Clinical Trials: Discerning Hype From Substance

Thomas R. Fleming, PhD

The interest in being able to interpret and report results in clinical trials as being favorable is pervasive throughout health care research. This important source of bias needs to be recognized, and approaches need to be implemented to effectively address it. The prespecified primary analyses of the primary and secondary end points of a clinical trial should be clearly specified when disseminating results in press releases and journal publications. There should be a focus on these analyses when interpreting the results. A substantial risk for biased conclusions is produced by conducting exploratory analyses with an intention to establish that the benefit-to-risk profile of the experimental intervention is favorable, rather than to determine whether it is. In exploratory analyses, *P* values will be misleading when the actual sampling context is not presented to allow for proper interpretation, and the effect sizes of outcomes having particularly favorable estimates are probably overestimated because of "random high" bias. Performing exploratory analyses should be viewed as generating hypotheses that usually require reassessment in prospectively conducted confirmatory trials. Awareness of these issues will meaningfully improve our ability to be guided by substance, not hype, in making evidence-based decisions about medical care.

Ann Intern Med. 2010;153:400-406. www.annals.org
For author affiliation, see end of text.

There is a compelling need for broader access to effective health care and for patients and caregivers to not only have a choice, but an informed choice. An important component of the strategy to address these challenges should be to increase the influence of evidence-based medicine when deciding whether and how best to intervene for the treatment or prevention of diseases. In turn, the clinical research to provide such evidence should be driven by the passion to enhance the quality of health care, in which research integrity is not meaningfully compromised by financial and professional conflicts of interest.

We must recognize the sources of bias and understand their influence in order to discern hype from substance when pursuing evidence-based medical care (1, 2). A major source of bias in health care research comes from the interest to find evidence suggesting that an intervention for treatment or prevention of disease has a favorable benefit to-risk profile. In industry-sponsored clinical trials, obtaining positive results provides substantial financial benefits to companies and both financial and professional benefits to their employees. There is almost an obsession to find statistically significant *P* values—that is, the magical 2-sided *P* value of 0.05 or less, with effects in the favorable direction, or more concisely, a 1-sided *P* value of 0.025 or less—for establishing benefit. Government sponsors also have considerable bias toward positive results. Leverage for federal research funding is improved by claims of success in enhancing the quality of health care. The National Institutes of Health program officers informed me that positive results from a National Institutes of Health–sponsored trial establishing a safe and effective regimen to reduce mother-to-child transmission of HIV in developing country settings (3, 4) were also very influential in efforts to increase National Institutes of Health funding for prevention research. Many journal editors also have a preference toward publishing articles presenting positive results, thereby increasing academic researchers' interest in obtaining positive results because publications are of integral importance to their professional reputation, timing of promotion, and salary (5–10). Caregivers who seek more therapeutic options that they can offer to patients have an interest in viewing results to be positive. These various interests are appropriate. However, they induce conflicts of interest that can adversely affect health care if they are pursued in a manner that induces substantial bias in the evaluation of treatment effects.

In this article, I discuss the risks of obtaining misleading results when exploratory analyses are conducted with the intention to identify positive evidence for interventions and consider measures to reduce these risks. Specific attention is given to the importance of understanding sampling context when interpreting *P* values, the influence of "random high" bias, the importance of confirmatory trials, and the psychological influence of the driving goal to establish benefit.

See also:

Print
Key Summary Points 401

Web-Only
Conversion of graphics into slides

The Importance of the Sampling Context in Interpreting Data

Consider clinical trials intended to provide a reliable evaluation of the benefit-to-risk profile of an intervention, such as registration trials in a regulatory setting. In these trials, there should be clear specification of and focus on the prespecified primary analyses of the primary and secondary end points. It is common practice to conduct hundreds of exploratory analyses by performing supportive analyses of primary end points, analyses of additional outcome measures of biological activity as well as clinical efficacy, analyses with and without multivariate adjustments,

analyses overall and within subgroups, and analyses repeated over calendar time on accruing data. Although exploring the data is useful, it becomes particularly problematic when exploratory analyses are conducted with the intention to find evidence more favorable than what was provided by the prespecified primary analysis. The extreme *P* values that are obtained from extensive exploration of the data can be very misleading if the multiplicity of these analyses is not properly addressed.

Insight about the interpretation of exploratory analyses is provided by an experience when I visited a hospital nursery nearly 40 years ago. The nursery was the central gathering place for the hospital's newborns in those days, and I was surprised that there were 20 babies of 1 sex and only 2 of the other. As anyone might do, I computed a *P* value for the likelihood that an imbalance this extreme would have occurred by chance if in truth there were an equal sex distribution in the population at birth. The 2-sided *P* value was 0.0001, so the probability of an imbalance this extreme occurring by chance was 1 in 10 000. The *P* value is correct even though the sample size is very small, so what could explain this paradox of obtaining compelling evidence against a hypothesis that nevertheless is known to be true? The explanation is that I did not walk into the hospital with the intention to gather prospective data to assess and report on this hypothesis. Rather, the data generated the hypothesis. Had the unusual sex ratio not occurred at that hospital at that time, I would be commenting now on some other equally rare coincidental observation made at some other time. On the other hand, had I walked into the hospital with the intention to assess a prespecified primary hypothesis about sex balance in newborns, evidence against such balance would have been far more persuasive. The important insight is that a *P* value is interpretable only when you understand the sampling context from which it is derived.

A North Central Cancer Treatment Group (NCCTG) colon cancer trial illustrates the risk in clinical research of being misled by exploratory analyses. This trial provided borderline-significant evidence of benefit, in which adjuvant chemotherapy using levamisole alone or the combination of 5-fluorouracil plus levamisole reduced the mortality rate by approximately 30% after resection of Dukes stage III colon cancer (11) (**Figure 1**). Exploratory analyses suggested that the benefit was meaningful only in women and younger patients. In such settings, a confirmatory trial is needed for the overall result. Should eligibility in that confirmatory trial be restricted to women and younger patients to achieve enrichment and enhanced sensitivity? Although the answer should be influenced by the biological plausibility that either sex or age is a true treatment-effect modifier, we should not query clinicians about this biological plausibility after showing them the data. Just as biostatisticians can obtain low *P* values by doing exploratory analyses, similarly clinicians can readily provide biological explanations for why a factor may be a treatment-effect modifier by calling on their clinical insights. More reliable insight about the biological plausibility for effect modifiers is obtained by asking clinicians to predict, before they are shown results, which factors will be treatment-effect modifiers.

A recent industry-sponsored trial underscores this point. When the sponsor was disappointed that treatment had less effect rather than the predicted greater effect on the prespecified targeted subgroup, a group of key opinion leaders was assembled and developed a white paper to defend that the unexpected reverse relationship observed could be scientifically motivated. Shortly thereafter, the trial's biostatistician identified a coding error that, when corrected, contradicted their retrospectively developed theory by revealing that the greater treatment effect was in fact in the originally targeted subgroup.

The confirmatory Cancer Intergroup Trial (CIT) was conducted with inclusive enrollment by sex and age because there was no biological justification that these factors should affect efficacy (12). It confirmed that a regimen of 5-fluorouracil plus levamisole provides a 30% reduction in mortality (**Figure 1**). Of interest, age and sex again seemed to be treatment-effect modifiers (**Table 1**); surprisingly, the benefit of 5-fluorouracil plus levamisole occurred predominantly in men and older patients.

Key Summary Points

The desire to report favorable results is an important source of bias in clinical trials.

The goal of research should be to determine whether the experimental intervention has a favorable benefit-to-risk profile rather than to prove that it does.

Trial reports should clearly specify and focus on analyses of the primary and secondary end points.

In exploratory analyses, an understanding of the sampling context is necessary to interpret *P* values, and "random high" bias leads to overestimation of outcomes having particularly favorable estimates.

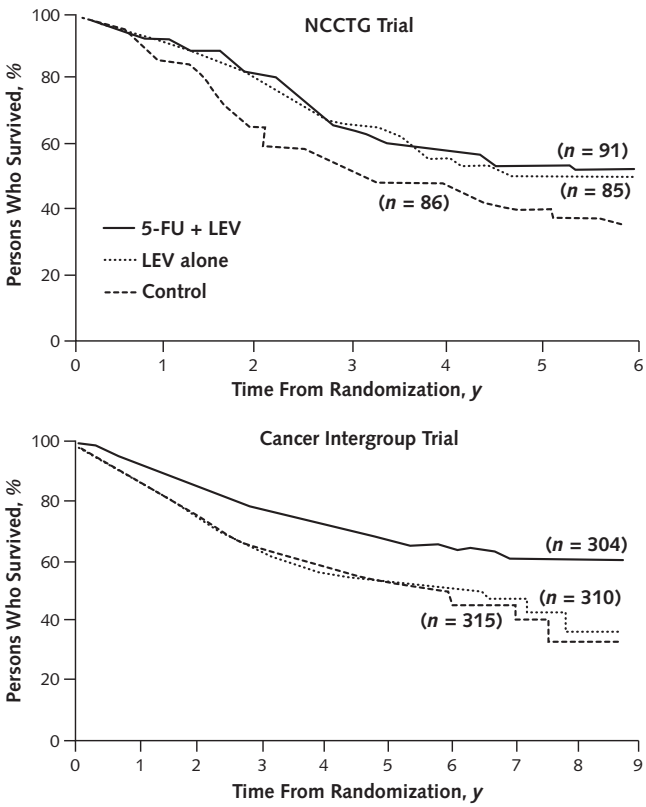
Exploratory analyses should be labeled as hypotheses-generating and requiring confirmation.

Meta-analyses that include hypothesis-generating data may be biased.

Random High Bias and Regression to the Mean

Estimates of treatment effect as well as *P* values should be interpreted with caution when data are explored. In exploratory analyses, particularly when conducted in search of favorable evidence, the effect sizes of outcomes having particularly favorable estimates are probably overestimated because of random high bias that can in later experience with the treatment lead to "regression to the mean." To explain this bias, it

Figure 1. **Survival effects in colorectal cancer of surgical adjuvant therapy when using LEV alone or the combination of 5-FU and LEV in the NCCTG trial and the Cancer Intergroup Trial.**



Adapted from reference 34. 5-FU = 5-fluorouracil; LEV = levamisole; NCCTG = North Central Cancer Treatment Group.

should be recognized that the true effect size on any outcome measure is always unknown and is estimated with variability. Suppose many exploratory analyses are conducted with the goal of obtaining evidence of benefit. When choosing among many measures, each estimated with variability, that which yields the best result will tend to be an overestimate of truth. This random high bias increases as the number of analyses and the variability of each increase.

Random high bias happens whether the search for extreme results from several statistical estimates occurs in medical research settings or elsewhere. Major League Baseball, with its tradition of extensive use of statistics, provides a classic illustration. A first-year player with the most outstanding performance is recognized as the Rookie of the Year. Although the career-long performance for the typical player will, on average, improve after his first year, there is a paradox that the Rookie of the Year does not show such improvement. For the most recent 30 players who were awarded Rookie of the Year, 80% regressed or statistically performed less impressively in their careers after their rookie seasons (based on the classic measures of batting average for hitters and earned run average for starting pitchers). This paradox is explained, not by the effects of pressures or distractions due to recognition as being the best, but rather by random high bias.

In the setting of the Rookie of the Year, to have the best performance among a large number of players, the winner of the award will not only be a good player but will also tend to be someone who overachieved in his rookie year; hence, such players will tend to regress to their true level of performance over subsequent years. In medical research, when exploratory analyses are done with the goal to achieve positive results, analyses that seem to be most favorable will tend to be random overestimates of true treatment effect on those measures. The likelihood of random high bias would be greatly reduced if our goal were the pursuit of truth rather than the achievement of positive results, because our focus would not systematically or preferentially be on the positive outliers. Note also that the prespecified primary analysis of the prespecified primary end point does not have this source of random high bias because it is reported whether it is an over- or underestimate of truth.

Morton and Torgerson (13) observe that "Regression to the mean affects all aspects of health care." An important setting, as noted by Bland and Altman (14), arises when treatments are evaluated to normalize extreme levels of baseline measures. When these measures are particularly variable, such as $CD4^+$ cell counts, blood pressure, serum cholesterol levels, body mass index, or bone mineral density, normalization of values that were extreme due to natural variation may be misinterpreted as benefit from treatment (13–19).

Placebo-controlled trials that evaluated single-fraction (20) and multiple-fraction (21) preoperative radiation in patients with rectal cancer provide an example of random high bias. Evidence suggested little or no survival benefit in prespecified primary analyses. However, apparent benefit was found in each trial in exploratory subgroup analyses. After finding evidence of benefit (2-sided $P = 0.014$) of preoperative irradiation in the subset of patients with Dukes stage III rectal cancer in the single-fraction trial (20), the authors

Table 1. **Effects of 5-Fluorouracil Plus Levamisole on Patient Survival Presented Overall and Within Subgroups, by Sex and Age***

Analysis Group, by Sex and Age	Hazard Ratio of Mortality	
	North Central Cancer Treatment Group Trial (*n* = 177†)	Cancer Intergroup Trial (*n* = 619†)
All patients	0.72	0.67
Women	0.57	0.85
Men	0.91	0.50
Young	0.60	0.77
Old	0.87	0.59

* The table provides the hazard ratios for mortality for 5-fluorouracil plus levamisole relative to control groups.
† Overall sample size in the clinical trial.

stated, "Thus, there can be few arguments against its universal use." This claim, based on analyses that had random high bias, was convincingly refuted by the much larger placebo-controlled confirmatory trial of 824 patients conducted by the Medical Research Council (22). This trial found persuasive evidence of no effect from either single-fraction or multiple-fraction preoperative radiation regimens, either overall or in the subgroup of patients with Dukes stage III rectal cancer (23). The trial also found no benefit in patients with abdominoperineal resections—the exploratory subgroup that had favorable results in the trial (21) that evaluated the multiple-fraction regimen.

Random high bias may provide misleading results more than once during a clinical development program. A series of trials was conducted to evaluate the effect of abetimus sodium on the occurrence of renal flare in patients with systemic lupus erythematosus. No effect was found in the primary analysis of time to renal flare (2-sided $P = 0.51$) in an initial trial, but a positive signal on this end point (2-sided $P = 0.007$) was obtained in the exploratory high-affinity subgroup (24). A second trial was conducted to evaluate the effect on time to renal flare in these high-affinity patients. The prespecified primary analysis was negative and was influenced by a lack of long-term differences (after 18 months) between abetimus sodium and placebo groups in the Kaplan–Meier estimates of time to renal flare. However, an exploratory analysis that truncated follow-up at 12 months after randomization suggested a transient benefit (25). A third trial, conducted in high-affinity patients and with follow-up truncated at 12 months in its prespecified primary analysis of time to renal flare, was negative, and the trial was terminated because of futility at an interim efficacy analysis (26). Although the prespecified primary analysis in each of these 3 trials was negative, random high bias in exploratory analyses led to the appearance of benefit in the exploratory high-affinity subgroup in the initial trial and in the exploratory 12-month analysis in the second trial. These biased analyses contributed meaningfully to the justification to conduct additional large, phase 3 randomized trials evaluating abetimus sodium, in a clinical setting in which it probably does not have favorable effects.

The Importance of Confirmatory Trials

Exploratory analyses, when properly viewed, are hypothesis-generating exercises. Hypotheses generated by particularly favorable results usually require formal evaluation in subsequent adequate and well-controlled trials (27–29). Such studies have been called *confirmatory trials* by the International Conference on Harmonisation guidelines (27).

Confirmatory trials, such as the CIT (12), also are needed in many settings in which prespecified primary analyses achieve only borderline significance (29). Although the CIT in **Figure 1** established benefit for 5-fluorouracil plus levamisole, it is noteworthy that the trial indicated that levamisole is not effective. This might be surprising given that the previous NCCTG trial provided borderline-significant evidence of benefit for each adjuvant regimen against the control. Yet, an approximate (1-sided) P value of 0.025 does not mean that there is a 97.5% chance that treatment works. We should be careful to distinguish the frequentist concept (that is, the P value actually means that 2.5% of all clinical trials of inactive treatments would show results at least this favorable) from the Bayesian concept of a posterior probability (that is, the probability that the treatment truly is effective given the data from the trial as well as the pretrial likelihood of treatment benefit). When the NCCTG trial was initiated in the mid-1970s, many colon adjuvant trials had previously been conducted, and no therapy had been proven effective. Hence, the pretrial likelihood that a new treatment would be effective was very low. If this pretrial likelihood is approximately 4%, then **Figure 2** reveals the probability that a regimen truly is effective is only 60%, even after an initial trial has shown benefit on the primary end point with a 1-sided P value of 0.025. Consistent with this insight, for the 2 positive results in the NCCTG trial, the larger, more reliable confirmatory CIT indicated that the favorable outcome for 5-fluorouracil–levamisole was a true-positive result and that for levamisole alone was a false-positive result. Of interest, **Figure 2** also reveals that if a regimen is found to have a beneficial effect on the primary end point with a 1-sided P value of 0.025 in the confirmatory trial as well, then the probability that the regimen truly is effective is 98%. Although these calculations involve oversimplifications, such as dichotomizing into truly effective versus truly

Figure 2. **Using pretrial likelihood regarding whether a treatment is truly effective in an initial trial and a confirmatory trial.**

Initial Trial*

Result of the trial	Truth: Positive	Truth: Negative		
Positive	36	24	60	⇒ A "positive trial" will be a "true positive" with probability: 36 / 60 = 0.60
Negative	4	936	940	
	40	960	1000	

Confirmatory Trial*

Result of the trial	Truth: Positive	Truth: Negative		
Positive	540	10	550	⇒ A "positive trial" will be a "true positive" with probability: 540 / 550 = 0.98
Negative	60	390	450	
	600	400	1000	

When a trial is positive, the probability that the treatment is truly effective depends on the pretrial likelihood that the treatment is effective. In both scenarios, the false-positive error rate is 0.025 (corresponding to requiring a 2-sided P value ≤ 0.05 in the favorable direction in order for the trial to be positive) and the false-negative error rate is 0.10 (corresponding to having 90% statistical power).

* For the initial trial, suppose the pretrial likelihood that the treatment is effective is 4% (i.e., 40/1000). For the confirmatory trial, suppose the pretrial likelihood that the treatment is effective is 60% (i.e., 600/1000).

ineffective regimens, the importance of confirmatory trials is clearly illustrated for such settings. Suppose the CIT had been considered unnecessary for the regulatory approval of the agent levamisole (noting that 5-fluorouracil was already available in the marketplace). Although we would have avoided the delay of 5 years required to conduct the confirmatory trial, levamisole alone might have been the regimen used in clinical practice because of cost and toxicity issues. A quote from Artemus Ward reinforces the importance of confirmatory research: "It isn't so much the things we don't know that get us in trouble. It's the things we know that aren't so."

Suppose confirmatory evidence were sought for the maternity ward data by visiting a second hospital. Even if the sex split in that maternity ward were 11 versus 11, one might be tempted to combine the data from both hospitals because that analysis yields 31 versus 13, with a corresponding 2-sided *P* value of 0.0096. However, such an analysis would also be biased because it includes the hypothesis-generating data. Although this recognition seems like common sense in the maternity ward setting, common sense is not that common. Many study sponsors have made this exact error. Often when an original trial did not yield positive results, a favorable result for an exploratory end point or subgroup led to a confirmatory trial targeting that end point or subgroup. Usually, the latter trial revealed disappointing effects, explainable by random high bias in the exploratory results in the original trial. Typically, sponsors then conducted a meta-analysis that included the hypothesis-generating data and proceeded to seek regulatory approval when low *P* values were obtained.

In this type of situation, if the favorable exploratory result in the original trial is a subgroup analysis in women and the confirmatory trial with unfavorable results was conducted only in women, a meta-analysis of women from the 2 trials would be biased. An unbiased estimate of the effect in women could be obtained only by restricting the analysis to the second trial. Any meta-analysis of the women's data from the 2 trials would tend to overestimate the effect of an inactive treatment. This tendency to inflate the false-positive error could be diminished, but not eliminated, by an analysis that included both the men's and women's data from the 2 trials. To more fully address the risk for bias, the prespecified primary analyses of the first trial and, if conducted, the second trial must be reported independent of the positivity of the trial results.

The Psychological Influence of the Driving Goal to Establish Benefit

A review of many clinical trial protocols revealed that a substantial fraction stated the protocol objective to be, "To establish the experimental intervention is safe and effective," rather than providing the scientifically unbiased wording, "To determine whether the experimental intervention is safe and effective." Thus, authors of protocols often are not objective even when writing the protocol objective. This distinction is profound because the goal of establishing an effective treatment leads to giving preferential attention to positive results and to random high bias in estimation of treatment effect. Another indication of the lack of objectivity among investigators whose driving goal is to conclude benefit is their definition of a successful clinical trial often is stated to be "one that achieves a positive result" rather than "one that reliably answers the questions the trial was designed to address."

The psychological influence of the driving goal to establish benefit is sufficiently subtle that it often is not recognized by investigators who have this goal. One sponsor went to great lengths to persuade regulatory authorities that the appearance of lesser benefit in men than women in their clinical data was due to chance, in order to avoid a restriction to women in the indication for the intervention; yet in the sponsor's interest to be persuasive about the positivity of the results, it did not recognize the logical inconsistency when the key summary statement indicated that the agent not only provided overall benefit but had an even more impressive effect in women.

Illustrating Substantial Consequences of Misleading Exploratory Analyses

The substantial consequences from the commitment to obtain positive results are illustrated in the setting of idiopathic pulmonary fibrosis (IPF), a life-threatening disease with no known effective therapy. A placebo-controlled randomized trial of 330 patients was conducted to evaluate Actimmune (InterMune, Brisbane, California) (GIPF-001 [30]). The primary end point was progression-free survival, a composite largely driven by a decrease in forced vital capacity or an increase in alveolar–arterial gradient.

Efficacy results were unfavorable based on prespecified primary and secondary end points. The trial revealed that Actimmune provided a small and nonsignificant 5.5% absolute reduction on progression-free survival (**Table 2**). All 10 prespecified secondary end points were negative, although survival results trended in the positive direction. The final analysis of survival data was to include events occurring through the time of the close-out visits because of an anticipated small number of deaths. With the inclusion of 2 additional deaths that occurred between the prespecified cutoff for the primary analysis of the primary end point (15 June 2002) and the release of data to the sponsor (19 August 2002), 18 patients who died had been taking Actimmune and 28 had been taking placebo, with a corresponding nominal 2-sided *P* value of 0.15. The safety profile was judged to be acceptable, with increases in serious adverse events from pneumonia.

Nine days after receiving study results, the sponsor issued a press release announcing, "Phase III data demonstrating survival benefit of Actimmune in IPF . . . (it) reduces mortality by 70% in patients with mild to moderate disease, ($P = 0.004$) . . . the mortality benefit is very compelling and represents a major breakthrough in this difficult disease" (31). The press release projected a $400 to $500 million per year market for Actimmune, an agent already approved for marketing in

Table 2. **Results of the GIPF-001 Trial Providing a Placebo-Controlled Evaluation of Actimmune in Idiopathic Pulmonary Fibrosis***

Treatment Group	Sample Size, *n*	Progression, *n* (%)	Progression or Death, *n* (%)	Death, *n*†
Actimmune	162	68 (42.0)	75 (46.3)	16 (18)
Placebo	168	75 (44.6)	87 (51.8)	28 (28)
Hazard ratio‡	–	0.942	0.894	–
2-sided *P* value	–	–	0.53	0.084 (0.15)
Safety Profile	**Sample Size, *n***	**Pulmonary SAEs**	**Pneumonia SAEs**	**Vascular Disorders**
Actimmune	162	41	20	7
Placebo	168	34	8	1

SAE = serious adverse event.
* The trial had 90% power to detect a difference of 40% versus 20% in the probability of progression or death at 1 y. Data are shown for events on Actimmune (InterMune, Brisbane, California) versus placebo occurring by 15 June 2002, the prespecified date for the primary analysis of the composite primary end point, progression or death.
† Data are also shown in parentheses when including events that occur by 19 August 2002, the date of the last data monitoring committee meeting when results were released to the sponsor.
‡ Derived by using a Cox regression analysis of time-to-event data.

the orphan indication, chronic granulomatous disease. After this press release, off-label use of Actimmune in IPF and the company's stock value increased sharply.

The sponsor obtained the favorable results reported in the press release by conducting an analysis that was exploratory at 3 levels: The end point, overall survival, was the seventh secondary end point in the statistical analysis plan; the analysis ignored the 2 additional deaths of participants who took Actimmune, which occurred after 15 June 2002; and the patients were from the exploratory subgroup that had mild-to-moderate disease (**Table 3**). This subgroup had not been prespecified in the trial's protocol or detailed statistical analysis plan. The analysis reported in the press release excluded many more known deaths in the Actimmune group than in the placebo group (12 vs. 7 deaths). Of importance, the press release also did not provide an ability to adequately understand the sampling context for the reported analysis.

In 2004, the sponsor initiated a confirmatory placebo-controlled, randomized trial to assess the effect of Actimmune on survival in patients with IPF (32). All 826 patients in this trial had mild-to-moderate disease—more than 3-fold the number of such patients in the original trial. This confirmatory trial was terminated when an interim analysis in early 2007 revealed that the O'Brien–Fleming boundary for lack of benefit (33) was crossed, indicating that these data were statistically inconsistent with meaningful benefit. At this analysis, 80 (14.5%) of 551 patients who took Actimmune had died, compared with 35 (12.7%) of 275 who took placebo. Lack of benefit was established in the very set of patients in which there had been claims that Actimmune provided a very compelling mortality benefit that represented a major breakthrough in this difficult disease.

Summary

The interest in being able to report favorable results is pervasive throughout health care research. Steps can be taken to address the risk for bias induced by this. For results from exploratory analyses, it should be recognized that *P* values will mislead when the actual sampling context is not presented to allow for proper interpretation, and the effect sizes of outcomes that have particularly favorable estimates are probably overestimated because of random high bias, especially when they arise in exploratory analyses conducted in search of favorable evidence. Hence, there should be a clear specification of and focus on the prespecified primary analyses of the primary and secondary end points when submitting results for peer review and when disseminating results in press releases and journal publications. Protocols should have at most 3 or 4 prespecified secondary analyses to further address multiplicity. Journal editors and reviewers should be able to understand the sampling context by being provided access (in this Internet era) to the study protocol; statistical analysis plan; and, if it exists, clinical study report that was prepared for regulatory authorities. In most instances, hypotheses-generating results from exploratory analyses should be assessed in prospectively conducted confirmatory trials. Point estimates and confidence intervals are preferable to reporting *P* values when findings about treatment effects from exploratory analyses are presented. It should be recognized that bias will persist if meta-analyses include the hypothesis-generating trial. All clinical trials should be registered with ClinicalTrials.gov to reduce publication bias (5, 8). Furthermore, the criteria used by journal editors and reviewers in evaluating manuscripts should be based on the importance of the questions that the studies are designed to address and the quality of study conduct rather than on the level of positivity of study results (7).

Table 3. **Survival Data in the GIPF-001 Trial for Actimmune Versus Placebo in Idiopathic Pulmonary Fibrosis***

Treatment Group	Overall Study Group		Mild-to-Moderate Disease Subgroup	
	Participants, *n*	Deaths, *n*	Participants, *n*	Deaths, *n*
Actimmune	162	16	126	6
Placebo	168	28	128	21
Hazard ratio	0.59		0.29	
2-sided *P* value	0.084		0.004	

* Results of the GIPF-001 trial when a clinical cutoff date (15 June 2002) was used. Deaths occurred by the cutoff date and corresponded with the prespecified time for primary analysis of death or progression primary end point. The hazard ratio was derived by using a Cox regression analysis of time-to-event data. The mild-to-moderate disease subgroup is defined as FVC of 55% or more at baseline. Twelve deaths from the Actimmune subgroup (10 in patients with advanced disease and 2 occurring from 15 June 2002 to 19 August 2002) and 7 deaths from the placebo group (all in patients with advanced disease) were excluded. Actimmune is manufactured by InterMune (Brisbane, California).

Exploratory analyses should provide an opportunity for enhanced insight. However, if these exploratory analyses are conducted with an intention to establish that the experimental intervention has a favorable benefit-to-risk profile, rather than to determine whether it does, there is substantial risk for obtaining meaningfully biased conclusions. Indeed, as often stated, if you torture data long enough, they will confess.

From the University of Washington, Seattle, Washington.

Acknowledgment: The author thanks Williamson Bradford and Steven Porter for their supportive role in dissemination of scientifically reliable insights about clinical trials evaluating Actimmune in IPF.

Grant Support: By the National Institutes of Health, National Institute of Allergy and Infectious Diseases (R37 AI 29168).

Potential Conflicts of Interest: Disclosures can be viewed at www.acponline.org/authors/icmje/conflictofinterestforms.do?msNum=M10-1264.

Requests for Single Reprints: Thomas R. Fleming, PhD, Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195-7232; e-mail, tfleming@u.washington.edu.

References

1. **Laine C, Goodman SN, Griswold ME, Sox HC.** Reproducible research: moving toward research the public can really trust. Ann Intern Med. 2007;146:450-3. [PMID: 17339612]
2. **Peng RD, Dominici F, Zeger SL.** Reproducible epidemiologic research. Am J Epidemiol. 2006;163:783-9. [PMID: 16510544]
3. **Guay LA, Musoke P, Fleming T, Bagenda D, Allen M, Nakabiito C, et al.** Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: HIVNET 012 randomised trial. Lancet. 1999;354:795-802. [PMID: 10485720]
4. **Jackson JB, Musoke P, Fleming T, Guay LA, Bagenda D, Allen M, et al.** Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: 18-month follow-up of the HIVNET 012 randomised trial. Lancet. 2003;362:859-68. [PMID: 13678973]
5. **Dickersin K.** The existence of publication bias and risk factors for its occurrence. JAMA. 1990;263:1385-9. [PMID: 2406472]
6. **Sharp DW.** What can and should be done to reduce publication bias? The perspective of an editor. JAMA. 1990;263:1390-1. [PMID: 2304218]
7. **Chalmers TC, Frank CS, Reitman D.** Minimizing the three stages of publication bias. JAMA. 1990;263:1392-5. [PMID: 2406473]
8. **Chalmers I.** Underreporting research is scientific misconduct. JAMA. 1990;263:1405-8. [PMID: 2304220]
9. **Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR.** Publication bias in clinical research. Lancet. 1991;337:867-72. [PMID: 1672966]
10. **Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K.** Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev. 2009:MR000006. [PMID: 19160345]
11. **Laurie JA, Moertel CG, Fleming TR, Wieand HS, Leigh JE, Rubin J, et al.** Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic. J Clin Oncol. 1989;7:1447-56. [PMID: 2778478]
12. **Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Tangen CM, et al.** Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: a final report. Ann Intern Med. 1995;122:321-6. [PMID: 7847642]
13. **Morton V, Torgerson DJ.** Effect of regression to the mean on decision making in health care. BMJ. 2003;326:1083-4. [PMID: 12750214]
14. **Bland JM, Altman DG.** Some examples of regression towards the mean. BMJ. 1994;309:780. [PMID: 7950567]
15. **Bland JM, Altman DG.** Regression towards the mean. BMJ. 1994;308:1499. [PMID: 8019287]
16. **Cummings SR, Palermo L, Browner W, Marcus R, Wallace R, Pearson J, et al.** Monitoring osteoporosis therapy with bone densitometry: misleading changes and regression to the mean. Fracture Intervention Trial Research Group. JAMA. 2000;283:1318-21. [PMID: 10714731]
17. **Barnett AG, van der Pols JC, Dobson AJ.** Regression to the mean: what it is and how to deal with it. Int J Epidemiol. 2005;34:215-20. [PMID: 15333621]
18. **Suissa S.** Lung function decline in COPD trials: bias from regression to the mean [Editorial]. Eur Respir J. 2008;32:829-31. [PMID: 18827150]
19. **Atkinson G, Taylor CE, Jones H.** Inter-individual variability in the improvement of physiological risk factors for disease: gene polymorphisms or simply regression to the mean? [Letter]. J Physiol. 2010;588:1023-4; author reply 1025. [PMID: 20231148]
20. **Rider WD, Palmer JA, Mahoney LJ, Robertson CT.** Preoperative irradiation in operable cancer of the rectum: report of the Toronto trial. Can J Surg. 1977;20:335-8. [PMID: 871980]
21. **Roswit B, Higgins GA, Keehn RJ.** Preoperative irradiation for carcinoma of the rectum and rectosigmoid colon: report of a National Veterans Administration randomized study. Cancer. 1975;35:1597-602. [PMID: 1148993]
22. **Medical Research Council Working Party.** The evaluation of low dose preoperative X-ray therapy in the management of operable rectal cancer; results of a randomly controlled trial. Br J Surg. 1984;71:21-5. [PMID: 6360300]
23. **Fleming TR, Watelet LF.** Approaches to monitoring clinical trials. J Natl Cancer Inst. 1989;81:188-93. [PMID: 2642969]
24. **Alarcón-Segovia D, Tumlin JA, Furie RA, McKay JD, Cardiel MH, Strand V, et al; LJP 394 Investigator Consortium.** LJP 394 for the prevention of renal flare in patients with systemic lupus erythematosus: results from a randomized, double-blind, placebo-controlled study. Arthritis Rheum. 2003;48:442-54. [PMID: 12571854]
25. **Cardiel MH, Tumlin JA, Furie RA, Wallace DJ, Joh T, Linnik MD; LJP 394-90-09 Investigator Consortium.** Abetimus sodium for renal flare in systemic lupus erythematosus: results of a randomized, controlled phase III trial. Arthritis Rheum. 2008;58:2470-80. [PMID: 18668592]
26. Study of LJP 394 in Lupus Patients With History of Renal Disease (ASPEN). Accessed at http://clinicaltrials.gov/ct2/show/NCT00089804 on 10 August 2010.
27. **U.S. Department of Health and Human Services; U.S. Food and Drug Administration.** International Conference on Harmonisation-Efficacy: Guidance for Industry, E9 Statistical Principles for Clinical Trial. 1998. Accessed at www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances /ucm073137.pdf on 11 August 2010.
28. **Donahue RMJ.** Confirmatory trials. In: D'Agostino R, Sullivan L, Massaro J, eds. Wiley Encyclopedia of Clinical Trials. Hoboken, NJ: J Wiley; 2007.
29. **Parmar MK, Ungerleider RS, Simon R.** Assessing whether to perform a confirmatory randomized clinical trial. J Natl Cancer Inst. 1996;88:1645-51. [PMID: 8931608]
30. **Raghu G, Brown KK, Bradford WZ, Starko K, Noble PW, Schwartz DA, et al; Idiopathic Pulmonary Fibrosis Study Group.** A placebo-controlled trial of interferon gamma-1b in patients with idiopathic pulmonary fibrosis. N Engl J Med. 2004;350:125-33. [PMID: 14711911]
31. **InterMune.** InterMune announces phase III data demonstrating survival benefit of Actimmune in IPF [news release]. Brisbane, CA: InterMune; 28 August 2002.
32. **King TE Jr, Albera C, Bradford WZ, Costabel U, Hormel P, Lancaster L, et al; INSPIRE Study Group.** Effect of interferon gamma-1b on survival in patients with idiopathic pulmonary fibrosis (INSPIRE): a multicentre, randomised, placebo-controlled trial. Lancet. 2009;374:222-8. [PMID: 19570573]
33. **Emerson SS, Fleming TR.** Symmetric group sequential test designs. Biometrics. 1989;45:905-23. [PMID: 2675998]
34. **Fleming TR, Sharples K, McCall J, Moore A, Rodgers A, Stewart R.** Maintaining confidentiality of interim data to enhance trial integrity and credibility. Clin Trials. 2008;5:157-67. [PMID: 18375654]