A New Method for a One-Shot Unblinded Sample Size Reassessment in Two-Group Trials : How & When ?

Elodie Blondiaux[1] and Eric Derobert[1]

[1] Department of Statistics, Sanofi-Aventis Research,
1, avenue Pierre Brossolette,
91385 Chilly-Mazarin Cedex, France
(e-mails: elodie.blondiaux@sanofi-aventis.com ; eric.derobert@sanofi-aventis.com)

Abstract. In two-group clinical trials, it has become usual, when performing a single interim unblinded sample size reassessment, to prevent the inflation of the Type-I error by using the weighted statistics proposed by Cui, Hung, Wang (1999). Keeping unchanged the targeted between-treatment difference, we used this statistics and followed the driving principle of considering that a sample size is optimal when the derivative of the power with respect to the sample size has reached some implicit value. Then we developed new strategies in order to express a) an optimally reassessed sample size as a function of the Z-statistic obtained at the first stage (at any interim analysis time, i.e. at any information fraction f $(0<f<1)$), b) an optimal information fraction f_{opt} for the interim analysis, obtained as a function of (α, β), the Type-I and Type-II error rates. We adapted preliminary "conceptual" findings into a finally proposed "pragmatic" strategy. This pragmatic strategy was compared to two adaptive design methods of the literature (Cui, Hung, Wang, 1999; Denne, 2001) and to one adaptive method sometimes proposed, based on the current trend. We used the information fraction $f = 0.5$ and our optimal $f_{opt} \approx 0.9$ (found for usual values of α and β). When the interim analysis is performed at $f=0.5$, our method provides good results. When the interim analysis is performed with 90% of the data information, methods become roughly comparable. Our finding of an optimal time for reassessment is in fact useful for all methods. Results suggest that a sample size reassessment is more beneficial when performed close to the planned end of a trial, allowing a study with borderline interim results to be saved.

Keywords. Adaptive designs, Conditional power, Sample size reassessment, Information fraction, Futility.

1 Introduction

In the recent years, several methods have been developed for modifying some features of two-group clinical trials, based on the unblinded look at the data collected until some interim time-point. The required sample size is the main feature which can be reassessed in this way. The major issue in such adaptive designs is how to prevent the inflation of the global Type-I error: this had been sometimes obtained by astutely combining appropriate rules for futility and reassessment (see for example Shun (2001)). However it is much more flexible and rigorous to use either the weighted statistics proposed by Cui, Hung, Wang (1999) or the method based on the conditional rejection error probability proposed by Schäfer-Müller (2001). These two methods are in fact equivalent in the case of a single interim look and reassessment (Vandemeulebroecke, 2006) considered in this work. A connected question is how to determine an appropriate time to perform the unblinded reassessment.

Based on the weighted Cui-Hung-Wang statistics, we develop new strategies in order to determine how and when an interim sample size reassessment should be performed. Section 2 presents the proposed model and argues about keeping the original targeted difference as the basis for reassessment. It also introduces the basic idea derived from sample size calculations for a fixed design, which drives all the further calculations. Section 3 shows the results obtained with the so-called "conceptual" and "pragmatic" strategies. Section 4 compares our results with those obtained with two adaptive design methods of the literature (Cui, Hung, Wang, 1999; Denne, 2001) and to one adaptive method sometimes proposed, based on the current trend.

2 Model, Bettings and Driving Principle

2.1 The Model

In the comparison of two treatment groups (typically: an experimental drug vs a control or a placebo), with 1:1 randomization, normal model and unit variance $\sigma^2=1$, we want to test H0 : $\delta=0$ vs Ha : $\delta>0$, where δ is the true unknown treatment difference. The total sample size originally planned N is: $N = 4(u_\alpha + u_\beta)^2 / \Delta^2$, where α and β are the Type-I and Type-II error rates, Δ is the targeted treatment difference, u_a is the $(1\text{-}a)$ quantile of the standard normal distribution.

We consider a two-stage procedure with one interim analysis performed at the information fraction $f \in]0; 1[$, where we consider the statistics: $Z_f = d_{obs} \sqrt{N.f} \;/\; 2$;
d_{obs} is the observed treatment effect at the interim analysis (first stage).

We develop a method for adapting the original sample size N into a new total sample size N_{new} at this interim time-point, and use, at the final analysis, the Cui-Hung-Wang statistics: $Z_t = \sqrt{f}\;\; Z_f + \sqrt{1 - f}\;\; Z_s$, (with $Z_s = d_{sup} \sqrt{N_{new} - N.f} \;/\; 2$; d_{sup} is the treatment effect to be observed on the post-interim data (second stage)).
This test statistics can be used regardless of the way chosen for modifying the sample size.

2.2 Bettings: Targeted vs Observed difference

Unlike some approaches to adaptive designs, here we opt to maintain the aim of showing the originally targeted between-treatments difference Δ. Indeed we assume that, in a Phase III trial, we know enough on the treatment and on the disease to choose a reasonable and realistic target.

Some adaptive methods change the original objective. They assume that the treatment difference d_{obs} observed at the interim analysis has become the most plausible value for the treatment effect, and bet on it. However, we consider that d_{obs} remains only an estimation and should not be taken for granted. By definition, an estimator is subjected to variability: even if d_{obs} is the most plausible value of the treatment effect, there are many other reasonable values. Furthermore, if the targeted Δ is in the confidence interval of d_{obs}, the interest of betting on a new target is questionable. For example, if $d_{obs} = \Delta/2$, based on half of the data ($f=0.5$ is a common practice in interim analyses), its 95% confidence interval would be $[-0.49\,\Delta : 1.49\,\Delta]$ if $\alpha=0.025$ and $\beta=0.20$ or $[-0.36\,\Delta : 1.36\,\Delta]$ if $\alpha=0.025$ and $\beta=0.10$. Then, why consider $\Delta/2$ as the new target, whereas the true effect can always be Δ (as well as 0 !) and could be either an unexpectedly high effect or a deleterious one ? Moreover, from a regulatory point of view, a change of target could be possibly challenged: resuming the example above, it is quite uncertain whether $\Delta/2$ (or some other fraction of Δ) would still be clinically interesting.

Actually this “targeted vs observed” debate should perhaps not be as agitated as it is sometimes.
In fact, because $Z_f = d_{obs} \sqrt{N.f} \;/\; 2$ can also be written $Z_f = (d_{obs} \;/\; \Delta).(u_\alpha + u_\beta).\sqrt{f}$,
any reassessment rule considered as a function of $(f, \alpha, \beta, Z_f, \Delta)$ could be equally viewed as a function of $(f, \alpha, \beta, Z_f, d_{obs})$ and vice versa. Differences occur of course in construction and interpretation. But ultimately, the operating properties of the methods (whatever the way they are built) is really what does count.

2.3 What is, implicitly, an optimal sample size ?

As there is no universal definition of an optimal sample size, we must first define the optimality criterion to be used. Here we develop the idea that the pre-specified power $(1\text{-}\beta)$, associated with the initially planned sample size N, would not be considered as meaningfully increased if one extra patient was included. This gives a “limit value” for the derivative of the power (considered as a function of the sample size N), which will serve defining an optimality criterion when finding how and when the sample size should be reassessed.

More precisely, if we call $n= r.N$ any possible sample size, then a standardized “limit value” LP is obtained when $r=1$ (i.e. when the calculated sample size N controls the targeted power). It holds that:

$$\frac{\partial(1-\beta)}{\partial r} = \frac{\partial(1-\beta)}{\partial u_\beta}.\frac{\partial u_\beta}{\partial r} = \frac{exp(-(u_\beta)^2/2)}{\sqrt{2\pi}}.\frac{u_\alpha + u_\beta}{2r} = \frac{(u_\alpha + u_\beta)}{2r\sqrt{2\pi}}.exp(\frac{-(u_\beta)^2}{2})$$

and subsequently : $LP = \dfrac{(u_\alpha + u_\beta)}{2\sqrt{2\pi}}.exp(\dfrac{-(u_\beta)^2}{2})$ (LP depends only on α and β).

Coming now to the adaptive strategy of reassessment, for every possible value of f and Z_f, this rate LP will be compared to the increase of the conditional power for each new patient in the second stage of the adaptive design. We will stop to include patients when the derivative of the conditional power will reach down the defined bound and the corresponding number of patients N_{new} will be the optimal sample size.

Based on the same principle, we will also define an optimal time f_{opt} for the interim analysis by using LP for comparing the average increase of the sample size and the average increase of the unconditional power obtained for each value of f.

3 Conceptual and pragmatic strategies of reassessment

3.1 Conceptual strategy

The mathematical properties of the above driving principle allowed us first developing a "conceptual" strategy, and led us to elicit a natural futility region for every triplet (α, β, f) which provides a conditional power (at the interim analysis) $CP(z_f)=Prob\{Z_t>u_\alpha \mid Z_f=z_f\}$ at least equal to 50%.

Defining $q=N_{new}/N$, it holds that: $CP(z_f) = Prob\{N(0,1) < \dfrac{\sqrt{f}\,.z_f - u_\alpha}{\sqrt{1-f}} + \sqrt{q-f}\,.(u_\alpha + u_\beta)\}$ [1]

Then equating $\partial CP/\partial q = LP$, an expression of the Z_f interim Z-statistic depending on N_{new} is derived and allows an indirect calculation of N_{new} :

$$z_f = \frac{u_\alpha}{\sqrt{f}} - \frac{\sqrt{1-f}}{\sqrt{f}}.\left[\sqrt{\frac{N_{new}}{N} - f}\;.(u_\alpha + u_\beta) - \sqrt{u_\beta^2 - \ln\left(\frac{N_{new}}{N} - f\right)}\right] \quad [2]$$

This equation has a unique solution for N_{new} on condition that:

$$z_f \geq z_0 = \frac{u_\alpha}{\sqrt{f}} - \frac{\sqrt{1-f}}{\sqrt{f}}.(u_\alpha + u_\beta).\exp\left(\frac{u_\beta^2}{2}\right)$$

$z_f=z_0$ is the natural bound for futility ; it leads to the maximum $N_{new}=N[f+\exp(u_\beta^2)]$.

[Note: when z_f goes under z_0, $CP(z_f)<50\%$ and solutions for N_{new} exist (they are obtained from another Equation than Equation[2]). But paradoxically, the sample size reassessment would then become less important than $N[f+\exp(u_\beta^2)]$ despite the worst interim results: this has been the reason for considering z_0 as the natural bound for futility.]

The pre-specified power $(1-\beta)$ and the conditional power $CP(z_f)$ are linked by the remarkable relation:

$u_{1-CP(z_f)} = \sqrt{u_\beta^2 - \ln(q-f)}$ [3]

Unfortunately, for usual values like α=0.025, β=0.10, f=0.5, the futility bound is z_0=-4.60 (with N_{new}=5.67 N). A complementary futility approach (not described here) led to z_0=-1.83 and N_{new}=3.26 N.

Therefore this conceptual strategy does not seem reasonable in practical situations. Indeed, an obstinate blind application of the above would mean that having observed, from half of the data, disappointing values like d_{obs}=-2.01Δ (z_0=-4.60) or d_{obs}=-0.80Δ (z_0=-1.83), would lead to pay such a big price as recruiting 3 to 6 times the number of patients originally planned. Extra constraints must definitely be added to the method in order that it meets practical requirements.

3.2 Pragmatic strategy

Constraints on the reassessed sample size

On the one hand, it follows from Equation[3] that, in the conceptual strategy, the conditional power is equal to the pre-specified power iff $q=1+f$. Therefore we chose $N(1+f)$ as the maximum value of N_{new} $(q\leq 1+f)$. Because $f \in]0; 1[$, a final reassessed sample size could never exceed 2 times the originally planned sample size.

On the other hand, the conceptual strategy allowed reducing the sample size. It was not impossible at all to be confronted to the final negative result of a promising but shortened study. The odds are that the popularity of a statistician having promoted this solution would also be shortened. In the pragmatic strategy, we therefore propose to force $N_{new} \geq N$ $(q \geq 1)$.

Finally: $N \leq N_{new} \leq N(1+f) \leq 2N$ [or $1 \leq q \leq (1+f) \leq 2 = q_{max}$].

Constraints on the futility rules

Two additional rules were considered:

i) stopping the trial if the first stage results are in the wrong direction ($z_f < 0$), then the futility bound should be $\geq z_{min(1)}=0$,

ii) as in the conceptual strategy, stopping the trial if the conditional power (under Δ) is less than 50% ; from Equation[1] and $q \leq 1+f$, it follows:

$$z_{min(2)} = \frac{u_\alpha - (u_\alpha + u_\beta).\sqrt{1-f}}{\sqrt{f}}$$

The complete reassessment rule of the pragmatic strategy

Let's note z_f the statistics observed at the first stage, and the special values $z_{1+f} = \{z_f \mid q=1+f\}$ and $z_1 = \{z_f \mid q=1\}$; from Equation [2], these values are:

$$z_{1+f} = \frac{u_\alpha}{\sqrt{\mathrm{f}}}.(1-\sqrt{1-\mathrm{f}}) \text{ and } z_1 = u_\alpha.\sqrt{\mathrm{f}} - \frac{1-\mathrm{f}}{\sqrt{\mathrm{f}}}.u_\beta + \sqrt{\frac{1-f}{f}}.\sqrt{u_\beta^{\,2} - \ln(1-\mathrm{f})}$$

If $z_f < max\{ z_{min(1)}, z_{min(2)}\} = max\left(0, \frac{u_\alpha - (u_\alpha + u_\beta).\sqrt{1-f}}{\sqrt{f}}\right)$, then stop for futility,

If $max\{ z_{min(1)}, z_{min(2)}\} \leq z_f \leq z_{1+f}$, then $q=1+f$,

If $z_1 \leq z_f \leq z_{1+f}$, then q is such as z_f verifies Equation [2] (a linear interpolation between 1 and $(1+f)$ may be an acceptable alternative for estimating q),

If $z_f \geq z_1$ then $q=1$.

Optimal time to perform the sample size reassessment

The LP criterion is used again when selecting the optimal information fraction f_{opt} for reassessment. For each possible value of f, we calculate (under Δ) the unconditional power $(1\text{-}\beta)_f$ and the average q_f . The optimal f_{opt} is then defined as the value f at which $[(1\text{-}\beta)_f - LP.q_f\,]$ is maximum. If ever this maximum value was less than $(1\text{-}\beta) - LP$ (the criterion calculated with the original power and q=1), it would mean that the fixed design is better than any adaptive one.

For usual values of α=0.0125, 0.025 and β=0.05, 0.10, 0.15, 0.20, we found that the optimal information fraction to perform the interim analysis is about 0.9.

4 Comparisons with methods proposed in the literature

4.1 Common constraints and considerations

Our pragmatic strategy was compared to two adaptive design methods of the literature (Cui, Hung, Wang, 1999; Denne, 2001) and to one adaptive method often proposed, based on the current trend. Comparisons were not easy because methods have different objectives and do not necessarily provide futility rules. Using such rules for our method and not for the other ones would have unfairly played in favour of our strategy. The same futility rules as in the pragmatic strategy were implemented to make all the methods comparable, i.e. $CP(z_f) \geq 50\%$ and $d_{obs} \geq 0$. Moreover, there is currently no golden standard for choosing the optimal information fraction for performing the reassessment. From force of habit, $f = 0.5$ is frequently chosen as a usual practice. Hence, $f = 0.5$ and our optimal $f_{opt} \approx 0.9$ were selected for the comparisons. Minimum and maximum sample sizes were also to be defined: we have chosen not to decrease the sample size and we imposed that it should not be more than 2 times larger than the original one (i.e. $N_{max}=2N$, or $q_{max}=2$).

4.2 Cui-Hung-Wang method

The Cui-Hung-Wang reassessment method is based on the ratio of the conditional power calculated under d_{obs} and under the original target:

If $CP(d_{obs})/CP(\Delta)<\gamma_I$ or $CP(d_{obs})/CP(\Delta)>\gamma_D$, a reassessment has to be done. Then,

$$N_{new} = \begin{cases} N.\left(\dfrac{\Delta}{d_{obs}}\right)^2, \text{if } N_{new} \leq N_{max} \\ N_{max}, \text{if } N.\left(\dfrac{\Delta}{d_{obs}}\right)^2 > N_{max} \end{cases}$$

Two parameters have to be defined: γ_I, γ_D . We chose $\gamma_I = 0.8$ and $\gamma_D=1$ which are the values selected in the Cui, Hung, Wang article.

4.3 Denne method vs d_{obs}-based method

The principle of the Denne method is to reassess the sample size in order to have a conditional power at least equal to the original power. Because originally, Denne does not propose either a maximum or a minimum sample size, and no futility bounds, we adapt his proposal according to the principles defined in 4.1. More precisely:

If $CP(N_{max}) < 50\%$ then stop for futility.

If $50\% \leq CP(N_{max}) \leq 1\text{-}\beta$ then $N_{new}=N_{max}$.

If $CP(N_{max}) \geq 1\text{-}\beta$ and $CP(N) \leq 1\text{-}\beta$ then $N_{new}=\{n \mid CP(n)= 1\text{-}\beta\}$.

If $CP(N) \geq 1\text{-}\beta$ then $N_{new}=N$.

Whereas the Denne method clearly claims that the conditional power should be calculated under the original target Δ, some statisticians advocate that the salvation can only come from using d_{obs} in this calculation. Therefore we experimented with the two scenarios, and called the second one "d_{obs}-based method". This second method leads to higher (always positive) futility bounds:

$$z_{min\ (d_{obs}\text{-based method})} = u_{\alpha}.\sqrt{f}\ /[\ f + \sqrt{(q_{max} - f\ ).(1 - f\ )}\ ]$$

4.4 Results

We chose to present comparisons (Fig. 1) in a very common setting: α=0.025 and β=0.10, with $N_{max}=2N$ (calculations combining also values α=0.0125 and β=0.20 show similar results). The times f for the interim analysis are f=0.5 and f=0.9. We notice that for f=0.9, all methods tend to provide better and closer results, so that our optimal f is favourable for all methods.

The Denne method provides results very close to the pragmatic strategy for both values of f. For very low values of the true δ, the pragmatic method uses nevertheless a little fewer patients.

The pragmatic strategy has lower power than Cui-Hung-Wang approach, but saves more patients when f=0.5. Choosing one of the two methods depends on the objectives of the trial (what is the cost/benefit ratio of showing delta values really lower than the targeted one ?).

The d_{obs}-based method requires fewer patients than ours but its power is lower than the power controlled by the fixed design when the true treatment effect is larger than 0.8Δ (when the interim analysis is performed at half of the data). The d_{obs}-based method futility bounds are larger than in the other strategies since the conditional power is computed under d_{obs} (for f=0.5, the futility bound is 1.01 – i.e. d_{obs}=0.44Δ – vs 0 for the other strategies); such a strategy leads clearly to stop many studies which would have been successful.

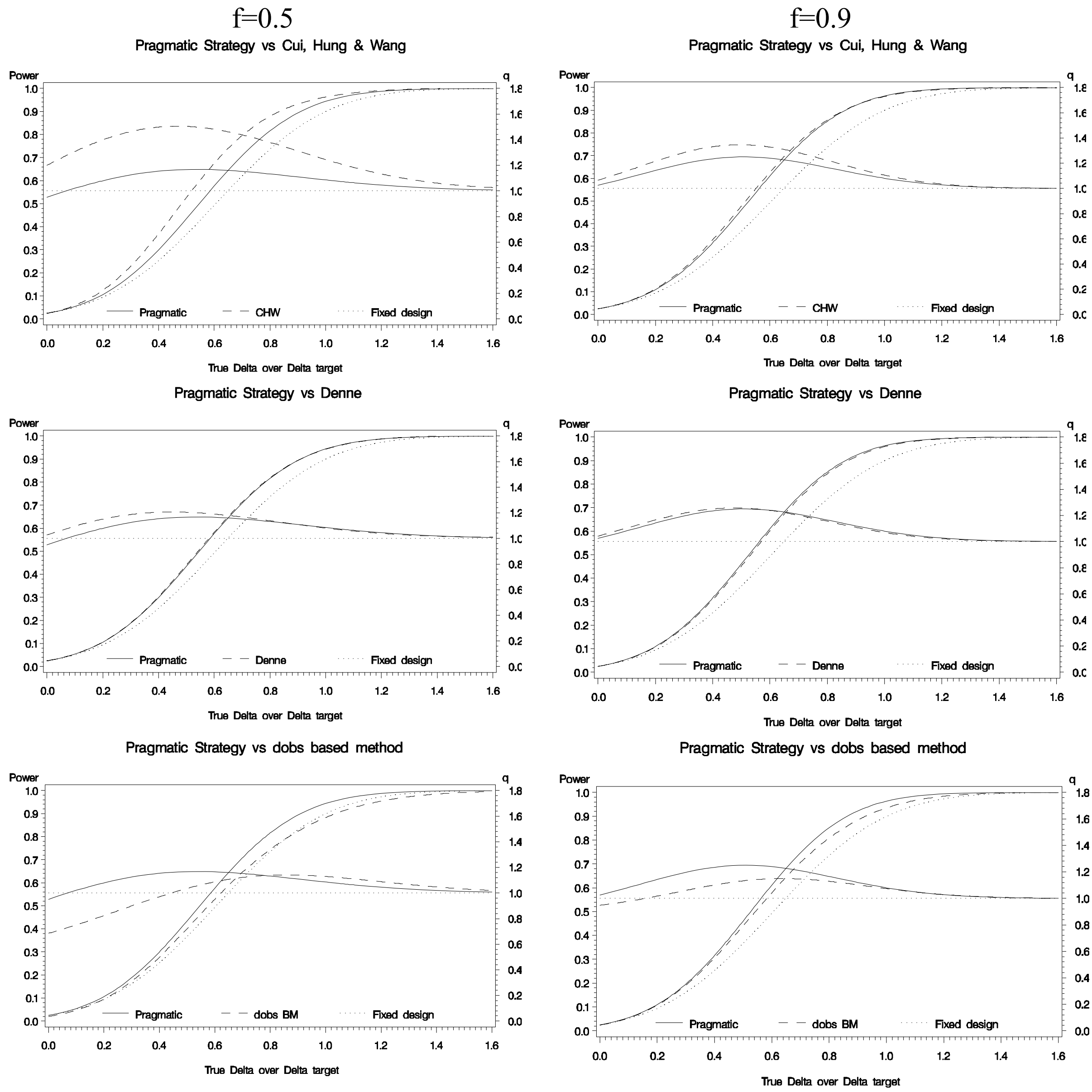

Fig. 1: Power and Standardized Average Sample Size q (N_{new}/N)
as a function of the ratio "True delta" / "Targeted delta" (δ/Δ)

5 Conclusion

Finally, it is shown that our pragmatic strategy can be a good solution in adaptive designs. It is also shown, more generally, that a sample size reassessment is more useful when performed close to the planned end of a trial, allowing a study with borderline interim results to be saved.

We thank gratefully Loïc Darchy for his help during our research, and Gérard Derzko for his support in the editorial work.

References

Cui L., Hung M.J. and Wang S.-J. (1999). Modification of sample size in group sequential clinical trial. *Biometrics*, **55**, 853–857.

Denne J.S. (2001). Sample size recalculation using conditional power. *Statistics in medicine*, **20**, 2645–2660.

Schäfer H. and Müller H.H. (2001). Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine*, **26**, 5422–5433.

Shun Z. (2001). Sample size reestimation in clinical trials. *Drug Information Journal*, **35**, 1409-1422.

Vandemeulebroecke M. (2006). An investigation of two-stages tests. *Statistica Sinica*, **16**, 933–951.