a new drug to the lowest number of patients while simultaneously testing a large dose range. In particular, the application of Bayesian principles to adaptive randomization that included both once daily and twice daily doses, predefined stopping rules, and additional limitations for the early stopping rules, e.g. no stopping until a minimum of 50 patients in best dose group and placebo were recruited and no stopping until at least 50% of maximum total number of patients enrolled. The additional constraints for stopping ensured that there was enough patients to adequately and robustly assess PoC and provide sufficient patients in the remaining arm for a robust assessment of other potential doses.

Other novel aspects in the design include the penalization of the efficacy result of patient dropping out for reason of adverse events and the definition of a futility difference.

All these measures resulted in the ability to take a clear-cut Go/No-Go decision. If the study was to be successful, no further dose finding study would be necessary and the most effective and safe dose arms could be taken to Phase III development (registration studies). A wide dose range was included in the study to adequately estimate the dose response curve and to circumvent the limitations described by Berry et al. (2010) regarding the trial inefficiencies caused by using a narrow or inappropriate dose range in a dose finding study. Using the wide dose range in this Bayesian response adaptive design ensured that the total number of patients needed for the combined PoC/DF study was about the same as a classically designed PoC study would have needed.

As previously discussed, generally, either a once daily dose regimen or a twice daily dose regimen is studied. The incorporation of a once daily dose together with a two daily dose range in this study is a novel approach to dose finding, as is the incorporation of a lower bound of the clinical effect that is deemed clinically non-significant. It can be argued that the stopping rules were conservative in that the study could have stopped earlier if for example, the stopping criteria could be activated earlier (prior to half of the patients being randomised). However, this approach ensured that a sufficient number of patients participated to make conclusive decisions regarding Go/No-Go. Together with the ongoing review of safety data by the independent DSMB, the safety of the patients was also duly investigated. Since the design and execution of this study other adaptive allocation approaches have been proposed, such as the mass weighted urn design described by Zhao (2015). Zhang and Rosenberger (2015) describe response adaptive randomization procedures from both frequentist and Bayesian perspectives and provide guidance for consideration when implementing an adaptive trial. He et al. (2015) also provides general guidance on practical applications for adaptive trial design and implementation. Adaptive trials should be considered whenever possible to optimize clinical trial design and reduce the unnecessary exposure of patients, taking into account the practical guidance and considerations for implementing such trials.

Acknowledgements The authors are grateful for the contributions by Scott and Don Berry of Berry Consultants LLC, Texas, USA who were involved in the clinical trial simulations and regulatory consultation process. Scott Berry (statistician) and Tessella, UK (interim analyses and interim and final reporting of Bayesian analyses) were the IDMC. Biostatistics, Medical Science and Clinical teams at Astellas Development responsible for the operational preparations and execution of the trial (protocol 3652-CL-0019 with Clinical trial.gov identifier: NCT01391338).

Conflict of Interest

The authors Jos Houbiers, Joost Melis and Olivier van Till are employees of Astellas. Reynaldo Martina was an employee of Astellas at the time of the design and execution of the trial and has since served as a consultant for Roche, Takeda and Dompe.

References

Berry, D. A. and Eick, S. G. (1995) Adaptive assignment versus balanced randomization in clinical trials: a decision analysis *Statistics in Medicine* **14**, 231-246.

Berry, D. Muller, P. Grieve, A. Smith, M. Park, T. Blazek, R. Mitchard, N. and Krams, M. (2002). Adaptive Bayesian designs for dose-ranging trials. *Case studies in Bayesian statistics (V. Springer)*, 99-181.

Berry, S. M. Carling, B. P. Lee, J. J. and Muller, P. (2011). Bayesian Adaptive Methods for Clinical Trials (Chapman & Hal/CRC Biostatistics Series).

Berry, S. M. Spinelli, W. Littman, G. S. Liang, J. Z. Fardipour, P. Berry, D. A. Lewis, R. J. and Krams, M. (2010). A Bayesian dose-finding trial with adaptive dose expansion to flexibly assess efficacy and safety of an investigational drug. *Clinical Trials* **7**(2), 121–135.

Chang, M. (2008). Adaptive Design Theory and Implementation Using SAS and R (Chapman & Hall/CRC).

Connor, J. T. Elm, J. J. Broglio, K. R. (2013). Bayesian adaptive trials offer advantages in comparative effectiveness trials: an example in status epilepticus. *Journal of Clinical Epidemiology* **66**, S130-S137.

Fisch, R. Jones, I. Jones, J. Kerman, J. Rosenkranz, G. K. and Schmidli, H. (2014). Bayesian design of proof-of-concept trials. *Therapeutic innovation & regulatory science* **49**(1), 155–162.

Gelman, A. Carlin, J. B. Stern, H. S. Rubin, D. B. (2004). Bayesian Data Analysis, second edition (Chapman & Hall/CRC).

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. Bayesian Analysis **1** (3), 515-533.

Grieve, A. P. and Krams, M. (2005). ASTIN: a Bayesian adaptive dose-response trial in acute stroke. *Clinical Trials* **2**, 340-351.

He, W. Pinheiro, J. Kuznetsova, O. M. (eds). Practical considerations for adaptive trial design and implementation. (Statistics for Biology and Health. Springer New York, NY).

Ivanova, A. Bolognese, J. A. and Perevozskaya, I. (2008). Adaptive dose finding based on *t*-statistic for dose-response trials. *Statistics in Medicine* **27** (10), 1581-1592.

Krams, M. Lees, K. R. Hacke, W. Grieve, A. P. Orgogozo, J-M. and Ford, G. A. (2003). Acute Stroke Therapy by Inhibition of Neutrophils (ASTIN): An Adaptive Dose-Response Study of UK-279,276 in Acute Ischemic Stroke. *Stroke* **34**, 2543–2548.

Litwin, M. S. McNaughton-Collins, M. Fowler Jr, F. J. Nickel, J. C. Calhoun, E. A. Pontari, M. A. Alexander, R. B. Farrar, J. T. and O'Leary, M. P. (1999). The National Institutes of Health Chronic Prostatitis Symptom Index: development and validation of a new outcome measure. Chronic Prostatitis Collaborative Research Network. *Journal of Urology* **162**, 369-375.

Nickel, J. C. Krieger, J. N. McNaughton-Collins, M. Anderson, R. U. Pontari, M. Shoskes, D. A. Litwin, M. S. Alexander, R. B. White, P. C. Berger, R. Nadler, R. O'Leary, M. Liong, M. L. Zeitlin, S. Chuai, S. Landis, J. R. Kusek, J. W. Nyberg, L. M. and Schaeffer, A. J. (2008). *New England Journal of Medicine* **359**, 2663-2673.

Nickel, J. C. Narayan, P. McKay, J. and Doyle, C. (2004). Treatment of chronic prostatitis/chronic pelvic pain syndrome with tamsulosin: a randomized double blind trial. *Journal of Urology* **171**, 1594-1597.

Riesenberg, R. Rosenthal, J. Moldauer, L. and Peterson, C. (2012). Results of proof-of-concept, dose finding, double-blind, placebo-controlled study of RX-10100 (Serdaxin) in subjectgs with major depressive disorder *Psychopharmacology* **221** (4), 601-610.

Sagkriotis, A. and Scholpp, J. (2008). Combining proof-of-concept with dose-finding: utilization of adaptive designs in migraine clinical trials. *Cephalalgia* **28** (8), 805–812.

Wagenlehner, F. M. Schneider, H. Ludwig, M. Schnitker, J. Brarhler, E. and Weidner, W. (2009). A pollen extract (Cernilton) in patients with inflammatory chronic prostatitis-chronic pelvic pain syndrome: a multicenter, randomized, prospective, double-blind, placebo-controlled phase 3 study. *European Urology* **56**, 544-551.

Wagenlehner, F. M. E. van Till, J. W. O. Houbiers, J. G. A. Martina, R. V. Cerneus, D. P. Melis, J. H. J. M. Majek, A. Vjaters, E. Urban, M. Ramonas, H. Shoskes, D. A. and Nickel, C. J. (2017). Fatty acid amide hydrolase inhibitor treatment in men with chronic prostatitis/chronic pelvic pain syndrome: an adaptive double-blind randomized controlled trial. Urology **103**, 191–197.

Whitehead, J. and Zhou, Y. (2001). Easy-to-implement Bayesian methods for dose escalation studies in healthy volunteers. *Biostatistics* **2** (1), 47-61.

Xie, F. Ji, Y. Tremmel, L. (2012). A Bayesian adaptive design for multi-dose, randomized, placebo-controlled phase I/II trials. *Contemporary Clinical Trials* **33** (4), 739-748.

Zhang, L. and Rosenberger, W. F. (2014). Response-adaptive randomization for clinical trials. In: He W. Pinheiro J. Kuznetsova O. (eds). Practical considerations for adaptive trial design and implementation. (Statistics for Biology and Health. Springer New York, NY).

Zhao, W. (2015). Mass weighted urn design- a new randomization algorithm for unequal allocations. *Contemporary Clinical Trials* **43**, 209-216.

Table 1 Simulation scenarios for selected cases. The top value of each scenario the is mean change from baseline in the NIH-CPSI total score and the bottom value is the mean change from baseline in the mean daily pain score

Scenario	Placebo	25mg (bid)	75mg (bid)	300mg (qd)	150mg (bid)	300mg (bid)
1	0	0	0	0	0	0
	0	0	0	0	0	0
1a	0	0	0	0	0	0
	0	-0.5	-1	-1.5	-1.5	-2
2	0	-0.5	-1	-1	-1	-1
	0	-0.3	-0.6	-0.6	-0.6	-0.6
3	0	-0.5	-1.5	-2	-2	-2
	0	-0.3	-0.9	-1.2	-1.2	-1.2
4	0	-1	-2	-2.5	-3	-3
	0	-0.6	-1.2	-1.5	-1.8	-1.8
5	0	-1	-2	-3	-4	-4
	0	-0.6	-1.2	-1.8	-2.4	-2.4
5a	0	-1	-2	-3	-4	-4
	0	-2.4	-2.4	-1.8	-1.2	-0.6
5b	0	-1	-2	-3	-4	-4
	0	0	0	0	0	0
6	0	-1	-2	-2	-2.5	-5
	0	-0.6	-1.2	-1.2	-1.5	-3
7	0	-3.5	-4	-4.5	-4.5	-5
	0	-2.1	-2.4	-2.7	-2.7	-3
8	0	-5	-5	-5	-5	-5
	0	-3	-3	-3	-3	--3
9	0	-3.5	-4.5	-5	-6	-6.5
	0	-1.2	-2.7	-3	-3.6	-3.9
10	0	-3.5	-4.5	-7	-5	-6
	0	-1.2	-2.7	-4.2	-3	-3.6
10a	0	-3.5	-4.5	-7	-5	-6
	0	-3	-3	-3	-3	-3
10b	0	-3.5	-4.5	-7	-5	-6
	0	-1	-2	-3	-4	-5
11	0	-0.5	-1	-6	-1.5	-1.5
	0	-0.3	-0.6	-3.6	-0.9	-0.9
12	0	-2	-4	-2	-6	-8
	0	-1.2	-2.4	-1.2	-3.6	-4.8
13	0	-2	-5	-5	-7	-5
	0	-1.2	-3	-3	-4.2	-3
14	0	-6	-5	-4	-3	-3
	0	-3.6	-3	-2.4	-1.8	-1.3

Table 2 Simulation results for selected scenarios.

Scenario	**Number of subjects**		**Probability**			
	Mean	SD	Final Success	Stop for Success	Stop for MaxN	Stop for futility
1	217.9	53.9	**0.049**	0.000	0.087	**0.914**
1a	217.5	54.5	0.047	0.000	0.085	0.915
2	259.9	73.2	0.245	0.001	0.322	0.678
3	303.9	68.6	0.583	0.003	0.645	0.352
4	330.1	49.9	0.846	0.021	0.845	0.134
5	334.1	42.3	**0.961**	0.119	0.843	0.038
5a	335.7	39.8	0.954	0.105	0.855	0.041
5b	336.3	39.0	0.941	0.091	0.860	0.050
6	315.7	56.9	0.983	0.331	0.650	0.019
7	312.9	51.4	0.997	0.447	0.550	0.003
8	304.0	48.1	0.998	0.636	0.364	0.0002
9	254.3	53.1	0.998	0.892	0.108	0.0003
10	247.2	45.6	1.000	0.937	0.063	0.000
10a	268.8	46.0	0.999	0.905	0.095	0.000
10b	261.3	53.0	1.000	0.897	0.103	0.000
11	275.1	64.6	0.997	0.716	0.280	0.004
12	202.0	25.7	1.000	0.997	0.003	0.000
13	282.1	48.2	0.999	0.825	0.175	0.0001
14	260.9	60.0	0.999	0.805	0.195	0.001

Table 3 Change from baseline in NIH-CPSI total score (FAS)

Analysis Set = FAS	Placebo (N=53)	25 mg bid (N=52)	75 mg bid (N=26)	300 mg qd (N=34)	150 mg bid (N=27)	300 mg bid (N=34)
N Baseline Mean (SE) EoT Mean (SE)	53 24.2 (0.71) 17.0 (1.05)	51 23.4 (0.73) 16.1 (1.07)	26 22.3 (0.98) 14.9 (1.46)	31 23.5 (0.78) 16.6 (1.12)	25 21.2 (0.71) 13.8 (1.37)	32 22.4 (0.90) 15.6 (1.30)
Change from Baseline Mean (SE)	-7.2 (0.94)	-7.3 (0.95)	-6.7 (1.22)	-6.7 (1.19)	-7.4 (1.38)	-6.8 (1.18)
Posterior θ_d (Std) 95% Credibility Interval	-7.3 (0.97) (-9.2,-5.4)	-7.0 (0.68) (-8.3,-5.7)	-6.9 (0.69) (-8.2,-5.5)	-6.5 (0.78) (-8.1,-5.0)	-6.8 (0.60) (-8.0,-5.6)	-6.7 (0.70) (-8.1,-5.3)
Posterior difference vs. placebo (Std) 95% Credibility Interval	--	0.3 (1.18) (-1.9,2.6)	0.4 (1.20) (-1.9,2.8)	0.7 (1.26) (-1.7,3.1)	0.5 (1.16) (-1.7,2.7)	0.6 (1.23) (-1.8,2.9)
Posterior probability (max effective dose)	--	0.328	0.189	0.157	0.129	0.196
Posterior probability (better than placebo)	--	0.399	0.363	0.280	0.335	0.316
Posterior prob (better by FD)	--	0.019	0.020	0.015	0.014	0.017
Posterior prob (better by CSD)	--	<0.001	<0.001	0.002	0.001	0.001

Std: Standard deviation; FD: Futility difference from placebo (-2); CSD: Clinically significant difference (-4)

θ_d: modeled change from baseline

Table 4 Change from baseline in NIH-CPSI pain domain score (FAS).

Analysis Set = FAS	Placebo (N=53)	25 mg bid (N=52)	75 mg bid (N=26)	300 mg qd (N=34)	150 mg bid (N=27)	300 mg bid (N=34)
N	53	51	26	31	25	32
Baseline Mean (SE)	12.4 (0.38)	12.5 (0.38)	11.8 (0.54)	12.0 (0.42)	11.1 (0.45)	11.5 (0.43)
EoT Mean (SE)	8.1 (0.51)	8.5 (0.60)	8.3 (0.84)	8.5 (0.64)	7.0 (0.72)	7.7 (0.67)
Change from Baseline Mean (SE)	-4.3 (0.51)	-4.0 (0.59)	-3.1 (0.66)	-3.5 (0.69)	-4.1 (0.72)	-3.8 (0.61)
Posterior θ_d (Std)	-4.4 (0.51)	-3.8 (0.45)	-3.5 (0.48)	-3.3 (0.53)	-3.7 (0.46)	-3.6 (0.51)
95% Credibility Interval	(-5.4,-3.4)	(-4.6,-2.9)	(-4.5,-2.6)	(-4.3,-2.3)	(-4.6,-2.8)	(-4.6,-2.6)
Posterior difference vs. placebo (Std)	--	0.6 (0.68)	0.9 (0.69)	1.1 (0.74)	0.8 (0.68)	0.8 (0.73)
95% Credibility Interval		(-0.7,2.0)	(-0.5,2.2)	(-0.4,2.6)	(-0.6,2.1)	(-0.6,2.2)
Posterior probability (max effective dose)	--	0.367	0.127	0.080	0.184	0.242
Posterior probability (better than placebo)	--	0.175	0.106	0.067	0.134	0.139

Std: Standard deviation

Table 5 Overview of Treatment Emergent Adverse Events (TEAEs).

	Placebo (N=56)	25 mg bid (N=53)	75 mg bid (N=28)	300 mg qd (N=35)	150 mg bid (N=28)	300 mg bid (N=38)
Incidence of TEAEs	23 (41.1%)	17 (32.1%)	12 (42.9%)	14 (40.0%)	15 (53.6%)	15 (39.5%)
Incidence of Drug-related TEAEs	10 (17.9%)	9 (17.0%)	3 (10.7%)	6 (17.1%)	5 (17.9%)	7 (18.4%)
Incidence of Deaths	0	0	0	0	0	0
Incidence of Serious TEAEs	1 (1.8%)	1 (1.9%)	2 (7.1%)	0	1 (3.6%)	0
Incidence of Serious Drug-Related TEAEs	0	0	0	0	0	0
Incidence of TEAEs Leading to Permanent Discontinuation of Study Drug	0	1 (1.9%)	2 (7.1%)	2 (5.7%)	2 (7.1%)	2 (5.3%)
Incidence of Drug-Related TEAEs Leading to Permanent Discontinuation of Study Drug	0	1 (1.9%)	1 (3.6%)	2 (5.7%)	1 (3.6%)	2 (5.3%)