IASLC

Statistical Considerations for Subgroup Analyses

Xiaofei Wang, PhD,[a,*] Steven Piantadosi, MD, PhD,[b] Jennifer Le-Rademacher, PhD,[c] Sumithra J. Mandrekar, PhD[c]

[a]Department of Biostatistics & Bioinformatics, Duke University, Durham, North Carolina
[b]Department of Surgery, Brigham and Women's Hospital, Boston, Massachusetts
[c]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota

Received 17 September 2020; revised 8 December 2020; accepted 12 December 2020
Available online - 26 December 2020

Introduction

Randomized clinical trials (RCTs) are conducted to evaluate the effect of an experimental treatment on outcomes of a target patient population. Eligibility criteria for large trials are often broad to ensure that the trial results can be generalized to a larger patient population. Subgroup analyses, either specified a priori or post hoc, are perfo rmed to evaluate the treatment effect specific to a subgroup of treated patients. Regardless of whether a subgroup analysis is specified a priori or post hoc, investigators must consider inflated false-positive rates, chance differences in observed treatment effects, low power for the comparisons of interest, and interpretation of the subgroup results. Subgroup analyses in clinical trials and observational studies have been discussed in regulatory agency guidelines,[1–3] and comprehensive review of this topic has been published in applied statistical journals.[4,5] This article reviews key statistical concepts associated with planning, conducting, and interpreting subgroup analyses in RCTs. It also highlights the pitfalls in conducting undisciplined subgroup analyses.

Defining Subgroups

In RCTs, subgroups are often defined by demographic variables, such as age, sex, and race. Subgroups can also be defined by variables that are prognostic of clinical outcomes or predictive of better treatment effect, such as disease severity, previous therapies, genotype, and biomarker status. Using the same definition of subgroup enables comparison of outcomes between similar subgroups across different clinical trials. If subgroups are defined by continuous variables, it is preferable to use well-established or published cutoffs. For example, it is common in cancer research to use age cutoffs of 40 years and 65 years to classify patients into age groups, such as the following: less than 40 years as adolescent and young adult, 40 to 65 years as adult, and greater than 65 years as older adult. When common cutoffs are not available for a continuous biomarker, cutoff points can be identified by data-driven approaches such as simple percentiles (e.g., median) or visualization with statistical graphs. When multiple variables are expected to contribute to the definition of a subgroup, a continuous prediction score calculated from a multivariable prediction model may be used to categorize patients into low, moderate, or high risk, indicating an ordinal increase in the risk of an adverse outcome or the severity for a disease condition. The cutoff points of a novel biomarker or a risk score are often chosen to maximize the difference in outcome or the difference in treatment benefits between the subgroups. When extensive preliminary data are available, statistical methods can be used to identify optimal boundaries to define subgroups on the basis of high-dimensional variables.

Caution is necessary when defining subgroups by a data-driven approach. For example, are the subgroups plausible given what is known about the clinical, pharmacologic, and biological mechanisms? Additional factors that affect the validity and the reproducibility of subgroup analyses are missing data and measurement error associated with the variables for searching and defining subgroups. Ill-defined and unreproducible subgroups could lead to biased and unreproducible estimates of treatment effects and hard-to-interpret findings. See Lipkovich et al.[6] for a

*Corresponding author.

Disclosure: *The authors declare no conflict of interest.*

Address for correspondence: Xiaofei Wang, PhD, Department of Biostatistics & Bioinformatics, Duke University, 2424 Erwin Rd, Durham, NC 27710. E-mail: xiaofei.wang@duke.edu

ISSN: 1556-0864
https://doi.org/10.1016/j.jtho.2020.12.008

comprehensive review of data-driven methods to define subgroups using clinical trial data. Although subgroups can be defined using continuous variables, the loss of power and efficiency from categorization of continuous variables should be considered when defining subgroups in this manner.[7]

Heterogeneity of Treatment Effect

The effect of a treatment may vary by baseline patient characteristics (e.g., sex, age, and race), tumor's molecular profile and genotype, and medical centers. Targeted agents or immunotherapy, for example, are expected to be effective for patients with certain genetic mutations or protein overexpression levels owing to their mechanism of action. In such cases, a favorable treatment effect in a targeted subgroup is anticipated whereas the treatment would likely have no effect or even a deleterious effect in the complementary subgroup, or in the unselected population.

The statistical term to describe differential treatment effects by subgroups is the interaction between the treatment and the subgroup variable. If there is no interaction between treatment and subgroup, the treatment effect is the same across subgroups, as in Figure 1*A*. If there exist differential treatment effect across subgroups, two types of treatment-by-subgroup interaction have been described—quantitative versus qualitative interaction. In quantitative interaction, the size of the treatment effect varies across subgroups, but the treatment effects in different subgroups are in the same direction. Figure 1*B* reveals an example of quantitative interaction in which although the magnitude of the treatment effect differs between the two subgroups, they are in the same direction. In such cases, the therapeutic implications are unchanged by the interaction. In qualitative interactions, the treatment benefit is in favor of the experimental treatment in one subgroup but is unfavorable or neutral for the other subgroup. Figure 1*C* reveals an example of qualitative interaction in which the treatment effects are in opposite directions. These circumstances carry important therapeutic consequences.

One example of qualitative interaction is from the Iressa Pan-Asia Study trial for NSCLC.[8] As illustrated in Figure 2, gefitinib (an EGFR inhibitor) was associated with better progression-free survival compared with control (carboplatin + paclitaxel) in EGFR mutants, whereas gefitinib was associated with worse progression-free survival compared with control in EGFR wild types. In this example, EGFR mutation is a predictive biomarker for gefitinib in NSCLC. A predictive biomarker, or more generically a predictive factor, gives information about the treatment effect of a new intervention relative to a control. In contrast, a prognostic biomarker (factor) provides information about the patient's overall cancer outcome, regardless of treatment. For example, the 21-gene recurrence score is associated with recurrence in tamoxifen-treated patients with node-negative, estrogen receptor–positive breast cancer.[9] It is important to point out that a predictive biomarker (factor) could have either a quantitative or a qualitative interaction with treatment. More discussion on

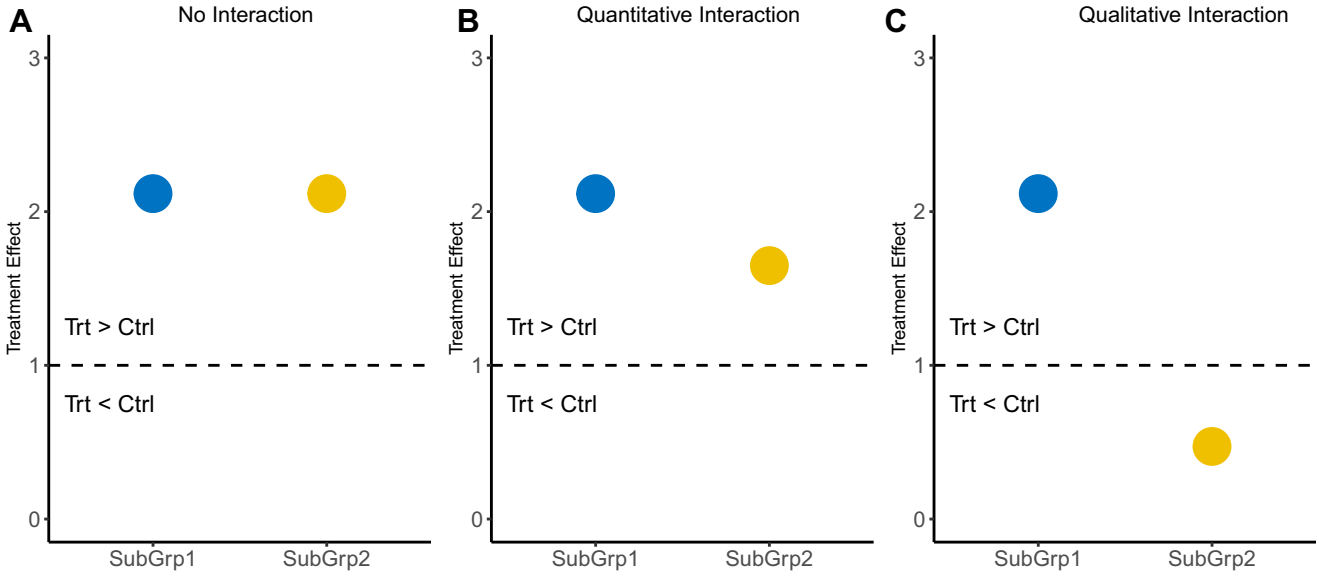

Figure 1. Types of treatment-by-subgroup interaction: (*A*) no interaction, (*B*) quantitative interaction, and (*C*) qualitative interaction. The y axis denotes the Trt effect, for example, the HR for Ctrl versus Trt. A Trt effect 1 indicates that the treatment is as effective as the Ctrl; a Trt effect greater than 1 indicates that the Trt is better than the Ctrl; and a Trt effect less than 1 indicates that the Trt is worse than the Ctrl. Ctrl, control; HR, hazard ratio; SubGrp, subgroup; Trt, treatment.

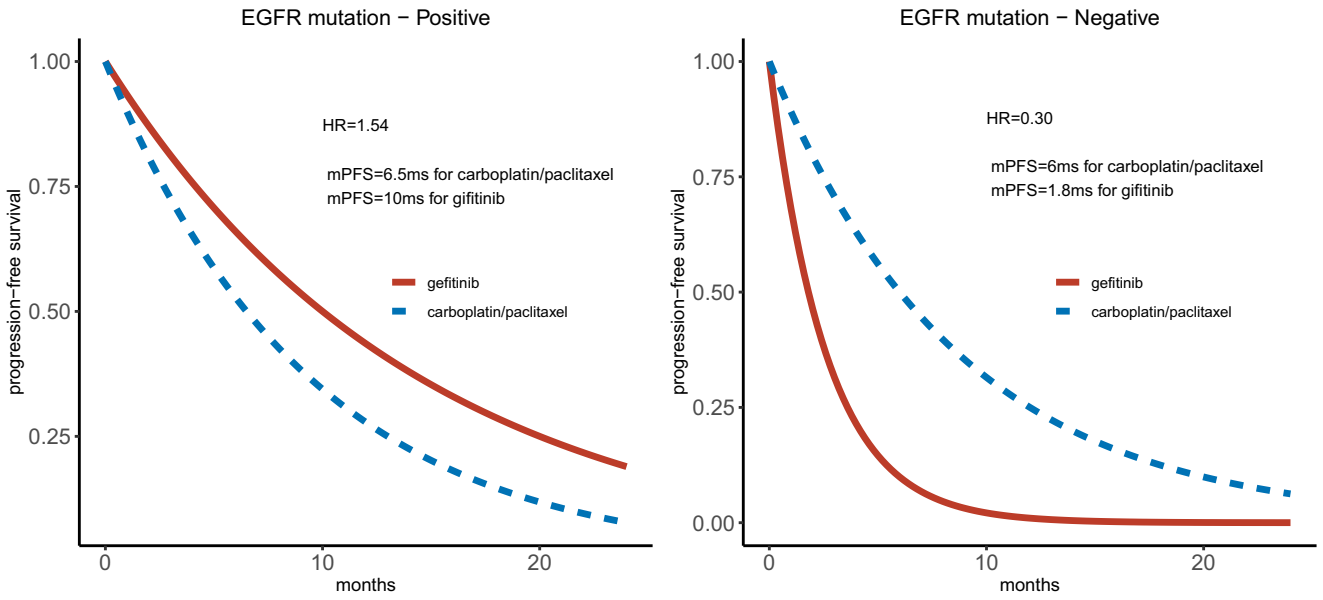

Figure 2. Schematic plot of qualitative interaction between treatment and EGFR mutation observed in Iressa Pan-Asia Study trial. HR, hazard ratio; mPFS, median PFS; PFS, progression-free survival.

prognostic and predictive biomarkers in cancer research can be found in Mandrekar and Sargent[10] and Ballman.[11]

Meta-analysis evaluates the heterogeneity of treatment effects reported from multiple clinical studies and synthesizes the estimates of the treatment effects. When heterogeneity is not significant, pooling the information from multiple studies is preferred as it usually increases power. More information on meta-analysis can be found in Higgins et al.[12]

Control of Type I Error Rate and Statistical Power

Statistical regression models are often used to test for treatment-by-subgroup interactions. The independent variables in these models include the main treatment effect, the variable defining the subgroups, and the treatment-by-subgroup interaction terms. Other prognostic or confounding variables of the outcomes can also be included in these models to adjust for their potential association with the outcomes of interest. The estimates of treatment-by-subgroup interactions from these models can be interpreted as the differential treatment effects, that is, the differences in the treatment effects among various subgroups. When there is sufficient statistical evidence to indicate an interaction, it is important to state the treatment effects by subgroups and not use the overall average treatment effect on all patients to characterize the effect of treatment across all subgroups. For RCTs designed to evaluate the effect of an experimental treatment in the overall patient population, statistical tests to detect a treatment-by-subgroup interaction are often underpowered. Therefore, failure to detect a significant interaction does not automatically imply the absence of treatment effect heterogeneity.

Forest plots are a useful tool to visualize the treatment effect across subgroups.[13,14] They usually present confidence intervals around the point estimate of subgroup treatment effect, and the size of the symbols used is proportional to the size of subgroups. The treatment effect estimates and standard errors used to create the forest plot should not be derived from fitting separate models for the subgroups, but from a model on the full set of data, also including the interaction term between the treatment and the subgroup variables if interaction is significant.

It is important to control for type I error rate when testing for treatment effects in multiple subgroups. The practice of conducting multiple statistical tests on a large number of subgroups to find a subgroup with a significant treatment effect runs the high risk of inflating the overall type I error rate. For example, the chance of making at least one false-positive claim for conducting 10 repeated tests at 5% significance level is 40% when there is no treatment effect. The Bonferroni's method addresses the issue of multiplicity by lowering the critical value (making it harder) to declare statistical significance and therefore controls the false-positive rate to its specified level. There are many other methods that can be used to control the overall type I error rate when subgroup analysis is of interest.

In the development of cancer-targeted agents and immunotherapy, RCTs are conducted with an interest in demonstrating the treatment effect in all patients or in a targeted subgroup. A simple gating approach is to sequentially test for the overall treatment effect

Table 1. Comparison of Confirmatory Versus Exploratory Subgroup Analyses

Attributes	Confirmatory Subgroup Analysis	Exploratory Subgroup Analysis
Purpose	Confirm heterogeneity of treatment effect across subgroups	Explore the possibility of treatment effect heterogeneity across subgroups
Prespecified and post hoc	Always prespecified	Prespecified or post hoc
Subgroup definition	Predefined	Predefined or to be discovered
End points	Always prespecified	Prespecified or post hoc
Null[a] and alternative[b] hypotheses	Prespecified	Prespecified or post hoc
Number of subgroup analyses	Limited	Moderate or large number
Multiplicity	Carefully chosen MTP for proper overall type I error rate[c] control	Control overall type I error rate[c] at higher level, for example, 10%, or without any control
Size of subgroups	Predetermined	Without planning
Power[d]	Low type II error rate[e] and sufficiently powered	High type II error rate[e] Very low power
Interpretation	Clear	Unclear, sometimes impossible to interpret correctly
Reporting	Mandatory	Selective

[a]Null hypothesis (H_0): a statement about the effect of the new treatment that the investigators hope to not be true. For example, the new treatment is worse than or as effective as the control.
[b]Alternative hypothesis (H_1): a statement about the effect for the new treatment that the investigators hope to be true. For example, the new treatment is superior or not the same as the control.
[c]Type I error rate: the probability of making a false claim that the new treatment is superior or not the same as the control (H_1 is true) when the new treatment is not as effective compared with the control (H_0 is true).
[d]Power: the probability of rejecting the null hypothesis (H_0) at a given size of effect for the new treatment (H_1 is true). Power is equal to 1 minus type II error rate.
[e]Type II error rate: the probability of making a claim that the new treatment is the same as the control (H_0 is true) when the new treatment is more effective than the control (H_1 is true).
MTP, multiple testing procedure.

before evaluating the effect within subgroups, and once the null hypothesis of the overall treatment effect fails to be rejected, no subgroup analysis should be conducted. One criticism of the gating approach is that one might miss the opportunity to identify a targeted subgroup of patients who may benefit from the new treatment on the basis of previous preclinical or early clinical observations. An alternative to sequential approach is the Bonferroni's method in which the type I error rate is split between the overall test and the subgroup-specific tests. In general, when a specified level of overall type I error rate is allocated to testing multiple hypotheses with respect to different patient subgroups or end points and each test is conducted at its allocated significance level, the overall type I error rate will not exceed its specified level regardless if these tests are dependent or not. The Bonferroni's approach is easy to implement but could be overly conservative, as it ignores the correlation between outcomes of patients in the subgroups and those of the overall patient group. More flexible multiple testing procedures have been proposed, and they take the correlation and the testing order into account and allow recycling the significance level after rejecting a hypothesis to test the remaining hypotheses at increased significance levels. Reviews of these more flexible methods, such as the fallback procedure[15] and the marker sequential test procedure,[16] can be found in Matsui et al.[17] and Dmitrienko et al.[18]

As mentioned, the power of treatment-by-subgroup interaction test is often low in a RCT in which the treatment effect of the overall population is of primary interest. For a test of treatment-by-biomarker interaction with equal numbers of subjects in all subgroups, the interaction test will have roughly four times the variance of an overall treatment effect. To compensate, the size of the trial would have to be increased accordingly which is seldom done. This increase may be mitigated if the interaction is of large size or qualitative, but tests for these can also have low power under certain situations.[19,20] Besides the size and nature of an interaction, the allocation ratio among subgroups also has an impact. Equal allocation yields the highest power. See Wang et al.[21] for further discussion of strategies to optimize the power of treatment-by-subgroup interactions in biomarker-stratified clinical trials.

Confirmatory Versus Exploratory Subgroup Analysis

Confirmatory subgroup analyses are intended to evaluate subgroup-specific treatment effects and require that the subgroup analyses be specified in the design of

the trial. In confirmatory cases, the subgroups must be clearly defined and the end points delineated with a small number of hypotheses about subgroup-specific treatment effects. A plan for control of the type I error rate and adequate power for testing of the overall and subgroup treatment effects must be prespecified. Prespecified hypotheses of confirmatory subgroup analysis are based on strong biological evidence. Findings from confirmatory subgroup analysis may be used in applications for regulatory approval of the new treatment in the subgroups. See Ondra et al.[22] for examples and reviews of confirmatory subgroup analysis in cancer clinical trials that test targeted agents and immunotherapy.

Unlike confirmatory subgroup analysis, exploratory subgroup analyses are intended to obtain preliminary evidence and to generate hypotheses for future investigation. Exploratory subgroup analyses are often conducted either in a post hoc manner or are prespecified at the design stage but without sufficient power to formally test for subgroup-specific treatment effects. Plan for prespecified exploratory subgroup analyses should include subgroup definition, end points, and how the subgroup analyses will be carried out. An example of prespecified exploratory subgroup analysis is the differential treatment effect between EGFR-positive and EGFR-negative patients in the Iressa Pan-Asia Study trial, achieved through testing the treatment-by-biomarker interaction.[8] Results of exploratory subgroup analyses should be clearly labeled as such and reported (in tables and forest plots) with point estimates and confidence intervals without p values. If the investigators are so inclined to the report of p values, they should be adjusted for multiplicity to control familywise type I error rate or false discovery rate. See Table 1 for a summary of the features of confirmatory and exploratory subgroup analyses. Lagakos[23] and Wang et al.[24] provide guidelines for presentation of subgroup analysis in medical journals.

Conclusions

Reporting of treatment effects on subgroups of patients is common in the medical literature. Although knowing how well a new treatment works in patients with a specific biomarker or a specific combination of disease stage and histology type are important for the patients and their clinicians to make informed treatment decisions, it is important to be cautious when interpreting the results of subgroup analyses. Common challenges with subgroup analyses include poor definitions, low statistical power, and inflated type I error owing to multiple hypotheses testing. The decision to conduct a subgroup analysis should depend on biological justification or on evidence of treatment heterogeneity from existing preliminary data. When a large number of unplanned subgroup analyses are conducted, the treatment may spuriously seem to be effective in one or more subgroups. One should always evaluate the validity of subgroup analysis results on the basis of biological plausibility, sample size for the subgroup, proper type I error control, and power.

Acknowledgments

The research work was partially supported by P01CA142538 (to Dr. Wang), R01AG066883 (to Dr. Wang), and P30CA15083 (Mayo Clinic Comprehensive Cancer Center grant; to Dr. Le-Rademacher and Dr. Mandrekar) from National Institutes of Health.

References

1. Lewis JA. Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. *Stat Med*. 1999;18:1903–1942.
2. Food and Drug Administration. Enrichment strategies for clinical trials support approval of human drugs and biological products. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enrichment-strategies-clinical-trials-support-approval-human-drugs-and-biological-products. Accessed January 15, 2021.
3. European Medicines Agency. Guideline on the investigation of subgroups in confirmatory clinical trials. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf. Accessed January 15, 2021.
4. Dmitrienko A, Muysers C, Fritsch A, Lipkovich I. General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *J Biopharm Stat*. 2016;26:71–98.
5. Alosh M, Huque MF, Bretz F, D'Agostino RB Sr. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Stat Med*. 2017;36:1334–1360.
6. Lipkovich I, Dmitrienko A, D'Agostino RB Sr. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat Med*. 2017;36:36–196.
7. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006;332:1080.
8. Mok TS, Wu YL, Thongprasert S, et al. Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. *N Engl J Med*. 2009;361:947–957.
9. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*. 2004;351:2817–2826.
10. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *J Clin Oncol*. 2009;27:4027–4034.
11. Ballman KV. Biomarker: predictive or prognostic? *J Clin Oncol*. 2015;33:3968–3971.

12. Higgins JP, Thomas J, Chandler J, et al., eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, United Kingdom: John Wiley & Sons; 2019.
13. Cuzick J. Forest plots and the interpretation of subgroups. *Lancet*. 2005;365:1308.
14. Ou F, Le-Rademacher JG, Ballman KV, Adjei AA, Mandrekar SJ. Guidelines for statistical reporting in medical journals. *J Thorac Oncol*. 2020;15:1722–1726.
15. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Stat Med*. 2009;28:586–604.
16. Freidlin B, Korn EL, Gray R. Marker sequential test (MaST) design. *Clin Trials*. 2014;11:19–27.
17. Matsui S, Choai Y, Nonaka T. Comparison of statistical analysis plans in randomize-all phase III trials with a predictive biomarker. *Clin Cancer Res*. 2014;20:2820–2830.
18. Dmitrienko A, Millen B, Lipkovich I. Multiplicity considerations in subgroup analysis. *Stat Med*. 2017;36:4446–4454.
19. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*. 1985;41:361–372.
20. Piantadosi S, Gail MH. A comparison of the power of two tests for qualitative interactions. *Stat Med*. 1993;12:1239–1248.
21. Wang X, Zhou J, Wang T, George SL. On enrichment strategies for biomarker stratified clinical trials. *J Biopharm Stat*. 2018;28:292–308.
22. Ondra T, Dmitrienko A, Friede T, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *J Biopharm Stat*. 2016;26:99–119.
23. Lagakos SW. The challenge of subgroup analyses-reporting without distorting [published correction appears in N Engl J Med. 2006;355:533]. *N Engl J Med*. 2006;354:1667–1669.
24. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med*. 2007;357:2189–2194.