

Reforming Subgroup Analysis

Anup Malani, Oliver Bembom and Mark van der Laan¹

Different people respond differently to drugs. Statisticians refer to this as heterogeneity in treatment response. For example, diclofenac sodium plus misoprostol (marketed as Anthrotec by Pfizer) is an effective treatment for osteoarthritis for patients who develop ulcers when using non-steroidal anti-inflammatory drugs (NSAIDs). But misoprostol is also documented to induce labor and used for medical elective abortion.² Therefore, while Anthrotec is generally effective for pain relief, it is contraindicated³ for pregnant women because of its abortifacient effects.

Or consider the case of the Phase III trial for motexafin gadolinium, which is sponsored by Pharmacyclics as Xcytrin and which we use as an example in this article. In that study of a new treatment for lung cancer patients with brain metastases, the average patient did not demonstrate benefit from treatment that was significant at the traditional confidence level of 0.05. However, this result is driven by patients who were previously treated for the primary and metastasized tumor. Among the patients whose primary tumor was still uncontrolled, the treatment reduced the median time till neurologic progression by more than a factor of two with a p-value of 0.002.

Nevertheless, the U.S. Food and Drug Agency (FDA) usually considers only average treatment effects when deciding whether to approve a drug. Given heterogeneity in treatment response, this approach can result in the approval of drugs with significant negative effects for identifiable subgroups (false positives) and in the non-approval of drugs with significant positive effects for identifiable subgroups (false negatives).

The FDA is not entirely deaf to these concerns. If a drug is approved based on average performance, *post hoc* subgroup analysis⁴ can be employed to justify a label that warns of

¹ University of Chicago, University of California-Berkeley, and University of California Berkeley, respectively. This paper was prepared for the AEI conference, “Oncology Drug Development: Rethinking FDA Oversight,” March 13-14, 2008. We thank Richard Miller of Pharmacyclics for access to the data from the Phase III Xcytrin trial. In the interest of full disclosure, two of this papers authors, Bembom and van der Laan work for Target Analytics, Inc., which was hired by Pharmacyclics to prepare a statistical analysis of the data from the Xcytrin trial.

² See Goldberg et al., Misoprostol and Pregnancy, 34(1) New Eng. J. Med. 38 (Jan. 4, 2001).

³ Arthrotec product insert at 1, available at FDA, Label and Approval History – Arthrotec, Labeling revision, Aug. 24, 2007 (<http://www.fda.gov/cder/foi/label/2007/020607s010lbl.pdf>) (checked Mar. 7, 2008).

⁴ Subgroup analysis is defined as statistical analysis that establishes the effect of a drug with one or more subgroups of the sample defined by baseline characteristics. *Post hoc* subgroup analysis is defined as subgroup analysis in the

negative effects for specified subgroups. Because the drug's sponsor has a financial conflict of interest, however, the FDA must conduct this analysis by itself or use an outside consultant.

To address false negatives, the FDA allows the sponsor to specify, before a Phase III trial, one or more subgroups among the target group for which it plans to undertake subgroup analysis. The sponsor may not, however, have enough information prior to Phase III trials to identify especially sensitive subgroups. Moreover, the FDA would require the sponsor to increase sample size so that its study is powered up for subgroup analysis. This additional cost may outstrip the financial resources available to many sponsors. Of course nothing stops the sponsor from conducting *post hoc* subgroup analysis following a trial that is not powered for that analysis and asking the FDA for permission to conduct a follow-up trial on that subgroup. But this approach is even more costly.

While sponsors may complain that the FDA's position imposes too high a cost on exploiting heterogeneity in treatment response, the agency's approach is not irrational. *Post hoc* subgroup analysis increases the risk of approving drugs that have no net beneficial effect. The more subgroups the sponsor analyzes, the more likely it is to find one that appears to benefit even if in fact there exists no subgroup that benefits. The sponsor has a financial incentive to ignore this risk. If the FDA knows the number of subgroups the sponsor has sampled in search of a positive response, the FDA can limit the false positives by employing multiple testing corrections. But the FDA will rarely be able to verify this number.

This paper asks whether there is a better way. In particular, we wonder if there is a process that allows approval based on *post hoc* subgroup without the cost of unnecessary additional trials or the risks of opportunistic behavior and spurious correlation. Based on our current statistical understanding, we suggest two compromises to achieve this result.

The FDA's current position is that whenever possible sponsors specify *a priori* the subgroups they plan to study before conducting a Phase III trial. However, there will be cases where sensitive subgroups are not known *a priori*. Our first proposal is that these subgroups be identified by use of an adaptive group sequential design trial. In standard trials, subjects are typically assigned to the treatment or control group according to a pre-set randomization scheme and remain in their assigned groups. In an adaptive design, new subjects can be randomized to treatment or control based on the baseline characteristics or the performance of subjects already enrolled in the trial. The goal is to identify subgroups based on data from early enrollees in order to power up analysis of those subgroups among later enrollees. Adaptive trials may require a larger sample size than standard trials, but they do not require as many total subjects as performing a subsequent trial based on *post hoc* analysis following an initial, failed trial.

case where the subgroups were not *a priori* specified in the statistical analysis plan submitted to the FDA prior to conducting a trial.

Our second proposal is that, in certain situations, it may be permissible to approve a drug on the basis of *post hoc* analysis if that analysis were done in a manner that eliminated the incentive for data dredging. For example, the analysis could be conducted by an independent consultant rather than the sponsor. To address concerns about whether the consultants is truly disinterested, we suggest two statistical algorithms that help the consultant generate subgroups to be analyzed by the sponsor. One algorithm provides the consultant access only to the trial data stripped of information on subjects' outcomes. The other algorithm gives the consultant access to all variables, but only for a portion of the trial sample. It permits the sponsor to submit subgroup results to the FDA only for the remainder of the trial sample. Either algorithm ensures that selection of subgroups is orthogonal to post hoc subgroup analysis by the sponsor.

To motivate the potential benefit of increased flexibility in the drug approval process and to demonstrate how our solutions to the data manipulation problem might be implemented we conduct a case study of Pharmacyclics' Xcytrin, a drug for the treatment of non-small cell lung cancer (NSCLC) patients with brain metastases. The company was unable to demonstrate efficacy in its confirmatory Phase III trial, but an analysis of the data suggests that the insignificant finding was due to an unexpected deviation from assumed (but not explicitly defined) protocol in certain trial centers.

The remainder of the paper has the following structure. Section 1 explains how the FDA currently handles heterogeneous treatment response. Section 2 discusses the problem of spurious correlation and opportunism associated with *post hoc* subgroup analysis. Section 3 presents trial designs and section 4 presents institutional arrangements and statistical analyses that allow identification of subgroups for which the drug is effective in a manner that limits false positives and the cost of trials. We acknowledge that this article is not the right venue to present formal statistical derivations and discuss in detail potential areas of statistical future research required to deal with the statistical challenges implied by our proposals. Finally, Section 5 examines data from the Xcytrin trial and uses this as an example to illustrate the statistical methods we propose.

1. FDA policy on heterogeneity in treatment response

The Food, Drug and Cosmetics Act requires that the FDA verify that a drug is safe and effective before it is approved for marketing as a therapeutic.⁵ FDA regulations require that a company applying for marketing approval conduct two Phase III trials to demonstrate efficacy and relatively tolerable side effects.⁶ Although regulations do not spell out exactly the evidentiary standard to which the FDA holds a new drug, the FDA has issued a guidance that

⁵ See § 102, 76 Stat. 780, 781

⁶ Peter Barton Hutt and Richard A. Merrill, FOOD AND DRUG LAW: CASES AND MATERIALS 527 n. 2 (2d ed. 1991).

emphasizes a drug should be evaluated based on the entire intent-to-treat population⁷ and that requires the rate of false positive findings (“Type I error”) be set to 5 percent.⁸ The implication – borne out by practice – is that the FDA judges the efficacy of a drug by the difference between average outcomes in the treatment and control arms of a trial.

The FDA understands that there is heterogeneity of treatment effects.⁹ But it only accommodates this heterogeneity in two limited ways. First, it encourages sponsors to specify prior to conducting a trial the subgroups they plan to analyze.¹⁰ When this is done, sponsors must implement corrections for multiple testing both in setting their initial sample size and in statistical analysis of data from a trial, though the agency recognizes that the corrections are less severe where subgroups overlap and therefore produce correlated test statistics.¹¹ The FDA guidance does not explicitly state that significant treatment effects among *a priori* specified subgroups can be the basis for drug approval, but it does not rule it out.

Second, the FDA acknowledges that it will not always be possible to identify subgroups *a priori* and that exploratory analysis may be required to identify subgroups.¹² The FDA’s approach to subsequent so-called *post hoc* subgroup analysis depends on the effect of treatment on the intent-to-treat or per-protocol population (“full trial population”) and whether the outcome at issue concerns efficacy or safety. In general, the FDA permits (and sometimes requires) *post hoc* subgroup analysis only to confirm the effects of a drug on the full trial population.

If the sponsor cannot demonstrate that the treatment is sufficiently effective to be approved for the full trial population, *post hoc* subgroup analysis cannot by itself be used to obtain approval for a subgroup.¹³ The FDA has not approved a single drug solely on the basis of *post hoc* subgroup analysis.¹⁴ The FDA does permit a sponsor to use *post hoc* subgroup analysis to justify a subsequent trial to confirm the findings of the subgroup analysis. But another trial can be very costly. And there is no indication that the FDA allows sponsors to

⁷ See Food and Drug Agency, International Conference on Harmonisation; Guidance on Statistical Principles for Clinical Trials, Availability, 63(179) Fed. Reg. 49583, 49593 (Sept. 16, 1998) (§5.2.1 Full Analysis Set).

⁸ Id. at 49291 (§3.5 Sample Size).

⁹ Id. at 49589 (§3.2 Multicenter Trials).

¹⁰ Id. at 49595 (§5.7 Subgroups, Interactions, and Covariates).

¹¹ Id. at 49587 (§2.2.5 Multiple Primary Variables)

¹² Id. at 49595 (§5.7).

¹³ Id. at 49595 (§5.7). See also John Powers et al., FDA Evaluation of Antimicrobials: Subgroup Analysis, Letter to Editor, 126(6) Chest 2298 (June 2005).

¹⁴ Aldo P. Maggioni, et al., FDA and CPMP Rulings on Subgroup Analyses, 107 Cardiology 97, 99 (2007).

combine the data from an initial trial with *post hoc* subgroup analysis with a subsequent confirmatory trial on a subgroup to establish a significant positive result for that subgroup.

Even if the sponsor does demonstrate efficacy or safety of treatment for the full trial population in the initial trial, the FDA may require the sponsor to demonstrate the drug is effective and safe for subgroups defined by the agency in order to validate the results for the full trial population. FDA guidelines identify subgroups defined by centers in multicenter trials as one such check on the consistency of the trial's main results,¹⁵ but clinically and biologically defined subgroups have also been suggested.¹⁶ If certain subgroups do not show efficacy or show side effects, the FDA may require that the drug label indicate it is indicated only for the subpopulations where it has been demonstrated both effective and safe. For example, following the Val-HeFT trial of 160 mg valsartan for patients with heart failure, the FDA only approved the drug for patients who are intolerant to ACE inhibitors. The reason was that in the full trial population the drug was superior to placebo only with respect to only one (combined mortality and morbidity) of two primary endpoints. (The other endpoint was mortality alone.) However, the drug was superior to both endpoints in the non-ACE subgroup.¹⁷ The FDA may also require the sponsor for a drug with an uneven or uncertain safety profile to conduct Phase IV post-approval trials. If the Phase IV trials reveal dangerous side effects, FDA has the ability to alter a drug's labeling to reflect those risks or to yank a drug from the market.

In short, the FDA takes a conservative approach with respect to subgroup analysis.¹⁸ If subgroups are identified prior to trial, they may positively influence approval so long as the sponsor powers the study to address multiple testing concerns. If subgroups cannot be identified prior to trial, *post hoc* subgroup analysis can only be used to negatively influence approval or to justify new, costly trials.¹⁹

It is worth noting that the FDA's approach – judging drugs largely on the basis of average treatment effects – implicitly assumes that doctors are very bad at matching the right patient subgroups to drugs.²⁰ The reason is that the average treatment effect of a drug y_1 (relative to

¹⁵ FDA, *supra* note ____, at 49589 (§3.2)

¹⁶ See Powers et al., *supra* note ____, at 2298.

¹⁷ See Maggioni et al., *supra* note ____, at 99.

¹⁸ The European Union's Committee for Proprietary Medicinal Products (CPMP) has a similar policy. See Maggioni et al., *supra* note ____, at 97.

¹⁹ See Richard Wunderink et al., FDA Evaluation of Antimicrobials: Subgroup Analysis, Letter to Editor, 126(6) Chest 2300 (June 2005) (highlighting asymmetric implications of *post hoc* subgroup analysis for FDA approval).

²⁰ See Anup Malani and Feifang Hu, The option value of new therapeutics 14, unpublished manuscript (2004).

control y_0) is equal to the positive treatment effect among subgroups with positive effects plus the negative treatment effects among those who have negative effects:

$$E(y_1 - y_0) = pE(y_1 - y_0 | y_1 > y_0) + (1 - p)E(y_1 - y_0 | y_1 < y_0)$$

where $p = \Pr(y_1 > y_0)$. Because the negative effects only occur if doctors blindly use the new drug to treat patients even though the drug not recommended for those patients, the FDA's rule implicitly assumes this is so.

2. Cost and benefits from *post hoc* subgroup analysis

The FDA's position on *post hoc* subgroup analysis is based on concerns about multiple testing. To illustrate, consider the following Statistics 101-type hypothetical. Suppose there is a population that can be divided into 10 subgroups. For example, if the population is all adults between ages 20 and 70, we can divide the group into 5-year age ranges. Suppose also that there is a drug which has no effect on either the full population or on any subgroup, though there is some random variation in observed outcomes either due to the drug or natural progression. For example, we might assume that the treatment effect can be described for either the full population or any subgroup as a standard normal distribution with mean zero and variance one. The probability that the drug will be proven effective on the full population with a confidence level of 95% is – by definition – just 5%.

But if the sponsor seeking approval for our hypothetical drug is permitted to separately test the drug against each of the 10 subgroups, the probability that he will be able to demonstrate efficacy for at least one subgroup is 0.4 ($= 1 - (.95^{10})$). This obviously raises the risk of Type I error or false positives, that is, the possibility of approving the drug even though it has not been demonstrated effective at the 95% confidence level.

If the FDA knows that the sponsor will test the drug against 10 subgroups, it can implement a multiple testing correction to eliminate spurious results. For example, if the outcomes in the 10 subgroups are known to be uncorrelated, then it can change the threshold p-value required for approval from $p = 0.05$ to $p = 0.05 / (\text{number of tests})$ or 0.005.²¹ This correction is known as the Bonferroni adjustment. It ensures the probability of observing even one subgroup with significant treatment effects is back to 5%. The proper p-value adjustment in the cases where outcomes in the subgroups are correlated is different, but this correlation and the adjustment may be derived from the data.²²

²¹ We are ignoring the 11th test on the full population to keep the numbers simple and because the full population result is positively correlated with each subgroup result.

²² Sandrine Dudoit and Mark van der Laan, *Multiple testing procedures with applications to genomics* (New York: Springer, 2008).

The problem becomes more complicated if the FDA does not know the number of subgroups against which the sponsor will test its drug. In that case the FDA cannot implement a multiple testing correction. This problem is made more serious by the fact the number of possible subgroups can get very large very quickly. Suppose the sponsor can also divide the adult population by gender (male/female) and by ethnicity (white/black/Hispanic/other). Now the available subgroups has jumped from 10 based solely on age to 80 ($= 10 \times 2 \times 4$) based on a combination of age, gender and ethnicity. If the sponsor can cherry-pick a subgroup in which to demonstrate efficacy, the probability it will be find able to find at least one with a p-value greater than 0.05 is 0.98 ($= 1 - (.95^{80})$)! The sponsor has a strong financial incentive to cherry-pick in this manner because the alternative may be not to obtain any return on its investment in the drug. We call this the problem of opportunistic behavior by sponsors.

The cost of the FDA's prudence – or more appropriately of opportunistic behavior – is either some rejection of drugs that are useful to certain subpopulations (“Type II error”) or a higher costs of drugs if the sponsor conducts a follow-on trial to confirm the results of *post hoc* subgroup analysis. To illustrate the problem of false negatives, suppose that the drug in our hypothetical actually has positive treatment effects in one of the 10 subgroups defined by age. The probability of approving the drug with proportional sampling across subgroups is virtually zero ($\approx 0.95^5 \times (0.05^5)$).

A natural question is whether *post hoc* subgroup analysis can ever provide sufficiently reliable information to warrant approval of a drug even in the absence of risks from spurious correlation and opportunistic behavior. After all, trials in which subgroups are not *a priori* specified are not powered to identify subgroup effects that meet the usual standards for Type I (5%) and Type 2 error (10-20%). Compounding this problem is that subgroups by definition have smaller sample size than the full trial population. But sample size calculations are based on *estimates* of the variance of treatment effects in the full trial population. Because those estimates themselves have variation, there is a positive probability that they are in fact too high, leaving excess power for identification of subgroup effects. As subgroups are a subset of the full trial population, they are correlated with that population. Thus analysis of the full trial population and one subgroup requires something less than a two-fold overestimate of variance to be powered to give reliable information. Moreover, since subgroups may be more homogenous than the full trial population, the subgroup may have smaller variation in the treatment effect. We shall demonstrate this possibility in our analysis of the Xcytrin trial.

3. Designing trials that enable subgroup analysis

In this and the next section we discuss proposals that offer a compromise between (1) false positives due to opportunistic behavior and (2) false negatives or the cost of additional trials due to the FDA's cautious approach to subgroup analysis. The aim is to extract more *reliable*

information on subgroup effects from trial data that can be used to approve drugs for use in subgroups with *little additional cost* from larger sample size in the initial trial or a new trial.

Before beginning, we should be clear that we agree with the FDA's position that if the sensitive subgroups are known prior to an initial Phase III trial, those subgroups may positively influence the approval decision so long as the trial is powered so as to be capable of identifying significant results for those subgroups.²³ For reasons mentioned earlier – correlation between subgroup outcomes and full trial population outcomes and greater homogeneity within subgroups – the additional sample size required to analyze two subgroups is not double that required for to analyze the single full trial population.

We also agree with FDA policy to use multiple testing adjustments to avoid spurious results from analysis of *a priori* specified subgroups. We agree with the literature that Bonferroni adjustments may be too conservative because of the assumption that subgroups are independent. Although the FDA recognizes the need for multiple testing adjustments in sample size calculation and analysis, there is reason to be concerned that the FDA does not apply those adjustments correctly across trials. For example, suppose a sponsor conducts an initial trial that does not show intent-to-treat effects but *post hoc* analysis reveals possible subgroup effects, and the sponsor conducts a second trial solely to confirm the subgroup effects. On the one hand, the second trial should be able to credit subgroup members in the first trial towards sample size requirements in the second trial. On the other hand, a significant result in the second trial may be spurious because it is itself a second test. Indeed, if one conducted 100 trials on a given subgroup, 5% would show significant effects for that subgroup even if its true effect is zero. We cannot find evidence from stated FDA policy or practice that the FDA makes such adjustments across trials. That said, it is likely the case that the extreme cost of Phase III trials limits the frequency of opportunistic behavior across multiple trials.

Beyond these minor points we discuss two bolder reforms. The first – the use of adaptive designs – should not come as a surprise. The remainder of this section sketches how adaptive designs help address subgroup effects and discuss the sample size costs of those designs. The second reform – deferred to the next section – requires a modified form of subgroup analysis to

²³ An even better approach may be to specify subgroups not by patient characteristics at baseline, but by an algorithm that has as inputs not just those characteristics but also outcomes recorded as the trial progresses. Suppose the sponsor suspects that treatment effects may depend on one of 10 genetic markers, but is not sure which one. Instead of picking one of the those markers before the trial begins, the sponsor could, for example, specify that after ϕ fraction of subjects have enrolled, it will correlate those markers with outcomes and pick as a subgroup those subjects possessing the marker with the highest correlation with outcomes. So long as ϕ is specified *a priori*, it is theoretically possible – though perhaps not easy – to derive a sample size to ensure this trial is properly powered. There may not be any penalty for multiple testing so that the critical p-value may remain 0.05. Nor is there a risk of opportunistic behavior by the sponsor since the FDA can implement the algorithm itself and verify the subgroup the sponsor has identified as correct.

be performed by an independent consultant. It would also allow the FDA to approve a drug without the expense of further trials.

A typical fixed design trial randomizes subjects between a treatment group and a control group. The randomization probabilities and sample size remain the same through the end of the trial. It is easy to calculate the proper sample size for this trial once so long as there is an estimate of the size d of a clinically relevant treatment effect and the variance σ^2 of the treatment effect.²⁴ If, however, the sponsor does not know the precise clinically relevant treatment effect or the variance of treatment effects, it can employ an adaptive design that uses data from the trial to estimate these parameters along the way and adjust sample size or treatment versus control allocation as appropriate.

Adaptive designs that use interim data to modify sample size fall in two categories. In one, called a sequential group approach, the sponsor starts with a trial that is conservatively large – using parameters at the lower end of the range for clinically relevant effects and at the higher end of the range for variation in treatment effects – but stops the trial early if interim data suggest that treatment effects are larger than clinically relevant or have smaller variance than hypothesized. The other design, called simply an adaptive approach, does the opposite. It starts with a trial that is liberally small and extends the trial if estimates of the treatment effect is smaller than the clinically relevant amount or estimates of the treatment effect variance are larger than hypothesized. Either adaptive design requires a larger sample size than a fixed design. Moreover, because the trial is updated after the sponsor “tests” the data by estimating treatment effects, the critical p-value may have to be reduced to account for multiple testing. The exact multiple testing penalty has been derived in statistical literature.²⁵

There are also adaptive designs intended to adjust the proportion of enrolled subjects assigned to the treatment group based on interim data analysis, or adjust other design settings such as interventions to improve compliance to the treatment protocol. If, for example, outcomes in the treatment group show higher variance relative to the control group than anticipated, then the sponsor may change group assignments so that more than half of subjects get treatment. So long as the estimate of the variance of treatment effects – the difference in outcomes in the treatment and control groups – does not increase, so that the sample size remains constant, the sponsor pays no multiple testing penalty for such an adaptive design.²⁶

²⁴ The formula for the size of each group is $2\sigma^2(z_{\alpha/2} + z_{\beta})^2 / 2d^2$ for two-sided tests of and usually $\alpha = 0.05$ and $\beta = 0.8$ or 0.9 .

²⁵ Cyrus R. Mehta and Nitin R. Patel, Adaptive, Group Sequential and Decision Theoretic Approaches to Sample Size Determination, 25 *Statistics in Medicine* 3250-3269 (2006).

²⁶ Even if the estimate of variance of treatment effects falls, the sponsor cannot stop the trial early. But if the estimate of overall variance of treatment effects rises, then the sample size increases and the sponsor must pay a

The FDA is open to use of adaptive designs. Its Critical Path initiative begun in 2004, seeks to identify biological and statistical innovations that can improve the efficiency of clinical trials and incorporate them into the drug development and approval process. That initiative has identified adaptive designs as one area on which to focus its attention. Indeed, the FDA is expected to release a guidance on adaptive designs to clarify its thinking.²⁷

None of these adaptive designs, however, are specifically intended to address subgroup effects. They are mainly directed at optimizing over power and cost for main group effects. That does not mean that no one has thought of applying adaptive designs to estimate subgroup effects. We know of no instances, however, where the FDA has approved an adaptive design to facilitate subgroup analysis, though the FDA has considered or allowed a number of trials with adaptive design.

How might an adaptive design be used for subgroup analysis? Consider a parallel-arm trial initially powered to test one hypothesis: the treatment effects for one large group are zero. At some interim point, the sponsor or the independent data monitoring committee (IDMC) examines the data to determine if there is a subgroup that ought to be studied. There are two types of data that might be used to identify subgroups: baseline characteristics alone or treatment outcomes. In the former case, the sponsor looks for abnormal variation in a relevant covariate. For example, if there is excess variation in treatment history or in the progression of symptoms, the full trial population can be divided into subgroups using a cut-off based on the extent of prior treatment or symptoms. In the latter case, the IDMC may look at the relationship between certain covariates and treatment effects. (The IDMC is used to ensure that the sponsor does not become unblinded.) If the data suggests, for example, that certain age or ethnic subpopulations are responding better to treatment, those groups can become target subgroups for the study.

After this interim analysis, the sponsor would have to revisit the objective of the trial. There are two choices. First, the sponsor could examine just one hypothesis but limit it to a subgroup identified by interim analysis as particularly sensitive to treatment. Specifically, the null hypothesis would become: the treatment effect for *one subgroup* is zero. We assume in this case that, after the interim analysis, the sponsor would discontinue enrollment of subjects that do not belong to this subgroup, lest they waste sample size. The sponsor's other choice is to examine two or more hypotheses based on the number of subgroups discovered through interim analysis. For example, if that analysis identified two subgroups based on ethnicity, the trial might test two hypotheses: the treatment effect for whites is zero and the treatment effect for non-whites is zero.

multiple testing penalty. The reason is that it was given a "real option" of testing and must pay a price for this option.

²⁷ Scott Gottlieb, Speech before 2006 Conference on Adaptive Trial Design, Washinton, DC, July 10, 2006 (available at <http://www.fda.gov/oc/speeches/2006/trialdesign0710.html>) (checked Mar. 7, 2008).

As with adaptive designs targeting sample size adjustments, adaptive designs targeting subgroups will require a larger sample size and appropriate adjustments of the statistical methodology. These can be derived, though we do not do so here. In particular, the design may require that the results be held to a lower critical p-value to account for the possibility of multiple testing. The table below summarizes the four basic options in a subgroup-identifying adaptive design and our speculation as to the multiple testing penalty. The rows indicate whether interim analysis employed outcome data or not. The columns indicate whether the sponsor added hypothesis tests after the interim analysis.

Data used to identify subgroups	Number of additional hypothesis tests added to study	
	Zero	One or more
Covariates (not outcomes)	No penalty	Penalty for adding one or more hypothesis
Outcomes	Must keep other subgroups in evaluated population, plus a penalty for using outcome data	Penalty for adding second hypothesis, must keep other subgroups in evaluated population, plus pay a penalty for using outcome data

If outcome data are not used to identify subgroups and no additional hypothesis tests are added to the study, then there is no need to impose a penalty, so long as the trial must proceed until the *a priori* specified sample size requirement is met. The reason is there was no testing of treatment effects in the interim analysis and the number of tests remain the same as when the trial began. Even though the sponsor may choose a subgroup with low variance with respect to covariate characteristics, since the data studied in the interim analysis is designed to be orthogonal to the data relevant for estimation of the treatment effect, there is in essence no actual testing occurring. The general idea is that one can extract a data set from the actual data set which contains no information about the treatment effect, and, as a consequence, no penalty needs to be applied for changing the hypothesis based on such data.

If outcome data were used to identify subgroups, there should be a penalty even if no additional hypothesis tests were added. The reason is that the sponsor was able to test whether treatment effects are significantly positive for a subgroup. Even with a subset of the full sample, it is highly likely that data mining would uncover at least one subgroup with significant treatment effects in the subsample. As a result, the sponsor would have been given the option to change the hypothesis test's scope based on treatment effects. It must pay a price for that.

This price is difficult to calculate since we may now know how many tests were performed to identify a subgroup. It helps if that the IDMC conducts the interim analysis because it has less incentive to engage in data mining and in any case is more likely truthfully to report the number of tests performed. But if the sponsor has a role on that committee or the IDMC is not otherwise truly independent, institutional design may not help with calculating the multiple testing adjustment. In that case, we speculate – though have not confirmed – that requiring the sponsor to include the excluded groups along with the newly targeted subgroup in the final analysis may address the problem that the FDA may not know the number of tests

performed in the interim analysis. For example, if interim analysis after 10% of the sample is enrolled reveals that only young subjects have a significant treatment effect, the sponsor may be required include even old subjects from the initial 10% sample in the final empirical analysis even though the modified hypothesis they are testing is that the treatment effect among young subjects is zero. Our logic is crudely that the larger the number of tests performed, the worse the relative performance of subgroups excluded from the modified hypothesis test, and the larger is the penalty to the sponsor of having to include the excluded subgroups in the final empirical analysis.

In the cases where the sponsor adds one or more hypothesis to the study, it must pay an additional price for multiple testing on top of the price it pays based on the data employed to conduct the interim analysis. The reason for this penalty is obvious – the number of hypothesis test has increased – and the size of the incremental penalty is straightforward to calculate.

Before concluding our discussion of adaptive designs, it is worth noting an important weakness of these designs. Because of ethical and profit considerations, they may not be optimal for identifying side effects of treatments. If interim analysis suggests a particular subgroup has worse side effects, both the sponsor and patient advocates will push to exclude that subgroup from further analysis. But doing so limits the amount of data we have on that subgroup and thus on the side effects of the drug.

4. Controlling opportunistic behavior

In this section we consider how it might be possible – starting from a fixed design – to use *post hoc* subgroup analysis to approve a drug without further trials. *Post hoc* subgroup analysis does not increase the risk of Type I error so long as the FDA makes appropriate multiple testing adjustments. But appropriate multiple testing adjustments require knowledge of the number of tests the sponsor has performed. Because of the financial incentive to have its drug approved, the sponsor cannot be relied upon to truthfully report the number of tests it has performed.²⁸ Indeed, the FDA can be confident that the sponsor probably conducted more tests than that for which the FDA plans to adjust. Therefore, there remains a residual risk of Type I error above 5%.

²⁸ We have considered the possibility that the FDA could specify prior to a phase III trial the exact subgroups the applicant can examine. This could be based on the subject-matter of the trial or on the FDA's knowledge of data from trials of competing drugs by other sponsors. There are two problems with this reform. First, the applicant could specify subgroups based on subject matter as well as the FDA and has a strong financial interest in doing so. We doubt there are valuable subgroups that the FDA could propose that the applicant has not already considered. Second, sponsors of the competing drugs are likely to object to the FDA's use of their trial data – which is treated as a trade secret, see Article 39.3 of the Trade-Related Intellectual Property (TRIPs) agreement – in this manner. They would have a reasonable argument that this competitively favors later new drug applicants over earlier ones. It would also subtly reduce the incentive to innovate quickly.

A critical assumption in this logic is that the sponsor is financially interested in having the drug approved. If *post hoc* subgroup analysis were performed by a truly independent agent, then the FDA could rely upon that agent's report of the number of tests it conducted and fully eliminate the risk of false positives by means of multiple testing adjustments. Of course the real question is whether the agent is truly independent, a topic to which we will turn in a moment.

There are two basic candidates for an independent agent: the FDA and an outside statistical consulting firm. Each has its strengths and weaknesses. The strength of using the FDA is that by doing the *post hoc* subgroup analysis itself, the FDA knows immediately the number of tests conducted. There is no need to rely on the absence of any other motive, as will be the case with an independent consulting firm. There are two weaknesses of the FDA. It has limited resources that make it difficult to conduct even the current level of scrutiny of new drug applications (NDAs). Moreover, the FDA is subject to political pressure. It has been criticized for being influenced both by drug companies and by political backlash following approval of unsafe drugs. These pressures are unlikely to perfectly offset to create an unbiased decisionmaker. As a result, the FDA may conduct too much subgroup analysis – at the cost of Type I error – or too little subgroup analysis – at the cost of Type II error or more costly approval.

The alternative is an independent statistical consulting firm. Many already exist to help sponsors design and analyze data from trials.²⁹ The strength of consulting firms is, perhaps, more statistical expertise than the FDA. Unlike the FDA, which has limited resources and no need to compete, these firms have every reason to specialize and innovate because it may make it more likely they are selected to perform subgroup analysis.

The main weakness of the consulting firm approach is that these firms may not be truly independent. Sponsors are repeat players. A consulting firm may have an incentive to give a favorable analysis so as to secure repeat business from sponsors. That repeat business may be for subgroup analysis or some other statistical service. This is a lesson well learned from the corporate accounting scandals from earlier this decade. Perhaps the indirect influence of sponsors can be addressed by requiring the FDA to select the independent consultant to perform *post hoc* analysis, by blinding the sponsor to the independent firm selected, and by banning firms that perform *post hoc* analysis from providing other statistical services to sponsors. We wonder, however, whether the agency will always be able to keep the identity of the consulting firm secret, even after the analysis is completed and the FDA has made its regulatory decision concerning a sponsor's drug. Moreover, restricting the consulting firms' scope of business will limit their ability to attract talent and incentive to innovate in the area of subgroup analysis since it comes at the cost of other lines of business.

²⁹ See, e.g., Cytel Statistical Software and Services, founded by Cyrus R. Mehta and Nitin R. Patel, and Target Analytics, Inc., run by Mark van der Laan.

A second weakness of using consulting firms is that “independence of the sponsor” is not the same thing as “motivated to reduce Type II error.” True independence only guarantees the consulting firm will not be swayed by the profit interests of the sponsor. It does not guarantee that the consulting firm extracts the most and reliable data from *post hoc* analysis after it is chosen to perform that analysis. This problem is one which economists call moral hazard. Independence merely substitutes the sponsor's interests with the those of the consulting firm. Most likely this is cost minimization, which may imply too much Type I or Type II error, whichever minimizes the consulting firm's labor expense.³⁰

To address the problem that neither the FDA nor the outside consultant may be truly independent of the sponsor, we propose two statistical methods to limit either agent's ability to skew the analysis in favor of the sponsor.³¹ For convenience, we shall speak as if the consulting firm has been chosen to conduct the analysis.

The first approach would provide the consultant with all the data from the trial except variables that identify outcomes and ask it to identify subgroups based on baseline characteristics that exhibit “remarkable and relevant variation” in the Phase III trial data. (This is similar to one of the approaches used to identify subgroups for the adaptive design trials discussed in the last section.) The consultant would not be asked to perform the *post hoc* subgroup analysis; that could be conducted by the sponsor, though the FDA would rely upon positive treatment effects only for the subgroups identified by the consultant. Whatever positive subgroup results the sponsor reports, the FDA would apply a multiple testing adjustment based on all the subgroups reported by the independent consultant.

In order to identify remarkable variation, the consultant needs to have a sense of what normal variation would be. It could estimate normal variation in baseline characteristics from Phase II trials or prior studies in the literature. The consultant would have to be sensitive to exclusion and inclusion criteria, which can affect the applicability of prior data to the current Phase II trial sample. Moreover, the consultant would have to keep in mind that any subgroup it

³⁰ The independent consulting firm must also be concerned about not doing too many tests. Each useless test it performs increases the multiple testing adjustment for any positive finding. Minimizing Type II error requires internalizing this negative externality. Since Type II error is unobservable, the FDA cannot directly incentivize the consulting firm to do so. And the FDA certainly should not give the firm an incentive keyed to drug approval, because then it would have incentives like the sponsor and replace Type II error with Type I error.

³¹ These methods do not address other problems, such as the limited resources of the FDA or the insufficient motivation of independent consulting firms. If the statistical methods we discuss help ensure that the consultant truly cannot manipulate the data to increase Type I error, then one might address the problem of a consultant's motivation by giving it stock in the sponsor. We do not advocate this because it is too radical and would be politically infeasible. That said, giving the consultant some sponsor stock is not the same as allowing the sponsor to conduct the entire *post hoc* subgroup analysis because the statistical methods we propose in the main text require that the consultant not have access to certain data that the sponsor already has, or could easily obtain.

identifies should be defined by variables that are plausibly relevant (from our current biological understanding of the disease targeted by the sponsor's drug and the pharmacology of that drug) to the treatment effects of the drug.

The second statistical method we propose to independently identify subgroups requires splitting the data from a Phase III trial into two parts. One part would be called the exploratory subsample and the other the confirmatory subsample. Importantly, the sample would be split randomly. The consultant would only be given the exploratory subsample and be asked to conduct a full *post hoc* subgroup analysis on that subsample to identify subgroups that respond better to the drug.³² The sponsor would then be allowed to perform *post hoc* subgroup analysis on the confirmatory sample using only the subgroup identified from the exploratory subsample by the consultant. As before, the FDA would apply a multiple testing adjustment based on all the subgroups reported by the independent consultant.

Both statistical methods ensure that subgroups are identified independent of the interests of the sponsor. Since the first method does not give the consultant access to outcome data, it cannot choose subgroups to help or hinder the sponsor. Since the second method requires the sponsor to limit its subgroup analysis to a subsample that is orthogonal to the subsample analyzed by the consultant, the consultant's analysis cannot help the sponsor engage in data mining.

Neither method requires that the FDA to impose any additional multiple testing penalty beyond one based on the total number of subgroups identified by the consultant. The biggest advantage, however, may be that the FDA could ask the consultant also to identify subgroups that might have a negative reaction to or severe side effects from the sponsor's drug even though the average subject in the full trial population has a positive reaction to the drug. This would help reduce Type I error associated with otherwise approved drugs.

Each statistical method also has its shortcomings. The weakness of the variation-in-covariates approach is that the subgroups with the most remarkable variation may not be perfectly correlated with the subgroups that have positive and significant treatment effects. Abnormal variation is just one factor that suggests differential treatment effects; it does not guarantee them. The main concern with split-sample approach is that the *post hoc* subgroup analysis, which would be underpowered even if performed on the whole trial sample, is particularly underpowered if performed on subsamples. This will increase the risk of Type II errors. This risk may be considered the cost of independence under this method.

³² The sponsor could not be asked to do this because it would likely be able to derive the confirmatory subsample from the exploratory subsample and the full sample, which it already possesses. This would allow it to choose subgroups ostensibly on the exploratory subsample but truly on the full sample. The result would be almost the same as *post hoc* subgroup analysis by the sponsor.

5. Application to Xcytrin

[**Note to readers and editor:** This section is incomplete. We shall describe the initial Phase trial results for Xcytrin. We plan to demonstrate (1) that subgroup analysis does not require a full Bonferroni adjustment because subgroups statistics are correlated and (2) that subgroup variation is less than full trial population variation so sample size adjustments are not as large as commonly thought. We also plan to apply the two statistical algorithms suggested in Section 4 to ensure the independence of statistical consultants to the Xcytrin data. The question we ask is whether the algorithms would have yielded the same results as full post hoc subgroup analysis.

The presentation at the AEI conference will hopefully include some empirical analysis of the Xcytrin data.]

6. Conclusion

[Highlight questions for future statistical research.]