A simulation method based on interim observed data to assess conditional power in a clinical trial

Lin Pan[1], Jill Stankowski[1], Joseph Massaro[2]

1 ICON Plc, North Wales, PA

2 School of Public Health, Boston University, Boston, MA; Harvard Clinical Research Institute, Boston, MA

The designs of current clinical trials often involve interim analysis for assessment of futility (low conditional power) to reject the primary null hypothesis at the protocol-specified sample size) or for sample size adjustment to maintain desired conditional power. In a single-arm study, Visual Analog Scale (VAS) improvement from baseline was measured at three post-baseline visits for each patient and the VAS change from baseline scores at the last visit was the primary efficacy endpoint. The primary analysis compared the mean of this endpoint to a pre-specified performance goal using Mixed Model Repeated Measures. A simulation method was proposed for assessing conditional power in this setting, conditioned on the interim observed data and under the assumption that the observed interim sample characteristics are the true population characteristics. Specifically, once the interim data were observed, multiple post-interim datasets were simulated from a population with the same characteristics as the interim observed data. Complete simulated data sets were then composed of the observed interim dataset appended to each simulated post-interim data set. The proportion of complete simulated data sets for which the null hypothesis was rejected was the simulated conditional power.

Introduction

The design of current clinical trials often involves at least one interim analysis. The information time (e.g., percentage of planned sample size) at which the interim analyses will be carried out is pre-specified in the protocol. By evaluating observed data in the mid-study from an ongoing trial, the result of interim analysis has the potential for modifying or adapting the conduct of the study. The interim analysis can include a review of safety and/or efficacy data. For efficacy, an interim assessment of the efficacy null hypothesis is carried out and may include further assessments of (1) whether the study may be stopped after interim analysis because of overwhelming evidence of the efficacy of the experimental treatment in the interim observed results; (2) whether the study should be stopped for futility, or i.e., because of low conditional power (CP) where CP is defined as the probability that experimental treatment will provide statistically significant beneficial results at the final protocol-specified sample size, conditioned on the interim observed results; (3) whether a sample size increase (beyond the protocol-specified sample size) is warranted to maintain the CP at a pre-specified value (e.g., 80%) for rejecting the efficacy null hypothesis at a desired level, and (4) continue as is.

For some clinical trials, a futility stopping criterion in terms of CP, assessed at interim analysis, is pre-specified. When calculated conditional power does not reach the criterion (e.g., 10%), study will stop for futility. This paper discusses using interim

analysis for assessment of futility and sample size re-calculation in a recent single-arm trial.

In a two-group randomized clinical trial, when the observations arise from normal distribution, the CP of a two-sided test of H_0: $\mu_1 - \mu_2 = 0$ vs. H_1: $\mu_1 - \mu_2 \neq 0$ after obtaining n_1 observations from each group can be explicitly calculated[1]. Following the CP calculation, once the additional observations, if any, have been obtained, the null hypothesis is tested using all accumulated data[2]. In this paper, we propose a simulation based technique to assess CP at interim analysis for a single-arm trial where the outcome is collected at several time points and a mixed model for repeated measure (MMRM) analysis is used to assess if the mean outcome at the end of follow-up exceeds a pre-specified threshold. Details are described in the following sections.

Method and Materials

This was a single-arm device study where an investigational device was implanted in the knee of patients with knee fractures in order to reduce pain. The Visual Analog Scale (VAS), a measurement of pain ranging from 0 to 100 with 0 meaning “no pain” and 100 meaning “worst imaginable pain” was measured before implantation as the baseline measurement and at three post-baseline visits (Visit 1-3 post implantation) for each subject. VAS score change from baseline at last visit (Visit 3) was the primary effectiveness endpoint. The primary analysis compared the mean of this endpoint to a pre-specified performance goal μ_0 obtained from historical controls obtained through literature review. Specifically, the primary null and alternative hypotheses are:

$$H_0: \mu_3 \leq \mu_0 = 53.8 \text{ vs. } H_1: \mu_3 > \mu_0 = 53.8$$

where μ_3 is the true unknown mean VAS change from baseline to last visit (Visit 3) and 53.8 is the pre-specified performance goal (indicating a mean improvement from baseline in VAS of 53.8 at last visit). An evaluable sample size of 68 patients yields at least 80% power to reject the null hypothesis in favor of the alternative under the assumptions that $\mu_3 = 58$ and the standard deviation of the change in VAS from baseline to Visit 3 VAS is 13. A mixed model repeated measures (MMRM) model will be executed for the dependent variable of “change from baseline at each visit”, with the categorical main effect of visit (Visits 1 - 3) and the continuous baseline VAS as the independent variables and assuming an unstructured within-patient covariance matrix. From this model incorporating all visits, the estimate of the mean change from baseline to Visit 3 and its standard error will be obtained, and from this a one-sample t-test will be generated to test the above null hypothesis of interest at a one-sided 0.05 level of significance.

As pre-specified in the protocol, an interim analysis on change from baseline VAS at Visit 3 was carried out when the 37th enrolled patient had been treated and followed until Visit 3. Subject with non-missing baseline and any post baseline VAS scores were included in the interim analysis. The purpose of this interim analysis was to potentially stop the trial for futility or potentially increase the sample size if the mean VAS improvement from baseline at Visit 3 was large but not as large as anticipated in the original sample size calculations specified in the protocol. Specifically, the study was to be stopped for futility if the CP with a final sample size of 68 patients was <10% and the sample size was to be increased to maintain CP of 80% if the CP at the interim analysis for the planned evaluable sample size of 68 subjects was between 50% and 80%; otherwise the study was to continue as is. This algorithm to increase sample size to maintain a CP of 80%, when the CP based on the original protocol-specified sample size is between 50% and 80%, does not require an adjustment to the final one-sided 0.05 level of significance as long as the sample size increase required to yield a CP of 80% is <105% of the original protocol-specified final sample size, as outlined in Chan, DeMets and Lan (2004)[3].

The one-sample t-test resulting from the above MMRM model was carried out on the interim dataset. Under the assumption that VAS score is normally distributed and the observed interim sample characteristics represents the true population characteristics, the following simulation algorithm was used to determine CP to reject the above null hypothesis in this setting at the protocol-planned final sample size: multiple post interim datasets were simulated using SAS version 9.4 from a population with the same characteristics (mean, standard deviation, pairwise within-patient correlation between time points) as the interim observed data. Full analysis data sets were composed by appending the observed interim data to each simulated post interim dataset. Within each full analysis data set, the MMRM model was performed and the statistical testing for comparing mean VAS change from baseline at Visit 3 with pre-specified value of 53.8 was assessed. The proportion of simulated full analysis data sets for which the null hypothesis is rejected was the simulated conditional power.

The SAS simulation program is given in the appendix. The first major step requires the user to run PROC MIXED on the interim data to obtain estimates of the characteristics of the interim data. It is assumed that the interim dataset is set up with one record per patient with variables for VAS at baseline, and change in VAS at visits 1, 2, and 3. The program transposes this dataset to have one record per post-baseline visit (with VAS as the main variable of interest at each post-baseline visit) prior to carrying out the PROC MIXED. The subject's Baseline VAS is also included as a variable in each record.

In the data step following the first PROC MIXED (see "**data** POST_INTERIM_CHARACTERISTICS(type=corr);), the user inputs the observed

characteristics (sample size, mean, standard deviation of baseline VAS and of the change from baseline VAS at each post-baseline visit; correlation of baseline VAS with each change in VAS at each visit; and the correlation of change in VAS between the three post-baseline time points) of the interim data in a “type=CORR” SAS dataset. The only remaining input required by the user is in the PROC SIMNORMAL statement; this is the procedure that simulates the post-interim baseline data. The user needs to input the number of post-interim subjects to be enrolled and the random seed in the “NUMREAL=” and “SEED=” options. The program is easily modifiable if there are <3 or >3 time points and for any continuous outcome. In order to assess the impact of missing VAS data on the results of the CP calculations at the time of the interim analysis, simulations were performed under three scenarios:

1. We generated an interim data set where all subjects have complete interim data (no missing data at both baseline and each post baseline visit) and there are no subjects with missing data in simulated post-interim datasets.
2. We generated an interim data set where some subjects have missing interim data (under missing completely at random (MCAR), missing rate at baseline and Visits 1-3 are 0%, 10%, 20%, and 25% accordingly); we carry out the simulation assuming there are no missing data in simulated post-interim datasets.
3. We generated an interim data set where some subjects have missing interim data (under MCAR assumption, missing rate at baseline and Visits 1-3 are 0%, 10%, 20%, and 25% accordingly); we carry out the simulation assuming same missing pattern in simulated post-interim datasets.

Note that no imputation of missing data was carried out prior to conducting any analyses, but the MMRM approach is an adequate method of handling missing data if the missing data mechanism is MCAR or missing at random (MCR).

The simulated conditional powers obtained in three scenarios were summarized and compared.

Results

At the time of interim analysis, a total of 37 subjects had surgery with investigational device implanted in the knee. All 37 subjects had non-missing VAS measurements at baseline and any post baseline visit. The protocol planned sample size in this study was 68 evaluable subjects. For purposes of this manuscript and for purposes of confidentiality, we modified the interim observed data to match the scenarios 1-3 above.

SAS version 9.4 programs were created to implement the simulation and assess the simulated conditional power under three scenarios. Again, the SAS code is presented in the appendix.

Table 1 presents estimates of the least square mean and standard error for VAS change from baseline to Visit 3 in the interim observed data, obtained from MMRM model under three scenarios discussed previously. Under each scenario, 1000 post interim samples with the sample size given in Table 1 were simulated with SAS PROC SIMNORMAL from a population with the same characteristics as the interim data; each post-interim sample was combined with the observed interim sample to obtain 1000 final datasets with a sample of size 68. Then the conditional power was calculated by counting the proportion of these 1000 combined datasets for which the null hypothesis of the study, i.e. VAS improvement from baseline is less than or equal to the pre-specified value (53.8), is rejected.

Table 1. Conditional Power Assessed with Simulated Data Under Three Scenarios[c]

Scenario	N	n	M	LS Mean (SE)[a]	Visits	ρ	μ_0[b]	CP
1	37	37	68-37=31	56.9 (2.10)	V1-V2	0.07	53.8	71.1%
					V1-V3	0.26		
					V2-V3	0.15		
2	37	28	68-28=40	55.2 (1.74)	V1-V2	0.25	53.8	24.2%
					V1-V3	0.36		
					V2-V3	-0.05		
3	37	28	(68-28)/0.75=53	55.2 (1.74)	V1-V2	0.25	53.8	25.5%
					V1-V3	0.36		
					V2-V3	-0.05		

Abbreviations: N=number of subjects included in the MMRM model at interim analysis; n=number of evaluable subjects (those with non-missing change from baseline to Visit 3; note however that all N subjects are included in the MMRM model); M=number of post interim subjects simulated in each post interim dataset with non-missing baseline and any post baseline data; LS Mean= Estimate of the Least Square Means of the change from baseline to last visit in the interim observed data; SE=Standard Error of the LS Mean; ρ=pair-wise correlation coefficient; CP=Conditional Power

a: Obtained from MMRM model with interim observed data, with the categorical main effect of visit (Visit 1, 2, 3) as the independent variable and the baseline value of the score as a covariate. Parameters for the MMRM were estimated using a direct likelihood approach as implemented in the SAS procedure PROC MIXED. An unstructured covariance matrix was assumed.

b: Performance goal of 53.8 (the pre-specified value used in the null hypothesis, is derived from historical control data through a literature review

c: To protect trial confidentiality and in order to allow data to match the desired scenarios above, interim results, data presented in this paper was modified from the actual data

The simulated CP obtained under first scenario was 71.1% which was between 50% and 80%, therefore, sample size re-assessment was performed. In order to achieve the desired 80% CP, the number of subjects in post-interim datasets was increased by increasing “numreal=” in the option of PROC SIMNORMAL procedure until a CP of 80% power was achieved. The number of subjects in the simulated post-interim datasets needed to be 52 to achieve 80% CP at the end of the trial. Final sample size increased from 68 to 89 (the sample size increase of 21 is only 24% of the original sample size of

68 and hence is allowable since it is below the 105% threshold allowed by the Chen, DeMets and Lan method sample size increase method). For the second and third scenarios, the trial will continue as it is for CP is between 10% and 50%.

Summary and Discussion

A simulation method is proposed for calculating conditional power in a single-arm trial on for a continuous outcome, conditioned on the interim observed data and under the assumption that the observed interim sample characteristics are the true population characteristics. In this scenario, continuous outcome data are collected from subjects at baseline and 3 post-baseline visits. The primary endpoint is the change from baseline at the last visit, but all visits are used in an MMRM model to estimate the change from baseline in the last visit. The simulations require assumptions of the mean and standard deviation of the change from baseline and assumptions of the within-patient correlations for the post-interim data (all of these assumptions are set to the MMRM-estimated values on the observed interim data). Simulations were carried out in SAS; a copy of the simulation program is provided below. This can be easily modified to an analogous single-arm trial for any continuous variable with <3 or with >3 time points.

We provided results of the simulation under three different missing data scenarios. As is expected, missing data in the observed interim data and post interim data caused biased point VAS estimate, which further affect the assessment on conditional power and sample size reassessment during interim analysis.

References

1. Sample size recalculation using conditional power. J. Denne, Statistics in Medicine, 2001 20: 2645-2660.
2. Sample size re-estimation: recent developments and practical considerations. A. Gould, Statistics in Medicine, 2001 20:2625-2643.
3. Chen YHJ, DeMets DL, Lan KKG (2004). “Increasing the sample size when the interim result is promising.” Statistics in Medicine. 23: 1023-1038

Appendix

SAS codes for conditional power assessment through simulation for a single arm with three time points.

```
*******************************************************************************
*******************************************************************************
* PROGRAM NAME:      CONDITIONAL_POWER_SINGLE_ARM_THREE_TIMEPOINTS
* PROGARM PURPOSE:   CALCULATES CONDITIONAL POWER GIVEN AN INTERIM OBSERVED DATASET.
* SAS VERSION:       9.4
* SITUATION 1:       37 SUBJECTS WITH COMPLETE BASELINE, POST VISIT 1-3 DATA
*------------------------------------------------------------------------------;

/* RUN PROC MIXED ON THE INTERIM OBSERVED DATA, PRIMARILY TO CALCULATE  */
/* AN ESTIMATE OF THE CORRELATION MATRIX BETWEEN VISITS AND THE ESTIMATE    */
/* OF THE INTERIM OBSERVED MEAN AND STANDARD ERROR.  THESE CHARACTERSISTICS */
/* WILL BE LATER USED TO SIMULATE 1000 POST-INTERIM VAS DATASETS FROM A     */
/* POPULATION WITH THE SAME CHARACTERISTICS OF THE INTERIM OBSERVED DATA,   */
/* IN ORDER TO EVENTUALLY CALCULATED CP UNDER THE ASSUMPTION THAT THE       */
/* INTERIM DATASET CHARACTERISTICS ARE THE SAME CHARACTERISTICS AS THE      */
/* POPULATION.                                                              */
data INTERIM_DATA_STACKED;
      set INTERIM_DATA;
      visit=1; CHGvas=CHGvas1; output;
      visit=2; CHGvas=CHGvas2; output;
      visit=3; CHGvas=CHGvas3; output;
      keep subjid baseVAS visit CHGvas;
run;
/* THE RCORR OPTION BELOW YIELDS AN ESTIMATE OF THE CORRELATION MATRIX,    */
/* THE LSMEANS STATEMENT YIELDS AN ESTIMATE OF THE MEAN CHANGE FROM        */
/* BASELINE VAS AND ITS STANDARD ERROR AT EACH OF VISITS 1-3.              */
proc mixed data=INTERIM_DATA_STACKED method=ml;
    class visit subjid;
    model CHGvas = visit baseVAS/ ddfm=kr;
    repeated visit / type=un subject=subjid r rcorr;
    lsmeans visit / diff adjust=GT2 adjdfe=row;
run;
quit;
/* OBTAIN AN ESTIMATE OFF THE BASELINE MEAN VAS AND THE CORRELATION OF      */
/* BASELINE VAS WITH EACH CHANGE FROM BASELINE VAS.                         */
proc corr data=INTERIM_DATA;
      var baseVAS;
      with CHGvas1-CHGvas3;
run;

/* NOW TAKE THE OPERATIONAL CHARACTERISTICS ESTIMATED FROM ABOVE STATEMENTS,*/
/* AND ASSUME THAT THESE CHARACTERISTICS DEFINE THE POPULATION FROM WHICH   */
/* THE POST-INTERIM SAMPLE IS TAKEN.  WE WILL SIMULATE 1000 POST-INTERIM    */
/* SAMPLES FROM THIS POPULATION.  FOR EACH SIMULATED SAMPLE, WE WILL        */
/* COMBINE IT WITH THE ABOVE INTERIM OBSERVED SAMPLE. WE WILL THEN          */
/* CALCULATE THE PROPORTION OF THESE 1000 COMBINED DATASETS FOR WHICH THE   */
/* NULL HYPOTHESIS OF THE STUDY IS REJECTED.  THIS IS OUR CONDITIONAL       */
/* POWER (CP) UNDER THE ASSUMPTION THAT THE CHARACTERISTICS OF THE          */
/* INTERIM OBSERVED DATASET ARE THE TRUE CHARACTERISTICS OF THE POPULATION. */

/* TO START OUT, CREATE A DATASET CONTAINING THE CHARACTERISTICS OF THE     */
/* POPULATION.  FOR THE DATASET BELOW:
/* THE FIRST ROW OF DATA IS THE INTERIM OBSERVED MEAN OF BASELINE VAS        */
/* (ESTIMATED FROM THE ABOVE PROC CORR) AND OF THE CHANGE FROM BASELINE VAS  */
/* TO VISITS 1 - 3 (ESTIMATED FROM THE LSMEANS STATEMENT OF THE ABOVE PROC   */
/* MIXED STATEMENT. THE SECOND ROW OF DATA IS THE INTERIM OBSERVED           */
/* STANDARD DEVIATION OF BASELINE VAS ESTIMATED FROM THE ABOVE PROC CORR)    */
```

```
/* AND OF THE CHANGE IN BASELINE VAS TO VISITS 1, 2 AND 3 (ESTIMATED         */
/* BY TAKING THE ESTIMATED STANDARD ERRORS FROM THE LSMEANS STATEMENT        */
/* OF THE ABOVE PROC MIXED AND MULTIPLYING THEM BY THE SQUARE ROOT           */
/* OF THE SAMPLE SIZE AT EACH VISIT). THE THIRD ROW ARE THE PLANNED EVALUABLE*/
/* SAMPLE SIZES AT EACH VISIT POST-INTERIM.  ROWS 4-7 ARE THE ESTIMATED      */
/* CORRELATION MATRIX FROM THE INTERIM OBSERVED DATA.                         */
data POST_INTERIM_CHARACTERISTICS(type=corr);
      input _TYPE_ $ 1-4 _NAME_ $ 6-12 BASEVAS CHGVAS1 CHGVAS2 CHGVAS3;
cards;
MEAN          81.3 78.1 60.3 56.9
STD           19.6 11.1 11.8 12.8
N             31 31 31 31
CORR BASEVAS  1.00 0.44 0.49 0.15
CORR CHGVAS1  0.44 1.00 0.07  0.26
CORR CHGVAS2  0.49 0.07 1.00  0.15
CORR CHGVAS3  0.15 0.26 0.15  1.00
run;

/* SIMULATE THE 1000 POST-INTERIM ANALYSIS DATASETS. EACH DATASET WILL      */
/* BE COMBINED WITH THE INTERIM OBSERVED DATASET IN ORDER TO CREATE 1000    */
/* COMPLETE SIMULATED CLINICAL TRIALS.  FOR EACH TRIAL, A PROC MIXED        */
/* ANALYSIS WILL BE CARRIED OUT TO TEST THE STUDY'S NULL HYPOTHESIS.        */
/* THE PROPORTION OF SIMULATED DATASETS FOR WHICH THE NULL HYPOTHESIS IS    */
/* REJECTED IS THE SIMULATED CONDITIONAL POWER.                             */
%MACRO DOIT2;
data POST_INTERIM_CHARACTERISTICS(type=corr);
      %DO I=1 %TO 1000;
            simulation=&I;
            set POST_INTERIM_CHARACTERISTICS;
            OUTPUT;
      %END;
run;
%MEND DOIT2;
%DOIT2;
proc sort data=POST_INTERIM_CHARACTERISTICS;
      by simulation;
run;
/* SIMULATE A MULTIVARIATE NORMAL DISTRIBUTION FOR BASELINE AND CHANGE FROM         */
/* BASELINE VAS TO VISITS 1 - 3 UNDER THE ABOVE CHARACTERISTICS FOR EACH SIMULATION.*/
proc simnormal data=POST_INTERIM_CHARACTERISTICS numreal=31 seed=315893282
out=POST_INTERIM_DATA;
      by simulation;
      var baseVAS CHGvas1-CHGvas3;
run;
data POST_INTERIM_DATA;
      set POST_INTERIM_DATA;
      subjid=rnum+100000; /* MAKE SURE SUBJIDSFOR POST-INTERIM DATA ARE NOT  */
                          /* THE SAME AS SUBJECT IDS FROM INTERIM DATA.      */
      drop rnum;
run;
/* COMBINE EACH SIMULATED DATASET WITH THE INTERIM OBSERVED DATASET.       */
/* THIS IS DONE BY FIRST MAKING 1000 COPIES OF THE INTERIM OBSERVED DATA,  */
/* AND THEN ADDING ONE COPY TO EACH OF THE 1000 SIMULATED POST-INTERIM     */
/* DATASETS.                                                               */
%MACRO DOIT3;
data INTERIM_DATA;
      %DO I=1 %TO 1000;
            simulation=&I;
            set INTERIM_DATA;
            subjid=rnum;
            OUTPUT;
      %END;
      drop rnum;
```

```
run;
%MEND DOIT3;
%DOIT3;

data FINAL;
      set INTERIM_DATA POST_INTERIM_DATA;
run;
proc sort data=FINAL NODUPKEY;
      by simulation subjid;
run;

/* ARRANGE THE DATA SO THAT PROC MIXED CAN BE CARRIED OUT FOR EACH SIMULATED */
/* DATASET.                                                                  */
data FINAL_STACKED;
      set FINAL;
      visit=1; CHGvas=CHGvas1; output;
      visit=2; CHGvas=CHGvas2; output;
      visit=3; CHGvas=CHGvas3; output;
      keep simulation subjid baseVAS visit CHGvas;
run;

ods select none;
/* THE LSMESTIMATE STATEMENT BELOW CONDUCTS A TEST OF THE ABOVE NULL  */
/* HYPOTHESIS OF INTEREST.                                           */
proc mixed data=FINAL_STACKED method=ml;
      by simulation;
    class visit subjid;
    model CHGvas = visit baseVAS/ ddfm=kr;
    repeated visit / type=un subject=subjid r rcorr;
    lsmeans visit / diff adjust=GT2 adjdfe=row;
      lsmestimate visit 0 0 1/upper testvalue=53.8;
      ods output lsmestimates=lsmestimates;
run;
quit;
ods select all;

/* DETERMINE THE PROPORTION OF SIMULATIONS FOR WHICH THE NULL HYPOTHESIS */
/* IS REJECTED.  THIS IS THE CONDITIONAL POWER.                          */
data FinalResult;
      set lsmestimates;
      if probt<=0.05 then reject=1;
      else reject=0;
run;
proc sort data=FinalResult;
      by descending reject;
run;
proc freq data=FinalResult order=data;
      table reject;
run;
```