Drug Saf (2015) 38:233–243
DOI 10.1007/s40264-015-0268-x

LEADING ARTICLE

Drug Safety Meta-Analysis: Promises and Pitfalls

Michael A. Stoto

Published online: 17 February 2015

Abstract Meta-analysis has increasingly been used to identify adverse effects of drugs and vaccines, but the results have often been controversial. In one respect, meta-analysis is an especially appropriate tool in these settings. Efficacy studies are often too small to reliably assess risks that become important when a medication is in widespread use, so meta-analysis, which is a statistically efficient way to pool evidence from similar studies, seems like a natural approach. But, as the examples in this paper illustrate, different syntheses can come to qualitatively different conclusions, and the results of any one analysis are usually not as precise as they seem to be. There are three reasons for this: the adverse events of interest are rare, standard meta-analysis methods may not be appropriate for the clinical and methodological heterogeneity that is common in these studies, and adverse effects are not always completely or consistently reported. To address these problems, analysts should explore heterogeneity and use random-effects or more complex statistical methods, and use multiple statistical models to see how dependent the results are to the choice of models.

M. A. Stoto (✉)
Department of Health Systems Administration, School of
Nursing and Health Studies, Georgetown University, St. Mary's
Hall, Room 236, 3700 Reservoir Road NW, Washington,
DC 20057, USA
e-mail: stotom@georgetown.edu

Key Points

Standard meta-analyses methods tend to overstate the precision of the estimates of the risk of adverse effects of drugs and vaccines.

Analysts should explore heterogeneity and use random-effects or more complex statistical methods, and use multiple statistical models to see how dependent the results are to the choice of models.

1 Introduction

In recent years, meta-analysis has increasingly been used to identify adverse effects of drugs and vaccines, but the results have often been controversial. In 2004, for instance, 3 months after Merck announced the withdrawal of rofecoxib because of increased cardiovascular risk in patients taking the drug for more than 18 months, Jüni et al. [1] published the results of a cumulative meta-analysis indicating that the risks could have been apparent—and that rofecoxib should have been withdrawn—several years earlier. Three years later, Nissen and Wolski [2] published a meta-analysis indicating that rosiglitazone was associated with a significant increase in the risk of myocardial infarction. A number of subsequent analyses, however, came to both similar and different conclusions. Another series of meta-analyses by the US Food and Drug Administration (FDA) and other authors [3–5] addressed the risk of suicidal behaviour and ideation in children and adolescents who use antidepressants, but came to different conclusions.

Drug safety concerns are complex, and there are many reasons why authors come to different conclusions. But one thing that these three examples have in common is their reliance on meta-analysis of efficacy studies for drug safety determinations. In one respect, meta-analysis is an especially appropriate tool in these settings. Efficacy studies are generally designed and powered to see whether medications produce a desired therapeutic effect, and this can often be done with hundreds of patients. Study samples that are sufficient for this purpose are too small to reliably assess risks that might occur in one per thousand patients or fewer, although such risks become important when a medication is in widespread use. So meta-analysis, which is a statistically efficient way to pool evidence from similar studies, seems like a natural approach.

However, even when meta-analysis is based on a thorough systematic review, the available studies are usually not simple replicates of one another using the same design or performed in similar populations, thus contradicting a basic statistical assumption of meta-analytic methodology. Published efficacy studies often differ from one another in terms of which serious adverse events are reported, which makes it difficult to pool the results in a meta-analysis. And even when studies are reasonably similar in design and execution, there are few adverse events, which means that standard statistical methods for meta-analysis may be problematic. As a result, standard meta-analytic methods tend to overstate the precision of the estimates of the risk of adverse effects.

The examples also demonstrate how the qualitative results of a drug safety meta-analysis can depend on statistical assumptions that are difficult to empirically verify. In particular, large sample approximations that produce similar results through use of different methods in other settings can come to different conclusions with the sparse data typically available for a safety meta-analysis.

Building on the three examples mentioned above, the objective of this paper is to illustrate both the potential contributions of meta-analysis of drug safety studies and the pitfalls that can occur. Following an in-depth review of the meta-analyses informing these three issues, the paper concludes with a methodological discussion of the problems commonly encountered in this setting and approaches to overcome them. The goal of this analysis is not to resolve the controversies about these three examples, which requires consideration of many issues in each that go beyond meta-analysis per se. Rather, the goal is to inform researchers who undertake safety meta-analyses in the future to improve the quality and usefulness of their analyses. In particular, these examples demonstrate the importance of exploring heterogeneity and using random-effects analysis or more complex statistical methods. The examples also demonstrate the value of using multiple statistical models to see how dependent the results are on the choice of models.

2 Rofecoxib

Rofecoxib is a cyclo-oxygenase (COX)-2 non-steroidal anti-inflammatory drug (NSAID), a class of drugs that provides analgesic (pain-killing) and antipyretic (fever-reducing) effects. Because they produce fewer adverse gastrointestinal effects than the COX-1 variant, COX-2 NSAIDs have dominated the market since their approval by the FDA in 1999. In September 2004, Merck withdrew rofecoxib from the market because of concerns about increased cardiovascular risk associated with long-term, high-dosage use. This followed disclosures that Merck had withheld information about the risks associated with rofecoxib for years. Three months later, Jüni et al. [1] published the results of a cumulative meta-analysis indicating that the risks could have been apparent—and that rofecoxib should have been withdrawn—several years earlier.

Jüni et al.' analysis was based on a search of the Cochrane Controlled Trials Register, as well as the Medline, Embase and Cumulative Index to Nursing and Allied Health (CINAHL) databases, for randomized, controlled trials that compared rofecoxib, administered at 12.5–50 mg daily, with other NSAIDs or placebo in adult patients with chronic musculoskeletal disorders. They identified 18 randomized, controlled trials that met the predefined inclusion criteria. The analysis of the primary endpoint—myocardial infarction—was based on 64 events from 16 comparisons between rofecoxib and control, with 52 events in the rofecoxib groups and 12 in the control groups. On the basis of this analysis, the combined relative risk (RR) was 2.24 [95 % confidence interval (CI) 1.24–4.02], with little evidence of between-trial heterogeneity ($I^2 = 0$ %; P value for heterogeneity = 0.82). Jüni et al.' [1] search also identified 14 additional randomized, controlled trials that did not report cardiovascular endpoints, either because they were not recorded or because no statistical differences were found. One can only wonder how the inclusion of the data from these studies would have affected the result of the meta-analysis. Jüni et al. [1] also conducted a cumulative meta-analysis showing that an increased risk of myocardial infarction became evident in 2000, and they concluded that rofecoxib should have been withdrawn several years earlier than was done by Merck in 2004.

Merck's voluntary withdrawal of rofecoxib was based on a small placebo-controlled trial, which was designed for a different purpose, and this led some to conclude that excess risk is primarily a problem of high-dose, long-duration use. To address this issue, Jüni et al. presented the results of several subgroup analyses. As shown in Table 1, there was little evidence that the RRs differed depending on the dose of rofecoxib or the duration of the trials. The estimated RR for myocardial infarction, however, was greater in trials with an external endpoint adjudication

Table 1 Relative risk (RR) of myocardial infarction comparing rofecoxib with control, from stratified meta-analyses

	RR	95 % CI	*P* value for interaction
All comparisons	2.24	1.24–4.02	
Type of control			0.41
Placebo	1.04	0.34–3.12	
Non-naproxen NSAID	1.55	0.55–4.36	
Naproxen	2.93	1.36–6.33	
Daily dose			0.69
12.5 mg	2.71	0.99–7.44	
25 mg	1.37	0.52–3.61	
50 mg	2.83	1.24–6.43	
Trial duration			0.82
≥6 months	2.17	1.03–4.59	
<6 months	2.33	0.90–6.03	
Concealment of allocation			0.96
Adequate	2.04	0.32–12.93	
Unclear	2.26	1.22–4.19	
External endpoint adjudication committee			0.011
Yes	3.88	1.88–8.02	
No or unclear	0.79	0.29–2.13	

Reproduced from Jüni et al. [1], with permission

CI confidence interval, *NSAID* non-steroidal anti-inflammatory drug

committee than in trials without such a committee (RR = 3.9; $P = 0.011$). This suggests that independent endpoint adjudication committees should be the rule, and exceptions to this rule should be justified [1].

An increased risk of myocardial infarction had been observed in 2000 in the Vioxx Gastrointestinal Outcomes Research (VIGOR) study [6], which was by far the largest study in the meta-analysis and the one primarily responsible for the finding that rofecoxib (trade name Vioxx) caused an excess risk of myocardial infarction. However, this risk had been attributed to a cardioprotective effect of naproxen, which was used as the comparator, rather than a cardiotoxic effect of rofecoxib per se. In their primary analysis (see Table 1), Jüni et al. found that the RR estimates varied depending on whether rofecoxib had been compared with placebo, an NSAID other than naproxen, or naproxen, but the 95 % CIs were wide and a test of interaction was not significant ($P = 0.41$). To address this further, Jüni et al. conducted a meta-analysis of eight case–control studies and three retrospective cohort studies. Using data primarily from large administrative or clinical databases, they found almost identical results regardless of whether naproxen or another NSAID was used as the comparator. In both cases, the combined estimate was 0.86 (95 % CI 0.75–0.99), but there was considerable between-study heterogeneity (I^2 values of 68 and 43 %, respectively). A meta-regression analysis found that most of the heterogeneity was explained by the funding source, with studies funded by Merck indicating larger cardioprotective effects of naproxen. Thus, Jüni et al. [1] concluded that if a protective effect of naproxen exists, it is probably small and not large enough to explain the findings of the VIGOR study.

Kearney et al. took this analysis further by extending the analysis beyond rofecoxib to other COX-2 inhibitors (testing a class effect) and beyond myocardial infarction to other vascular events. The authors searched Medline, Embase, FDA records and data on file from the manufacturers Novartis, Pfizer and Merck. They identified 138 randomized, controlled trials, involving a total of 145,373 participants, that included a comparison between a COX-2 inhibitor and placebo or a traditional NSAID, were of at least 4 weeks' duration and provided information on serious vascular events (defined as myocardial infarction, stroke or vascular death) [7].

Combining data from 121 randomized, placebo-controlled trials, Kearney et al. found an almost twofold proportional increase in myocardial infarction (RR = 1.86, 95 % CI 1.33–2.59; $P = 0.0003$), roughly corresponding to Jüni et al.' RR of 2.24. On the other hand, there was no significant difference in the risk of stroke (RR = 1.02, 95 % CI 0.71–1.47; $P = 0.9$). The observed increased risks of all vascular events (RR = 1.42, 95 % CI 1.13–1.78; $P = 0.003$) and vascular death (RR = 1.49, 95 % CI 0.97–2.29; $P = 0.07$) were marginally significant. Kearney et al. [7] thus concluded that, as a class, COX-2 inhibitors are associated with a moderately increased risk of vascular events, largely attributable to a twofold-increased risk of myocardial infarction.

Of the 121 placebo-controlled trials identified by Kearney et al., nine were long-term trials with 1 year or longer of scheduled treatment (mean 139 weeks) and 112 were shorter trials (mean 11 weeks). Around two thirds of the vascular events had occurred in the nine long-term trials, in which allocation to COX-2 inhibitor treatment was associated with a 45 % increase in vascular events (RR = 1.45, 95 % CI 1.12–1.89; $P = 0.005$). Noting that some researchers have suggested that the hazard emerges only after a year or 18 months, combining short- and long-term trials might underestimate the effects of long-term exposure to a COX-2 inhibitor. However, since the summary rate ratio for the long-term trials was similar to that for the short- and long-term trials combined (RR = 1.42), Kearney et al. [7] concluded that time dependence is not an issue.

On the question of comparators, Kearney et al. found a doubling in the risk of myocardial infarction when COX-2 inhibitors were compared with naproxen (RR = 2.04, 95 % CI 1.41–2.96; $P = 0.0002$) and no effect when COX-2 inhibitors were compared with non-naproxen NSAIDs (RR = 1.20, 95 % CI 0.85–1.68; $P = 0.3$). There was notable heterogeneity in these studies, resolved in large part by the naproxen/non-naproxen NSAID subgroup analysis. Thus, rather than indicating a protective effect of naproxen, this analysis suggested that the risk of myocardial infarction associated with COX-2 inhibitors could not be explained by using naproxen as a comparator. Even when COX-2 inhibitors were compared with naproxen, Kearney et al.' [7] analysis suggested that there was a twofold-increased risk of myocardial infarction associated with COX-2 inhibitors.

2.1 Discussion

Reviewing the two meta-analyses, it becomes clear that there is an excess risk of myocardial infarction associated with rofecoxib (the original question) as well as with other COX-2 inhibitors. But there are reasons to think that the estimate of this risk in Jüni et al.' [1] study (RR = 2.24, 95 % CI 1.24–4.02) might have either over- or underestimated the actual risk. On the overestimate side, one might expect that if COX-2 inhibitors increased the risk of myocardial infarction, they would have a similar effect on stroke and other vascular events, but Kearney et al.' [7] analysis did not support this. Is it possible that the myocardial infarction relationship emerged simply by chance after a search through many possible adverse effects? Another possibility is that the excess risk was due to the use of naproxen as the comparator in VIGOR and other studies, although both Jüni et al.' [1] and Kearney et al.' [7] analyses addressed this hypothesis and did not support it. Similarly, it is possible that a risk that only emerges after months of use of rofecoxib would not be apparent in the short-term studies that made up the bulk of the database, but Kearney et al.' [7] analysis did not support this hypothesis either. On the underestimate side, the biggest factor was Jüni et al.' [1] finding that the RR for myocardial infarction was more than four times higher in studies with external endpoint adjudication committees than in those without such committees (RR = 3.88 versus 0.79). In my estimation, the totality of the data suggests that the RR for myocardial infarction associated with rofecoxib is higher than Jüni et al.' overall estimate of 2.24.

This estimate, however, is based on the 18 randomized, controlled trials that reported cardiovascular endpoints; 14 additional trials did not report them. Since these were primarily short-term studies of analgesic and antipyretic effects, not reporting cardiovascular events is understandable, and it is likely that there were no events, but this does complicate inferences about the long-term risks. This and the lingering uncertainty about the possible over- and underestimation in the previous paragraph suggest that the uncertainty in the RR for myocardial infarction associated with rofecoxib is more substantial than was suggested by Jüni et al.' [1] 95 % CI (1.24–4.02). This is not to suggest that Jüni et al.' believed that this CI fully represented the uncertainty in the estimates, and indeed their subgroup analysis (reproduced in Table 1) and other aspects of their analysis reflected their recognition of the potential for heterogeneity. Rather, the point is that, especially in drug safety studies, the results of a simple summary meta-analysis often understate the uncertainty in the estimates. Because issues other than the risk of adverse effects must be taken into account, Jüni et al.' contention that "rofecoxib should have been withdrawn several years earlier" than it was [1] is an oversimplification of a complex situation.

3 Rosiglitazone

Rosiglitazone is an oral antidiabetic drug in the thiazolidinedione class. First approved by the FDA in 1999, rosiglitazone was marketed by GlaxoSmithKline (GSK) as a stand-alone drug or for use in combination with metformin or with glimepiride. By 2006, the annual sales reached approximately $2.5 billion. In June 2007, Nissen and Wolski [2] published a meta-analysis in the *New England Journal of Medicine*, which concluded that rosiglitazone was associated with a significant increase in the risk of myocardial infarction [odds ratio (OR) 1.43, 95 % CI 1.03–1.98; $P = 0.03$] and an increase in the risk of death from cardiovascular causes, which had borderline significance (OR 1.64, 95 % CI 0.98–2.74; $P = 0.06$). These conclusions were based on a meta-analysis using the Peto method, chosen because a test for heterogeneity had a P value of greater than 0.10.

Soon after these results were published, other analyses questioned a number of the assumptions and details of this analysis. One issue was the choice of the Peto method for the meta-analytic calculations. Nissen and Wolski [2] cited three papers giving some support for this choice for rare events [8–10], but calculations using the Mantel–Haenszel method, a common alternative, yield a substantially different result. For myocardial infarction, the OR is 1.28 (versus 1.43 with the Peto method), the 95 % CI is 0.96–1.72 (versus 1.03–1.98) and the P value is 0.11 (versus 0.03). As a result, a statistically significant increase in risk becomes non-significant with the Mantel–Haenszel method, at least at the standard 5 % level. For cardiovascular death, the P value is 0.23 (versus 0.06) [2].

The Peto and Mantel–Haenszel methods are both 'fixed-effects' meta-analytic methods, which are appropriate when the studies to be combined are essentially all replicates of one another and are similar in design, methods and the populations studied. Nissen and Wolski [2] justified this choice because the standard Cochran's Q test for heterogeneity was not significant. But this test is known to be underpowered [11].

Furthermore, as Diamond et al. [12] noted, a review of the studies included in the Nissen and Wolski analysis suggests that the studies are substantively quite heterogeneous. For instance, the populations studied varied markedly; although most studies were in patients with diabetes, one of the 42 trials in the analysis targeted Alzheimer patients and another two studies targeted psoriasis patients, for whom the potential risk of adverse cardiovascular events is likely different. One study included patients with congestive heart failure, for whom rosiglitazone is contraindicated. If this single study were excluded from the analysis, the estimated OR would drop from 1.43 to 1.39 and the 95 % CI (0.99–1.94) would no longer exclude 1.00. The comparison groups also varied markedly among the studies. In ten studies, rosiglitazone was compared with a placebo; in 28 studies, it was assessed as add-on therapy to a sulfonylurea, metformin, or insulin; and four studies compared rosiglitazone with a sulfonylurea or metformin. The studies included phase 2, 3 and 4 designs, and the lengths of follow-up ranged from 12 to 208 weeks [12]. Any statistical analysis that took this heterogeneity into account would likely have a larger CI than Nissen and Wolski found, and the results would not probably no longer be statistically significant in the conventional sense.

Another issue was the treatment of studies that had zero events in the treatment group or the control group. The Peto method was chosen because it could accommodate studies with no events in either group, but Nissen and Wolski [2] excluded six of 48 otherwise eligible studies that had no events in either group. To address this deficiency, Diamond et al. [12] explored the use of two kinds of continuity correction for zero-event cells. The constant correction (CC) simply adds 0.5 to all cells of the 2 × 2 contingency table of the study selected for correction. The other is treatment arm correction (TAC), which adds values proportional to the reciprocal of the size of the opposite treatment group. Diamond et al. then compared the results of using these approaches within standard inverse variance and Mantel–Haenszel meta-analyses with the results of the Peto method calculation. The results, presented in Table 2, suggested that only the Peto method would yield a CI for the OR for myocardial infarction that would exclude the null value of 1.0. Diamond et al. [12] also noted that the random-effects version of the inverse variance method (also known as the DerSimonian–Laird method) does not differ from the fixed-effects version, which is consistent with Cochran's Q test for heterogeneity not being significant.

Table 2 Meta-analytic odds ratios (ORs) for myocardial infarction and cardiovascular death

Meta-analytic method	Myocardial infarction			Cardiovascular death		
	k	OR	95 % CI	*k*	OR	95 % CI
Fixed, Peto	38	1.43	1.03–1.98	23	1.64	0.98–2.74
Fixed, IV (TAC)[a]	38	1.34	0.97–1.84	23	1.46	0.88–2.42
Fixed, IV (CC)[a]	38	1.29	0.94–1.76	23	1.31	0.80–2.13
Fixed, MH (TAC)	38	1.36	1.00–1.84	23	1.51	0.94–2.44
Fixed, MH (CC)	38	1.28	0.95–1.72	23	1.33	0.83–2.13
Fixed, MH (TAC+)	42	1.35	1.00–1.82	42	1.39	0.91–2.13
Fixed, MH (CC+)	42	1.26	0.93–1.69	42	1.17	0.77–1.77

Reproduced from Diamond et al. [12], with permission

CC constant correction for continuity, *CC+* constant correction for continuity that includes all zero-total-event studies, *CI* confidence interval, *IV* inverse variance, *k* number of studies, *MH* Mantel–Haenszel, *TAC* treatment arm correction for continuity, *TAC+* treatment arm correction for continuity that includes all zero-total-event studies

[a] The results from analogous random-effects analyses are identical

In another approach to the problem of zero-event studies, Tian et al. [13] developed a non-parametric procedure to make valid inferences about the parameter of interest with all available data without artificial continuity corrections. Using the non-parametric method and the risk difference (RD) as the effect size, Tian et al. were able to include 48 rather than 23 studies and estimated the RD to be 0.06 % (95 % CI −0.13 to 0.23; $P = 0.83$). This compares with 0.16 % (95 % CI 0.00–0.31; $P = 0.05$) for the standard Mantel–Haenszel analysis. For the myocardial infarction endpoint, Tian et al. [13] estimated the RD to be 0.18 % (95 % CI −0.08 to 0.38; $P = 0.27$), based on 48 studies. This compares with 0.22 % (95 % CI 0.02–0.42; $P = 0.03$) for the standard Mantel–Haenszel analysis based on 38 studies.

Localio et al. [14] explored a broader range of approaches, using both ORs and RDs for myocardial infarction. The first section of Table 3 shows that the P values testing the hypothesis that OR = 1.0 ranged from 0.025 to 0.141, depending on the statistical method used. The comparable results for the RD, shown in the second section of the table, ranged from <0.001 to 0.244 [14]. In other words, whether the excess risk was statistically significant depended heavily on which method was used to perform the meta-analysis.

Three years after their original publication, Nissen and Wolski [16] published a follow-up analysis using similar methods to those used in their original study but also using alternative analyses to enable inclusion of trials with no cardiovascular events. In addition, more studies were available for analysis, bringing the total number to 56. One approach was to group smaller studies with others that had

Table 3 Estimates of the effect of rosiglitazone on the risk of myocardial infarction

Meta-analytic method	OR	95 % CI	*P* value
ORs: effect of Avandia on risk of myocardial infarction [$n = 42$ trials]			
Peto	1.43	1.03–1.98	0.030
MH fixed (0.5)	1.30	0.96–1.75	0.090
MH fixed (0.0)	1.45	1.05–2.01	0.025
Random [D&L] (0.5)	1.31	0.95–1.79	0.095
Random [D&L] (0.0)	1.31	0.91–1.89	0.141
Conditional logistic	1.45	1.05–2.01	0.025
Exact stratified	1.45	1.03–2.04	0.030
MH fixed[a]	1.45	1.05–2.01	0.026
Random [D&L][a]	1.33	0.93–1.91	0.123
Random intercept/slope	1.37	0.99–1.90	0.059
Bayesian[a,b]	1.46	1.02–2.21	
RDs: effect of Avandia on risk of myocardial infarction [$n = 42$ trials]			
MH fixed[c]	21.7	2.5–40.9	0.026
MH fixed	17.0	−4.0 to 28.0	0.114
MH fixed[a]	22.0	4.0–41.0	0.027
MH random	18.0	7.3–28.7	<0.001
MH random	10.0	−7.0 to 26.0	0.243
MH random[a]	3.0	−2.0 to 7.0	0.244
Exact	–	−3.9 to 47.6	0.215
Permutation of conditional linear regression (1,000 iterations)	21.7	7.3 to 43.8+	0.023

Reproduced by courtesy of Localio et al. [14]

CI confidence interval, *D&L* DerSimonian–Laird method, *MH* Mantel–Haenszel, *OR* odds ratio, *RD* risk difference

[a] Required continuity correction using Sweeting's opposite treatment arm correction with total = 0.01, to all 42 studies

[b] Warn et al. [15]

[c] Required continuity correction using Sweeting's opposite treatment arm correction with total = 0.0 or 0.01

the same randomization ratios and pool all of the cardiovascular outcomes as if they were in a single study; larger trials were considered individually. The results of this alternative analysis were very similar to the primary results obtained using the Peto method. For myocardial infarction, the OR was 1.28 (95 % CI 1.01–1.62), compared with 1.28 (95 % CI 1.02–1.63) obtained using the Peto method. Nissen and Wolski [16] also re-did the analysis excluding the RECORD trial, about which concerns had been raised, and this had only a minor effect on the results for myocardial infarction (OR = 1.38; 95 % CI 1.01–1.87).

3.1 Discussion

Taking into account all of the evidence that is now available, including much not discussed in this paper, it seems clear that the FDA's ultimate decision to allow rosiglitazone to remain on the market only under a restricted access programme [17] was appropriate. But was the finding of excess risk in the original Nissen and Wolski [2] paper really as unambiguous as it seemed? There are a number of reasons why it might not have been.

First, because the studies were powered to evaluate the efficacy of rosiglitazone—not its safety—many of the studies reported zero events in one or both arms. Nissen and Wolski [2] argued that the Peto method was best in this situation, and the Cochrane Handbook makes the point that the Peto method has greater statistical power, which, in a safety analysis, might be more important than taking heterogeneity into account [18]. But other analysts reasonably chose different approaches to deal with this problem and came to markedly different conclusions about whether the risk of myocardial infarction was elevated, according to the conventional significance level. As noted above, the *P* values ranged from <0.001 to 0.244.

Second, although the standard Cochran's *Q* test for detecting heterogeneity was not significant, it is known to be underpowered, and a review of the studies included in the meta-analysis suggests that there was a good deal of substantive heterogeneity regarding the populations studied, comparison groups and length of follow-up. As a result, the Peto and other fixed-effects meta-analytical methods likely underestimate the uncertainty of the estimates and thus make the results seem more significant.

On the other hand, although this was not much discussed in the existing reviews, most of the small studies included in the meta-analysis did not have an external endpoint adjudication committee [11]. If these committees worked as they did for rofecoxib, perhaps the adjudicated studies would have found a larger, and more significant, OR.

For a variety of reasons, therefore, the statistical results—and whether they meet conventional levels of significance—vary substantially from one analysis to another. This is because they each make assumptions that cannot be directly verified. Leaving out studies with zero events, for instance, is equivalent to assuming that they are similar to the studies included in the analysis. Choosing a fixed-effects model for the meta-analysis is equivalent to assuming that the only variation among studies is that which results from randomization. Beyond this, there are also many technical statistical assumptions that differ among the models used.

Thus, while the excess risk of myocardial infarctions associated with rosiglitazone is clear, the *P* value in the original Nissen and Wolski [2] paper ($P = 0.03$) seems like an overstatement of the strength of the evidence. Perhaps conventional significance levels should not be the guiding factor when there are important safety concerns, but clarity about the strength of the evidence is also important.

4 Antidepressants

Starting with a June 2003 report to the FDA, it has been suspected that children and adolescents, particularly those with major depressive disorder (MDD), are at risk of suicide-related adverse events (SREs) when treated with antidepressant medication. One of the FDA's responses to this concern has been a meta-analysis on suicidality in paediatric patients treated with antidepressant drugs, conducted by Hammad et al. [3]. The authors identified 23 randomized, placebo-controlled trials conducted in nine drug company-supported programmes evaluating the effectiveness of antidepressants in paediatric patients, and one multicentre trial funded by the National Institute of Mental Health.

Going beyond existing reports of these studies, the FDA authors asked the manufacturers to search their records to identify adverse events that might potentially represent suicidal ideation or behaviour, possible SREs. Because the adverse events captured using this approach varied substantially in the level of detail provided and in their nature, the FDA arranged to have all potential SRE narratives independently and blindly classified into relevant categories by a group of ten paediatric suicidology experts to provide as much assurance as possible that the SREs had been appropriately classified. SREs classified as suicide attempts, preparatory actions towards imminent suicidal behaviour and suicidal ideation were the focus of the meta-analysis. Hammad et al. used the Mantel–Haenszel method to calculate the RR and RD, with a continuity correction (adding 0.5 to all cells in studies with no event in one or both arms). Because Cochran's *Q* test was not significant, the authors used a random-effects model only as a sensitivity analysis [3].

Hammad et al. found that the use of antidepressant drugs in paediatric patients was associated with a modestly increased risk of suicidality (there were no completed suicides in any of the studies). For selective serotonin reuptake inhibitors (SSRIs) in depression trials, the RR was 1.66 (95 % CI 1.02–2.68), and for all drugs, the RR was even higher at 1.95 (95 % CI 1.28–2.98) [3].

Bridge et al. [4] followed up on the FDA's analysis with a meta-analysis, which both sought to compare risk with efficacy and separated out the risk by indication [MDD versus obsessive–compulsive disorder (OCD) versus anxiety] and age group (<12 versus ≥12 years). A further search identified 27 studies, four more than in the analysis by Hammad et al. [3]. To estimate efficacy, Bridge et al. calculated the pooled response rate (the RD) and the number needed to treat (NNT), using Hedges' g statistic for continuous measures. For the risk of suicidality, they used the RD and the number needed to harm (NNH). Random-effects models were used throughout, and subgroup analyses were performed by indication and age [4].

The results of this analysis suggested that, relative to placebo, antidepressants are efficacious for paediatric MDD, OCD and non-OCD anxiety disorders, although the effects are strongest in non-OCD anxiety disorders (RD = 37 %, 95 % CI 23–52; $P < 0.001$), intermediate in OCD (RD = 20 %, 95 % CI 13–27; $P < 0.001$) and more modest in MDD (RD = 11 %, 95 % CI 7–15; $P < 0.001$). In terms of suicidality, Bridge et al. estimated the overall RD as being 0.7 % (95 % CI 0.1–1.3). The estimated RD varied by indication, with RD values of 0.9 % (95 % CI −0.1 to 1.9) for MDD, 0.5 % (95 % CI −1.2 to 2.2) for OCD and 0.7 % (95 % CI −0.4 to 1.8) for anxiety. While the overall RD was significantly greater than zero, none of the three subgroups individually achieved significance. Bridge et al. concluded that the benefits of antidepressants appear to be much greater than the risks of suicidal ideation/suicide attempt across indications, although comparison of benefit and risk varied as a function of the indication, age, chronicity and study conditions. With regard to the three different indications, MDD has the lowest efficacy (RD = 11 %) and the highest risk of suicidality (RD = 0.9 %) [4].

Kaizar et al. [5] noted that the FDA's meta-analytic assumption that different drug formulations and psychiatric diagnoses were equivalent in their effect on suicidality could have underestimated the variance of the risk estimate. To address this, they developed a Bayesian hierarchical meta-analytic model that allows for an additional level of variability beyond a typical random-effects model. Their model distinguished between variability among studies using different formulations of the same drug and variability among different drugs in the same class. The model also allowed a similar analysis of drug class, study length and diagnosis, but the data were not sufficient to study each of these factors simultaneously. This model facilitated extensive sensitivity analyses and also allowed for the inclusion of studies with zero-event cells.

Kaizar et al.' analysis focused on four subsets of the available data:

- *Subset A:* This included studies that examined the risk of SSRIs in patients with MDD. They were of particular interest, since the antidepressant drugs were originally approved for this indication in adults and the authors expected the patients in these studies to be at higher risk of suicide or suicidal behaviour/ideation.
- *Subset B:* This included studies that examined the effects of antidepressants (both SSRIs and atypical antidepressants) on MDD, thus estimating the risk in MDD patients.
- *Subset C:* This included studies that examined the effects of antidepressants (both SSRIs and atypical antidepressants) on OCD, anxiety or ADHD.
- *Subset D:* This included all studies that examined the effects of antidepressants (both SSRIs and atypical antidepressants) on any diagnosis (MDD, OCD, anxiety or ADHD), thus estimating the overall risk [5].

Kaizar et al. began by replicating the FDA's fixed- and random-effects analyses with their Bayesian hierarchical model by setting certain variance parameters equal to 0. For subsets A, B and D, Kaizar et al.' results were similar to those of the FDA analysis [3], indicating strong evidence of an association between antidepressant use and suicidality. The subset of the data that did not include patients with an MDD diagnosis (subset C) did not similarly support a link between antidepressant use and suicidality. Their random-effects analytic results were similar, and standard goodness-of-fit statistics could not distinguish between the fixed- and random-effects models [5].

Application of the full Bayesian model, however, yielded substantially different results. Focusing first on a model incorporating variance among different drug formulations, Kaizar et al. found a significantly elevated risk of suicidality in subset B (all MDD studies: OR = 2.123, 95 % CI 1.11–3.65) and subset D (all studies: OR = 2.18, 95 % CI 1.17–3.60). For the SSRI and non-MDD study subsets (A and C), the 95 % credible intervals (Bayesian equivalents of CIs) for the drug effect did contain 1.00, so they were not significant in the conventional sense. In a similar analysis using drug class as the third-level variable, Kaizar et al. found support for a link between SSRI-class antidepressants and suicidality (OR = 2.21, 95 % CI 1.30–3.36) but not for other drug classes or study lengths. Taken together, these results suggested that the positive results in the FDA analysis could be attributed to studies where the diagnosis was MDD and where the antidepressant was an SSRI, and not other indications or drug types [5].

Figure 1 illustrates the sensitivity analysis that Kaizar et al.' Bayesian hierarchical model makes possible. The central symbols indicate the posterior means, and the horizontal lines span the 95 % credible intervals. Each panel shows the impact of six different sets of assumptions that decision makers might make about the risk of suicidality, apart from the studies in the meta-analysis. They range from a largely uninformative prior (prior mean OR = 0 and variance for the log OR = 9.0) to a very specific one consistent with a high risk (prior mean OR = 7.4 and variance for the log OR = 0.49). The 'Overall Effect' panel shows that the estimate for the OR and CI depend on the prior beliefs that the analyst has about the risk of suicidality; in other words, the results are highly dependent on assumptions that cannot be directly assessed. Similar results were found for study length and drug class. The 'MDD' and 'Other Diagnoses' panels, in contradistinction, show that when the studies are divided into those where MDD and other conditions were the diagnosis, the results are quite stable, clearly indicating a risk of suicidality in patients with MDD and not with other diagnoses [5], as one might expect from behavioural science theory.

On the basis of the totality of these analyses, Kaizar et al. concluded that there is an association between antidepressant use and an increased risk of suicidality in studies where the diagnosis was MDD (OR 2.3, 95 % CI 1.3–3.8) or where the antidepressant was an SSRI (OR 2.2, 95 % CI 1.3–3.6), but not in other studies. These results were insensitive to model perturbations, but the robustness of the FDA's meta-analysis to model assumptions was less clear. Thus, they concluded that because of model specification issues, the evidence supporting a causal link between antidepressant use and suicidality in children is weak [5].

4.1 Discussion

Three meta-analyses analysed essentially the same set of studies but came to markedly different conclusions. The FDA analysis by Hammad et al. found a significant risk of suicidality associated with SSRIs in the treatment of MDD (RR = 1.66, 95 % CI 1.02–2.68). For all drugs and all indications, they estimated a larger, more significant risk (RR = 1.95, 95 % CI 1.28–2.98) [3]. Using a more complex statistical model, Kaizar et al. [5] concluded that there was an association between antidepressant use and an increased risk of suicidality in studies where the diagnosis was MDD (OR 2.3, 95 % CI 1.3–3.8) or where the antidepressant was an SSRI (OR 2.2, 95 % CI 1.3–3.6), but not in other studies. Note that the results reported by Kaizar et al. were for treatment of MDD regardless of drug type or for use of an SSRI regardless of indication, and both estimates were substantially larger than Hammad et al.' estimated risk of SSRIs used in the treatment of MDD. If

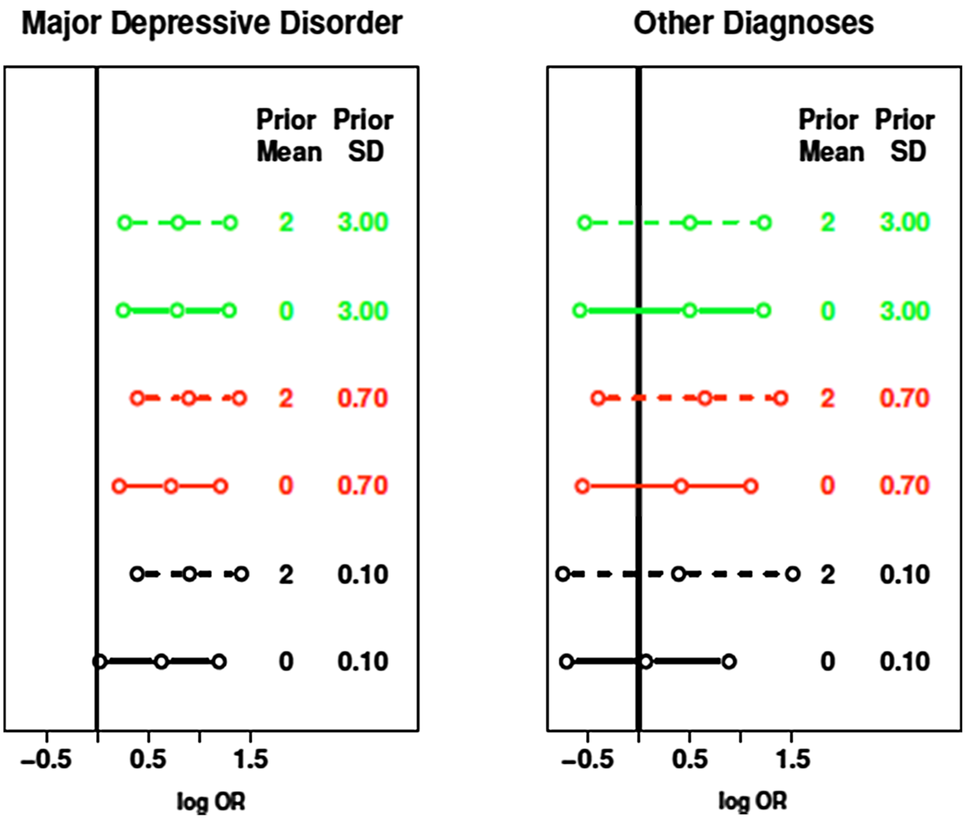

Fig. 1 Sensitivity analysis for the prior mean and variance on the overall effect for major depressive disorders (MDD) and other diagnoses using the full Bayesian model incorporating variance among different psychiatric diagnoses. The *central symbols* indicate the posterior means, and the *horizontal lines* span the 95 % credible intervals. Adapted and reproduced from Kaizar et al. [5], with permission

enough studies were available, the approach used by Kaizar et al. could be used to simultaneously look at indication and drug type, but unfortunately that was not possible in this analysis.

Bridge et al. addressed benefits as well as risks and concluded that the benefits of antidepressants appear to be much greater than the risks of suicidality across indications, although comparison of benefit and risk varies as a function of indication, age, chronicity and study conditions. Of the three indications, treatment for MDD has the lowest efficacy (RD = 11 %) and the highest risk of suicidality (RD = 0.9) [4]. Although Bridge et al. found a significantly elevated risk of suicidality for all indications taken together, neither MDD nor either of the other diagnoses alone has a significantly increased risk [3–5].

Taken as a group, therefore, these three studies established that suicidality is a matter of concern with children's and adolescents' use of antidepressants. They suggested that the problem may be associated with children diagnosed with MDD and/or treated with SSRIs, but could not be more specific. As in the rosiglitazone analysis, the answer depends on which statistical model is used. And, since their underlying assumptions varied but could not be empirically verified, the meta-analyses were unable to resolve this important policy issue.

5 Conclusions

The three examples discussed in this review illustrate both the promise, and the potential pitfalls, of using meta-analysis to assess drug safety. Randomized, controlled trials are usually designed to test efficacy and hence often do not have sample sizes that are sufficient to assess the risk of rare adverse effects. So, in principle, meta-analysis provides a way to pool the results of multiple studies to obtain more precise estimates of the risk. But, as these examples show, different syntheses of the same studies can come to qualitatively different conclusions, and the results of any one analysis are usually not as precise as they seem to be. There are three major reasons for these problems.

First, the adverse events of interest are rare. As a consequence, the results of different statistical models for meta-analysis that would be similar with large samples vary markedly. This includes differences in meta-analytic approaches' ability to handle studies with zero events, as is especially apparent in the rosiglitazone example. The antidepressant example illustrates a different aspect of this problem; the inability to disentangle the effects of the indication (MDD versus other diagnoses) and drug type (SSRIs versus others) may be due in part to the small number of suicide-related events.

Second, because the available studies may have been designed for a variety of different primary efficacy questions, clinical heterogeneity is to be expected. As a result, standard meta-analytic methods may not be appropriate for the clinical and methodological heterogeneity that is common in these studies. In the rofecoxib example, for instance, the comparators included a variety of NSAIDs, including naproxen, which might have protected against cardiovascular effects. In the rosiglitazone example, the trials differed markedly in terms of the populations studied (including one that included patients for whom rosiglitazone was contraindicated), comparison groups and lengths of follow-up. Some meta-analytical models are better than others in handling this heterogeneity, but the standard Cochran's Q test is underpowered to detect heterogeneity, and the results are more sensitive to untestable assumptions.

Finally, because trials are designed primarily for efficacy endpoints, adverse effects are not always completely or consistently reported. The best example of the impact of this problem can be seen in the difference in the risk of myocardial infarction between rofecoxib studies with and without external endpoint adjudication committees. To avoid this problem, the FDA commissioned such a review for all of the antidepressant studies in its analysis. But, in other situations, and especially for less common or serious outcomes, one wonders whether studies that report no adverse events in either arm have truly experienced none or simply have not been looking for them. The examples also demonstrate how the qualitative results of a drug safety meta-analysis can depend on statistical assumptions that are difficult to empirically verify. In particular, large sample approximations that produce similar results in different methods in other settings can come to different conclusions with the sparse data typically available for a safety meta-analysis.

So what can be done? At a minimum, these examples suggest that random-effects models rather than fixed-effects models for meta-analysis should be considered [11]. Beyond that, it is important to explore heterogeneity through subgroup analyses and meta-regression, and perhaps to use more sophisticated statistical methods for meta-analysis that take into account the small number of events and heterogeneity in the underlying studies. Kaizar et al.' [5] model for antidepressants illustrates both the benefits and challenges of this approach. But, even in this example, the model is not able to distinguish the effects of the indication and drug type, as noted above. In some circumstances, collecting and analysing individual patient-level data can be useful, but this approach can also be challenging [11]. The problem is that every statistical model relies on assumptions about the data, and many of these assumptions are difficult or impossible to verify empirically. The result is best seen in Localio et al.' [14] summary of the different models for the rosiglitazone analysis; the P values testing the hypothesis of an elevated risk of myocardial infarction ranged from <0.001 to 0.244.

Dealing with differences in the design, methods and adverse event reporting among existing studies can be more challenging. One way to deal with this, and the statistical issues in the preceding paragraph, is for analysts to use multiple statistical models to do their meta-analysis as a sensitivity analysis to see just how dependent the results are on the choice of models. One can also conduct exploratory subgroup analyses, including looking at how the results depend on precisely which studies are included in the analysis. Awareness that the results of any one analysis are not as precise as they seem to be is an important step in the right direction. And, as with all statistical analyses, the results need to be considered in the context of other information about biological plausibility, mechanisms of action, parallel indications and outcomes, and other information about the potential risk of a drug.

Since part of the problem is inconsistent reporting of adverse events, it can also be useful to seek additional information, and perhaps even conduct an external review, as the FDA did for its antidepressant analysis [3]. Prospectively, regulatory authorities anticipating the need for a meta-analysis of adverse effects might coordinate the designs of primary trials to ensure that adverse events are ascertained and reported in an objective and consistent way.

Acknowledgments The author is grateful for comments received from John Kraemer and Ira Shoulson.

Funding and conflict of interest No sources of funding were used in the preparation of this article. Michael A. Stoto has no conflicts of interest that are directly relevant to the content of this article.

References

1. Jüni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. Lancet. 2004;364(9450):2021–9.
2. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med. 2007;356(24):2457–71.
3. Hammad TA, Laughren T, Racoosin J. Suicidality in pediatric patients treated with antidepressant drugs. Arch Gen Psychiatry. 2006;63(3):332–9.
4. Bridge JA, Iyengar S, Salary CB, Barbe RP, Birmaher B, Pincus HA, Ren L, Brent DA. Clinical response and risk for reported suicidal ideation and suicide attempts in pediatric antidepressant treatment: a meta-analysis of randomized controlled trials. JAMA. 2007;297(15):1683–96.
5. Kaizar EE, Greenhouse JB, Seltman H, Kelleher K. Do antidepressants cause suicidality in children? A Bayesian meta-analysis. Clin Trials. 2006;3(2):73–98.

6. Bombardier C, Laine L, Reicin A, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. N Engl J Med. 2000;343(21):1520–8.
7. Kearney PM, Baigent C, Godwin J, Halls H, Emberson JR, Patrono C. Do selective cyclo-oxygenase-2 inhibitors and traditional non-steroidal anti-inflammatory drugs increase the risk of atherothrombosis? Meta-analysis of randomised trials. BMJ. 2006;332(7553):1302–8.
8. Bradburn MJ, Deeks JJ, Berlin JA, Localio RA. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Stat Med. 2007;26(1):53–77.
9. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med. 2004;23(9):1351–75.
10. Sutton AJ, Cooper NJ, Lambert PC, Jones DR, Abrams KR, Sweeting MJ. Meta-analysis of rare and adverse event data. Expert Rev Pharmacoecon Outcomes Res. 2002;2(4):367–79.
11. Berlin JA, Crowe BJ, Whalen E, Xia HA, Koro CE, Kuebler J. Meta-analysis of clinical trial safety data in a drug development program: answers to frequently asked questions. Clin Trials. 2013;10(1):20–31.
12. Diamond GA, Bax L, Kaul S. Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. Ann Intern Med. 2007;147(8):578–81.
13. Tian L, Cai T, Pfeffer MA, Piankov N, Cremieux PY, Wei LJ. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 × 2 tables with all available data but without artificial continuity correction. Biostatistics. 2009;10(2):275–81.
14. Localio R, Cornell J, Mulrow C. Much ado about Avandia: the meta-analysis of rare events in the service of health policy. In: 7th International Conference on Health Policy Statistics: Philadelphia; 2008.
15. Warn DE, Thompson SG, Spiegelhalter DJ. Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. Stat Med. 2002;21(11):1601–23.
16. Nissen SE, Wolski K. Rosiglitazone revisited: an updated meta-analysis of risk for myocardial infarction and cardiovascular mortality. Arch Intern Med. 2010;170(14):1191–201.
17. Rosen CJ. Revisiting the rosiglitazone story—lessons learned. N Engl J Med. 2010;363(9):803–6.
18. Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions version 5.1.0. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.