Identifying and Addressing Safety Signals in Clinical Trials: Some Issues and Challenges

Thomas R. Fleming

Abstract Reliable evidence is needed from clinical research about whether the interventions used in clinical practice are safe as well as effective. Regarding risk, safety is not established by failure to establish excess risk, such as obtaining confidence intervals for the relative risk of safety events that include unity. Absence of evidence is not evidence of absence. Rather, safety is established if available data about safety are sufficiently favorable and reliable to rule out the threshold for unacceptable risk, where this threshold should be determined by considering the strength of the evidence for efficacy.

Important insights about safety usually will be provided before marketing through Phase 1, 2, and 3 clinical trials. These insights, especially regarding risks associated with long-term use of the intervention and risks of rare but clinically compelling events, are enhanced by post-marketing active and passive surveillance, and especially by large, long-term randomized trials that provide the most reliable approach for identifying and addressing safety signals. The integrity of these randomized trials is enhanced by preventing irregularities in the quality of trial conduct that would reduce their sensitivity to detecting clinically meaningful safety risks caused by the experimental regimen.

After considering approaches to identifying and addressing safety risks and discussing performance standards to improve the quality of conduct of safety trials, we will consider further the vulnerability to undetected safety risks when evidence for efficacy has been limited to documentation of effects on surrogate endpoints such as biomarkers, and then discuss important considerations regarding cardiovascular safety trials conducted in the setting of type 2 diabetes mellitus.

T.R. Fleming (✉)
Department of Biostatistics, University of Washington, Seattle, WA, USA
e-mail: tfleming@uw.edu

T.R. Fleming and B.S. Weir (eds.), *Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials*, Lecture Notes in Statistics 1205,
DOI 10.1007/978-1-4614-5245-4_9, © Springer Science+Business Media New York 2013

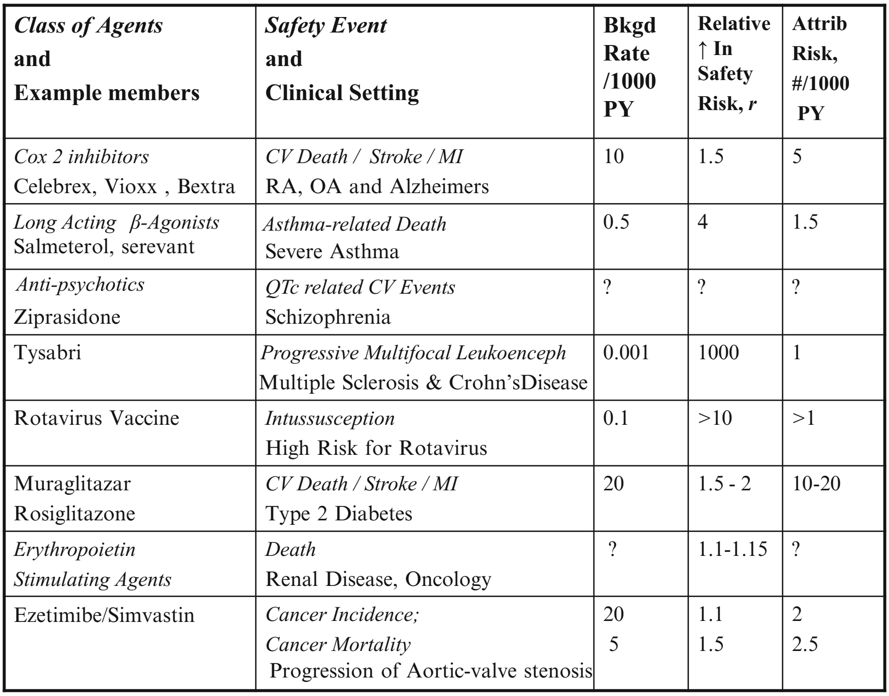
1 Introduction

When interventions are sufficiently potent to provide clinically significant benefits, it is plausible they have unintended effects that could meaningfully alter their benefit-to-risk profile. Hence, it is important to have reliable evidence about whether these interventions are safe as well as effective.

There are many examples where unintended effects would be problematic. Potent immunosuppressive treatments given to rheumatoid arthritis patients to treat symptoms and prevent the progression of disease may provide a clinically important increase in the risk of malignancy or opportunistic infections. Cox-2 inhibitors that provide analgesic relief to osteoarthritis and rheumatoid arthritis patients without inducing risk of gastrointestinal ulceration may have off-target effects such as increasing blood pressure that could induce increased risk of macrovascular complications, including cardiovascular death, myocardial infarction, or stroke. Agents that provide important antipsychotic symptomatic benefits may adversely impact cardiovascular markers, such as increasing weight or low density lipoprotein cholesterol, or decreasing high density lipoprotein cholesterol, where these effects could lead to increased risk of diabetes or major cardiovascular complications.

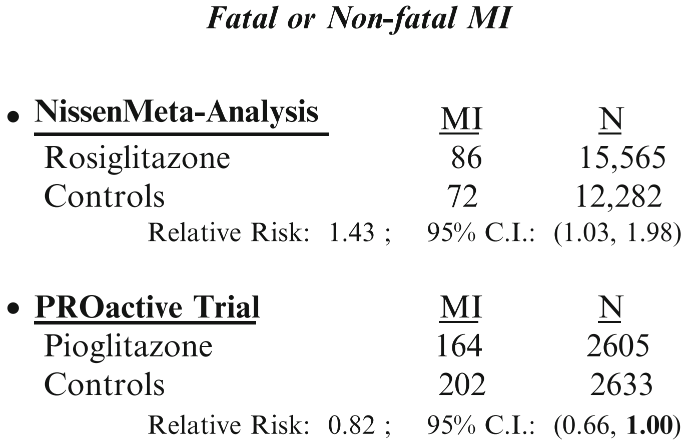
Figure 1 presents several settings where the benefit-to-risk profile of an intervention could be or in fact has been established to be meaningfully altered by harmful effects on measures of irreversible morbidity or mortality. One key setting arises where agents have been established to provide symptomatic benefits, such as analgesic, anti-asthmatic, or antipsychotic treatments. These benefits could be offset by harmful effects on the risk of major cardiovascular events, asthma-related death, torsade de pointes, or sudden death [1–3]. A second key setting arises where agents have been established to have favorable effects on a biomarker thought to be a surrogate for (and hence to provide indirect evidence of favorable effects on) measures of major morbidity or mortality. Examples include erythropoietin-stimulating agents that were accepted for use in renal disease or oncology patients based on normalization of hematocrit, or many therapeutic agents accepted for use in patients with type 2 diabetes mellitus based on reductions in levels of H_bA_{1c}, or selective estrogen receptor modulators considered for long-term use to prevent or treat osteoporosis based on their short-term effects on radiologic fractures or bone mineral density [4–13]. When evidence about clinical efficacy is available only indirectly though evidence about effects on biomarkers, judgment about whether the intervention's benefit-to-risk profile is favorable is strongly influenced when even just a signal for adverse effects on major morbidity/mortality emerges. For example, evidence that some selective estrogen receptor modulators have been associated with increased risk of venous thromboembolism or stroke, and evidence providing a signal that rosiglitazone increases the risk of fatal or non-fatal myocardial infarction in type 2 diabetes mellitus, as shown in Fig. 2, then raised concerns about their continued use [14].

In the remainder of this article, we will provide an overview of observational and clinical trial-based approaches to identifying and addressing safety risks



Class of Agents **and** **Example members**	***Safety Event*** **and** **Clinical Setting**	**Bkgd Rate /1000 PY**	**Relative ↑ In Safety Risk, *r***	**Attrib Risk, #/1000 PY**
Cox 2 inhibitors Celebrex, Vioxx , Bextra	*CV Death / Stroke / MI* RA, OA and Alzheimers	10	1.5	5
Long Acting β-Agonists Salmeterol, serevant	*Asthma-related Death* Severe Asthma	0.5	4	1.5
Anti-psychotics Ziprasidone	*QTc related CV Events* Schizophrenia	?	?	?
Tysabri	*Progressive Multifocal Leukoenceph* Multiple Sclerosis & Crohn'sDisease	0.001	1000	1
Rotavirus Vaccine	*Intussusception* High Risk for Rotavirus	0.1	>10	>1
Muraglitazar Rosiglitazone	*CV Death / Stroke / MI* Type 2 Diabetes	20	1.5 - 2	10-20
Erythropoietin Stimulating Agents	*Death* Renal Disease, Oncology	?	1.1-1.15	?
Ezetimibe/Simvastin	*Cancer Incidence;* *Cancer Mortality* Progression of Aortic-valve stenosis	20 5	1.1 1.5	2 2.5

Fig. 1 Illustrations where safety events meaningfully alter the benefit-to-risk profile of an intervention that has established effects on symptom endpoints or on biomarkers as replacement endpoints for direct measures of patient benefit. *r* is relative risk, *PY* is person years



Fatal or Non-fatal MI

- **NissenMeta-Analysis**

	MI	N
Rosiglitazone	86	15,565
Controls	72	12,282

Relative Risk: 1.43 ; 95% C.I.: (1.03, 1.98)

- **PROactive Trial**

	MI	N
Pioglitazone	164	2605
Controls	202	2633

Relative Risk: 0.82 ; 95% C.I.: (0.66, **1.00**)

Fig. 2 Cardiovascular safety of thiazolidinediones in patients with type 2 diabetes mellitus. *MI* is myocardial infarction, *N* is sample size per arm. With unity as the upper limit of the confidence interval, the PROactive trial rules out pioglitazone has adverse effects on MI

during both pre- and post-marketing phases. We then will discuss in greater detail how randomized clinical trials can be designed to reliably determine whether an intervention provides an unacceptable increase in safety risks, and will discuss performance standards that should be in place to improve the quality of conduct of such trials. We will consider the vulnerability to undetected safety risks that arises when evidence for efficacy has been limited to documentation of effects on surrogate endpoints such as biomarkers, and then discuss important considerations regarding cardiovascular safety trials conducted in the setting of type 2 diabetes mellitus.

2 Approaches to Pre- and Post-Marketing Evaluation of Safety

2.1 Pre-marketing Safety Evaluation

Enhanced insights about safety of interventions are provided by each phase of the clinical development plan. Phase 1 clinical trials often give insights about dose-limiting toxicities in addition to information about metabolism, bioavailability or acceptability of the experimental regimens. Often, in a Phase 1 trial, after a dose escalation stage, approximately 15–25 patients are given what is thought to be the "maximum tolerated dose" of the regimen in order to determine whether we can rule out that the true rate of important toxicities is at least 15–20%. To be specific, suppose 17 (respectively, 23) patients receive the experimental regimen at a targeted dose and schedule, and no events of a specified category are seen. Then with this preliminary evidence about that safety risk, based on the upper 97.5% one-sided confidence interval, one could rule out that the true probability of events of that specified category is 20% (retrospectively, 15%).

Phase 2 trials usually are designed to provide enhanced insights about safety, in addition to "proof of concept" information regarding efficacy. If 35 (respectively, 72) patients receive the intervention at a targeted dose and schedule and no events of a specific category would be seen, then one could rule out (at a 2.5% false positive error rate) that the true rate of that event is at least 10% (respectively, 5%).

While Phase 1 and 2 trials provide useful insights about safety and efficacy, much more reliable evidence usually is needed before a product could be used in non-research clinical settings. Such information typically is provided by prospective randomized Phase 3 trials. However, even when these Phase 3 trials yield sufficient evidence to justify marketing a product, there may be safety signals or inadequate insights about rare but clinically important events, or about safety in the (usually) broader population of people who will eventually receive the drug once its marketed, or there may be a need for an adequate understanding about the long-term safety profile, especially when there could be long-term use of an intervention in a chronic disease setting. These considerations motivate the need for evidence about safety from a Phase 4 post-marketing extended evaluation.

2.2 Post-marketing Safety Evaluation

There are several approaches for "post-marketing" evaluation of safety that may be implemented to pursue existing signals or as surveillance for new safety signals: passive surveillance systems, active surveillance systems, and larger or longer term randomized clinical trials. We will consider the strengths and weaknesses of these approaches.

Passive surveillance for safety risks is provided by an adverse event reporting system based on caregivers' *voluntary* submission of MedWatch forms for serious adverse events they believe might be related to a drug or biologic. Advantages of this approach include obtaining timely information from a reporting procedure that is uniformly implemented by caregivers. However, this information often is difficult to interpret due to (1) lack of a comparator group that usually would be needed to assess whether events are occurring more frequently than expected based on the patient's clinical condition, (2) lack of a "denominator" (i.e., the number of patients who has received the treatment), (3) inaccurate information about the "numerator" (i.e., the number with the safety event) due to underreporting caused by the voluntary nature of data submission, and (4) lack of information on important confounders.

Active surveillance systems, based on use of large prospective cohorts and linked databases, can provide better insights about the numerator and denominator. This system could be prospective, created by a sponsor's pharmaco-vigilance program or by linking automated databases from health maintenance organizations that provide access to data from hundreds of thousands of patients [15, 16]. This creates the possibility of having sensitivity to safety events that occur at a rate of less than 1 per 1,000 person years of exposure. Even with these improved features, active surveillance systems still have important shortcomings, including lack of randomization, inadequate information about important confounders, and inconsistent levels of sensitivity (i.e., whether all targeted events are being captured) and specificity (i.e., whether reported events truly are targeted events). Additional weaknesses of active surveillance systems include lack of a stable population (e.g., due to enroll/disenrollment), exposure misclassification (e.g., databases measure "prescriptions filled" not actual "drug taken"), and often constraints around combining data across multiple HMO sites due to privacy concerns (which can limit analytic possibilities).

The strengths of randomized trials when seeking a reliable assessment of efficacy also apply when evaluating safety. Caregivers and patients usually do not start or stop interventions "at random" in clinical practice. Therefore, when comparing a cohort of patients receiving an intervention with a non-randomized control group, estimates of the effect of treatment on either efficacy or safety measures will be confounded by imbalances between these two groups in important prognostic variables. Statistical procedures used to adjust for these imbalances have limited usefulness, since known and recorded covariates are only the "tip of the iceberg" of the factors that are confounding the estimates of treatment effect.

This confounding may not meaningfully compromise sensitivity to safety risks when the safety assessment is focused on the detection of very large effects on the risk of events, such as when treatment induces a tenfold or larger increase in important safety events. Examples of such large increases in Fig. 1 include the greater than tenfold increase in intussusception induced by the rotavirus vaccine, and the 1,000-fold increase in progressive multifocal leukoencephalopathy induced by natalizumab in patients with Crohn's disease or multiple sclerosis [17, 18]. In spite of their vulnerability to confounding, active and passive surveillance approaches might be sufficient in these settings because these approaches are able to detect such large increases in risk, and because these are settings where it might not be necessary to detect small to moderate increases in risk. To be specific, for the rotovirus vaccine, a doubling of an endpoint of moderate clinical relevance such as intussusception would be offset by the benefits of an effective vaccine and, for the setting of natalizumab, inducing a doubling of progressive multifocal leukoencephalopathy would correspond to an absolute increase of only one case per million person years of treatment, due to the extremely low risk of this event in patients who do not receive natalizumab.

The value of having safety data from randomized trials is high when interventions would have unacceptable safety profiles even if they induce only 1.1- to 2-fold increases in the relative risk of clinically important safety outcomes. This is true in several of the examples in Fig. 1. In these settings, passive and active surveillance could not reliably discern between interventions having no increase in risk and those inducing these small yet unacceptable increases in the risk of clinically important safety outcomes. One approach to reliably address this issue is to conduct large randomized clinical trials designed to evaluate effects on important prespecified safety measures. Through randomization, we can eliminate confounding due to systematically occurring imbalances between treatment and control groups in important prognostic factors. In these trials, to preserve the integrity of randomization, all patients in the treatment and control groups should be uniformly followed until a fixed time post-randomization or until a prespecified calendar time. Stopping follow-up after a patient stops randomized treatment or changes supportive care will lead to risk of substantial bias due to informative missingness.

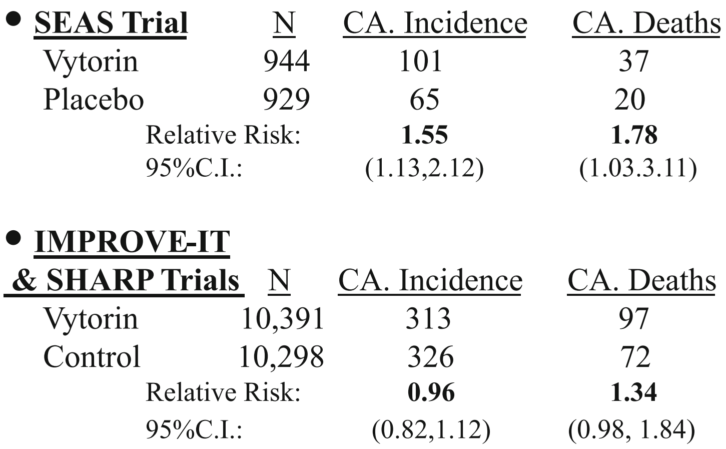
The duration of follow-up in randomized safety trials should be influenced by the anticipated duration of use of the intervention in clinical care and by the likelihood that adverse effects of the treatment would be seen immediately or on a delayed basis. These safety trials often should be designed to follow patients for duration sufficient to provide sensitivity to safety risks that depend on cumulative exposure, such as cardiovascular risks arising with Adriamycin. Uniform long-term follow-up also is necessary to have sensitivity to safety risks that would be seen only after longer follow-up, such as risks for induction of malignancy. In clinical trials that have group sequential boundaries for monitoring efficacy data, such efficacy boundaries should be adequately conservative regarding early termination in settings where sensitivity to long-term safety risks is important [19].

One of the important advantages of having prospective cohorts for assessing safety risks, such as those arising in randomized clinical trials, is that procedures can be put in place to ensure sensitivity and specificity, by ensuring systematic and reliable capture of all targeted safety events, and by ensuring timely adjudication of these events, a process that would be particularly important in open label safety trials.

While the reliability of evidence is greatly enhanced when randomized trials are conducted to address safety risks, whether conducted in pre- or post-marketing settings, such trials typically will require large sample sizes and long timeframes to be completed. Suppose it is intended to have 90% power to rule out a threefold increase in safety risks when the intervention truly provides no increase in risk, using a statistical test having a (one-sided) 2.5% false positive risk of declaring safety when the intervention truly does induce the threefold increase in risk. When ruling out a difference of 1 vs. 3 events per 1,000 person years, as in the setting of evaluating the effect of long acting beta-agonists on asthma-related death or intubation, or evaluating the effect of type 2 diabetes mellitus drugs on pancreatitis, a randomized trial would need 20,000 person years follow-up. When the background rate of targeted safety events is 10 events per 1,000 person years, as in the adult setting when evaluating the effect of Cox-2 inhibitors or ADHD drugs on the risk of the composite endpoint, "cardio-vascular death, stroke or myocardial infarction," then assessing a threefold increase of 10 vs. 30 events per 1,000 person years would require a randomized trial with follow-up of only 2,000 person years. However, it likely would not be adequate to simply rule out 20 excess events per 1,000 person years when these events are major morbidity/mortality outcomes, especially when efficacy relates to symptom benefit or effects that have only been established on a biomarker. When the background rate is 10 events per 1,000 person years, to rule out an increase of 5 events, the randomized trial again would require 20,000 person years of follow-up. In the setting of type 2 diabetes mellitus where the background rate is 20 events per 1,000 person years, ruling out an increase of 5 events would require a randomized trial having 40,000 person years follow-up.

2.3 Interpreting and Addressing Safety Signals

The SEAS trial evaluating the efficacy of ezetimibe/simvastatin on slowing progression of aortic-valve stenosis illustrates the challenges in interpreting and addressing safety signals discovered in exploratory analyses [20]. While the SEAS trial was not designed to provide confirmatory evidence about malignancy, Fig. 3 presents evidence from exploratory analyses in SEAS suggesting ezetimibe/simvastatin provides an estimated 55% increase in cancer incidence, and a 78% increase cancer-related deaths, where 95% confidence intervals for relative risk exclude unity. While this meets a traditional standard for "statistical significance," the exploratory nature of this result should lead to caution about making inferential statements. Furthermore, the increase in risk likely is overestimated due to random high bias [21]. Such bias occurs because there are both true signal and random noise in every estimate of treatment effect, and when many analyses are conducted, the results that



• **SEAS Trial**	N	CA. Incidence	CA. Deaths
Vytorin	944	101	37
Placebo	929	65	20
Relative Risk:		**1.55**	**1.78**
95%C.I.:		(1.13,2.12)	(1.03.3.11)

• **IMPROVE-IT & SHARP Trials**	N	CA. Incidence	CA. Deaths
Vytorin	10,391	313	97
Control	10,298	326	72
Relative Risk:		**0.96**	**1.34**
95%C.I.:		(0.82,1.12)	(0.98, 1.84)

Fig. 3 Cancer risk with ezetimibe/simvastatin (vytorin) in SEAS trial evaluating effect in slowing progression of aortic-valve stenosis, and in the confirmatory IMPROVE-IT and SHARP trials

appear to be most extreme tend to be at least partially due to random overestimates of the true effect**.**

Usually, conducting analyses such as these would be regarded as hypothesis generation rather than hypothesis confirmation. For efficacy or safety signals discovered in exploratory analyses to be viewed to be reliable, several important criteria may need to be simultaneously satisfied. *First*, the likelihood this safety signal could be explained by chance should be low, even when taking into account the sampling context of the exploratory analysis. For example, when three cases of progressive multifocal leukoencephalopathy occurred in several thousand patients receiving natalizumab, this event rate that is 1,000-fold higher rate than would be expected in natural history almost surely could not be attributed to chance, even when taking into account the exploratory context for this discovery. *Second*, there should be a biologically plausible linkage between the safety event and the intervention, based on its known mechanisms of action. For example, ezetimibe blocks the absorption of phytosterols and other phytonutrients linked to protection against cancer, which provides some biologic plausibility that the drug could have an effect on the growth of cancer cells [22–24]. *Third*, there should be independent evidence to confirm the observed association. While additional evidence regarding a safety signal could be provided within the trial generating that signal, such as obtaining data about whether there are increases seen not only in cancer mortality but also in cancer incidence, the ideal independent evidence would be provided by separate prospective trials.

After the discovery of cancer risks in the SEAS trial, the ongoing IMPROVE-IT and SHARP trials were used as an independent source of confirmatory evidence to better understand this finding [25, 26]. Figure 3 gives the data from these two trials and provides evidence that the findings of excess cancer risks from the exploratory analysis of SEAS data were overestimates, consistent with random high bias that results from such exploration of data. The estimates of the ezetimibe/simvastatin to placebo relative risk for cancer incidence from IMPROVE-IT and SHARP were 0.96 for cancer incidence and 1.34 for cancer death, where the 95% confidence intervals included unity in both instances. Peto et al. [27] stated, "*The available results from these 3 trials do not provide credible evidence of any adverse effect of Ezetimibe on rates of cancer*." However, safety is not established by obtaining confidence intervals for relative risk that include unity. That is, failure to establish excess risk is not sufficient to justify safety. Rather, safety is established by ruling out any level of excess risk that would be unacceptable [28]. The IMPROVE-IT and SHARP trials yield an estimated 34% increase in risk of cancer death, where the upper limit of the 95% confidence interval is 1.84. Since the data are consistent with ezetimibe/simvastatin causing as much as an 84% increase in cancer incidence, further study of the effect of this agent on cancer risk is needed.

3 Safety Trials Assessing Whether Unacceptable Excess Risk Can Be Ruled Out

Suppose there is an important signal from either pre- or post-marketing studies that an intervention induces unacceptable safety risks. This may arise when a therapy has been proven to have beneficial effects on direct outcomes of how a patient feels or functions, or on biomarkers that are surrogates for major morbidity/mortality outcomes, and yet a signal exists for adverse effects on measures of major morbidity/mortality. Examples of such measures are fulminant hepatic failure, asthma-related death, cardiovascular death, stroke, myocardial infarction, or risk of malignancy. If only small to moderate relative increases (such as relative risks in the range of 1.1–4.0) in the risk of these major safety events would result in an unfavorable benefit-to-risk profile of the intervention, then a prospective randomized safety trial may be needed. In this section, we will provide insights into the design of such a trial, with an illustration provided from the setting of use of cox-2 inhibitors in patients with osteoarthritis or rheumatoid arthritis.

Cox-2 inhibitors, such as celecoxib, provide analgesic benefit and, relative to NSAIDS, have reduced risk of inducing gastrointestinal ulceration. However, data from at least 50,000 patients who participated in randomized trials conducted across multiple disease settings have suggested that the class of cox-2 inhibitors meaningfully increases the risk of macrovascular complications [1]. This resulted in withdrawal of the drugs rofecoxib and valdecoxib from the market. Celecoxib remained on the market under agreement that a large prospective randomized safety trial would be conducted in a timely manner.

In the PRECISION trial, patients with osteoarthritis or rheumatoid arthritis were randomized to receive either celecoxib, or one of two control regimens, ibuprofen or naproxen [28, 29]. The size of the trial was based on the need to reliably estimate the relative risk for the composite endpoint, “cardiovascular death, stroke or myocardial infarction,” for each pairwise comparison of celecoxib with a control regimen. Even though celecoxib provides important analgesic benefit and, relative to nonselective NSAIDS, an important reduction in the risk of gastrointestinal ulceration, it was judged it would be clinically unacceptable for the agent to induce more than 3 or 4 excess “cardiovascular deaths, stroke or myocardial infarction” events per 1,000 person years. Since the rate of this composite endpoint is approximately 10 events per 1,000 person years on the control arm, the PRECISION trial then was designed to determine whether a 1.333 relative increase in the rate of this composite could be ruled out. For each pairwise comparison of celecoxib with a control regimen, 508 patients would need to experience the composite endpoint for the trial to have 90% power to rule out a 1.333 relative risk when celecoxib truly provides no increase in the risk of the composite endpoint, when using a statistical test having a (one-sided) 2.5% false positive risk of declaring safety when celecoxib truly does induce a 33.3% increase in relative risk. Given the annual event rate of 10 composite endpoints per 1,000 person years follow-up, it follows that approximately 20,000 patients would need to be randomized for each pairwise comparison in the PRECISION trial and be followed for approximately 30 months in order to yield 508 events.

Figure 4 provides the interpretations for some of the possible results from the PRECISION safety trial. Focus is on the relative risk of the “CV death/MI/stroke” composite endpoint for the comparison of celecoxib to the naproxen control arm. The least favorable result that successfully rules out that celecoxib truly induces a 33% increase in the composite endpoint is given in scenario B, where the estimated relative risk is 1.12, corresponding approximately to celecoxib inducing an estimated increase of 1.2 “CV death/MI/stroke” events per 1,000 person years. In contrast, in scenario D, celecoxib has “inferior” cardiovascular safety to naproxen since the data with an estimated 1.9 additional “CV death/MI/stroke” events per 1,000 person years provide statistically significant evidence that rules out a 1.0 relative risk.

Scenario E presents an informative situation of a statistically overpowered trial having more than 1,000 (rather than 508) patients who have primary endpoint events. In such a trial, if the data yield an estimated 15% increase in the relative risk of primary endpoints, then that trial simultaneously would establish inferiority (by ruling out the 1.0 relative risk, corresponding to no increase) while establishing noninferiority of safety risk (by ruling out the 1.33 relative increase). This paradox is explained by recognizing that cardiovascular safety can be worse on celecoxib than on naproxen (i.e., corresponding to ruling out the relative risk of unity) while not being unacceptably worse (i.e., corresponding to ruling out the relative risk is greater than 1.33). Importantly, ruling out the 1.33 margin in scenario B does not allow one to conclude that celecoxib is at least as safe as naproxen with respect to cardiovascular risk; scenario A is the only one in Fig. 4 where such a conclusion holds. Rather, the conclusion justified by the data in scenario B is that the data rule out that the safety profile of celecoxib is unacceptably worse than that of naproxen. To justify that “positive” conclusion in scenario B, it is necessary that any increase

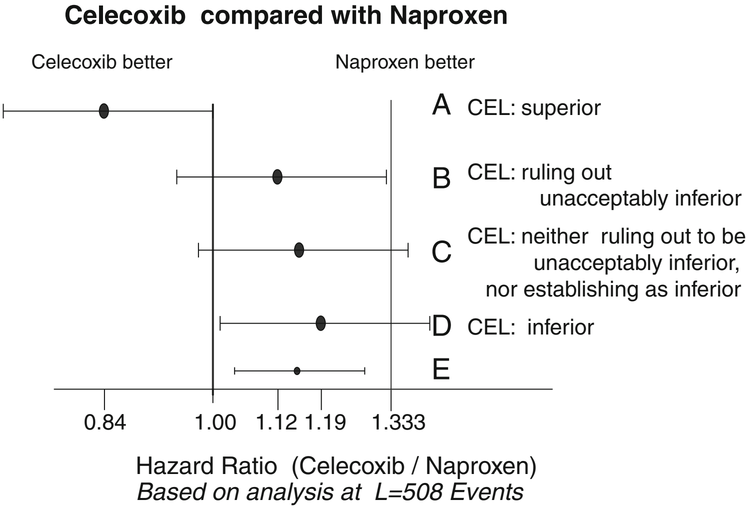


Fig. 4 Possible outcomes in the PRECISION trial for the celecoxib to naproxen hazard ratio for the safety endpoint, "cardio-vascular death/myocardial infarction/stroke." *Black circles* are point estimates, and the *horizontal lines* are the 95% confidence intervals CEL is celecoxib

in cardiovascular safety that is less than a 33% relative increase must be clinically acceptable in the context of the relative safety, acceptability, convenience, and cost of these two regimens.

4 Study Performance Standards for Safety Trials

For safety trials such as PRECISION that are designed to provide interpretable and reliable evidence regarding harm, irregularities in the quality of trial conduct can reduce the sensitivity to detecting clinically meaningful true differences in safety risks between the experimental and control regimens. These irregularities include failure to achieve timely enrollment of the targeted population in which excess risk from the experimental regimen is most plausible, failure to enroll patients who are at sufficiently high risk for the targeted number of primary safety events to be achieved, violations in eligibility criteria, lack of adherence to the experimental and control regimen at a level that matches best achievable in a real-world setting, cross-ins from one regimen to the other, and lack of achieving long-term retention in nearly all patients who undergo randomization [28, 30, 31]. These irregularities could lead to an increased likelihood of falsely ruling out unacceptable thresholds of risk in settings where the experimental regimen truly provides increased risks of clinically important harmful effects. As stated in [32]: "Many flaws on the design or conduct of the trial will tend to bias the results toward a conclusion of equivalence."

To address the concerns about these risks to trial integrity, performance standards regarding quality of trial conduct should be established. These should be specified either in the study protocol or in a separate performance standards document that is developed before the initiation of the clinical trial. For each of these performance standards, there should be specification of both targeted levels and minimally acceptable levels of performance, and specification of creative strategies that are being developed to maximize the ability to achieve these targeted levels. For illustration, consider the performance standard regarding retention. The targeted level of performance might be that no more than 2% of the randomized patients have incomplete capture of the primary endpoint information for reasons related to informative missingness, while the minimally acceptable level of performance might be an upper limit of 10% of patients with this irregularity. The creative approaches to achieve the targeted levels of retention might be similar to those specified in Fleming [30]. Finally, monitoring procedures should be in place during the trial to evaluate whether these performance standards are being met, with a plan for correction actions or even for trial termination if minimally acceptable levels of performance are not met.

5 Vulnerability to Undetected Safety Risks When Relying on Biomarkers as Surrogate Endpoints

In order to reduce the size and duration of Phase 3 clinical trials, it often is proposed to use biomarkers as replacement endpoints for measures that directly assess how a patient feels, functions, or survives [33–37]. For example, in type 2 diabetes mellitus, agents given to improve long-term microvascular or macrovascular disease endpoints might be assessed simply by evaluating their effect on changes in H_bA_{1c} at 6 months post randomization or, in end-stage renal disease, erythropoiesis-stimulating agents (ESAs) given to reduce the risk of myocardial infarction or death may be assessed simply by evaluating their effect on hematocrit [4–9].

Many agents have received regulatory approval for marketing based on small trials having short-term follow-up, where only the effects on a replacement endpoint such as a short-term intermediate endpoint or a biomarker have been assessed. For example, natalizumab received accelerated approval based on its effects on 1-year relapse rate in multiple sclerosis patients [18], while full regulatory approval was given to ESAs in chemotherapy-induced anemia and in hemodialysis in end stage renal disease based on effects on hematocrit, and full regulatory approval has been given to new agents in type 2 diabetes mellitus based on effects on HbA1c at 6 months post-randomization.

Strategies to use replacement endpoints such as biomarkers to reduce the size and duration of clinical trials lead to widespread marketing of interventions not only when efficacy data are limited, but also when safety data are limited. Hence, it should not be surprising in such settings when important off-target effects are discovered only after the product has already been widely used in clinical practice. In post-marketing settings, it was discovered that natalizumab induces progressive multi-focal leukoencephalopathy, ESAs induce thrombosis, and rosiglitazone appears to negatively impact the risk of macrovascular events in type 2 diabetes mellitus [37].

The hazards of reliance on replacement endpoints for regulatory approval are even worse when one recognizes that interventions are assessed based on their benefit-to-risk profile. The stronger the efficacy evidence, the greater the resilience regarding uncertainties about safety. Therefore, when post-marketing safety risks are detected in settings where efficacy has been established only on short-term replacement measures, there have been significant controversies about whether the product should continue to be used in the clinical practice. Natalizumab was removed from the market for some time, and the indications for use of ESAs and rosiglitazone have been substantially reduced. These controversial circumstances could have been avoided had trials been conducted pre- or post-marketing that provided reliable evidence about efficacy and safety. For example, the uncertainty about the clinical utility of natalizumab could have been addressed had long-term trials been conducted to assess its effects on measures of irreversible morbidity for multiple sclerosis patients, such as a delay in time to walking with a cane (i.e., Expanded Disability Status Scale Score = 6) or being wheelchair bound (i.e., Expanded Disability Status Scale Score = 7). Similarly, the uncertainty about the clinical utility of rosiglitazone in type 2 diabetes mellitus could have been avoided had this agent been studied in a trial similar to PROactive, a large-scale long-term clinical trial conducted in a high-risk patient population to reliably evaluate effects of another thiazolidinedione, pioglitazone, on measures of major morbidity and mortality (see Fig. 2) [38].

6 Illustration: Considerations for Cardiovascular Safety Trials in Type 2 Diabetes Mellitus

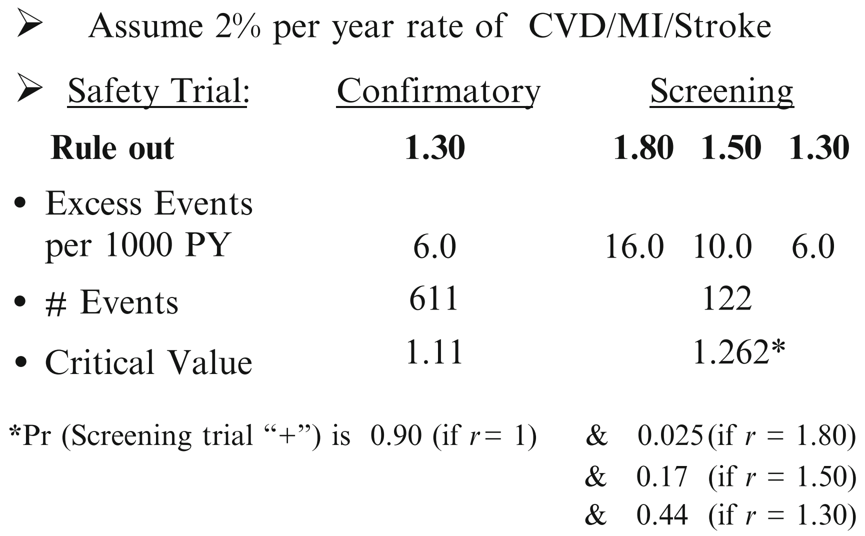
For many years, experimental regimens were approved for use in clinical practice in type 2 diabetes mellitus (T2DM) patients even though efficacy was evaluated only through an assessment of the effect of treatment on the biomarker, "H_bA_{1c} changes at 6 months." As discussed in the previous section, one of the consequences of this approach is the lack of reliable insights about important safety risks on measures of major morbidity and mortality.

A principal goal of effective management of T2DM is the reduction in risk of long-term complications that result in major morbidity and mortality. Clinical trials such as the DCCT [39] and the UKPDS 33 [40] provide evidence that sustained intensive blood-glucose control by using either sulfonylureas or insulin substantially decreases the long-term risk of microvascular complications, such as diabetic retinopathy, nephropathy, and neuropathy. However, sustained blood-glucose control has not been established to reduce the long-term risk of macrovascular disease, such as stroke, myocardial infarction, or death due to cardiovascular disease. Furthermore, while clinical trials of muraglitazar and rosiglitazone had established their positive effects on glucose control (specifically, on the outcome "HbA1c at 6 months"), recent overviews of available clinical trials strongly suggest these agents lead to increased major morbidity and mortality due to adverse effects on macrovascular complications [14, 41].

In late 2008, FDA's Division of Metabolic and Endocrine Drug Products, in response to advice received in July of that year from the FDA Endocrinologic and Metabolic Drugs Advisory Committee, determined that small- and short-term studies evaluating effects on H_bA_{1c} do not provide sufficiently reliable evidence about the benefit-to-risk profile of experimental products in patients with T2DM [42, 43]. The Agency took a significant step forward in its efforts to protect public health by deciding that long-term cardio-vascular (CV) safety trials would be required as part of the evaluation of new interventions in this clinical setting.

In these CV safety trials, patients with T2DM would be randomized to the experimental regimen against a standard of care therapeutic intervention. To ensure sensitivity to CV safety risks induced by the experimental regimen, evidence establishing CV safety should be required for any ancillary agents that would be administered more frequently in the standard of care control relative to the experimental arm of the trial. Examples of ancillary agents with such evidence include pioglitazone, metformin, sulfonylureas, or insulin. The trial would need to be of sufficient size and duration to allow reliable estimation of the relative risk for the composite CV endpoint, "cardiovascular death, stroke or myocardial infarction." Due to enhanced beneficial effects on microvascular complications provided by interventions that enhance glucose control, it was judged that a new regimen with improved effects on HbA1c would be acceptable as long as it induces no more than 5 or 6 excess "cardiovascular death, stroke or myocardial infarction" events per 1,000 person years. Since the rate of this composite endpoint is expected to be approximately 20 events per 1,000 person years on the control arm, the CV safety trial would be designed to determine whether a 1.3 relative increase in the rate of this composite could be ruled out. In the CV safety trial, 611 patients would need to experience the composite endpoint to have 90% power to rule out a 1.3 relative risk when experimental regimen truly provides no increase in the risk of the composite endpoint, when using a statistical test having a (one-sided) 2.5% false positive risk of declaring safety when the experimental truly does induce a 30% increase in relative risk. This CV safety trial would successfully rule out a relative risk of 1.3 if the estimated (experimental to control) relative risk, r, of "CV death, stroke, myocardial infarction" events is less than 1.11, corresponding to an estimate of approximately 2.2 excess major CV events per 1,000 person years. Given the annual event rate of 20 composite endpoints per 1,000 person years follow-up, it follows that approximately 6,000 patients would need to be randomized and followed for an average of approximately 60 months in order to yield 611 events.

Due to the size and duration of the CV safety trial, it was proposed at the July 2008 meeting of the FDA Endocrinologic and Metabolic Drugs Advisory



➢ Assume 2% per year rate of CVD/MI/Stroke

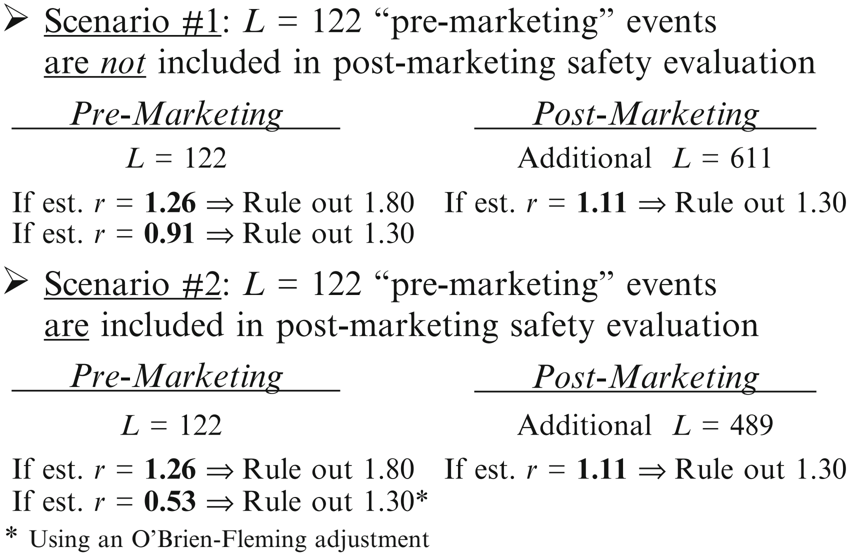
➢ Safety Trial:	Confirmatory	Screening		
Rule out	**1.30**	**1.80**	**1.50**	**1.30**
• Excess Events per 1000 PY	6.0	16.0	10.0	6.0
• # Events	611	122		
• Critical Value	1.11	1.262*		

*Pr (Screening trial "+") is 0.90 (if $r = 1$) & 0.025 (if $r = 1.80$)
& 0.17 (if $r = 1.50$)
& 0.44 (if $r = 1.30$)

Fig. 5 Illustration of sizes and properties of "screening" and "confirmatory" cardiovascular safety trials in type 2 diabetes mellitus patients. *CVD* is cardiovascular death, *MI* is myocardial infarction, *PY* is person years

Committee that the trial could be conducted in a pre-marketing "screening" stage and a post-marketing "confirmatory" stage [42]. If the data in the "screening" stage rule out that the true (experimental to control) relative risk, r, of "CV death, stroke, myocardial infarction" events is at least 1.8, then marketing approval followed by the post-marketing "confirmatory" stage would be considered. With 122 events in the screening stage, this would be achieved if the estimated relative risk is less than 1.262, corresponding to an estimate of approximately 5 excess major CV events per 1,000 person years. This "screening" stage would have 90% power for achieving favorable results that justify marketing approval for regimens with true relative risk, r, equal to unity, while having 97.5%, 83%, and 56% probability for screening out experimental regimens with true relative risk, r, equal to 1.80, 1.50, and 1.30, respectively, (see Fig. 5). To obtain 122 events in the screening stage, if approximately 2,500 patients would be randomized, they would need to be followed for an average of approximately 24 months.

The "confirmatory" stage can be separate from the "screening" stage, as in scenario #1 in Fig. 6 or can include the 122 events from the "screening" stage, as in scenario #2 in Fig. 6. In both scenarios, the 122 event dataset in the "screening" stage is used to test the "null hypothesis" that $r = 1.80$, at a traditional false positive error rate 0.025, where satisfying this CV safety criterion would allow marketing of the product. These data also can be used to test the "null hypothesis" that $r = 1.30$, the defined criterion for establishing CV safety. In scenario #1, the "null hypothesis" that $r = 1.30$ is tested at a traditional 0.025 false positive error rate. If the estimated relative risk is less than 0.91, a post-marketing "confirmatory" stage would not be



- Scenario #1: L = 122 "pre-marketing" events are *not* included in post-marketing safety evaluation

Pre-Marketing	*Post-Marketing*
L = 122	Additional L = 611
If est. r = **1.26** ⇒ Rule out 1.80	If est. r = **1.11** ⇒ Rule out 1.30
If est. r = **0.91** ⇒ Rule out 1.30	

- Scenario #2: L = 122 "pre-marketing" events are included in post-marketing safety evaluation

Pre-Marketing	*Post-Marketing*
L = 122	Additional L = 489
If est. r = **1.26** ⇒ Rule out 1.80	If est. r = **1.11** ⇒ Rule out 1.30
If est. r = **0.53** ⇒ Rule out 1.30*	

* Using an O'Brien-Fleming adjustment

Fig. 6 Issues regarding "interim analyses" of "confirmatory" CV safety trials in type 2 diabetes mellitus; L represents targeted number of events, r represents relative risk

needed; however, if the estimate is greater than 0.91, the "screening" stage data could not be used in the "confirmatory" stage due to multiplicity issues. In contrast, scenario #2 allows pooling of the "screening" and "confirmatory" stage data to test the "null hypothesis" that $r = 1.30$, by making a multiplicity adjustment (such as using an O'Brien-Fleming design [44], as illustrated in Fig. 6) when testing that hypothesis at the time of the analysis of the 122 event "screening" stage data.

7 Conclusions

It is important to have reliable evidence about whether the interventions used in clinical practice are safe as well as effective. Substantial insights about safety usually will be provided before marketing through Phase 1, 2, and 3 clinical trials. These insights, especially regarding risks associated with long-term use of the intervention and risks of rare but clinically compelling events, are enhanced by post-marketing active and passive surveillance, and especially by large, long-term randomized trials that provide the most reliable approach for identifying and addressing safety signals. To enhance the integrity of these randomized trials, it is important to prevent irregularities in the quality of trial conduct that reduce the sensitivity to detecting clinically meaningful true differences in safety risks between the experimental and control regimens.

The discovery and evaluation of the signals for cancer risks from use of ezetimibe/simvastatin to slow progression of aortic-valve stenosis illustrate important opportunities and challenges in the evaluation of safety, (see Fig. 3). Exploratory analyses of the SEAS trial provided a signal for increased risk of cancer incidence and cancer death [20]. However, due to the exploratory nature of the process leading to the discovery of this finding, *p*-values representing the strength of evidence for this safety signal are difficult to interpret, and estimates of the excess risk have random high bias. The IMPROVE-IT and SHARP trials provided important confirmatory evidence about the signal for cancer risk from exetimibe/simvastatin [25, 26]. While the confidence intervals for relative risk of cancer events from the meta-analysis of data from these two trials include unity, the failure to establish excess risk is not sufficient to justify safety. Rather, safety is established by ruling out any level of excess risk that would be unacceptable [28]. Since the data are consistent with ezetimibe/simvastin causing as much as an 84% increase in cancer incidence, further study of the effect of this agent on cancer risk is needed.

The use of the IMPROVE-IT and SHARP trials to address the safety signal from SEAS was appropriate. However, since these 2 trials were ongoing, the SEAS trial data should have been presented to the Data Monitoring Committees from IMPROVE-IT and SHARP, and the interim data from these 2 trials regarding cancer risks should have been released only when their Data Monitoring Committees judged the trials provided reliable answers to the questions they were designed to address. Unfortunately, rather than following this process, it appears the safety data from these two trials were prematurely released to the sponsor and others who were seeking rapid access to data informative about the safety signal from SEAS. Interim data prematurely released from IMPROVE-IT and SHARP regarding cancer risks are difficult to interpret due to the lack of a prespecified sampling context for such post-hoc interim analyses, and the lack of full access to peer-reviewed summaries of data from the two trials to address whether performance standards for safety trials have been met. Furthermore, the release of such interim data compromises the integrity of IMPROVE-IT and SHARP regarding the effect of ezetimibe/simvastin on the primary outcome measures these trials were designed to address.

A primary goal of clinical research is to obtain a timely and reliable assessment of the benefit-to-risk profile of an intervention. Benefit is established by clinical data providing substantial evidence of efficacy. Regarding risk, as discussed with ezetimibe/simvastin, safety is not established by failure to establish excess risk, such as obtaining confidence intervals for the relative risk of safety events that include unity. Absence of evidence is not evidence of absence. Rather, safety is established by determining the threshold for unacceptable risk, where this threshold should be dependent upon the strength of the evidence for efficacy, and then by obtaining safety data that rule out that threshold.

Acknowledgements The source of financial support for research described in this article is an NIH/NIAID grant entitled "Statistical Issues in AIDS Research" (R37 AI 29168). The author acknowledges New England Journal of Medicine as the source for Fig. 4. The author also acknowledges Jennifer Clark Nelson and Daniel L. Gillen for providing important insights.

References

1. Meeting Transcript of the joint Advisory Committee meeting of FDA's Arthritis and Drug Safety and Risk Management Advisory Committees, February 16–18, 2005. http://www.fda.gov/ohrms/dockets/ac/05/transcripts/2005-4090T1.htmhttp://www.fda.gov/ohrms/dockets/ac/05/transcripts/2005-4090T2.htmhttp://www.fda.gov/ohrms/dockets/ac/05/transcripts/2005-4090T3.htm
2. Nelson HS, Weiss ST, Bleecker ER, Yancey MS, Dorinsky PM, SMART Study Group (2006) The Salmeterol Multicenter Asthma Research Trial. Chest 129:15–26
3. Strom BL, Faich GA, Reynolds RF, Eng SM, D'Agostino RB, Ruskin JN, Kane JM (2008) The Ziprasidone Observational Study of Cardiac Outcomes (ZODIAC): design and baseline subject characteristics. J Clin Psychiatry 69(1):114–121
4. Pfeffer MA, Burdmann EA, Chen CY, Cooper ME, de Zeeuw D, Eckardt KU, Feyzi JM, Ivanovich P, Kewalramani R, Levey AS, Lewis EF, McGill JB, McMurray JJ, Parfrey P, Parving HH, Remuzzi G, Singh AK, Solomon SD, Toto R, for the TREAT Investigators (2009) A trial of darbepoetin alfa in type II diabetes and chronic kidney disease. N Engl J Med 361: 2019–2032
5. Singh AK, Szczech L, Tang KL, Barnhart H, Sapp S, Wolfson M, Reddan D, for the CHOIR Investigators (2006) Correction of anemia with epoetin alfa in chronic kidney disease. N Engl J Med 355:2085–2098
6. Drüeke TB, Locatelli F, Clyne N, Eckardt KU, Macdougall IC, Tsakiris D, Burger HU, Scherhag A, for the CREATE Investigators (2006) Normalization of hemoglobin level in patients with chronic kidney disease and anemia. N Engl J Med 355:2071–2084
7. Besarab A, Bolton WK, Browne JK, Egrie JC, Nissenson AR, Okamoto DM, Schwab SJ, Goodkin DA (1998) The effects of normal as compared with low hematocrit values in patients with cardiac disease who are receiving hemodialysis and epoetin. N Engl J Med 339:584–590
8. Center for Drug Evaluation and Research. Approval package: Avandia (rosiglitazone maleate) tablets. Company: SmithKline Beecham Pharmaceuticals. Application no. 21–071. Approval date: 5/25/1999. (Accessed 15 May 2007, at http://www.fda.gov/cder/foi/nda/99/21071_Avandia.htm)
9. The Action to Control Cardiovascular Risk in Diabetes Study Group (2008) Effects of intensive glucose lowering in type 2 diabetes. N Engl J Med 358:2545–2559
10. Ettinger B, Black DM, Mitlak BH et al (1999) Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene: results from a 3-year randomized clinical trial. JAMA 282(7):637–645
11. Wooltorton E (2006) Osteoporosis treatment: raloxifene (Evista) and stroke mortality. CMAJ 175(2):147–148
12. Stefanick ML (2006) Risk–benefit profiles of raloxifene for women. N Engl J Med 355: 190–192
13. Mosca L, Grady D, Barrett-Connor E, Collins P, Wenger N, Abramson BL, Paganini-Hill A, Geiger MJ, Dowsett SA, Amewou-Atisso M, Kornitzer M (2009) Effect of raloxifene on stroke and venous thromboembolism according to subgroups in postmenopausal women at increased risk of coronary heart disease. Stroke 40:147–155
14. Nissen SE, Wolski K (2007) Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med 356:2457–2471
15. Baggs J, Gee J, Lewis E et al (2011) The Vaccine Safety Datalink: a model for monitoring immunization safety. Pediatrics 127:S45–S53

16. Behrman RE, Benner JS, Brown JS et al (2011) Developing the sentinel system — a national resource for evidence development. N Engl J Med 364(6):498–499
17. Murphy TM, Gargiullo PM, Massoudi MS, Nelson DB, Jumaan AO, Okoro CA, Zanardi LR, Setia S, Fair E, LeBaron CW, Schwartz B, Wharton M, Livingood JR, for the Rotavirus Intussusception Investigation Team (2001) Intussusception among infants given an oral rotavirus vaccine. N Engl J Med 344:564–572
18. Assche GV, Van Ranst M, Sciot R, Dubois B, Vermeire S, Noman M, Verbeeck J, Geboes K, Robberecht W, Rutgeerts P (2005) Progressive multifocal leukoencephalopathy after Natalizumab therapy for Crohn's Disease. N Engl J Med 353:362–368
19. Emerson S, Kittelson J, Gillen D (2007) Frequentist evaluation of group sequential clinical trial designs. Stat Med 26(28):5047–5080
20. Rossebo AB, Pedersen TR, Boman K, Brudi P, Chambers JB, Egstrup K, Gerdts E, Gohlke-Bärwolf C, Holme I, Kesäniemi YA, Malbecq W, Nienaber CA, Ray S, Skjærpe T, Wachtell K, Willenheimer R, for the SEAS Investigators (2008) Intensive lipid lowering with Simvastatin and Ezetimibe in aortic stenosis. N Engl J Med 359:1343–1356
21. Fleming TR (2010) Clinical trials: discerning hype from substance. Ann Intern Med 153: 400–406
22. Bradford PG, Awad AB (2007) Phytosterols as anticancer compounds. Mol Nutr Food Res 51:161–170
23. Assmann G, Kannenbert F, Ramey DR, Musliner TA, Gutkin SW, Veltri EP (2008) Effects of ezetimibe, simvastatin, atorvastatin, and ezetimibe-statin therapies on non-cholesterol sterols in patients with primary hypercholesterolemia. Curr Med Res Opin 24:249–259
24. Imanaka H, Koide H, Shimizu S et al (2008) Chemoprevention of tumor metastasis by liposomal β-sitosterol intake. Biol Pharm Bull 31:400–404
25. Cannon CP, Guigliano RP, Blaxing MA et al (2005) Rationale and design of IMPROVE-IT (IMProved Reduction of Outcomes: Vytorin Efficacy International Trial): comparison of ezetimibe/simvastatin versus simvastatin monotherapy on cardiovascular outcomes in patients with acute coronary syndrome. Am Heart J 149:464–473
26. Baigent C, Landry M (2003) Study of heart and renal protection (SHARP). Kidney Int 63(Suppl 84):S207–S210
27. Peto R, Emberson J, Landray M et al (2008) Analyses of cancer data from three ezetimibe trials. N Engl J Med 359(13):1357–1366. doi:10.1056/NEJMsa0806603
28. Fleming TR (2008) Identifying and addressing safety signals in clinical trials. N Engl J Med 359:1400–1402
29. Becker MC, Wang TH, Wisniewski L, Wolski K, Libby P, Lüscher TF, Borer JS, Mascette AM, Husni ME, Solomon DH, Graham DY, Yeomans ND, Krum H, Ruschitzka F, Lincoff AM, Nissen SE, for the PRECISION Investigators (2009) Rationale, design, and governance of Prospective Randomized Evaluation of Celecoxib Integrated Safety versus Ibuprofen Or Naproxen (PRECISION), a cardiovascular end point trial of nonsteroidal antiinflammatory agents in patients with arthritis. Am Heart J 157:606–612
30. Fleming TR (2011) Addressing missing data in clinical trials. Ann Intern Med 154:113–117
31. Fleming TR, Odem-Davis K, Rothmann MD, Shen YL (2011) Some essential considerations in the design and conduct of non-inferiority trials. Clin Trials 8:432–439
32. ICH E-9—International conference on harmonisation: statistical principles for clinical trials, published in the Federal Register of May 9, 1997, (62 FR 25712)
33. Fleming TR, DeMets DL (1996) Surrogate end points in clinical trials: are we being misled? Ann Intern Med 125(7):605–613
34. Fleming TR (2005) Surrogate endpoints and FDA's accelerated approval process. Health Aff 24(1):67–78
35. Temple RJ (1995) A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT (eds) Clinical measurement in drug evaluation. Wiley, New York
36. IOM (2010) Evaluation of biomarkers and surrogate endpoints in chronic disease. Washington DC, National Academies Press. http://www.iom.edu/Reports/2010/Evaluation-of-Biomarkers-and-Surrogate-Endpoints-in-Chronic-Disease.aspx

37. Fleming TR, Powers JH (2012) Biomarkers and surrogate endpoints in clinical trials. Stat Med doi:10.1002/sim.5403, 2012
38. Dormandy JA, Charbonnel B, Eckland EJA et al (2005) Secondary prevention of macrovascular events in patients with type 2 diabetes: a randomized trial of pioglitazone. The PROactive Study (PROspective pioglitAzone Clinical Trial In macroVascular Events). Lancet 366:1279–1289
39. The DCCT Research Group (1993) The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 329:977–986
40. UKPDS Group (1998) Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet 352:837–853
41. Nissen SE, Wolski K, Topol EJ (2005) Effect of muraglitazar on death and major adverse cardiovascular events in patients with type 2 diabetes mellitus. JAMA 294:2581–2586
42. Meeting Transcript of the FDA Endocrinologic and Metabolic Drugs Advisory Committee and Drug Safety and Risk Management Advisory Committee, September 24, 2012/www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/Endocrinologicand MetabolicDrugsAdvisoryCommittee/UCM 222628.pdf and UCM222629.pdf
43. Guidance for industry. diabetes mellitus — evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071627.pdf
44. O'Brien PC, Fleming TR (1979) A multi-stage procedure for clinical trials. Biometrics 35: 549–556