Accounting for Heterogeneous Treatment Effects in the FDA Approval Process

Anup Malani
Oliver Bembom
Mark van der Laan*

Introduction

Only one out of every five drugs that begin clinical trials emerges successfully from that testing.[1] This high failure rate explains a substantial portion of the exploding costs of health care in America. Although it costs roughly $300 million to conduct a full battery of clinical trials on a drug, these costs can only be billed to drugs that are successful enough to be approved by the U.S. Food and Drug Administration (FDA). Dividing the cost of conducting trials by the number of successful drugs inflates the cost of clinical testing to $800 million per approved drug.[2] In this manner, the trial failure rate magnifies the already growing cost of conducting clinical trials.[3] As a result, the cost of prescription drugs has grown twice as fast as the cost of the other two major components of health care costs, hospital stays and physician care, helping to push the overall cost of healthcare to 16% of GDP.[4]

There are two reasons for why so many drugs might fail clinical trials. First, most drugs might not work. Perhaps scientists have plucked the low-hanging pharmacological fruit and drug companies now face diminishing returns on their research

[*] University of Chicago Law School, Resources for the Future and National Bureau of Economic Research; Department of Statistics, University of California-Berkeley; and Department of Statistics, University of California-Berkeley, respectively. We thank Richard Miller for access to the data from the Phase III motexafin gadolinium (MgD) trials. We thank Glenn Cohen, Adam Cox, Jack Calfee, Richard Epstein, Jake Gersen, Dan Ho, Jonathan Masur, Richard Miller, Lars Noah, Ben Roin, participants at the AEI conference, "Oncology Drug Development: Rethinking FDA Oversight," March 13-14, 2008, and workshop participants at the University of Chicago Law School for helpful comments. Malani thanks the Milton and Miriam Handler Foundation for financial support and E. Phoebe Holtzman for research assistance. In the interest of full disclosure, two of the authors, Bembom and van der Laan work for Target Analytics, Inc., which was hired by an investor in Pharmacyclics to prepare a statistical analysis of the data from the MgD trials. The data analysis for this particular paper was conducted after Pharmacyclics' application for FDA approval for MgD was rejected by the agency in Dec. 2007. B. Jungbauer, *Pharmacyclics' Xcytrin Gets FDA "Not Approvable" For NSCLC Patients With Brain Metastases*, The Pink Sheet (Dec. 28, 2007).

[1] More precisely, out of every five drugs for which an Investigational New Drug (IND) application is filed with the U.S. Food and Drug Administration (FDA) to begin clinical testing, only one results in a New Drug Application (NDA) for marketing approval from the FDA. Christopher P. Adams and Van V. Brantner, *Estimating The Cost Of New Drug Development: Is It Really $802 Million?*, 25 Health Aff. 420, 422 (Exhibit 1) (2006).

[2] Joseph A. DiMasi, Ronald W. Hansen, Henry G. Grabowski, *The price of innovation: new estimates of drug development costs*, 22 J. Health Econ. 151, 166 (2003). The reason one cannot simply divide the $200 million estimate by the probability of success during clinical testing (0.2 = 1/5) is that not all drugs complete a full battery of clinical tests.

[3] *Id.* at 167.

[4] The cost of prescription drugs grew, on average, by 12% from 1996-2006. By contrast, hospital and physician costs grew by roughly 6% per year during that period. Kaiser Family Foundation, *Prescription Drug Trends* 1 (2008), *available at* http://www.kff.org/rxdrugs/upload/3057_07.pdf (last visited Mar. 1, 2010).

investments. Second, clinical trials might fail to identify all the drugs that work. If the first explanation is correct, there is little that can be done to reduce the cost of prescription drugs. However, the first explanation depends on the second explanation being incorrect. If trials cannot identify all safe and effective drugs, then one cannot know if we have run out of such drugs. This paper suggests that the high failure rate of drugs is partly due to the failure of trials to identify good drugs and proposes a regulatory solution to the problem.

The problem with clinical trials is not so much how they are designed, but rather how the information they produce is evaluated. The FDA uses an "average patient standard" to determine whether a drug is safe and effective. That is, the FDA compares how the average trial participant on a drug compares to the average trial participant on placebo. If the former does better than the latter, then the drug is approved. Otherwise the drug is not approved. If a drug works in some patients but not others, however, the average patient standard may reject drugs that are effective for some patients. These drugs may be effective for a subset of trial participants, just not the average participant. In short, the average patient standard suffers from false negatives.

It is common for drugs to have different effects in different patients, a phenomenon statisticians call "heterogeneity in treatment effects". For example, Anthrotec (Pfizer) is an effective treatment for osteoarthritis for patients who develop ulcers when using certain common pain medications.[5] But a key ingredient in Anthrotec – misoprostol – is documented to induce labor and used for medical elective abortion.[6] Therefore, while Anthrotec is generally effective for pain relief, it is contraindicated[7] for pregnant women because of its abortifacient effects. Another example is isoniazid, a widely used chemotherapy drug used to treat patients with tuberculosis. Different patients metabolize the drug at different rates and a patient's rate of metabolism of the drug can be discovered through genetic testing. It has been demonstrated that the drug is effective in patients who metabolize the drug at a slow rate but not among patients who metabolize it at a fast rate.[8]

The FDA is aware of the risk of false negatives under the average patient standard. The reason they stick with the standard is that they think the alternative is to allow drug companies to mine data from clinical trials to find subgroups of participants who show benefits from the drug. This alternative – which is called *post hoc* subgroup analysis[9] – just replaces the risk of false negatives with false positives. The more subgroups a drug company examines, the more likely they are to find one that shows improvement on their drug, whether or not their drug actually works. Statisticians call this the risk of spurious correlation from "multiple testing". Drug companies have a financial incentive to ignore – and even promote – this risk. They

[5] Anthrotec is diclofenac sodium plus misoprostol. Osteoarthritis is pain and inflammation caused by the breakdown of cartilage in a patient's joints. The pain medications that cause ulcers are non-steroidal anti-inflammatory drugs (NSAIDs). The drug is made by Pfizer.

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

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

[8] Gordon A. Ellard and Patricia T. Gammon, *Pharmacokinetics of isoniazid metabolism in man*, 4 J. Pharmokinetics & Pharmacodynamics 83, 83-84 (1976).

[9] To be precise, subgroup analysis is statistical evaluation of the effect of a drug on one or more subgroups of the subjects in a clinical trial. *Post hoc* subgroup analysis, in particular, is the estimation of treatment effects in subgroups that were not specified prior to the start of the trial. Instead the statistical analysis is performed on subgroups *identified from the data after the trial had begun or was completed.*

have invested millions in lab research and development; they can only recoup this if their drug is approved.[10]

It seems patients are stuck between a rock and a hard place: too many false negatives with the FDA's average patient standard or too many false positives with drug company data mining. This paper offers a third option – actually, a series of proposals – that reduces false negatives without increasing false positives. This will increase the rate of approval only for effective drugs, and reduce the waste of research expenditures.

The problem we identify with clinical trial is not their design but the way the FDA and drug companies interpret the data from trials. Our first proposal, however, is to change the design of trials to reduce errors from the average patient standard. To understand the change, some additional background on FDA policy is required. While the FDA rules out *post hoc* subgroup analysis due to the risk of false positives, it does allow companies to identify subgroups of patients they plan to study before they conduct a trial. That is, companies are permitted to select subgroups of patients *from the overall population* to enroll, but they may not select subgroups of patients *from those already enrolled in a trial* to demonstrate that a drug works. The reason is that the latter approach lets drug companies "peek" at the answers before giving the test. But the FDA's compromise does not really address the false negative risk with the average patient standard. Drug companies may not know which subgroups of patients will benefit from a drug before testing the drug on a broad population of patients.

Therefore, our first proposal is to identify treatment-sensitive subgroups using an "adaptive group sequential design" trial. In standard trials, patients 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 patients can be randomized to treatment or control based on the performance of patients previously enrolled in the trial. If a certain subgroup of participants seems to be doing better than others based on early returns, the drug company could selectively enroll more patients from that subgroup in the general population as the trial progresses. The goal is to identify treatment-sensitive subgroups based on data from early enrollees in order to power up analysis of those subgroups among later enrollees.

Our second proposal – actually a pair of proposals – focuses on fixing *post hoc* subgroup analysis to reduce data mining by drug companies. These proposals combine institutional design – evaluation of trial data by an independent auditor – with statistical tools to reinforce the institutional design – specifically, to ensure the auditor is truly independent.

Specifically, our proposal would permit *post hoc* subgroup analysis so long as it were performed by an independent auditor rather than the drug company. We have learned from the Enron and MCI Worldcom scandals, however, that independent auditors are not truly independent.[11] They may be indirectly compromised by career concerns. They may seek to provide optimistic assessments of firms in the hope that they will subsequently be employed by those firms for other auditor work. The same would be true of an outside statistician auditing data from a company's clinical

[10] If the FDA knew the number of subgroups the sponsor sampled in search of a positive response the FDA could limit the risk of spurious correlation by employing statistical corrections for multiple testing, such as raising the p-value required to demonstrate statistical significance. Unfortunately, the FDA will rarely be able to verify this number. (We will discuss corrections for multiple testing at greater length in the text accompanying note 42.)

[11] *See* Joel S. Demski, *Corporate Conflicts of Interest*, 17 J. Econ. Persp. 51, 57 (2003).

trial. To address this concern, our second proposal suggests statistical algorithms that prevent the independent auditor from identifying subgroups in a manner that would financially benefit drug companies but would not help patients. These algorithms may be thought of as veil of ignorance rules[12] because they blind the auditor to information that is required for the auditor to help the drug company.

Our most promising algorithm is what we call "split-sample analysis." This algorithm would give the auditor access only to a random subsample of the trial data as selected by the FDA. The consultant would be asked to identify subgroups with *post hoc* subgroup analysis on this "exploratory" sample. The drug company would then be permitted to seek drug approval for a subgroup identified by the auditor under two conditions. First, it can only obtain approval for a subgroup if the drug was significantly effective and safe for that subgroup in the remainder of the trial sample, which we call the "confirmatory" sample. Second, statistical significance[13] will be judged according to a higher standard to account for the risk of spurious correlation due to multiple testing bias. The larger the number of subgroups identified by the auditor, the stricter the standard of statistical significance the company would have to meet.[14]

This algorithm protects against false positives in two ways. Because the auditor does not have access to the confirmatory sample, it cannot help the drug company out by choosing subgroups that respond positively only in the confirmatory sample. Nor can the consultant help the drug company simply by identifying a large number of subgroups based on the exploratory sample because the sponsor pays a multiple-testing penalty in the confirmatory subsample for each additional subgroup that the auditor identifies.

To illustrate the potential benefit of increased flexibility in the drug approval process and to demonstrate how our statistical algorithms could address the data manipulation problem, we conducted a case study of motexafin gadolinium (MGd), a drug for lung cancer patients whose tumors have spread to their brain and are threatening to induce dementia. In a key clinical trial, the drug's maker Pharmacyclics was unable to show that the average participant experienced a statistically significant benefit from the drug.[15] This failure may have been driven, however, by poor results among patients who had already taken chemotherapy for their cancer. Among the patients who were newly diagnosed and had not yet received any chemotherapy,[16] the drug appeared to significantly delay the onset of dementia. Despite this possibility, the FDA rejected the drug.[17] Mimicking the role of outside auditors, we apply our proposed statistical algorithms to correct for bias due to any opportunistic behavior by the drug company. We find that one of our algorithms would surely have validated the company's claims and the other – split-sample analysis – would have validated them with probability 0.11. That is,

[12] For a discussion of this class of rules, *see* Adrian Vermeule, *Veil of Ignorance Rules in Constitutional Law*, 111 Yale L. J. 399 (2001).

[13] Statistical significance is conventionally defined as having a p-value less than or equal to 0.05. The p-value of a statistical estimate is one minus the probability that the estimate is different from zero, which signifies no treatment effect. In other words, if $p < 0.05$, then it can be said that we are more than 95% sure that the estimated treatment effect is different than zero.

[14] Our proposal would let the drug company choose the number of subgroups the auditor should identify to ensure the auditor did not "tank" the drug company by simply identifying too many subgroups.

[15] *See infra* text accompanying notes 80 and 82.

[16] Their tumors tended to be more resilient as they had already survived previous treatment.

[17] *See* Jungbauer, *supra* note 1.

there is at least a one-in-ten chance that a truly independent audit would have led to the approval of MGd.

Before we turn to the meat of the paper, we should clarify three things about our proposals. First, while we advocate a change in how the FDA evaluates data from clinical trials, we do not advocate lowering the FDA's standard for approving drugs. Our proposals would all continue to require that a drug be shown safe and effective with 95% confidence in order to be approved.

Second, our proposals can be used to eliminate false positives as well as false negatives. Although a drug may not have side effects for the average patient, it may have prohibitive side effects for a subgroup of patients. Yet such a drug might be approved under the FDA's average patient standard. Our proposals for identifying subgroups of patients who benefit from a drug can equally be used to identify subgroups of patients who would be harmed by a drug. While there may be little reason to worry that a drug company will exaggerate the harm from a drug, some have expressed concern that the FDA is overly risk averse or antagonistic to drug companies and might exaggerate the harm from a drug.[18] Our proposals can be used to ensure that an outside auditor is truly independent of FDA influence and thus does not spuriously identify subgroups harmed by a drug.

Finally, our proposals – especially the split-sample algorithm – highlight the potential that statistical veil of ignorance rules can promote auditor independence. This is true for whether the target of an audit is a clinical trial, a tax filing, or a report to shareholders. In short, our algorithms may have utility in other areas where law enforcement is concerned with dishonest reporting by private companies.

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 our proposals for controlling opportunism by drug companies. Section 4 examines data from the MGd trial to illustrate the new statistical algorithms we propose. Finally, the conclusion discusses, *inter alia*, how our proposals might be used to root out harmful drugs.

1. *FDA policy on heterogeneity in treatment response*

The Food, Drug and Cosmetics Act requires that a company seeking marketing approval for a new drug – also called the drug sponsor – demonstrate to the FDA that the drug is safe and effective.[19] This demonstration typically requires three phases of clinical testing. In Phase I, the sponsor conducts a small clinical study, usually on healthy individuals, to demonstrate that a clinically useful dose is not toxic for patients. In Phase II, the sponsor conducts a medium-sized study on sick patients to provide evidence that the drug has some clinical benefit in humans and warrants further clinical testing. In Phase III, the sponsor conducts two large-scale, controlled clinical trials, again in sick patients, to demonstrate that the drug is both effective and has relatively tolerable side effects.[20]

[18] *See, e.g.*, Henry I. Miller and David R. Henderson, *The FDA's Risky Risk-Aversion, Policy Review* 5 (Oct. & Nov. 2007).

[19] 21 U.S.C. § 355(b), (d).

[20] Peter Barton Hutt and Richard A. Merrill, *Food and Drug Law: Cases and Materials* 527 n. 2 (2d ed. 1991). For a more detailed analysis of the two-trial requirement, *see* Jennifer Kulynych, *Will FDA Relinquish the "Gold Standard" for New Drug Approval? Redefining "Substantial Evidence in the FDA Modernization Act of 1997*, 54 Food & Drug L. J. 127, 129-130 (1999) (explaining that a second trial is due to the scientific requirement of replication, that the requirement is occasionally waived, and that biologics are less likely to face this requirement).

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 emphasizes a drug should be evaluated based on all the patients that enroll in a trial[21] and that requires the rate of false positive findings be set to 5 percent.[22] 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 across patient subgroups.[23] But it only accommodates this heterogeneity in a limited way. To begin with, it encourages sponsors to specify prior to conducting a trial – or in statistic parlance, specify *a priori* – the subgroups they plan to analyze.[24] When this is done, sponsors must account for the risk of spurious correlation due to multiple testing bias when setting their initial sample size and when analyzing data from a trial.[25] 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.

In many cases, however, the sponsor may not have enough information prior to trials to identify subgroups that may be especially sensitive to treatment. Even if it did, the FDA would require the sponsor to increase sample size so that its study gathers sufficient statistical information – or "power" – to accurately estimate treatment effects in those subgroups.[26] This additional cost may outstrip the financial resources available to many sponsors.

The FDA acknowledges that it will not always be possible to identify subgroups *a priori* and that exploratory analysis of trial data may be required to identify subgroups.[27] The FDA's approach to subsequent so-called *post hoc* subgroup analysis[28] depends on the average treatment effect in the full trial population and whether the outcome at issue concerns efficacy or safety.

[21] This is called the intent-to-treat population. It includes even the subjects that drop out. *See* Food and Drug Agency, *International Conference on Harmonisation; Guidance on Statistical Principles for Clinical Trials, Availability*, 63(179) Fed. Reg. 49,583, 49,593 (Sept. 16, 1998) (henceforth "Statistical Guidance" (§5.2.1 Full Analysis Set). The alternative is to evaluate the drug on the per-protocol population. In this case the drug is evaluated only among subjects who enroll, do not drop out, and follow all trial procedures.

[22] *Id.* at 49,291 (§3.5 Sample Size).

[23] *Id.* at 49,589 (§3.2 Multicenter Trials).

[24] *Id.* at 49,595 (§5.7 Subgroups, Interactions, and Covariates).

[25] *Id.* at 49,587 (§2.2.5 Multiple Primary Variables). The agency recognizes that the corrections are less severe where subgroups overlap and therefore produce correlated test statistics. If two subgroups are mutually exclusive, the outcomes across groups are statistically independent. The appropriate adjustment for multiple testing bias is the Bonferroni adjustment, which is described in Section 2. If the two subgroups overlap in part, the outcomes of the groups will be positively correlated. In that case the failure of a test on one group will contribute to the failure of a test for the overlapping second group. This reduces the risk of spurious results from statistical tests on the second group. Therefore, something weaker than the Bonferroni adjustment is required to correct for multiple testing bias.

[26] The basic formula for how large the sample *in each subgroup* must be to ascertain a treatment effect is $n=2\sigma^2 (z_{\alpha/2}+z_\beta)^2/d^2$. In this formula, *d* is how sensitive the researcher wants the estimate to be. In other words, the formula gives the sample size required to identify treatment effects that are at least as large as *d*. The formula also depends on σ, the variance of the treatment effect from the drug. The larger the variance, the more the noise in the data and thus the larger the sample size required to identify a treatment effect as small as *d*. Usually the researcher estimates σ from previous studies of the drug or the disease. The crucial statistical parameters are $z_{\alpha/2}$, which determines defines the confidence level of the analysis, and z_β, which the power of the analysis. The higher the confidence level of the analysis, the lower is the chance of a false positive. A confidence level of 95% (i.e., $\alpha = 0.05$) means that the probability that a significant result is false is just 5%. The critical value for this level of confidence in a two-sided test with a normal distribution is $z_{\alpha/2} = 1.96$. The higher is the power of an analysis, the lower is the chance of a false negative. A power level of 80% ($\beta = 0.8$) means that the probability that a drug with a positive treatment effect is mistakenly reported as having an insignificant treatment effect is 20%. The critical value for this level of power is $z_\beta = 0.842$. As is apparent from the formula, higher confidence levels and powers require a larger sample size.

[27] Food & Drug Agency, *supra* note 21,. at 49,595 (§5.7).

[28] *Supra* note 9 (for a precise definition of *post hoc* subgroup analysis).

If the sponsor cannot demonstrate that the average treatment effect is sufficient that the drug will be approved for the full trial population, *post hoc* subgroup analysis by itself cannot be used to obtain approval for a subgroup.[29] The FDA has not approved a single drug solely on the basis of *post hoc* subgroup analysis.[30] The FDA does permit a sponsor to use *post hoc* subgroup analysis to identify a subgroup and then to conduct a subsequent trial on that subgroup to confirm the findings of the subgroup analysis. But another trial can be very costly.[31] And there is no indication that the FDA allows sponsors to combine or "pool" the data from the subgroup in the initial trial with data on the subgroup from the subsequent confirmatory trial to establish a significant positive result for the subgroup across the two trials.[32] Such a combination would mitigate the sample size requirement – and hence costs – for the confirmatory trial.

Even if the sponsor is able to demonstrate that its drug is on average safe and effective 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,[33] but clinically and biologically defined subgroups have also been suggested.[34] If certain subgroups do not show efficacy or show side effects, the FDA may require that the drug label state 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 the two outcomes studied.[35] (The other outcome was mortality alone.) However, the drug was superior on both outcomes in the non-ACE subgroup.[36] 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 and asymmetric approach with respect to subgroup analysis.[37] 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.[38]

[29] *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).

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

[31] It is difficult to estimate from published data the cost of individual trials but it is possible to estimate the cost of different phases of clinical testing. The out-of-pocket plus opportunity costs of phase I are $45.7 million, phase II are $65.1 million, and phase III are $205.5 million. *See* DiMasi, et al., *supra* note 3, at 162 (table 1) and 165 (table 3). If phase III involves just two trials, then the cost of per phase III trial is over $100 million.

[32] We examine this approach in *infra* note 59.

[33] FDA, *Statistical Guidance*, *supra* note 34, at 49589 (§3.2)

[34] *See* Powers et al., *supra* note 40, at 2298.

[35] The primary endpoint of a trial is the outcome on which the treatment is judged.

[36] *See* Maggioni et al., *supra* note 41, at 99.

[37] The European Union's Committee for Proprietary Medicinal Products (CPMP) has a similar policy. *Id.* at 97.

[38] *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).

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.[39] To understand why, observe that the average treatment effect of a drug y_1 (relative to control y_0) is equal to its positive treatment effect among patient subgroups that benefit from the drug (i.e., $y_1 > y_0$) plus its *negative* treatment effects among those who do not (i.e., $y_1 < y_0$):

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

where $p = \Pr(y_1 - y_0)$ is the fraction of people who benefit from the drug. However, the drug only harms patients among whom it is contraindicated if doctors give those patients the drug. The problem is that the FDA's average-effects rule mathematically assumes this occurs, i.e., that doctors give the drug to every patient even if it harms those patients. If the FDA had more faith in doctors, it would instead estimate the value of a drug solely by its positive effects among the subgroup that benefits from the drug.[40] The FDA would not have to worry about harming patients for whom the drug is inappropriate because doctors would not give these patients the drug.

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

The FDA's conservative position on *post hoc* subgroup analysis is based on concerns that multiple testing creates a risk of spurious correlation, which can result in false positive drug approvals.[41] To illustrate, consider the following Statistics 101-type hypothetical. Suppose there is a population that can be divided into 10 mutually exclusive subgroups. For example, if the trial population ranges in age uniformly from 20 to 70, we can divide the group into 10 equally sized five-year age bins: 20-24, 25-29, etc. Suppose also that there is a drug that 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 for each subgroup is a normally distributed random variable with mean equal to zero and variance equal to one. The probability that the drug will be proven effective on the full population with a confidence level of 95% is 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 ($= 1 - 0.95^{10}$). This obviously raises the risk of 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 or penalty 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 $p = 0.05$ / (number of tests) or 0.005.[42] This correction is known as the Bonferroni adjustment. It ensures the probability of observing even one subgroup with significant

[39] *See* Anup Malani and Feifang Hu, *The option value of new therapeutics* 14, unpublished manuscript (2004). The bad matching may be because there is no way to determine which patients benefit and which do not, because doctors cannot or do not distinguish between patients that benefit and those that do not, or because doctors give the drug to patients that they know will not benefit from it.

[40] This value is larger than the average effect of the drug in clinical trials. Because $E(y_1 - y_0|y_1 < y_0) < 0$, $pE(y_1 - y_0|y_1 > y_0) > pE(y_1 - y_0|y_1 > y_0) + (1 - p)E(y_1 - y_0|y_1 < y_0)$.

[41] *See* Kulynych, *supra* note 33, at 141, John A. Lewis, *Statistical Issues in the Regulation of Medicine*, 14 Stat. in Med. 127, 132 (1995).

[42] 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.

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.[43]

The problem is that the FDA does not know the number of subgroups against which the sponsor will test its drug and cannot trust drug companies to honestly report this number. Therefore, the FDA cannot in practice impose a proper multiple testing penalty. This problem becomes exponentially more severe the larger the number of possible subgroups. Suppose the sponsor can also divide the adult population by gender (male/female) and by ethnicity (white/black/Hispanic/other) and the drug still has no treatment effect for any subgroup. Now the available subgroups has jumped from 10 based solely on age to 80 (= 10 × 2 × 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 less than 0.05 is 0.98 (= $1 - 0.95^{80}$)! 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.[44]

The FDA's response to this risk is to base its drug approval decision solely on average treatment effects for the full trial population. This bars any approval based on *post hoc* subgroup analysis without further clinical trials. But this response swings the pendulum of error too far in the opposite direction. Instead of approving some drugs with no treatment effect (false positives), the FDA's conservative policy rejects some drugs with positive treatment effects for some subgroups (false negatives) or increases the 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 based on the results for the full trial population is virtually zero.[45]

Before we can suggest a compromise solution, a good question to ask is whether *post hoc* subgroup analysis can ever provide sufficiently reliable information to warrant approval of a drug, even ignoring the risks from spurious correlation and opportunistic behavior. After all, when subgroups are not specified before a trial begins, the trial is not powered – i.e., does not have sufficient sample size and thus does not generate sufficient statistical information – to estimate subgroup effects in a manner that meets the usual standards for confidence (5%) and power (20%).[46] Compounding this problem is that subgroups by definition have smaller sample size than the full trial population.

We do not think, however, that these concerns render *post hoc* analysis useless. There are a number of technical reasons why such analysis may yield useful infor-

[43] Sandrine Dudoit and Mark van der Laan, *Multiple testing procedures with applications to genomics* (New York: Springer, 2008). *See also* the discussion in note 38.

[44] We do not dispute that *ex ante* drug sponsors have an incentive to produce drugs that actually work. Productive drugs surely generate more revenue than unproductive ones. The problem is *post hoc* subgroup analysis we examine occurs after a drug sponsor has found that the average patient does not benefit from its drug. Therefore it is facing a loss equal to the cost of its drug development expenses. The only reason it has to avoid spurious correlation due to multiple testing is litigation risk from failure-to-warn product liability suits, but these suits can be foreclosed by appropriate warnings. Moreover, they do not cover the risk of a drug being less effective than alternative treatment. Finally, it is unlikely that the average ineffective drug has expected litigation costs larger than the incurred cost of development.

[45] One would need five or more subgroups that do not benefit to "show" a benefit. This is roughly equivalent to the probability that five or more successes out of ten draws from a binomial distribution with p=0.05. The formula and value are $\Pr(x \geq 5) = 1 - \sum_{i=0}^{5} \binom{n}{i} 0.05^{i} 0.95^{n-i} = 0.00000275$.

[46] *See* Salim Yusuf et al., *Analysis and Interpretation of Treatment Effects in Subgroups of patients in Randomized Clinical Trials*, 266 JAMA 93, 94 (1991).

mation even without a larger sample. First, sample size calculations are based on *estimates* of the variance of treatment effects in the full trial population. Because those estimates themselves have sample variation, there is a positive probability that they are in fact too high, leaving samples larger than required for accurate identification of subgroup effects. Second, because 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. Third, because subgroups may be more homogenous than the full trial population, the subgroup may have smaller variation in treatment effects. This diminished variation means that doubling the number of subgroups does not double the required sample size for the trial. We shall demonstrate this in our analysis of the MGd trial.

3. *Rehabilitating post hoc subgroup analysis*

In this section we discuss two 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. Our 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 as *little additional cost* as possible from larger sample size in the initial trial or a new trial.

Our first proposal – the use of adaptive designs – should not come as a surprise. It has been advocated by biostatisticians and regulators as a way to reduce sample size or limit harm to trial participants even in the absence of varying treatment effects across patient subgroups.[47] The remainder of this section sketches how adaptive designs help address subgroup effects and discuss the sample size costs of those designs. The second proposal – deferred to the next section – requires a modified form of subgroup analysis to be performed by an outside consultant. It would also allow the FDA to approve a drug without the expense of further trials.

a. *What the FDA should continue to do*

Before explaining our reform proposals, we want to highlight two things that the FDA gets right in its current policy towards subgroup analysis.[48] First, the FDA

[47] *See, e.g.*, Donald A. Berry, *Bayesian clinical trials*, 5 Nature Reviews - Drug Discovery 27 (2006) (arguing that adaptive design can be employed to lower sample size and improve treatment outcomes for enrolled patients); Scott Gottlieb, *Speech before 2006 Conference on Adaptive Trial Design*, Washington, DC (July 10, 2006), *available at* http://www.fda.gov/oc/speeches/2006/trialdesign0710.html (last visited Jan. 28, 2009) (observing that adaptive designs can ends trials of drugs with severe side effects more quickly).

[48] However, we are concerned that the FDA makes other mistakes when aggregating information *across* clinical trials. Although the FDA may correctly apply multiple-testing corrections within trials, there is reason to be concerned that the FDA does not apply those corrections correctly across trials. 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. This risk of multiple testing across trials is partly addressed by the fact that a company must inform the FDA of every trial it conducts to support an IND and by the fact that many journals will not publish an article reporting the results of a trial that has not been reported to a trial registry such as clinicaltrials.gov. *See* 21 C.F.R. § 312.23 (2008); Catherine De Angelis, et al., *Clinical trial registration: a statement from the International Committee of Medical Journal Editors*, 141 Ann. Intern. Med. 477 (2004). We cannot, however, find evidence from stated FDA policy or practice that the FDA understands and properly addresses these concerns. That said, it is likely the case that the extreme cost of Phase III trials limits the frequency of opportunistic behavior across multiple trials.

is correct that, if the identity of sensitive subgroups is known prior to a trial, the sponsor should set the sample size for a trial so that the trial is able to estimate significant results for these subgroups.[49] 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 to analyze the single full trial population. Therefore, the costs of powering a trial to test *a priori* specified subgroups is less than proportional to the number of groups, as is often assumed.

Second, the FDA is also correct to use multiple-testing adjustments to avoid spurious results from analysis of *a priori* specified subgroups. The FDA is aware that Bonferroni adjustments may be too conservative because of the assumption that subgroups are independent. Because some patients fall in multiple subgroups or share biological features of members in other subgroups, treatment results in one subgroup may be related to effects in another subgroup. Thus, separate tests on two subgroups are less than two bites at the apple. Applying a Bonferroni adjustment in this case would result in an overcorrection for the risk of spurious correlation.

b. *Adaptive trials proposal*

i. *Background on adaptive trials*

The prototypical clinical trial is a fixed design trial. In this design, patients are randomized between a treatment group and a control group. The total number of patients enrolled in the trial – the sample size – and the fraction of patients assigned to the treatment group are fixed before the trial begins and remain the same until it ends. The sample size required to run such a trial depends on the minimal size of clinically-relevant treatment effects the sponsor wants to be able to identify and the variance of the treatment effect from the drug.[50]

The problem is that the sponsor may not know these parameters. Indeed, one of the purposes of the trial is to estimate these parameters. One solution is to use results from prior studies of the sponsor's drug or of related drugs. When such studies are not available or are unreliable, the sponsor can use what is called an adaptive design trial. Such a trial begins without firm or completely reliable estimates of the parameters above. Instead, the trial employs real time data gathered from early-enrolling patients to refine estimates of the parameters and adjust sample size or treatment allocation based on the new estimates.

There are two types of adaptive designs that use interim data to modify sample size while the trial is in progress. 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

[49] An even better approach may be to specify subgroups not by patient characteristics at baseline, but by an algorithm that has inputs of not only 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 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 f is specified before the trial begins, 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.

[50] Food & Drug Admin., *supra* note 21.

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 deliberately small and extends the trial if estimates of the treatment effect are 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.[51]

There are also adaptive designs intended to adjust the proportion of enrolled subjects assigned to the treatment group based on interim data analysis. 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.[52]

The FDA is open to the 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.[53]

ii. *Adaptive design for subgroup analysis*

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 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 designs.

How might an adaptive design be used for subgroup analysis? Consider a two-arm trial (treatment and control)[54] with sample size set to test just one hypothesis: the average treatment effect for all enrolled patients is zero. At some interim point, the sponsor or the independent data monitoring committee (IDMC) examines the data to determine if there is a subgroup of patients on which the trial should focus because they may be particularly responsive to treatment. There are two types of data that might be used to identify subgroups: baseline characteristics measured before the start of a trial alone or treatment outcomes measured during the course of the trial. In the first case, the sponsor looks for abnormal variation in a relevant covariate. For example, if there is much more variation than expected in treatment history or in the pre-trial progression of symptoms, the full trial population can be divided into subgroups using a cut-off based on the extent of prior treatment

[51] 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).

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

[53] Gottlieb, *supra* note 47.

[54] This is also called a parallel-armed trial.

or symptoms. In the second case, the IDMC may look at the relationship between certain covariates and treatment effects (The IDMC is used rather than the sponsor to ensure that the sponsor does not become un-blinded). 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.

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, which can be reformulated, although we do not do so here. In particular, the design may require that the sponsor pay a penalty, i.e., that the results be held to a more stringent or lower critical p-value before they are declared statistically significant, to account for the possibility of multiple testing. We explore these penalties in the Appendix.

Before we conclude our discussion of adaptive designs, it is worth noting an important weakness of these designs. Adaptive designs may not be optimal for identifying side effects of treatments because of ethical and profit considerations. If interim analysis suggests a particular subgroup may have 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.

c. *Proposal for independent post hoc subgroup analysis*

In this subsection we consider how it might be possible – working with a fixed, non-adaptive design trial – to use *post hoc* subgroup analysis to approve a drug without further trials. *Post hoc* subgroup analysis does not increase the risk of false positive drug approvals so long as the FDA makes appropriate multiple-testing corrections, but these corrections require knowledge of the number of tests the sponsor has performed. The sponsor cannot be relied upon to truthfully report the number of tests it has performed because of the financial incentive to have its drug approved.[55] Indeed, the FDA can be confident that the sponsor probably conducted

[55] We have considered the possibility that the FDA could specify, prior to a phase III trial, the exact subgroups the sponsor may 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 sponsor, as well as the FDA could specify subgroups based on subject matter in which they have a strong financial interest in doing so. We doubt there are valuable subgroups that the FDA could propose that the sponsor will not have 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 new drug applicants over earlier ones. It would also subtly reduce the incentive to innovate quickly.

more tests than that for which the FDA plans to adjust. Therefore, there remains a residual risk of false positive above 5%.

i. *Choosing an independent agent*

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 auditor, 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 auditor is truly independent, a topic to which we will turn in a moment.

There are two basic candidates for an independent auditor: 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 outside consulting firm. There are two weaknesses of the FDA. One, it has limited resources that make it difficult to maintain even the current level of scrutiny of new drug applications (NDAs).[56] Second, 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.[57] These pressures are unlikely to perfectly offset or to create an unbiased decision-maker. As a result, the FDA may conduct too much subgroup analysis – at the cost of false positives – or too little subgroup analysis – at the cost of false negatives or more costly approval.

The alternative is an outside statistical consulting firm. Many already exist to help sponsors design and analyze data from trials.[58] 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.[59] Perhaps the indirect influence of sponsors can be addressed by requiring the FDA to select the outside consultant to perform *post hoc* analysis, by blinding the sponsor to the outside 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

[56] *See* Institute of Medicine, *The Future of Drug Safety: Promoting and Protecting the Health of the Public* 193 (2007) (Drawing inspiration from PDUFA, one possible solution is to charge companies that seek drug approval for a patient subgroup higher user fees to fund subgroup analysis conducted by the FDA).

[57] *See* Gardiner Harris, *F.D.A. is Faulted for Drug-Safety Process*, New York Times (Sept. 20, 2006), *available at* http://www.nytimes.com/2006/09/22/business/22fdacnd.html?ex=1316577600&en=04c9d9824b892f3b&ei=5088&partner=rssnyt&emc=rss (last visited Jan. 30, 2009); Avery Johnson and Ron Winslow, *Drug Makers Say FDA Safety Focus Is Slowing New-Medicine Pipeline*, Wall Street Journal (June 30, 2008), *available at* http://online.wsj.com/article/SB121476772560213981. html?mod=hps_us_whats_news (last visited Jan. 30, 2009).

[58] *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.

[59] *See* Demski, *supra* note 11, at 57.

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.

A second weakness of using consulting firms is that "independence of the sponsor" is not the same thing as "motivated to reduce false positives." 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 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 those of the consulting firm. Most likely this is cost minimization, which may imply too many false negatives or false positives, whichever minimize the consulting firm's labor expense.[60]

ii. *Statistical methods to guarantee independence*

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.[61] For convenience, we shall speak as if the consulting firm has been chosen to conduct the analysis.

No-outcome data analysis. The first approach would provide the consultant with all the data from the trial *except variables that identify treatment assignments and health outcomes* and ask it to identify subgroups based on baseline characteristics that exhibit "remarkable and relevant variation" in the trial data.[62] (This is similar to one of the approaches used to identify subgroups for the adaptive design trials discussed in the last subsection.) 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 outside consultant.

[60] The outside 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 false negatives requires internalizing this negative externality. Since false negatives are unobservable, the FDA cannot directly incentivize the consulting firm to do so. The FDA certainly should not give the firm an incentive keyed to drug approval, because then it would incentivize the sponsor to replace false negatives with false positives.

[61] These methods do not address other problems such as the limited resources of the FDA or the insufficient motivation of outside consulting firms. If the statistical methods we discuss help ensure that the consultant truly cannot manipulate the data to increase false positives, then one might address the problem of a consultant's motivation by giving it stock in the sponsor (This is identical to extracting the outcome or a random subsample of data from the data archives of the drug sponsor. We consider granting the consultant stock instead because it is virtually impossible to separate the sponsor from knowledge of its data.). 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.

[62] While we focus on identifying subgroups by examining baseline characteristics with remarkable *variation*, it is equally valid to ask the consultant to identify subgroups by examining baseline characteristics which have a *distribution* which is remarkable in any significant way. This instruction would allow subgroups to be identified by characteristics with, for example, surprising high or low means or skew. The statistical method we present in the main text focuses on the second moment of the baseline characteristic only for purposes of illustration.

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 trials of the sponsor's drug 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 trial sample. Moreover, the consultant would have to keep in mind that any subgroup it 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.

Split-sample analysis. The second statistical method we propose to ensure that the consultant's choice of subgroups is not influenced by the drug company requires splitting the data from a trial into two parts. One part would be called the exploratory subsample and the other part the confirmatory subsample. Importantly, the FDA must split the sample to ensure the drug company has no influence, and it should be split randomly to ensure the samples are statistically independent. 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.[63] The sponsor would then be allowed to perform *post hoc* subgroup analysis on the confirmatory sample using only the subgroups identified from the exploratory subsample by the consultant. As before, the FDA would apply a multiple-testing penalty based on all the subgroups reported by the outside consultant. If, after such penalty, the confirmatory subsample validates the positive subgroup effects from the exploratory subsample, the FDA could approve the drug only for those subgroups.

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 statistically independent of the subsample analyzed by the consultant, the consultant's analysis cannot help the sponsor engage in data mining. Moreover, neither method requires that the FDA impose any additional multiple-testing penalty beyond one based on the total number of subgroups identified by the consultant.

Each statistical method also has its shortcomings. The weakness of the no-outcome data 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 the split-sample approach is the *post hoc* subgroup analysis, which would be underpowered even if performed on the whole trial sample, and is particularly underpowered if performed on subsamples. This will increase the risk of false negatives. This risk may be considered the cost of independence under this method. In short, the two statistical algorithms reduce, but do not eliminate, false negatives.

4. *An illustration with motexafin gadolinium*

In this section we illustrate our two statistical algorithms for ensuring independent *post hoc* subgroup analysis by applying them to a real world example: motexafin gandolium (MGd) for patients with brain metastases from solid lung tumor. MGd is sponsored by Pharmacyclics (ticker PYCY), a small biotech company that branded

[63] 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.

the drug as Xcytrin. We first provide some background on clinical testing of the drug and then discuss *post hoc* subgroup analysis of the testing results.

a. *Background on MGd*

Tumorous cancers in one part of the body often spread – or metastasize – to other parts of the body. In up to 24% of all cancer patients, they spread to the brain.[64] The risk is especially severe with lung cancer, where up to 50% of patients experience brain metastasis[65] and metastasis occurs earlier than with other cancers.[66] Most patients with brain metastases die. Median survival on whole brain radiation therapy, the typical conventional treatment, is only 4 months. For those who do manage to survive, however, there is a major risk of neurological impairment.[67]

MGd is a drug that demonstrated the ability to increase the radiation response of tumor cells in preclinical studies. Pharmacyclics sought to market the drug as a treatment for brain metastases. The company filed an investigational new drug (IND) application with the FDA to begin clinical testing of the drug in human patients. After a successful Phase I/II study,[68] the company began a Phase III study (called trial 9801) that enrolled patients with any type of cancerous tumor who developed brain metastases. Subjects were randomized to either whole brain radiation therapy (WBRT) alone (the control arm) or MGd and WBRT (treatment arm). Unfortunately, this study did not find a statistically significant treatment effect with respect to median survival or time to neurological impairment.[69]

One bright spot in trial 9801, however, was that patients specifically with lung cancer did experience a statistically significant extension of time to neurological impairment.[70] So Pharmacyclics conducted a second Phase III trial (called Trial 0211) targeting only lung cancer patients. Unfortunately, this second trial was unable to validate the results from the initial trial. This is illustrated in Table 2, which summarizes the results from trial 0211 and from lung cancer patients in trial 9801. According to the first panel, whereas the relative hazard rate for neurological impairment[71] was 0.61 (p = 0.05) in the initial trial, it was merely 0.78 and not significantly different from 1 (p = 0.1) in the second trial.

[64] J.B. Posner, *Neurological complications of cancer* (1995).

[65] M. Stuschke, W. Eberhardt, C. Pottgen, et al., *Prophylactic cranial irradiation in locally advanced non-small-cell lung cancer after multimodality treatment: Long-term follow-up and investigations of late neuropsychological effects*, 17 J. Clin. Oncol. 2700-2709 (1999); T.J. Robnett, M. Machtay, J.P. Stevenson, et al., *Factor affecting the risk of brain metastases after definitive chemoradiation for locally advanced non-small-cell lung carcinoma*, 19 J. Clin. Oncol. 1344-1349 (2001).

[66] Posner, *supra* note 74.

[67] Minesh P. Mehta, Patrick Rodrigus, C.H.J. Terhaard, et al., *Survival and Neurological Outcomes in a Randomized Trial of MotexafinGadolinium and Whole-Brain Radiation Therapy in Brain Metastases*, 21 J. Clin. Oncol. 2529 (2003).

[68] The sing-armed study found a 72% radiologic response rate. P. Carde, R. Timmerman, M.P. Mehta, et al., *Multicenter phase Ib/II trial of the radiation enhancer motexafin gadolinium in patients with brain metastases*, 19 J. Clin. Oncol. 2074-2083 (2001).

[69] Median survival was 5.2 months on treatment versus 4.9 months on control (p = 0.48). Median time to impairment of neurological function was 9.5 months on treatment versus 8.3 months on control (p = 0.95); Mehta et al., *supra* note 77, at 2533 (Fig. 2, panel C).

[70] The median patient on WBRT and MgD did not experience neurological impairment in 24 months, while the median patient on WBRT alone experienced impairment at 7.4 months (p = 0.048); Mehta et al., *supra* note 77, at 2533 (Fig. 2, panel C).

[71] Neurological impairment was judged by a battery of standardized neurocognitive tests. The tests were scored by blinded graders. Patients were said to be impaired if the composite score was at least 1.5 standard deviations worse than the mean of the test's age-adjusted distribution. Christina A. Meyers, et al., *Neurocognitive Function and Progression in Patients With Brain Metastases Treated With Whole-Brain Radiation and Motexafin Gadolinium: Results of a Randomized Phase III Trial*, 22 J. Clin. Oncology 157, 158 (2004) (The hazard rate for impairment is the rate at which patients are judged impaired, i.e., it is the fraction of additional patients judged impaired each month. The relative hazard rate is the ratio of the hazard rate in the treatment group to the rate in the control group).

Table 2. Subgroup treatment effects in trial 0211 and trial 9801.

	Trial 0211			Trial 9801		
Group	n	RH	raw p-value	n	RH	raw p-value
All	554	0.78	0.1	251	0.61	0.05
PT controlled?						
Yes	140	1.71	0.059	93	0.94	0.86
No	414	0.56	0.002	158	0.48	0.03
Newly diagnosed?						
Yes	259	0.59	0.032	109	0.47	0.046
No	295	0.92	0.69	142	0.74	0.37
Time from BM to Tx						
Tx ≤ 2 wks	274	0.6	0.022	119	0.78	0.5
2 < Tx ≤ 4 wks	161	0.78	0.41	69	0.63	0.29
Tx > 4 wks	119	1.23	0.5	63	0.33	0.09
Prior chemotherapy						
No	315	0.67	0.06	155	0.57	0.07
Yes	239	0.91	0.66	96	0.72	0.42
Trail B score						
Low	258	0.53	0.012	121	0.76	0.44
High	254	1.02	0.92	96	0.7	0.41
Country						
USA	185	0.39	0.0048	123	0.76	0.45
Netherlands	11	5.6	0.14	54	0.38	0.074
Canada	163	0.72	0.26	46	0.4	0.14
UK	0			21	1.11	0.89
France	117	1.49	0.21	7	0.82	0.89
Germany	47	0.61	0.26	0		

Notes. RH = relative hazard for MGd plus whole brain radiation therapy (WBRT) versus WBRT. Raw p-value does not adjust for multiple testing. Trial 9801 was original Phase III trial. Trial 0211 was second Phase III trial. PT = primary tumor. BM = brain metastases. Tx = treatment in treatment or control group.

Trying to explain the discrepancy between the trials and to salvage MGd for a new drug application (NDA), Pharmacyclics conducted a *post hoc* subgroup analysis. According to the company, this analysis revealed a problem at some of the study centers in France. Although the trial protocol required that subjects be randomized to treatment as soon as they were diagnosed with brain metastases, the French centers waited several weeks or more after diagnosis before randomizing subjects to treatment.[72] In the interim, the centers gave subjects chemotherapy[73]

[72] Ordinarily, a drug sponsor is responsible for ensuring that its study centers follow trial protocols. The company attributes the problem in this case to recruitment difficulties. Whereas the trial 9801 had 401 total patients, of which only 251 had lung cancer, trial 0211 required 554 lung cancer patients to have the statistical power to validate the positive results for lung cancer patients from the initial trial. This required the second trial to recruit patients from 90 treatment centers throughout the world, more than double the 40 centers involved in the initial Phase III trial. The company argues that it is difficult to precisely enforce the protocol with so many centers involved in a study. Personal communication with Richard Miller, former CEO of Pharmacyclics, Mar. 14, 2008.

[73] Chemotherapy is not thought to be a reliable treatment for brain metastases because chemotherapy relies upon drugs delivered by blood, and the brain tumor is therefore somewhat protected from chemotherapy drugs by the blood-brain barrier.

(hence these subjects are labeled "controlled" patients in the data). Moreover, subjects were ultimately randomized in trial 0211 only if their brain tumors progressed despite chemotherapy, i.e., their tumors were resistant to treatment.[74] So this was a self-selected group of tumors. Whereas "uncontrolled" subjects randomized immediately had a mix of resistant and non-resistant brain tumors, the controlled subjects ultimately randomized in the problematic French centers largely had tumors resistant to WBRT. This placed MGd at a disadvantage in these centers.[75]

Not surprisingly, these subjects also did not show benefits from MGd. Excluding these late-randomizing centers from the analysis revealed that the drug had a statistically significant effect on delay until onset of neurological impairment. As reported in the second of panel Table 2, the relative hazard rate for uncontrolled subjects on MGd in the 0211 trial was 0.56 (p = 0.02). In other words, while MGd proved effective among subjects with an uncontrolled brain tumor, this effect was masked by including subjects with a controlled brain tumor in the main analysis. Pharmacyclics filed an NDA with the FDA relying on this subgroup analysis. But the FDA did not credit the company's explanation and finally rejected its NDA in December 2007.[76]

Of course this is the company's explanation for its unsuccessful final Phase III trial, however, and it had a financial stake in getting MGd approved for some subgroup. Our aim is to scrutinize these claims by taking the role of an outside consultant and checking whether analysis of data in a manner that is independent of the financial interests of the sponsor identifies the same sensitive subgroups that the sponsor identified, namely the subjects with uncontrolled tumors.

b. *No-outcome data analysis*

Our first analysis examines the final Phase III study (trial 0211) data stripped of outcome variables. The idea is that an independent statistical consultant without outcome data would not be able to select subgroups that would financially benefit the company because it does not know whether any subgroups had better or worse outcomes than average among the full trial population. Instead, this consultant would identify subgroups by searching for baseline characteristics on which current trial subjects had excess variation relative to subjects in previous trials or in the population. These characteristics could then be used to define subgroups on which the drug company could perform *post hoc* subgroup analysis. If, and only if, that subgroup analysis suggested the subgroup responded positively to the drug should the FDA approve the drug for use in that subgroup.

To implement this algorithm, we compare the variation of certain medical characteristics in the trial 0211 sample with variation of those variables in the initial

[74] The implicit but reasonable assumption here is that a brain tumor resistant to chemotherapy is resistant to any other form of treatment.

[75] The selection story is a bit more complicated. The delay also screened out patients who had died from, inter alia, the brain metastases prior to randomization. Since early death is an indicator of a more severe brain tumor, this mortality screen likely selected for less severe brain tumors. It is probably the case that the selection on the basis of resistance to chemotherapy (which likely reduced the effect of MGd) was more significant than selection based on survival (which possibly increased the effect of MGd). The reason is that median survival following diagnosis with brain metastases is 4 months, so it is unlikely that mortality was a material screen in the first two weeks following diagnosis. Yet it is in these first two weeks that the company found a significant delay in neurological impairment among patients treated with MGd.

[76] Food & Drug Agency, *supra* note 21.

Phase III study (trial 9801) sample. The variables include features of the primary (lung) tumor, treatment of the primary tumor, features of the brain metastases, the treatment of the brain metastases before enrollment, and neurological impairment at baseline. Table 3 reports the ratio of variances of each variable across the two samples and the p-value for the hypothesis test, that the variances are equal across the samples after adjusting for multiple testing.

Table 3. Identification of subgroups in non-outcome analysis: ratio of variance for baseline characteristics in 0211 trial versus 9801 trial.

Covariate	Ratio	p-value
Extracranial metastases?	1.73	0.0006
USA?	0.89	0.0006
PT controlled?	0.81	0.0025
Canada?	1.39	0.0068
PT resected without recurrence?	0.50	0.0137
Baseline Trail B score	1.63	0.0158
PT is large cell carcinoma?	0.60	0.0267
Days from diagnosis to randomization	5.29	0.0267
PT treated, <=1 month follow-up?	1.72	0.0554
PT is non-small-cell carcinoma?	0.80	0.1200
RPA status	0.71	0.1519
Baseline weight	1.27	0.2171
Sex	0.98	0.2171
PT has squamous histology?	0.78	0.2249
PT treated, >1 month follow-up?	0.85	0.2249
PT treated and progressing?	1.24	0.2734
PT has other histology?	1.90	0.3143
Prior chemotherapy?	1.04	0.4045
Baseline delay score	0.90	0.4045
Baseline COWA score	0.88	0.4045
Baseline Trail A score	0.59	0.4045
Baseline height	1.11	0.4075
Multiple BM lesions?	0.89	0.4154
PT newly diagnosed and/or untreated?	1.01	0.5843
Age 65+?	0.96	0.6924
Karnofsky Performance Score >= 90?	1.01	0.7010
PT is adenocarcinoma?	0.99	0.8469
Caucasian?	1.07	0.8469
Baseline recall score	0.97	0.8469
Baseline recognition score	0.98	0.9619
Other race?	1.00	0.9931
Notes. P-value adjusts for multiple testing. PT = primary tumor (i.e., lung cancer). BM = brain metastasis.		

Eight subgroups stand out. The 0211 trial had excess variation in the variables: days from diagnosis of brain metastases to randomization, extracranial metastases, baseline Trail B score, and whether the study center was in Canada. The delay variable captures some of the company's concern that patients in some French centers received chemotherapy for a few weeks before being randomized and that the brain masses that survived this chemotherapy were more resilient. The trail B is a test of cognitive ability where the subject is asked to follow a "trail" on a sheet of paper with his pencil. A lower score is better: it indicates less time was required to follow the trail. It also indicates that the brain metastasis probably has not advanced beyond the point where it can be treated.

The 0211 trial also had insufficient variation in the variables: primary tumor ("PT") resected without recurrence, primary tumor is large cell carcinoma, primary tumor controlled, and study center is in the U.S. The PT resected variable is one of five categories into which a primary tumor is categorized at the time a patient is randomized.[77] The resected category indicates that the primary tumor was surgically and successfully treated and the patients only remaining concern is the brain metastases. The PT controlled variable is the complement of the variable that the company identified as being responsible for the failure of the 0211 trial.

Having identified eight subgroups with remarkable variation, we now check to see if MGd was particularly effective amongst these subgroups in the full 0211 data. For binary variables, subgroups are defined by their two states. For continuous variable subgroups in the 0211 data are defined by whether a characteristic lies above or below the median for that characteristic in the 9801 data. Table 4 summarizes our subgroup analyses with a multiple-testing adjustment that accounts for 16 tests (eight variables with excess variation and two subgroups for each variable). We find four subgroups have significant treatment effects, i.e., lower rate of neurological impairment: subjects with little delay before randomization, subjects at U.S. study centers, subjects with low (good) baseline Trail B scores, and subjects with uncontrolled brain metastases.

Table 4. Treatment effect in 0211 trial for subgroups identified by no-outcome data analysis.

	Relative hazard	p-value
Canada? No	0.82	0.3782
Canada? Yes	0.72	0.3782
Extracranial metastases? - Low	0.85	0.5467
Extracranial metastases? - High	0.71	0.2800
PT is large cell carcinoma? No	0.78	0.2800
PT is large cell carcinoma? Yes	0.90	0.8853
Days from diagnosis to randomization - Low	0.55	0.0325
Days from diagnosis to randomization - High	1.14	0.6646
USA? No	0.99	0.9500
USA? Yes	0.39	0.0325
Baseline Trail B score - Low	0.50	0.0480
Baseline Trail B score - High	0.96	0.8853
PT controlled? No	0.56	0.0320
PT controlled? Yes	1.71	0.1573
PT resected without recurrence? No	0.73	0.1536
PT resected without recurrence? Yes	3.07	0.3378
Notes. Relative hazard is for subjects on MGD and WBRT versus subjects on WBRT only. P-value adjusts for multiple testing.		

[77] The tumor may have been (1) "newly diagnosed," which means the primary tumor and the brain metastasis was diagnosed at the same time, (2) surgically removed or resected without recurrence, (3) treated for less than 4 weeks without clear indication that it has progressed, (4) treated for greater than 4 weeks with no sign of progression, or (5) treated for any amount of time with evidence of progression.

Thus the no-outcome data analysis would support Pharmacyclics' case for approval for subjects with uncontrolled brain tumors. The other subgroups that survive no-outcome data analysis are consistent with the company's theory that MGd works on less resistant tumors. Subjects randomized quickly and subjects in the U.S. are less likely to have received chemotherapy before randomization. As for subjects with low trail B scores, these are subjects whose tumors are not so far along enough to seriously impede subjects' cognitive capacity. It makes sense the treatment is also likely to work in these cases.

c. *Split-sample analysis*

Our second (and preferred) proposal is to split the data from trial 0211 into two subsamples, have an outside consultant identify subgroups via *post hoc* subgroup analysis on one (exploratory) sample, and then allow the drug sponsor to validate significant treatment effects for the other (confirmatory) sample. Only if a subgroup identified by the outside consultant demonstrates statistically significant effects – after a multiple-testing penalty – in the confirmatory sample should the FDA approve the drug for that subgroup. Because the outside consultant does not have access to the confirmatory sample, and because the FDA would only credit evidence of treatment effects from the confirmatory subsample, the consultant cannot rig its analysis to help the drug sponsor.

We begin our simulation of the split-sample analysis by randomly dividing the 0211 trial sample into a 20% exploratory sample and an 80% confirmatory sample.[78] The second step is to conduct a subgroup analysis of the exploratory sample where subgroups are defined according to baseline variables. The results of this analysis are presented in Table 5. The vertical panel labeled "Value A" gives the relative hazard of neurological impairment and the p-value for each variable at value 0 for binary variables and at the first quartile for continuous variables. The vertical panel labeled "Value B" gives the relative hazard of neurological impairment and the p-value for each variable at value 1 for binary variables and at the third quartile for continuous variables.[79] None of the p-values have been adjusted for multiple testing.[80] We choose a subgroup (now defined both by a given variable and a specific value for that variable) for the third step in our simulation if it has a p-value less than 0.05, i.e., if we can be 95% confident that the membership in the subgroup improves treatment effects. The two subgroups we identify in this manner are (1) enrollment at center other than one in France and (2) not having previously received the chemotherapy drug carboplatin. The relative hazard rate (into neurological impairment) for subjects outside France and on MGd is 0.32 (p = 0.05363) and for subjects not having received carboplatin and on MGd is 0.039 (p = 0.07833).

The last step in the split sample analysis is to estimate the treatment effects for these subgroups in the confirmatory sample. This analysis reveals that the relative hazard rate for MGd among subjects outside France is 0.676 (raw p = 0.038, adj. p = 0.076) and among subjects who had not previously been treated with carboplatin is 0.905 (raw p = 0.58, adj. p = 0.58). Even with the high correlation (r = 0.39)

[78] Our selection of a 20-80 split is arbitrary. Further statistical analysis is required to determine the optimal split of the sample. The larger the exploratory sample, the greater is the probability of identifying a subgroup that benefits, but the smaller it is the greater the probability that one will be able to confirm that it benefits.

[79] The first two columns of data present the ratio of relative hazards among the two subgroups defined for each variable and the p-value for this ratio.

[80] No multiple testing penalty is required when analyzing the exploratory sample. The only purpose in that sample is to identify certain subgroups, relative to other subgroups, that have a better response. Moreover, the size of the exploratory sample is too small to pass any of the usual statistical tests (e.g., p = 0.05), let alone ones that adjust for spurious correlation from multiple testing.

in membership across these subgroups, the effect of MGd outside France is not statistically significant after adjusting p-values for multiple tests (on two groups). The effect of the type of prior chemotherapy is not significant even without the multiple-testing adjustment.

So it appears that our spit sample analysis – unlike the no-outcome data analysis – fails to support Pharmacyclics' explanation for why the 0211 trial did not validate the 9801 trial. To be precise, however, all we have shown is that this particular sample we split randomly did not validate Pharmacyclics' claim. Perhaps another random split would validate their claim. Indeed, a better way to characterize the value of the split-sample analysis for eliminating alleged false negatives is to ask, what fraction of splits would validate Pharmacyclics' claim that MGd works in the subgroup of patients whose brain tumor was not controlled via chemotherapy?

Table 5. Identification of subgroups in split sample analysis: treatment effects in exploratory sample of 0211 trial.

		Raw	Value A				Value B			
	Factor	p-value	Value	n	RH	p-value	Value	n	RH	p-value
france	9.11	0.027	0.0	84	0.32	0.05363	1.0	24	2.94	0.17954
pr.carbo	10.87	0.048	0.0	83	0.39	0.07833	1.0	25	4.26	0.18103
ptstagen	0.21	0.059	3.0	--	1.72	0.42139	5.0	--	0.37	0.07388
neurbl20	12.67	0.088	0.0	98	0.63	0.28314	1.0	10	7.99	0.14457
ltmptbm	0.20	0.098	0.0	49	1.32	0.61953	1.0	59	0.27	0.09389
sympbl12	4.67	0.100	0.0	--	0.47	0.12400	2.0	--	2.18	0.36172
neurbl18	0.22	0.110	5.0	--	0.63	0.29587	5.0	--	0.63	0.29587
basetrla	2.08	0.130	-0.3	--	0.51	0.15601	2.7	--	1.06	0.90295
euroaus	3.86	0.130	0.0	64	0.34	0.11599	1.0	44	1.33	0.61980
metachro	4.25	0.130	0.0	58	0.28	0.11374	1.0	50	1.19	0.73841
prior	3.66	0.140	0.0	67	0.40	0.13240	1.0	41	1.48	0.53320
neurbl21	2.44	0.150	0.0	--	0.57	0.24937	1.0	--	1.40	0.54303
sympbl7	4.71	0.150	0.0	--	0.57	0.19931	0.0	--	0.57	0.19931
prantflg	5.46	0.160	0.0	86	0.54	0.20250	1.0	22	2.97	0.33156
bmlesn	5.17	0.170	0.0	25	0.20	0.14510	1.0	83	1.01	0.98116
sympbl2	1.96	0.170	0.0	--	0.47	0.16551	1.0	--	0.92	0.84827
f.prrad	0.26	0.180	0.0	84	1.02	0.95880	1.0	24	0.27	0.13672
pr.csca	3.23	0.180	0.0	68	0.42	0.15428	1.0	40	1.37	0.61691
karnofsk	1.80	0.190	80.0	--	0.51	0.16114	90.0	--	0.92	0.85675
ltmptdx	0.28	0.190	0.0	51	1.09	0.86437	1.0	57	0.31	0.14175
neversmk	4.81	0.210	0.0	101	0.58	0.23766	1.0	7	2.80	0.37554
prantday	1.46	0.210	0.0	--	0.55	0.21645	44.5	--	0.80	0.62516
pbm.all	4.57	0.220	0.0	100	0.59	0.24367	1.0	8	2.68	0.39687
txgt1.4	3.57	0.220	0.0	90	0.58	0.23659	1.0	18	2.06	0.43119
controll	3.46	0.230	0.0	86	0.57	0.22919	1.0	22	1.98	0.45638
motorlbl	0.38	0.240	5.0	--	0.71	0.45018	5.0	--	0.71	0.45018

Notes. Value A is 0 for binary variables and the first quartile for continuous variables. Value B is 1 for binary variables and the third quartile for continuous variables. Values are specified in the columns labeled value. Sample size is given in columns labeled "n". RH = relative hazard for MGd plus whole brain radiation therapy (WBRT) versus WBRT. P-value for Value A and Value B adjusts for multiple testing. Factor gives the ratio of variance at Value A and Value B. P-value for factor does not adjust for multiple testing.

To conduct this analysis we drew 100 splits of the 0211 trial data and ranked the subgroups in the exploratory stage in order of statistical significance of relative hazard for neurological impairment, just as Table 5 did for our initial split. The first row of Table 6 provides the distribution of p-values that the "uncontrolled" subgroup takes across the hundred splits, with a smaller p-value indicating a larger significance of that difference. We find that, in 35% of draws, uncontrolled has a p-value of less than 0.05. Thus in 35% of cases, the "uncontrolled" subgroup advances to the validation stage. Further, in 31% (= 11/35) of these cases, the effect among the uncontrolled subgroup is validated in the confirmatory sample after a conservative Bonferroni multiple-testing adjustment that accounts for the number of subgroups that emerge from the exploratory analysis in each sample split. See row 3 of Table 6. In other words, in only 11 % of cases, the split-sample analysis confirms Pharmacyclics' claim that MGd works so long as the patient's brain tumor is not previously treated with chemotherapy.

Table 6. Distribution of p-values for "uncontrolled" subgroup treatment effects, by sample, in various sample splits.

	P-value ranges					
	0-0.01	0.01-0.05	0.05-0.10	0.10-0.25	0.25-1	Total
Exploratory sample	16	19	11	19	35	100
Validation sample (raw p-values)	14	18	2	1	0	35
Validation sample (adj p-values)	2	9	9	10	4	35

Notes. First row gives distribution of p-values for "uncontrolled" subgroup across 100 sample splits after adjusting for multiple testing. Second and third rows gives distribution of p-values in the validation sample for the 35 splits where uncontrolled group is selected in exploratory sample, i.e., adjusted p-value for uncontrolled subgroup in exploratory sample is less than 0.05. P-values in second row do not adjust for multiple testing. P-values in the third row apply an overly conservative Bonferroni adjustment for multiple testing.

Because in the case of MGd we have not one but two Phase III trials, there is one other test we can do – in the spirit of split sample analysis – to verify Pharmacyclics' claim about uncontrolled patients. We can check whether the subgroup effects identified by the company, by the no-outcome data analysis, and by 11 % of the split-sample analyses above can be validated by the subsample of lung cancer patients in the 9801 trial. As can be seen in Table 2, the uncontrolled subgroup is indeed associated with statistically significant treatment effect (RH = 0.48, p = 0.03).

CONCLUSION

Roughly one in five drugs that enter clinical testing fails to prove that it is effective and safe. Even in phase III, the failure rate is 36%. Sometimes failure is just that: the drug has no value. But other times a drug is right for some patients and wrong for others. Denying approval for a drug that benefits some patients, but not the average patient, increases the costs of drug development but not the benefits. Ideally, one would like to salvage such a drug by allowing its use for the non-average patient who would benefit.

The challenge is bad behavior or moral hazard by drug sponsors. With *post hoc* subgroup analysis (or data dredging in less polite language), sponsors will nearly always be able to find a subgroup of patients who appear to benefit from a drug. Currently, the FDA addresses this risk of spurious correlation by requiring sponsors to validate their findings with additional clinical trials. But this may cost tens of millions of dollars, and in turn increase the price of drugs.

This paper offers a combination of institutional designs and statistical methods that can limit the risk of spurious findings – or false positives – from *post hoc* subgroup analysis without requiring additional, whole trials. Our proposal for adaptive trials allows the use of subgroups to revise the hypothesis tests in a trial with little additional sample size. Our proposal for independent statistical analysis, when combined with subgroup analysis, without outcome data or subgroup analysis validated on a split sample, can actually identify subgroups without increasing the risk of false positives or requiring additional sample size. In other words, our proposals offer an approach to reduce the rate of failure in clinical trials, without a higher risk of false positives or with minimal additional clinical testing costs.

While our proposals may be helpful, we recognize they are not panaceas. It is possible that a drug which fails the average patient standard used by the FDA may not in fact be helpful to any subgroup of patients. Our methods for identifying false negatives – drugs that have value for a subgroup of patients but not the average patient – may not identify every false negative. Finally, even if a drug is approved for the right subgroup, doctors may use it for the wrong subgroups or use it off label. These are forms of false positives that we cannot address.

The reason we believe our proposals are worth pursuing, however, is that the alternatives – using an average-patient standard or requiring additional trials – are worse. As we explained in Section 1, the average-patient standard implicitly assumes that doctors *always* give the drug to the wrong subgroup. While doctors may not be perfect at sorting patients to drugs, we do not believe they are as bad as the FDA's standard assumes. Moreover, trials – especially Phase III trials – are very expensive. Trials focusing on subgroups are even more costly. Because fewer patients are members of the subgroup than the full trial population, a trial focusing on a subgroup will take longer complete recruitment. This, in turn, increases the opportunity costs of the trial.

It is natural to wonder whether our proposal to eliminate false negatives in drug approval can also help eliminate false positives. That is, should our proposals be used to identify drugs that have a no effect or a side effect for a patient subgroup, even though they are effective and safe for the average patient and thus FDA-approvable? The answer is complicated. As we have mentioned, adaptive trials are not helpful for identifying side effects. With respect to our other proposals, the answer depends on whether there is reason to believe that the FDA has a skewed incentive to disapprove helpful drugs.

The central problem that motivates our proposals is moral hazard by drug sponsors. Sponsors have a financial incentive to find subgroups that show benefits from a drug whether or not the subgroups actually benefit from the drug. Likewise our proposals are only useful as against the FDA if it too suffers from moral hazard – though different in kind. If the FDA has an incentive to disapprove a drug for certain patients even short of statistically significant evidence that it harms them, perhaps because the agency is concerned about press criticism, then there is also a reason to fear the FDA will overuse *post hoc* subgroup analysis to identify subgroups that do not benefit from a drug.

If the FDA does not suffer such moral hazard, however, it should be allowed to conduct the *post hoc* analysis to eliminate false positives. Since the FDA is conducting the analysis, it knows the number of subgroups it has tested and thus can apply a multiple-testing penalty to its own analysis. The only reform that we can unambiguously recommend to address false negatives is that the FDA be sure to apply multiple testing penalties when it conducts or requires that drug sponsors check side effects in certain subgroups. Just as data dredging allows a drug company nearly always to find some subgroup of patients who appear to benefit from a drug, the more subgroups the FDA tests for side effects, the more likely the agency will find side effects when they do not actually exist.

We are aware that our paper raises a number of statistical questions, the answers to which would help the FDA refine regulations that allow certain *post hoc* subgroup analyses to inform approval decisions. For example: What is the proper multiple testing penalty for adaptive design trials? Can sponsors use moments other than the variance, such as the mean or skew, to identify subgroups in the no-outcome data analyses? What are the appropriate proportions (20-80 or something different) to use when dividing a sample into an exploratory and a confirmatory subsample? Should the multiple-testing penalty applied to tests on the confirmatory subsample account for the fact that subsample is smaller than the full sample? Under what conditions will the no-outcome data analysis eliminate more false negatives than the split-sample analysis. We leave these questions for future research by statisticians.

Modern medicine is becoming increasingly sensitive to the importance of heterogenous treatment effects. Evidence for this is the great push towards personalized medicine. Personalized medicine faces an important practical problem: although treatment may be tailored to a smaller group of patients, the sample size required to demonstrate that it is safe and effective remains the same. Either trials have to become more efficient at generating information or the FDA has to reduce its standard for judging drugs. Multiple testing offers a way to achieve the former, but only if the FDA can control false positives associated with data mining. By controlling these false positives, perhaps the statistical proposals can increase the productivity of clinical trials and reduce cost of developing personalized treatments.

APPENDIX

The table below summarizes the four basic options in a subgroup-identifying adaptive design and our speculation as to the appropriate 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.

Table 1. Multiple testing penalties in adaptive trials.

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 multiple testing penalty, so long as the trial must proceed until the sample size specified prior to starting the trial is achieved. 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, so long as the data employed in the interim analysis (patients' baseline characteristics) are unrelated to the data relevant for estimation of the treatment effect (patients' treatment assignment and health outcomes), there is in essence additional testing of treatment effects in the interim analysis. The general idea is, so long as one analyzes a subset of the final data set that contains no information about treatment effects and does not increase the *number* hypotheses to be tested, there is no multiple testing penalty for changing the nature of the hypothesis to be tested with the final data.

If outcome data were used to identify subgroups, there should be a multiple testing 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 during the interim analysis. Even with that subset of the final 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 not 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 to truthfully 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.[81]

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.

[81] 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. Suppose, for example, that interim analysis after 10% of the sample is enrolled reveals that only young patients have a significant treatment effect. We recommend, when the sponsor tests the hypothesis that the treatment effect among young patients in its final empirical analysis that the sponsor be required to use the entire sample and not just young subjects. Specifically, the sample tested should include the elderly patients from the initial 10% sample even though they are not nominally the subject of the hypothesis test. Our crude logic is 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 cost or penalty to the sponsor of having to include the excluded subgroups in the final empirical analysis. Including the elderly from the 10% sample automatically makes it less likely that the sponsor will be able to show that the drug works among young patients.