PAREMA1 Statistical Analysis Plan

1.1 Change log

Version	Changes
1.0	NA
1.1	The following changes were implemented prior to the interim analysis unblinding and readout: • Study power increased from 80% to greater than 85%. • Minimum enrolment number increased from 174 to 200. Rationale for this change is to increase study power. • Maximum enrolment number increased from 220 to 240. Rationale for this change is to increase study power and decrease risk of not observing a true difference in the primary outcomes. • Figure 1 has been updated with the changes described above. • Addition of an additional exploratory analysis data set that includes the training attack. Rationale for this change is to investigate the effect of training. • The number of attacks per patient in the full data set will be estimated via a simulation using as its input the attack frequency observed in the interim data set (see Appendix for details). • The dropout percentage in the full data set will be estimated via a simulation using as its input the risk of dropout in a given week during the interim period (see Appendix for details). • It has been specified that trial site staff will not be informed of the final enrolment target found via the interim analysis. Rationale for this change is to reduce the risk of partial unblinding that could otherwise occur.

1.2 Study Design

PAREMA1 is a randomized, controlled, double-blinded, parallel-group, group-sequential clinical study randomizing participants to either active or sham treatment, in a 1:1 ratio.

The clinical investigation will randomize at minimum 200 participants and at maximum 240.

In Stage 1 of the trial the participants will treat up to four attacks each. The first treated attack is a training attack and will not be used in the statistical analysis. All attacks reported after the training attacks are study attacks and will be used in the analysis. The reason for discarding the first attack is that it (cf. the pilot trial and post-market data) is not representative of normal use and effectiveness, the user having not yet become accustomed to correct use of the device.

1.3 Endpoints

The primary endpoint is absence of moderate or severe pain at 2 hours (AMSP2) which is an outcome that is valuable to patients and which has a good benefit risk/ratio when factoring in the low risk level of the treatment. The statistical power for AMSP2 in the study is above 95% (see section 1.5, below).

The study's most important secondary endpoint is Pain Freedom at 2 hours (PF2). Recognizing that this is the ideal outcome for patients, the study has been sized and designed with the aim of having a high statistical power for PF2.

1.4 Interim analysis procedure

After 60 subjects have provided data from at least one study attack in Stage 1, an interim analysis (IA) will be conducted. The IA will include all reported study attacks according to the ITT principle (see section 1.6 below).

Based on the data from the IA it will be decided how many additional patients to recruit into the study, and in turn how many patients will be included in total (N_t). This decision will be taken according to the decision procedure flowchart shown in Figure 1:

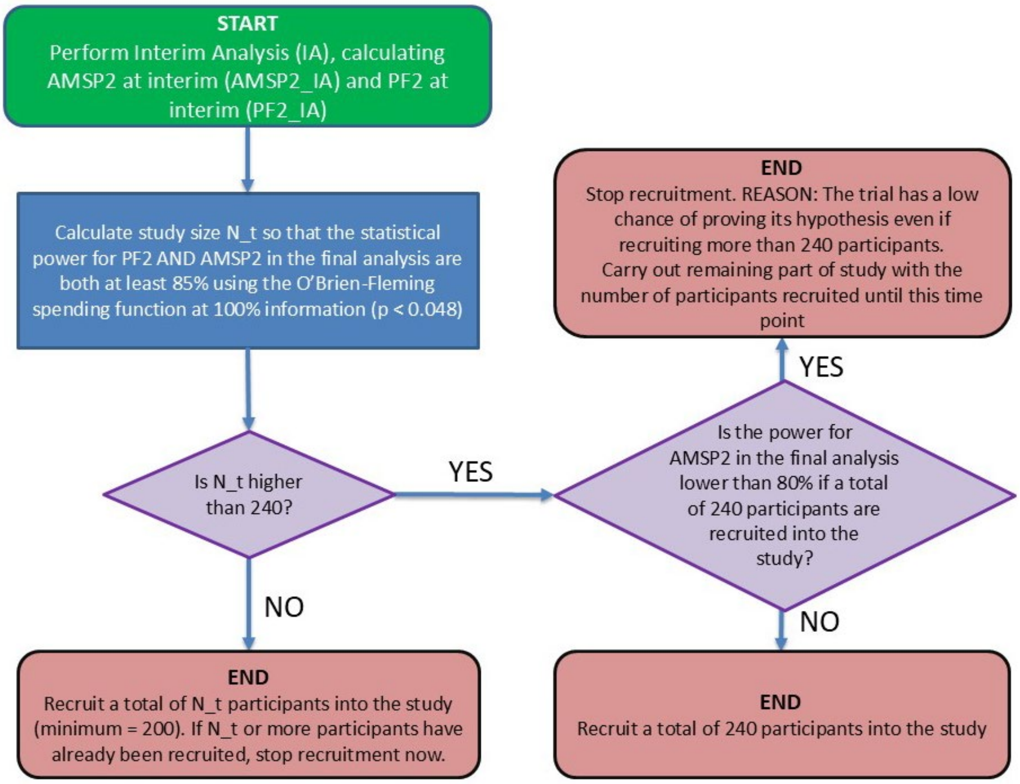
START
Perform Interim Analysis (IA), calculating AMSP2 at interim (AMSP2_IA) and PF2 at interim (PF2_IA)
Calculate study size N_t so that the statistical power for PF2 AND AMSP2 in the final analysis are both at least 85% using the O'Brien-Fleming spending function at 100% information (p < 0.048)
Is N_t higher than 240?
YES
NO
END
Recruit a total of N_t participants into the study (minimum = 200). If N_t or more participants have already been recruited, stop recruitment now.
Is the power for AMSP2 in the final analysis lower than 80% if a total of 240 participants are recruited into the study?
YES
END
Stop recruitment. REASON: The trial has a low chance of proving its hypothesis even if recruiting more than 240 participants.
Carry out remaining part of study with the number of participants recruited until this time point
NO
END
Recruit a total of 240 participants into the study

Figure 1: Interim Analysis decision procedure

As shown in Figure 1, the interim analysis performs hypothesis tests for AMSP2 and PF2, using the O'Brien Fleming Spending function in which the significance thresholds of the p-value are respectively 0.0052 at the interim analysis and 0.048 in the full analysis.

The interim analysis will use the following assumptions:

- The observed AMSP2 and PF2 at the interim for both treatment arms
- The observed Intraclass Correlation based on the interim data set
- The number of attacks per patient in the full data set will be estimated via a simulation using as its input the attack frequency observed in in the interim data set (see Appendix for details).
- The dropout percentage in the full data set will be estimated via a simulation using as its input the risk of dropout in a given week during the interim period (see Appendix I for details).

As shown, the total number of included participants will be capped at 240 since this is the largest study that will be feasible to conduct.

The interim analysis does not include a criterion for early stopping due to effectiveness.

Staff involved in the conduct of the trial and closing enrolment will remain blind to the exact result of the interim analysis, but will be informed of the final enrolment target (N_t). Study site staff will remain blind to the exact result of the interim analysis and final enrolment target (N_t).

1.5 Power and sample size calculation

1.5.1 Assumptions

The estimated study size of 200 to 240 included participants has been determined on the basis of data from the pilot study (Fuglsang et al., 2018) and post-market data collected in Denmark, Sweden and Germany from 2018 to 2020.

To estimate the sample size for this trial, it was assumed that 12% of subjects in the sham group and 30% of subjects in the active group would experience PF2. For AMSP2, a 26% response rate in the sham group and a 61% response rate in the active group was assumed.

The primary statistical test for this analysis will be a χ^2 significance test adjusted for clustered data (using the methods of Donner and Banting (Donner and Banting, 1988)) since subjects could contribute more than one migraine attack. The intraclass correlation coefficient (ICC) from preliminary data on this device was 0.486. This is a group-sequential design with one interim analysis, so the O'Brien Fleming Spending function will be used to set significant thresholds of 0.0052 at the interim analysis and 0.048 at the final analysis.

Power and sample sizes were estimated via simulation using the above parameters for a χ2 significance test adjusted for clustered data. These simulations assumed a third of subjects would have one migraine attack during the course of the study, a third of subjects would have two study attacks, and a third of subjects would have three study attacks. In the interim analysis, a more refined method of estimating number of attacks per participant will be employed, using the observed attack frequencies from the interim data (cf. section 1.4). Additionally, it was expected that 22% of subjects would drop out of the trial and 10% of attacks would have missing data at the 2 hour time point, for a total of 32% of randomized subjects that would not provide data usable for the statistical analysis of the primary or secondary endpoint.

1.5.2 Results

The results of the simulations described above are presented in Figure 2 and Table 1 below:

Total N (dropout inflated)	Total N (excluding dropouts)	PF2 Power	AMSP2 Power
100	68	54.3%	85.2%
110	76	60.2%	90.0%
120	82	64.1%	91.8%
130	90	67.1%	93.8%
140	96	71.4%	95.6%
150	102	73.7%	96.5%
160	110	76.7%	97.1%
170	116	79.0%	97.8%
180	124	82.2%	98.4%
190	130	83.9%	99.0%
200	136	85.5%	99.1%

Table 1: Power to detect differences in AMSP2 and PF2 at a given N, assuming 12% sham PF2 response, 30% active PF2 response, 26% sham AMSP2 response, 61% active AMSP2 response, 32% total dropout, and 1-3 attacks per subject

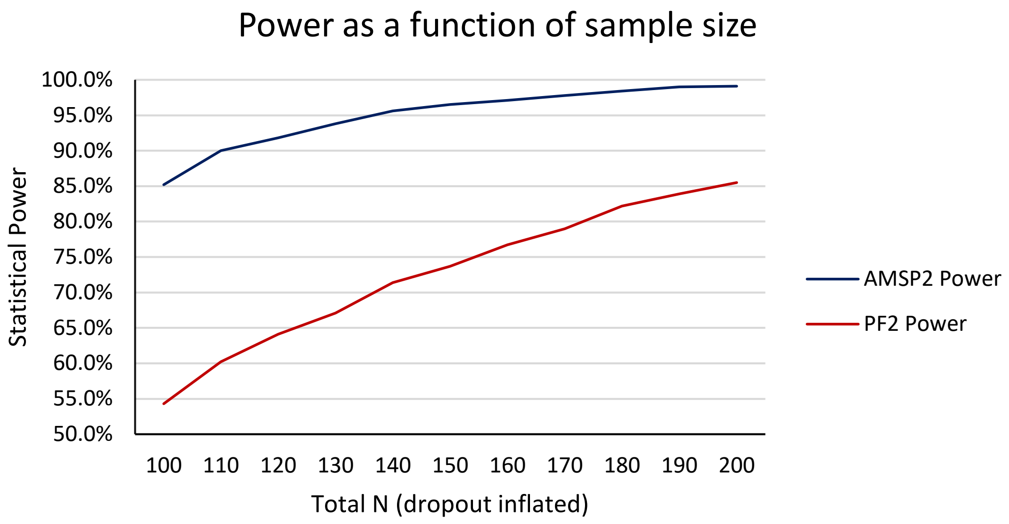

Figure 1: Power to detect differences in AMSP2 and PF2 as a function of sample size, assuming 12% sham PF2 response, 30% active PF2 response, 26% sham AMSP2 response, 61% active AMSP2 response, 32% total dropout, and 1-3 attacks per subject

By interpolation of the data in Table 2, it was found that an estimated 200 subjects would be required to achieve over 85% power to detect a difference in PF2 between active and sham groups. This would correspond to greater than 99% power to detect a difference in AMSP2 between groups.

1.6 Analysis sets

1.6.1 Intention-to-treat (ITT) analysis set

The ITT includes all study attacks from all participants who underwent randomization and who reported at least one study attack. The ITT analysis set will not exclude attacks with protocol deviations or use of rescue medication before the two-hour time point. The ITT data set will be used for analysis of the primary and secondary effectiveness end points.

If any subjects in the ITT population have a missing headache score at the two-hour time point, multiple imputation may be used to impute their headache score for the analysis of primary and secondary endpoints, as described in section 1.9.

1.6.2 As-treated analysis set

The as-treated analysis set includes all participants who received a study device, whether or not it was the actual randomized device. Subjects will be grouped by the device they actually used. This analysis set will be used for assessment of safety.

1.6.3 Per-protocol analysis set

The per-protocol analysis set includes all study attacks with a reported headache score at the two-hour time point, in which:

- The participant started using the study device within five minutes of the beginning of aura and before development of moderate or severe pain
- The participant did not use rescue drugs before the two-hour assessment point
- There were no changes in concomitant preventive treatment during the study that might potentially affect response.
- There were no other major protocol deviations

When comparing the ITT and Per Protocol analysis sets, the impact of protocol deviations and user errors can be investigated with the aim of identifying potential improvements to the investigational device and instructions for use. The ITT analysis set will also be used to analyse the impact of protocol deviations on the outcome parameters (e.g. use of rescue medication during the first two hours and starting the use of the device after the onset of headache).

1.6.4 Exploratory analysis set

The exploratory analysis data set will include both study attacks and training attacks, in order to investigate the effect of training. The exploratory data set will in all other respects be identical to the ITT analysis set, e.g. it will not exclude attacks with protocol deviations or use of rescue medication before the two-hour time point.

1.7 Outcomes and Analysis

1.7.1 Primary hypothesis

In line with earlier studies on aura-phase treatment (Bates D. et al., 1994; Olesen J. et al., 2004), the trial's **primary end point** will be absence of moderate or severe pain at 2 hours (AMSP2),), i.e. the percentage of study attacks in which the participant reported absence of headache of moderate/severe intensity at 2 hours post-treatment initiation. "Study attacks" are all attacks treated and reported during Stage 1, apart from the first attack which is a training attack and is not included in the analysis.

The intention-to-treat (ITT) analysis set will be used for the primary effectiveness hypothesis test.

The primary effectiveness hypothesis is:

H_0: $p_{Rehaler} = p_{sham}$

H_a: $p_{Rehaler} \neq p_{sham}$

Where:

$p_{Rehaler}$ = proportion of attacks in the Rehaler group with AMSP2

p_{sham} = proportion of attacks in the sham group with AMSP2

The primary effectiveness hypothesis will be tested using a two-sided adjusted χ^2 test for clustered binary data (Donner and Banting, 1988) at a 5% level of significance.

If Pearson's chi-squared test statistic is written as $\chi^2 = \sum_{i=1}^{G} \chi_i^2$ with G = the number of treatment groups, then the adjusted χ^2 approximately follows a chi-squared distribution with G-1 degrees of freedom and is calculated as follows (Donner and Banting, 1988):

$$\chi_A^2 = \sum_{i=1}^{G} \frac{\chi_i^2}{\overline{C_\iota}}$$

Where:

- $\overline{C_\iota} = \sum_{j=1}^{n_i} \frac{m_{ij} C_{ij}}{\sum m_{ij}}$
- $C_{ij} = 1 + (m_{ij} - 1)\rho$
- G is the number of treatment groups
- n_i is the number of individuals in group i
- ρ is the correlation coefficient between any two responses in the same individual
- m_{ij} denotes the number of observations for individual j in group i

The corresponding 95% confidence interval for the difference in proportion of attacks with AMSP2 between the Rehaler and sham groups will also be calculated as follows (Donner and Klar, 1993):

$$\text{95\% Confidence Interval: } (\hat{p}_{Rehaler} - \hat{p}_{sham}) \pm 1.96 * \widehat{SE}(\hat{p}_{Rehaler} - \hat{p}_{sham})$$

Where:

- $\hat{p}_{Rehaler}$ = observed proportion of attacks in the Rehaler group with AMSP2
- $\hat{p}_{sham}$ = observed proportion of attacks in the sham group with AMSP2
- $\widehat{SE}(\hat{p}_{Rehaler} - \hat{p}_{sham}) = \left(\frac{\overline{C_1}\hat{p}_{Rehaler}(1-\hat{p}_{Rehaler})}{n_1} + \frac{\overline{C_2}\hat{p}_{sham}(1-\hat{p}_{sham})}{n_2}\right)^{1/2}$
- $\overline{C_\iota} = \sum_{j=1}^{n_i} \frac{m_{ij} C_{ij}}{\sum m_{ij}}$
- $C_{ij} = 1 + (m_{ij} - 1)\rho$
- n_i is the number of individuals in group i
- ρ is the correlation coefficient between any two responses in the same individual
- m_{ij} denotes the number of observations for individual j in group i

1.7.2 Additional hypotheses

The following additional hypotheses will be tested for potential labelling claims, according to the step-down hierarchy procedure described in section 1.7.3 below:

Endpoint		Hypotheses
Pain Freedom at 2 hours (PF2)	H_0:	Proportion of attacks in the Rehaler group with PF2 = Proportion of attacks in the sham group with PF2
	H_a:	Proportion of attacks in the Rehaler group with PF2 ≠ Proportion of attacks in the sham group with PF2
Freedom from Most Bothersome Symptom at 2 hours (MBSF2)	H_0:	Proportion of attacks in the Rehaler group with MBSF2 = Proportion of attacks in the sham group with MBSF2
	H_a:	Proportion of attacks in the Rehaler group with MBSF2 ≠ Proportion of attacks in the sham group with MBSF2
Sustained Pain Freedom at 24 hours (SPF24)	H_0:	Proportion of attacks in the Rehaler group with SPF24 = Proportion of attacks in the sham group with SPF24
	H_a:	Proportion of attacks in the Rehaler group with SPF24 ≠ Proportion of attacks in the sham group with SPF24
Headache Score at 2 hours (HS2)	H_0:	Mean HS2 attack score in the Rehaler group = Mean HS2 attack score in the sham group
	H_a:	Mean HS2 attack score in the Rehaler group ≠ Mean HS2 attack score in the sham group

Most Bothersome Symptom Score at 2 hours (MBS2)	H_0:	Mean MBS2 attack score in the Rehaler group = Mean MBS2 attack score in the sham group
	H_a:	Mean MBS2 attack score in the Rehaler group ≠ Mean MBS2 attack score in the sham group
Functional Disability Score at 2 hours (FDS2)	H_0:	Mean FDS2 attack score in the Rehaler group = Mean FDS2 attack score in the sham group
	H_a:	Mean FDS2 attack score in the Rehaler group ≠ Mean FDS2 attack score in the sham group
Use of rescue medication from the 2 hours time point until 24 hours (Res24)	H_0:	Proportion of attacks in the Rehaler group with Res24 = Proportion of attacks in the sham group with Res24
	H_a:	Proportion of attacks in the Rehaler group with Res24 ≠ Proportion of attacks in the sham group with Res24
Participant Satisfaction at 48 hours (PS48)	H_0:	Mean PS48 attack score in the Rehaler group = Mean PS48 attack score in the sham group
	H_a:	Mean PS48 attack score in the Rehaler group ≠ Mean PS48 attack score in the sham group
Light Sensitivity Score at 2 hours (LSS2)	H_0:	Mean LSS2 attack score in the Rehaler group = Mean LSS2 attack score in the sham group
	H_a:	Mean LSS2 attack score in the Rehaler group ≠ Mean LSS2 attack score in the sham group
Nausea Score at 2 hours (NS2)	H_0:	Mean NS2 attack score in the Rehaler group = Mean NS2 attack score in the sham group
	H_a:	Mean NS2 attack score in the Rehaler group ≠ Mean NS2 attack score in the sham group
Sound Sensitivity Score at 2 hours (SSS2)	H_0:	Mean SSS2 attack score in the Rehaler group = Mean SSS2 attack score in the sham group
	H_a:	Mean SSS2 attack score in the Rehaler group ≠ Mean SSS2 attack score in the sham group
Freedom from Relapse at 48 hours (FR48)	H_0:	Proportion of attacks in the Rehaler group with FR48 = Proportion of attacks in the sham group with FR48
	H_a:	Proportion of attacks in the Rehaler group with FR48 ≠ Proportion of attacks in the sham group with FR48

Table 2: Secondary end point analysis sequence and hypotheses

1.7.3 Hypothesis testing method

All primary and secondary performance end points will be reported and hypothesis tested comparing the active and the sham device. The statistical analyses will test a null hypothesis assuming no difference in treatment effect between the active and sham device.

The hypothesis testing of the primary and secondary endpoints will be performed by an independent biostatistician who will be blinded to which of the two study groups used the active device and which used the sham device.

When analysing the Primary and Secondary Endpoints listed in the study protocol, the fixed-sequence step-down will be used for determining labelling claims of the product when marketed. This method starts by conducting a hypothesis test of the primary endpoint. If the hypothesis test for this endpoint is positive (i.e. the null hypothesis of equal effect of active and sham device was rejected), the following will result:

1. The endpoint will be considered eligible for including as a claim in the product labelling.
2. The first hypothesis on the list of additional hypotheses (See section 1.7.2 above) will be tested. If the null hypothesis is rejected, this process is repeated for the next hypothesis on the list, continuing until a hypothesis test is found to be negative, at which point the step-down process is stopped

The hypothesis testing of primary and secondary end points will be conducted using the methods listed in Table 3:

Endpoint type	Endpoint	Hypothesis testing method
Binary	Absence of moderate or severe pain at 2 hours (AMSP2) Pain Freedom at 2 hours (PF2)	Odds ratios and Pearson's χ^2 significance tests (two-tailed), comparing active and sham device. Threshold p value in interim analysis: 0.0052 (O'Brien-Fleming spending function) Threshold p value in full data set analysis: 0.048
	Freedom from Most Bothersome Symptom at 2 hours (MBSF2) Sustained Pain Freedom at 24 hours (SPF24) Use of rescue medication from the 2 hours' time point until 24 hours (Res24) Freedom from Relapse at 48 hours (FR48)	Odds ratios and Pearson's χ^2 significance tests (two-tailed), comparing active and sham device. Threshold p value: 0.05
Categorical	Headache Score at 2 hours (HS2) Most Bothersome Symptom Score at 2 hours (MBS2) Functional Disability Score at 2 hours (FDS2) Participant Satisfaction at 48 hours (PS48) Light Sensitivity Score at 2 hours (LSS2) Nausea Score at 2 hours (NS2) Sound Sensitivity Score at 2 hours (SSS2)	Ordinal logistic regression using the proportional odds model with device type (active or sham) as the exposure variable. Threshold p value: 0.05

Table 3: Hypothesis testing for primary and secondary endpoints

The following Stage 2 exploratory endpoints will be compared to corresponding baseline values reported at Site Visit 1:

- How many times in the past month has the participant taken respectively a prescription acute migraine drug or an over-the-counter acute analgesic, with the intention to treat migraine.
- Proportion of patients who used an opioid drug to treat migraine the past month
- Proportion of patients who have been hospitalized for migraine in the past month

All hypothesis tests will be performed on the ITT data set of attacks. In the resulting data set each attack will constitute a data point and each patient will constitute a cluster including from one to three data points. The statistical analysis will take clustering into account by using the adjusted χ^2 statistic for clustered binary data developed by Donner and Banting (Donner and Banting, 1988) and validated by others (Gonen et al., 2001; Jung et al., 2001). This method negates the risk of type I errors that is incurred if clustering is disregarded. In the event that changes are made to the statistical analysis plan, these changes will be submitted for approval to the Competent Authorities.

1.8 Homogeneity Across Study Sites

The homogeneity of the treatment effect across study sites will be evaluated using a generalized linear mixed model where device used and site are fixed effects and each patient is a random effect (cluster). Homogeneity of the treatment effect between subjects in the United States and subjects outside the United States will also be evaluated using a generalized linear mixed model. If heterogeneity is detected, exploratory analyses will be conducted, which may include adjusting analyses for baseline characteristics that significantly differ across sites.

1.9 Sensitivity Analysis for Missing Data

All attempts will be made to minimize missing data and to ensure the primary endpoint is powered such that an analysis of complete data will be considered the primary analysis for the study.

A sensitivity analysis of the primary effectiveness endpoint will be conducted as a tipping-point analysis, to examine the impact of missing data. It will only be conducted if the hypothesis test passes (i.e., rejects H_0). We will start by assuming that all missing primary effectiveness endpoints have values indicating the inferiority of Rehaler (i.e. all subjects missing data in the treatment group did not have AMSP2 and all subjects missing the primary endpoint in the control group did have AMSP2). If the primary effectiveness hypothesis is not passed, we will change one missing endpoint until the hypothesis test passes. This will enable determination of the amount of missing data that would alter the final result on the primary effectiveness endpoint.

If any study attacks in the ITT population have missing outcomes at 2 hours and: 1) the missingness is due to the patient's study condition or treatment (missing for cause) or 2) the analysis is underpowered using complete data (observed proportion of missingness is greater than assumed), five (5) multiple imputations will be used for the primary effectiveness endpoint. This consists of imputing values for each missing value as a set, analyzing the results for each set, and then pooling the results. The primary effectiveness endpoint is binary, so the imputation method will be a multiple logistic regression.

In addition, an analysis will be conducted in which if the hypothesis test of the primary effectiveness endpoint does not pass, the hypothesis test will be repeated using the per-protocol analysis population. This test is for sponsor's information only, and will not be used to support the study objective.

2 Appendix I: Interim analysis method for estimating full data set characteristics

2.1 Aim

The aim of the simulation is to, at the interim point, be able to estimate the following in the full data set:

1) Average number of reported study attacks per patient in the full data set
2) The number of patients who will drop out of study (N_{DO}), "drop out" defined here as not having reported at least one study attack by the overall completion of Stage 1. N_{DO} can be subdivided into:
    a) Participants that will have actively dropped out of, or been exited from, the study by the end of overall completion of Stage 1 and who by the time of dropout has NOT reported at least one study attack ("Intentional dropout")
    b) Participants that have not actively dropped out or been exited but who at the end of overall completion of Stage 1 has not yet reported a study attack ("Passive dropout")

There will be a number of participants that will have actively dropped out of, or been exited from, the study by the end of overall completion of Stage 1 but who by the time of dropout *have* reported at least one study attack. In the context of the sample size re-estimation, these patients will not count in the dropout percentage since they do contribute data to the end point analysis.

2.2 Method

A time-progressing simulation was implemented in Microsoft Excel, with the following steps:

Step 1: Input

Use interim data to calculate the following simulation input:

- The likelihood of a patient reporting an attack in a given week. This is calculated as the total number of reported attacks in the interim data set, divided by the summed number of patient weeks in study at the interim point.
- The likelihood of a patient dropping out of, or being exited from, the study in a given week. This is calculated as the total number of intentional dropouts in the interim data set, divided by the summed number of patient weeks in study at the interim point.

Simulation:

On a week by week basis going forward in time from the interim point, simulate occurrences of dropout and attack reports for each patient, according to the likelihoods in the interim data set. This entails adding new patients to the data set at the time in the future where they are projected to be enrolled.

Output:

At the end of the study simulation (i.e. the overall conclusion of Stage 1), calculate the number of dropouts (N_{DO}) and the average number of study attacks reported per patient.

Run 20 simulations and average results.

2.3 References

Bates D., Ashford E., Dawson R., Ensink F.-B.M., Gilhus N.B., Olesen J., Pilgrim A.J., Shevlin P., 1994. Subcutaneous sumatriptan during the migraine aura. NEUROLOGY 44, 1587–1592.

Donner, A., Banting, D., 1988. Analysis of site-specific data in dental studies. Journal of Dental Research 67, 1392–1395. https://doi.org/10.1177/00220345880670110601

Donner, A., Klar, N., 1993. Confidence interval construction for effect measures arising from cluster randomization trials. J Clin Epidemiol 46, 123–131. https://doi.org/10.1016/0895-4356(93)90050-b

Fuglsang, C.H., Johansen, T., Kaila, K., Kasch, H., Bach, F.W., 2018. Treatment of acute migraine by a partial rebreathing device: A randomized controlled pilot study. Cephalalgia 38, 1632–1643. https://doi.org/10.1177/0333102418797285

Gonen, M., Panageas, K., Larson, S., 2001. Statistical issues in analysis of diagnostic imaging experiments with multiple observations per patient. RADIOLOGY 221, 763–767. https://doi.org/10.1148/radiol.2212010280

Jung, S., Ahn, C., Donner, A., 2001. Evaluation of an adjusted chi-square statistic as applied to observational studies involving clustered binary data. STATISTICS IN MEDICINE 20, 2149–2161. https://doi.org/10.1002/sim.857

Olesen J., Diener H.C., Schoenen J., Hettiarachchi J., 2004. No effect of eletriptan administration during the aura phase of migraine. Eur. J. Neurol. 11, 671–677. https://doi.org/10.1111/j.1468-1331.2004.00914.x