

Meeting the demand for more sophisticated study designs. A proposal for a new type of clinical trial: the hybrid design

Guoxing G Soon,¹ Lei Nie,¹ Thomas Hammerstrom,¹ Wen Zeng,¹ Haitao Chu²

To cite: Soon GG, Nie L, Hammerstrom T, *et al*. Meeting the demand for more sophisticated study designs. A proposal for a new type of clinical trial: the hybrid design. *BMJ Open* 2011;**2**:e000156. doi:10.1136/bmjopen-2011-000156

► Prepublication history for this paper is available online. To view these files please visit the journal online (<http://bmjopen.bmj.com>).

Received 3 May 2011
Accepted 8 August 2011

This final article is available for use under the terms of the Creative Commons Attribution Non-Commercial 2.0 Licence; see <http://bmjopen.bmj.com>

ABSTRACT

Background: Treatment effect is traditionally assessed through either superiority or non-inferiority clinical trials. Investigators may find that because of safety concerns and/or wide variability across strata of the superiority margin of active controls over placebo, neither a superiority nor a non-inferiority trial design is ethical or practical in some disease populations. Prior knowledge may allow and drive study designers to consider more sophisticated designs for a clinical trial.

Design: In this paper, the authors propose hybrid designs which may combine a superiority design in one subgroup with a non-inferiority design in another subgroup or combine designs with different control regimens in different subgroups in one trial when a uniform design is unethical or impractical. The authors show how the hybrid design can be planned and how inferences can be made. Through two examples, the authors illustrate the scenarios where hybrid designs are useful while the conventional designs are not preferable.

Conclusion: The hybrid design is a useful alternative to current superiority and non-inferiority designs.

INTRODUCTION

Treatment-effect size is traditionally assessed through either superiority or non-inferiority randomised clinical trials. However, because of safety concerns and/or wide variability across strata of the superiority margin of active controls over placebo, neither a superiority nor a non-inferiority trial design is ethical or practical in some disease populations. For example, in the development of a direct-acting antiviral agent (DAA) for hepatitis C virus (HCV), the phase 3 development processes involved one trial in subjects who are treatment-naïve and one in subjects who are treatment-experienced. Each was a superiority trial at a two-sided significance level of 0.05 comparing the DAA with Pegylated-interferon plus Ribavirin.

ARTICLE SUMMARY

Article focus

- We propose hybrid designs for the trials when neither a superiority nor a non-inferiority trial design is ethical and practical.

Key messages

- The hybrid design is practical, flexible and feasible.
- We expect it to become a major alternative to the superiority and non-inferiority designs.

Strengths and limitations of this study

- Hybrid design provides a powerful and relatively simple solution to the difficult problem of active controls with varying efficacy and/or safety concern. The problem is becoming more common as more drugs become available.
- The design and analysis are moderately complex compared with the superiority and non-inferiority designs.

However, the trial for the treatment-experienced population is unethical because the previous null responders of Pegylated-interferon plus Ribavirin are unlikely to benefit from the second course of the same treatment and are likely to suffer from its side-effects. See next section for more details.

One possible solution is to conduct multiple trials with each trial focusing on one subgroup. In the previous example, this would require three trials instead of two: one trial in subjects who are treatment-naïve, the second trial in subjects who are previous null responders and the third trial in subjects who are partial responders or responder relapses (see next section for a clear definition for these terms). Traditional practice would be to conduct each trial at a two-sided significance level of 0.05.

The above design, referred to as a conventional design throughout the paper, consists of multiple traditional trials, each for one

¹Division of Biometrics IV, Office of Biostatistics/OTS/CDER/FDA, Silver Spring, Maryland, USA

²Division of Biostatistics, University of Minnesota at Twin Cities, Minneapolis, Minnesota, USA

Correspondence to

Dr Lei Nie;
lei.nie@fda.hhs.gov

subgroup. This conventional design is often neither necessary nor practical. It is unnecessary because it requires three trials in a situation where the regulatory practice is to require only two trials for drug approval (see next section for details). In fact, the ethic and feasibility problem shall not change this regulatory perspective of requiring of two trials. The major practical problem associated with this conventional design is its requirement of a sufficiently large sample size to power each of the multiple trials to a designated level (eg, a power of 80%) and, hence, an increased length of recruitment and cost. This is a particular problem when one of the trials requires exclusively recruitment from some minority subgroups.

For the above practical reasons, it is common in practice that trials enrol subjects with different reactions to treatments into the same trial and aim to demonstrate its overall efficacy. We give two examples. In example 1, subjects with or without lung disease react differently to Palivizumab, a prophylaxis in reduction in respiratory syncytial virus infection in high-risk infants.¹ In a recent Palivizumab-controlled non-inferiority trial to develop Motavizumab, a second generation of Palivizumab, subjects with different lung-disease statuses were enrolled into one trial.² In example 2, it is known that HIV-infected subjects with different phenotypic sensitivity scores (PSS) react differently to treatments; they were all enrolled into each of BENCHMRK-1 and BENCHMRK-2 to evaluate Raltegravir.^{3 4}

Consequently, the conventional design would often be undesirable. As part of efforts to improve drug development, we propose a hybrid design, for situations where the active controls have clinically significantly different margins of superiority to placebo in different strata of the population. In some of these situations, a simple design may be unethical; in others, a simple superiority design may require unrealistically potent new drugs to beat the active control in strata where the latter is most effective, while a simple non-inferiority design may require unrealistically narrow non-inferiority margins relative to the active control in strata where the latter is least effective. The hybrid design combines superiority designs for some subgroups and non-inferiority designs for other groups in one trial. Furthermore, the non-inferiority designs used in a hybrid design can be different with different active controls or different non-inferiority margins in different strata of the population. Consequently, in a hybrid design, different designs are allowed for different subgroups, while the intended inference for the new treatment is for the whole population in the trial.

The concept of hybrid design generally refers to new designs that are formed by combining features of established designs. In this paper, the hybrid design refers to a synthesis of a number of superiority and/or non-inferiority designs (hypotheses) for different disjoint subgroups in one trial. This concept has also been used in the area of epidemiology to refer to

a multilevel design which takes advantages of ecological and analytic studies when examining the association between exposure and a disease.⁵ It has also been used in the area of social science to refer to new designs, for example, the Solomon Four-Group Design,⁶ formed by combining existing experimental designs. In the area of multifactor experimental design for exploring response surface, hybrid designs, considered as alternatives to central composite designs, refer to a class of saturated or near-saturated second-order designs that allow experimenters to fit a quadratic model with a minimal or near-minimal number of runs.⁷

We utilise a Fisher combination test to make an overall inference from all subgroups. In addition, we propose a general method for sample-size determination and power calculation in a hybrid design. We demonstrated its usefulness and advantages over the conventional design. The hybrid design is naturally related to meta-analysis, which generally focuses on the data-analysis stage rather than the design stage.

MOTIVATION EXAMPLES

Approval of new drugs requires either two randomised clinical trials in which the new drug demonstrates efficacy at a two-sided level 0.05 or one larger trial in which efficacy is demonstrated at a level comparable with two trials each at 0.05.

The US FDA generally recommends such a development programme to demonstrate efficacy of the new drug in the broadest possible population. In several drug areas, for example, HCV and HIV, the current phase 3 development processes involve one trial in subjects who are treatment-naïve and another in subjects who are treatment-experienced. Both trials are active controlled trials.

In the treatment-naïve population, the active control will elicit an excellent response, and a non-inferiority trial will be conducted. Within the treatment-experienced population, it is not uncommon in our experience for there to be baseline covariates which identify strata with substantially different response rates on the active control. In such circumstances, the natural method of study may well be to conduct the conventional design consisting of multiple trials. Each trial targets one stratum. One would use a non-inferiority design in one trial and a superiority design in the other trial.

There is one practical problem with this conventional design: instead of the expected two phase 3 trials, one has three such trials. Because there is already one phase 3 trial showing efficacy at a two-sided level of 0.05 in the treatment-naïve population, the US FDA would approve a drug in which the evidence of efficacy from one treatment-experienced trial is 0.05. Therefore, sponsors (pharmaceutical companies) are reluctant to design development programmes using the conventional design which requires more than two phase 3 trials. Designing multiple trials to achieve a level of 0.05

demands more subjects and expense than is actually required to meet regulatory standards. Consequently, sponsors often design either one non-inferiority trial or one superiority trial in a population containing all strata where the active control is potent and the stratum where it is weak, even though neither design is truly appropriate to both strata. Below, we present two examples to illustrate the major problems of this practice.

Developing DAA for HCV

Pegylated-interferon plus Ribavirin was the standard of care (SOC) for HCV-infected subjects. However, SOC yields to a low response rate for infected subjects with HCV genotype 1.⁸ In addition, its side-effects affect almost all treated subjects.⁹ Concerns about the tolerability and efficacy of SOC have led to recent development of the DAA.

It was recommended that a superiority design be used to demonstrate the superiority of the DAA as an add-on to SOC comparison to placebo plus SOC.¹⁰ The subjects to be enrolled are HCV-treatment-naïve subjects (subjects received no prior therapy for HCV) or SOC-experienced subjects including null responders (subjects had a reduction in HCV RNA of $<2 \log_{10}$ at week 12 of the previous SOC), partial responders (subjects had a reduction in HCV RNA of $\geq 2 \log_{10}$ at week 12 of the previous SOC, but not achieving HCV RNA undetectable at the end of treatment with an SOC), and responder relapsers (subjects had HCV RNA undetectable at the end of treatment with an SOC, but HCV RNA detectable within 24 weeks of SOC treatment follow-up).

Following this guidance, some sponsors proposed the development of DAA using one trial for the treatment-naïve population and another for the treatment-experienced population. However, some regulatory reviewers have raised questions regarding the potential ethical problems. They found that the above recommended design is feasible for HCV-treatment-naïve subjects, partial responders and responder relapsers, but it is unethical for null responders because there is insufficient benefit with SOC¹¹ and because of safety concerns.

The conventional design may suggest separating null responders from the second trial and conducting a third single arm, a historical controlled trial for the null responders. The objective is to (1) demonstrate that the new treatment is efficacious in the trial for partial responders, and responder relapsers with a statistical significance at a level of 0.05; and (2) demonstrate that the new treatment is efficacious in the trial for null responders with a statistical significance at a level of 0.05.

This conventional design changed the original objective. Should there be no ethic concerns, the objective was to demonstrate that the new treatment is efficacious in the treatment-experienced population at a level of 0.05, with no need to show statistical significance in each of the two strata at level of 0.05. As we stated earlier, designing all three trials to achieve a level of 0.05 demands more subjects and expense than is actually required to meet regulatory standards.

As a new and alternative solution to the trial for subjects who are treatment-experienced, we propose using different designs for null responders and others (partial responders and responder relapsers) in one single trial. For example, we can enrol all null responders to the single-arm DAA plus SOC and enrol all others following the recommended design.¹⁰ For null responders, the design is to demonstrate that the new treatment is superior to the SOC based on historical experience on the response rate of SOC. Consequently, we used two different designs within the same trial according to their response status to the previous SOC treatment. The objective remains to demonstrate that the new treatment is efficacious in the trial for all treatment-experienced subjects with a statistical significance at a level of 0.05.

Developing new treatment for HIV

Subjects with HIV infection are usually treated with combinations of three or four drugs from among several classes: nucleoside reverse transcriptase inhibitors (NRTI), non-nucleoside reverse transcriptase inhibitors (NNRTIs), protease inhibitors (PIs), integrase, fusion or entry inhibitors (IIs, FIs, EIs) to increase the duration of viral suppression and final cure rate, and to reduce the risk of viral-resistance development.

Although many HIV drugs are currently available, development of new potent treatment from any inhibitor classes is urgently needed to overcome the increased genetic barrier of drug resistance. HIV viruses mutate rapidly, and some mutations lead to drug resistance of current HIV treatments. HIV's short life-cycle and high error rate cause the virus to mutate very rapidly, making it one of the hardest viruses to combat.

The typical design for evaluating a new treatment is to demonstrate the non-inferiority of the new treatment to an approved active control as an add-on to optimal background therapy (OBT) of several drugs. For example, an evaluation of a new NNRTI, such as Rilpivirine for antiviral-naïve subjects was carried out using a non-inferiority trial comparing the new treatment with an approved NNRTI, Efavirenz as an add-on to two NRTIs (eg, Tenofovir and Emtricitabine).

Consider a new integrase inhibitor tested against Raltegravir. The phase 3 development processes involves one trial in subjects who are treatment-naïve and a second in subjects who are treatment-experienced. Both trials are Raltegravir-controlled non-inferiority trials. The trial in subjects who are treatment-naïve is standard, so we focus on the second trial.

In order to define the non-inferiority margin in subjects who are treatment-experienced, we need to quantify the treatment benefit of Raltegravir relative to the placebo, as an add-on to OBT. Among treatment-experienced subjects, the subject's virus may be resistant to many or all candidate drugs for the OBT. The PSS is a measure of the number of drugs in an OBT to which the subject's virus is not yet resistant at baseline and is a good predictor of the duration of viral suppression by

the regimen. As expected, the benefits of Raltegravir over the placebo, as an add-on to an OBT, observed in the BENCHMRK trials varied in subjects with different PSS. The response-rate difference between Raltegravir and placebo changes from 49% in subjects with 0 PSS, to 32% in subjects with PSS between 1 and 2, to 10% in subjects with PSS ≥ 3 .^{3 4 12}

To illustrate the difficulty in conducting a single non-inferiority trial for the treatment-experienced subjects, we reproduced the subgroup analysis by PSS in table 1, based on the results presented by Cooper and colleagues.³

The overall response rates are 67% and 33% for Raltegravir and placebo, respectively, as an add-on to OBT. The treatment benefit of Raltegravir is 34% with a 95% CI of 26–42%. A fixed margin would define 13%, half of the low bound of the CI, as the non-inferiority margin. This margin seems to be too conservative for subjects with low PSS and too liberal for subjects with high PSS. After all, when such a high level of heterogeneity occurs, an overall treatment effect may not be relevant. Furthermore, if the trial has a high percentage of subjects with high PSS, the non-inferiority margin is inappropriate. Note that the percentage of subjects with high PSS is usually unknown at the design stage, which complicates the situation further. Consequently, designing a meaningful non-inferiority trial may be difficult.

Similar to the results presented in table 1, the response rate of OBT is generally affected by the PSS scores.¹² For subjects with low PSS, the response rate of OBT is generally low. For example, the response rate of OBT is <10% in subjects with 0 PSS. On the other hand, for subjects with high PSS, the response rate of OBT is generally high. For example, the response rate is more than 40% in subjects with PSS ≥ 3 .¹²

Based on the previously described information, we can design the trial based on our existing knowledge about HIV drugs and patient populations. We learnt that the OBT works well for subjects with PSS greater than 2, for example, and works less well for other subjects. Hence, we may consider the hybrid design. In one single trial, we evaluated the superiority of the new treatment as an add-on to OBT to placebo as a matched add-on to OBT in subjects with PSS ≥ 2 , and evaluated the non-inferiority of new treatment as to Raltegravir, as an add-on to OBT in subjects with PSS smaller than 2. For the superiority hypothesis part, as the OBT leads to a high response rate in subjects with high PSS, the superiority

trial design for these subjects is ethical. For the non-inferiority hypothesis part, because the difference in response rates in Raltegravir and placebo is 0.39 with an SE of 0.05, a statistical non-inferiority margin can be defined, for example, as 15%, a half of the lower bound of a 95% CI (0.30 to 0.49). Note that a superiority trial for subjects with low PSS is unethical, because OBT leads to a poor response for these subjects.

The conventional design, which involves two separate trials for treatment-experienced subjects with high and low PSS, requires many more subjects than the newly proposed hybrid design does. As we explained earlier, the sponsors are reluctant to take the conventional approach.

HYBRID DESIGN

We expect that more trials similar to those discussed above cannot be simply implemented as either simple superiority trials or simple non-inferiority trials owing to advanced prior knowledge—for example, side-effects and heterogeneity of efficacy margins of active control drugs within different subpopulations. We propose a hybrid design as an alternative. The hybrid design allows each subgroup to use a different design.

In general, we assume that there are k disjoint subgroups in a clinical trial. For these subgroups, k different designs addressing different hypotheses are used as components of a hybrid design. These designs could be superiority, non-inferiority designs, or historical controlled designs. Alternatively, they could all be superiority or non-inferiority designs but with different controls.

If the goal is to demonstrate the experimental treatment's superiority over the placebo in the i th subgroup, the i th null hypothesis is formatted as $H_{0i}: \theta_i=0$, where θ_i is the metric measuring the difference between the new treatment and the control for $i = 1, 2, \dots, k$. Note that the $H_{1i}: \theta_i>0$ implies the experimental treatment's superiority over the placebo. Therefore, for ease of presentation and without loss of generality, we consider here only a placebo control superiority design. If the goal is to demonstrate the non-inferiority of the experimental treatment relative to the active control in the i th subgroup, the i th null hypothesis is formatted as $H_{0i}: \theta_i=-\delta_i$, where θ_i is the metric measuring the difference between the new treatment and the control and $-\delta_i$ is the non-inferiority margin, that is a positive value of $\theta_i+\delta_i$ implies the non-inferiority of the new treatment over the active control. Because the non-inferiority margin δ is

Table 1 Subgroup analysis by phenotypic sensitivity scores (PSS)

	Treatment	Placebo	Difference (SE)
PSS=0	33/65 (51%)	1/44 (2%)	49% (0.066)
PSS=1	83/137 (61%)	20/69 (29%)	32% (0.069)
PSS=2	99/139 (71%)	24/62 (39%)	33% (0.073)
PSS=3	58/82 (71%)	28/46 (61%)	10% (0.088)

defined as the least difference with high confidence between the active control and the placebo measured in historical trials, a positive value of $\theta_i + \delta_i$ also implies the superiority of the new treatment over the placebo.

As a summary, although individual hypotheses, either superiority hypothesis or non-inferiority hypotheses, are different in all k subgroups, these hypothesis components share the same objective to demonstrate that an experimental treatment is better than a placebo. That is, the primary objective is to test the null hypothesis, $H_0: \theta = 0$, where θ is a metric measuring the difference between the new treatment and the placebo.

A p value was obtained for each subgroup. For a subgroup conducted as a superiority trial, a p value is obtained, considering the subgroup as a separate trial. For a subgroup conducted as a non-inferiority trial, a p value is obtained using the testing method described in, for example, Wang *et al.*¹³ again considering the subgroup as a separate trial. For a subgroup conducted as a historical controlled trial, a p value can be similarly obtained.

Similar to the meta-analysis, the Fisher combination test¹⁴ can be utilised to make statistical inferences for the hybrid designs. The Fisher combination test, which is based on the production of p values, was originally proposed to combine evidence from separate studies. It applies to analysing the hybrid design naturally.

In a trial where a hybrid design was used, a p value p_i was obtained for the i th subgroup. Some or all of them may not be significant at a prespecified level—for example, 0.05. This is expected because substudies are not powered to demonstrate the significance. In a hybrid design, we only want to determine whether the combined evidence reaches the prespecified significance level, using the Fisher combination test.¹⁴

The Fisher combination test was introduced to combine evidence from different subgroups and make inferences¹⁴ as follows. The null hypothesis that the experimental treatment is the same as the placebo, is rejected at a level α if

$$\prod_{i=1}^k p_i \leq \exp \left\{ -\frac{1}{2} \chi_{2k}^2 (1 - \alpha) \right\},$$

where $\chi_{2k}^2 (1 - \alpha)$ denotes the $(1 - \alpha)$ -quantile of the χ^2 distribution with $2k$ degrees of freedom.

The advantage of the Fisher combination test is that it relies only on p values of individual tests. The disadvantage of this test is as follows. The Fisher combination test does not quantify the treatment difference between treatment groups. In addition, the Fisher combination test may suffer a loss of power when the sample sizes are substantially unbalanced over treatments.¹⁵ While the Fisher test is recommended, some alternative ways^{16–20} can also be similarly modified to make inferences in the hybrid design. Further research may be needed to compare different methodologies for inference, which is beyond the scope of this paper.

POWER CALCULATION

Sample size determination can be based on the Fisher combination test. For the determination based on the Fisher combination test, we choose significance levels $\alpha_1, \dots, \alpha_k$ and n_1, \dots, n_k for individual subgroups and so that $\prod_{i=1}^k \alpha_i \leq C_\alpha$, where $C_\alpha = \exp \left\{ -\frac{1}{2} \chi_{2k}^2 (1 - \alpha) \right\}$, and α is the prespecified type I error rate.

The computation of power can be done based on the Fisher combination test through the following method. Let η_i be the test statistics used in the i th subgroup; Let $f_{H_{i0}}^{(i)}(\eta_i)$ and $F_{H_{i0}}^{(i)}(\eta_i)$ be the density function and cumulative distribution function of η_i , respectively, when the true values of parameters are those values given in the null hypothesis. Similarly, let $f_{H_{i1}}^{(i)}(\eta_i)$ and $F_{H_{i1}}^{(i)}(\eta_i)$ be the density function and cumulative distribution function of η_i when the true values of parameters are those values given in the alternative hypothesis. It is easy to show that the power can be computed through the following algorithm,

- ▶ Randomly generate $l=1, \dots, m$ independent samples of $\eta_1^{(l)}, \dots, \eta_k^{(l)}$ are using their distribution density function $f_{H_{i1}}^{(i)}(\eta_1), \dots, f_{H_{i1}}^{(k)}(\eta_k)$;
- ▶ For all $l=1, \dots, m$, compute $\prod_{i=1}^k \{1 - F_{H_{i0}}^{(i)}(\eta_i^{(l)})\} \leq C_\alpha$. The l th sample is considered as a success if $\prod_{i=1}^k \{1 - F_{H_{i0}}^{(i)}(\eta_i^{(l)})\} \leq C_\alpha$.
- ▶ The power is the percentage of samples which are success among m samples.

Implementation of this algorithm is simple. For example, we implement it in SAS, and the computation is fast even when k is relatively large. The numerical computation of power can also be carried out based on the Fisher combination test using the method described by Banik *et al.*¹⁵ In their method, the power is calculated through multiple integration. The drawback of this method is its computational challenge. The integration can be difficult, so it is impractical when k is not small.

We propose determining the sample based on the Fisher combination test for its simplicity, although other methods could also be feasible. We believe the sample size determination based on Fisher's combination test is generally reasonable. As was suggested previously,¹⁵ we expect the loss of power to be small compared with the optimal test, unless the sample sizes are substantially unbalanced over treatments, which should not appear in well-controlled clinical trials. The key is, in the planning stage, we typically try to avoid any substantial imbalance.

NUMERICAL STUDY

In this section, we consider an experiment to support a new DAA for the treatment of hepatitis C in treatment-experienced subjects. A conventional design of the experiment consists of two separate trials: a historical controlled single-arm trial for null responders and an SOC-controlled superiority trial of the DAA for partial responders and responder relapsers. We assume that the randomisation in the second trial is 1:1. In the hybrid design, we combine these two trials into one, using the

Table 2 Sample size for conventional and hybrid design

Assumed response rate for null responders		Assumed response rate partial responders or relapsers		Sample size	
Direct-acting antiviral agent (%)	Standard of care (historical) (%)	Direct-acting antiviral agent (%)	Standard of care (%)	Conventional design	Hybrid design
30	10	40	20	213=33+180	117=19+98
30	8	40	15	137=25+112	76=14+62
20	10	30	20	724=100+624	416=56+360
20	8	30	15	331=67+264	187=39+148

same controls and same superiority or non-inferiority comparison as in the conventional design.

We compare the total sample size needed for the hybrid design with that for the conventional design. For the hybrid design, the sample size is calculated so that the whole study has a power of 80% and a two-sided type I error rate of 0.05. On the other hand, for the conventional design, the sample size is calculated so that each trial has a power of 80% and a two-sided type I error rate of 0.05. The sample size for the conventional design is implemented in the SAS procedure PROC POWER, based on the Fisher exact test.

The results are given in table 2. As we expect, and as the results clearly indicate, the hybrid design typically requires a significantly smaller sample size. Let $p_{0,DAA}$ be the response rate for the null responses receiving DAA; let $p_{0,SOC}$ be the historical response rate of null responders received SOC treatments; let p_{DAA} and p_{SOC} be the assumed response rate for the partial responses and relapsers receiving DAA and SOC treatments respectively. When the $p_{0,DAA} = 30\%$, $p_{0,SOC} = 10\%$, and $p_{DAA} = 40\%$, $p_{SOC} = 20\%$, the conventional design suggests enrolling 33 subjects to the single-arm historical controlled trial for null responders, and 90 subjects to each treatment group in the active-controlled trial for partial responders and responder relapsers, so that the power of each trial is 80%. On the other hand, the hybrid design proposes enrolling 19 null responders and 49 partial responders and responder relapsers in one trial, so that the power of the trial could have a power of 80%. The result is consistent for all other cases evaluated. The hybrid design could be more economic, although the conventional design, if implemented, will generally provide more information about each subgroup. We believe that the hybrid is a useful alternative to conventional designs. Also refer to figure 1 for a visual comparison of two designs for

the case $p_{0,DAA} = 30\%$, $p_{0,SOC} = 8\%$, and $p_{DAA} = 40\%$, $p_{SOC} = 15\%$, table 2.

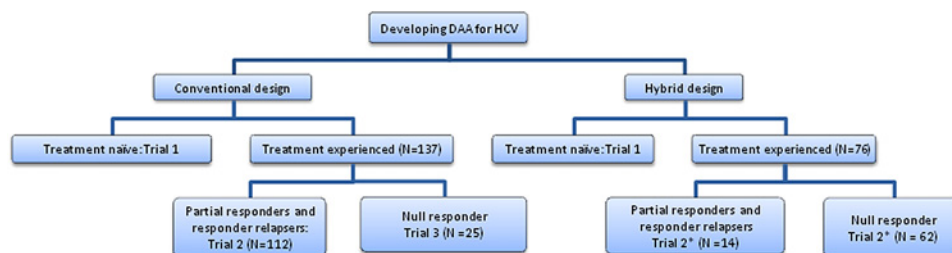
DISCUSSION

The hybrid design can be a good alternative to an active-controlled non-inferiority trial when the treatment heterogeneity of the active control has been observed in the phase 3 trial leading to the approval of that drug. Further problems with the non-inferiority margin can arise when the population of the planned active-control non-inferiority trial is different from that in the older trials comparing the active control with placebo. Nie and Soon attempted to solve the problem, but their method itself was complex, involving model fitting and model selection. Furthermore, the implementation of sample-size determination and power could be challenging, as Nie and Soon noted.²¹ The hybrid design offers a valuable alternative to simple non-inferiority trials. The latter only allows the same control arm and same non-inferiority margin in all strata.

Compared with a conventional design where separate trials were conducted for subgroups, the hybrid design can yield sufficient evidence of efficacy for regulatory approval with a smaller sample size. In addition, recruiting different subgroups for separate trials results in unnecessary overhead costs as well as practical difficulties. For example, sensitivity scores for the optimal background regimen in treatment-experienced subjects will not be known until a substantial amount of screening has been carried out, even though once the OBT has been selected and characterised, the sensitivity scores constitute a stratum for randomisation. The evaluation on these subgroups can be carried out in the hybrid design.

Cohort splitting has often been used in social science, medical topics and observational studies, but the hybrid design, which combines different designs into one trial,

Figure 1 Conventional design versus hybrid design to develop direct-acting antiviral agent (DAA). HCV, hepatitis C virus.



has never been introduced to clinical trial to date. Hybrid design is a combination of experimental design and analysis plan. Cohort-splitting studies most commonly compare efficacy in studies using the same endpoint—for example, superiority to placebo as measured in head-to-head comparisons in different cohorts. On the other hand, hybrid designs are intended to permit determinations of efficacy using different endpoints in different strata—for example, direct superiority to placebo in one stratum and non-inferiority to an active control in a different stratum.

As one referee pointed out, a CI approach could be more informative than the currently proposed p value approach. While further research on the CI approach is warranted, it is beyond the scope of this paper.

Acknowledgements The authors greatly appreciate the excellent comments and suggestions from Drs SA Julious, G Piaggio, M Sydes on earlier versions of the paper. The comments and suggestions significantly improved the quality of our paper.

Disclaimer Views expressed in this paper are the authors' professional opinions and should not be construed to represent FDA's view or policies.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None.

Contributors LN and GS initiated the paper. LN drafted the paper, and GS contributed to the conception and design of the paper. TH, WZ and HC participated in its design and helped draft the manuscript. LN, TH, GS and HT made major contributions in the revision. All authors approved the final version of the paper.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All data presented is freely available.

REFERENCES

1. The IMpact-RSV Study Group. Palivizumab, a humanized respiratory syncytial virus monoclonal antibody, reduces hospitalization from respiratory syncytial virus infection in high-risk infants. *Pediatrics* 1998;102:531–7.
2. Carbonell-Estrany X, Simoes EA, Dagan R, *et al.* Motavizumab for prophylaxis of respiratory syncytial virus in high-risk children: a noninferiority trial. *Pediatrics* 2010;125:e35–51.
3. Cooper DA, Steigbigel RT, Gatell JM, *et al.* Subgroup and resistance analyses of raltegravir for resistant HIV-1 infection. *N Engl J Med* 2008;359:355–65.
4. Steigbigel RT, Cooper DA, Kumar PN, *et al.* Raltegravir with optimized background therapy for resistant HIV-1 infection. *N Engl J Med* 2008;359:339–54.
5. Navidi W, Thomas D, Stram D, *et al.* Design and analysis of multilevel analytic studies with applications to a study of air pollution. *Environ Health Perspect* 1994;102(Suppl 8):25–32.
6. Solomon RL. An extension of control group design. *Psychol Bull* 1949;46:137–50.
7. Roquemore KG. Hybrid designs for quadratic response surfaces. *Technometrics* 1976;18:419–23.
8. Parfieniuk A, Jaroszewicz J, Flisiak R. Specifically targeted antiviral therapy for hepatitis C virus. *World J Gastroenterol* 2007;13:5673–81.
9. Fried MW. Side effects of therapy of hepatitis C and their management. *Hepatology* 2002;36(5 Suppl 1):S237–44.
10. FDA. *Guidance for Industry Chronic Hepatitis C Virus Infection: Developing Direct-Acting Antiviral Agents for Treatment*, FDA, Silver Spring, 2010.
11. Cheruvattath R, Rosati MJ, Gautam M, *et al.* Pegylated interferon and ribavirin failures: is retreatment an option? *Dig Dis Sci* 2007;52:732–6.
12. Struble K. Emerging Issues in HIV Clinical Trials for new ARV. *Clinical Trials for New ARVs Roundtable*, Washington, DC, 2010.
13. Wang SJ, Hung HMJ, Tsong Y. Utility and pitfalls of some statistical methods in active controlled clinical trials. *Control Clin Trials* 2002;23:15–28.
14. Fisher RA. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd, 1932.
15. Banik N, Kohne K, Bauer P. On the power of Fisher's combination test for two stage sampling in the presence of nuisance parameters. *Biom J* 1996;38:25–37.
16. Brannath W, Posch M, Bauer P. Recursive combination tests. *J Am Stat Assoc* 2002;97:236–44.
17. Burman CF, Sonesson C. Are flexible designs sound? *Biometrics* 2006;62:664–9; discussion 70–83.
18. Cui L, Hung HM, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* 1999;55:853–7.
19. Lehman W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999;55:1286–90.
20. Whitehead A. *Meta-analysis of Controlled Clinical Trials*. Hoboken, NJ: Wiley, 2002.
21. Nie L, Soon G. A covariate-adjustment regression model approach to noninferiority margin definition. *Stat Med* 2010;29:1107–13.