Meta-Analysis of “Sparse” Data: Perspectives From the Avandia Cases

Michael O. Finkelstein*       Bruce Levin†

* mofinkelstein@hotmail.com
† Columbia University, Bruce.Levin@Columbia.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/columbiabiostat/art18

Copyright ©2012 by the authors.
Meta-Analysis of “Sparse” Data: Perspectives From the Avandia Cases

Michael O. Finkelstein and Bruce Levin

Abstract

Combining the results of multiple small trials to increase accuracy and statistical power, a technique called meta-analysis has become well established and increasingly important in medical studies, particularly in connection with new drugs. When the data are sparse, as they are in many such cases, certain accepted practices, applied reflexively by researchers, may be misleading because they are biased and for other reasons. We illustrate some of the problems by examining a meta-analysis of the connection between the diabetes drug Avandia (rosiglitazone) and myocardial infarction that was strongly criticized as misleading, but led to thousands of lawsuits being filed against the manufacturer and the FDA acting to restrict access to the drug. Our scrutiny of the Avandia meta-analysis is particularly appropriate because it plays an important role in ongoing litigation, has been sharply criticized, and has been subject to a more searching review in court than meta-analyses of other drugs.
META-ANALYSIS OF “SPARSE” DATA: PERSPECTIVES FROM THE AVANDIA CASES

Michael O. Finkelstein
Bruce Levin*

ABSTRACT: Combining the results of multiple small trials to increase accuracy and statistical power, a technique called meta-analysis has become well established and increasingly important in medical studies, particularly in connection with new drugs. When the data are sparse, as they are in many such cases, certain accepted practices, applied reflexively by researchers, may be misleading because they are biased and for other reasons. We illustrate some of the problems by examining a meta-analysis of the connection between the diabetes drug Avandia (rosiglitazone) and myocardial infarction that was strongly criticized as misleading, but led to thousands of lawsuits being filed against the manufacturer and the FDA acting to restrict access to the drug. Our scrutiny of the Avandia meta-analysis is particularly appropriate because it plays an important role in ongoing litigation, has been sharply criticized, and has been subject to a more searching review in court than meta-analyses of other drugs.


When a drug company proposes to market a new drug in the United States, it is required by the U.S. Food and Drug Administration (FDA) to conduct clinical trials to demonstrate the drug’s safety and efficacy. These trials are done in phases, with small trials conducted first to establish preliminary safety and appropriate dose levels, and then successively larger trials for safety and efficacy. The result can be a number of small trials that by themselves do not have statistically significant results for adverse events but when combined may show a statistically significant effect. A need to combine results from

*Michael O. Finkelstein (mofinkelstein@hotmail.com) is a member of the New York Bar and the adjunct faculty of Columbia Law School. Bruce Levin (Bruce.Levin@Columbia.edu) is Professor and former Chair of Biostatistics at the Mailman School of Public Health, Columbia University. This article grew out of analyses done for a moot court in the seminar Statistics for Lawyers at Columbia Law School, co-taught by Mr. Finkelstein and Nathan Schachtman. We thank Emilia Bagiella and Roger Vaughan, professors at the Mailman School, who served as expert witnesses at the trial, and Mr. Schachtman for his valuable contributions to this article. The views expressed here are solely ours.
several studies also arises when postmarketing adverse events surface. Because these adverse outcomes are often rare, even large individual studies may lack the statistical power to definitively establish an association between the drug and the adverse effect. The process of combining studies to generate an overall estimate of effect size and to calculate confidence intervals for the estimate is known as meta-analysis. “It is a way of systematizing the time-honored approach of reviewing the literature, which is characteristic of science, and placing it in a standardized framework.”

The FDA needs a meta-analysis of all available clinical data on the safety and efficacy of new drugs to avoid being blindsided by a less structured narrative review that may leave out or minimize inconvenient results. For example, in December 2008, the FDA recommended that for certain new clinical trials of antidiabetic medications in the planning stage, “[s]ponsors should ensure that phase 2 and phase 3 clinical trials are appropriately designed and conducted so that a meta-analysis can be performed at the time of completion of these studies that appropriately accounts for important study design features and patient or study level covariates.”

The FDA’s embrace of meta-analysis reflects the fact that in the past thirty years meta-analysis has become a widely used tool in biostatistics. The number of published medical articles containing “meta-analysis” in the title, or in its keywords, has grown exponentially from 0 in 1980 to 2,324 in 2009. Given the growth of meta-analysis in medical research and other fields, such as education and criminology, it is not surprising to find expert witnesses in court proceedings relying on meta-analysis in their testimony. Before 1988, “meta-analysis” had not appeared in any court opinion; since then the term has appeared in at least some 50 opinions. In recent years, meta-analysis has played a role in litigation involving at least 22 drugs. In the Reference Manual on Scientific Evidence, published by the Federal Judicial Center for the guidance of judges, and frequently cited in judicial opinions, the authors of the chapter on epidemiology approve of meta-analysis in pooling randomized clinical trials. And the courts have usually brushed aside objections to meta-analysis, holding that testimony based on it would be sufficiently reliable to

1. FED. JUDICIAL CTR., REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 380 (2d ed. 2000).
2. FOOD & DRUG ADMIN., CTR. FOR DRUG EVALUATION & RESEARCH, GUIDANCE FOR INDUSTRY: DIABETES MELLITUS—EVALUATING CARDIOVASCULAR RISK IN NEW ANTIDIABETIC THERAPIES TO TREAT TYPE 2 DIABETES 3 (2008).
5. We are indebted to Nathan Schachtman for supplying us with a list of these drugs.
6. FED. JUDICIAL CTR., supra note 1, at 380–81 (noting, however, that use of meta-analysis becomes more problematic when applied to observational studies because of the methodological differences commonly encountered among such studies).
Meta-Analysis of “Sparse” Data

meet Daubert standards.\(^7\) Asserted defects are usually held to go to weight and not admissibility.

In this article we deal with certain issues that arise in meta-analyses where the effect being studied is relatively rare. This is an important class of cases because, in particular, initial trials of new drug interventions usually involve relatively small numbers of trial participants and such trials may not contain any events of interest or only a relatively small number. Widely practiced methods for doing meta-analysis when component studies have no events in either the treatment or control arms (or both) can generate misleading results in the sparse-data context. In addition, when the number of events is small, calculations of statistical significance or of confidence intervals lead to more marginal results that are more likely to be vitiated if the assumptions of the calculation are not fully satisfied. Thus, issues that may be minor and unacknowledged when there are ample data may become important when adverse events are rare. Although meta-analyses have been involved in a number of litigations, the issues we discuss here have not, with one minor exception to be noted, been addressed by the parties or the courts in judicial proceedings.

I. META-ANALYSES OF THE AVANDIA STUDIES

To explore some of these problems we use as our reference point a meta-analysis of certain cardiovascular (CV) risks from the drug rosiglitazone, marketed by GlaxoSmithKline (GSK) under the trade name Avandia. This drug was claimed by GSK to improve blood sugar levels in diabetics without having certain negative effects. It quickly became a huge seller for GSK. The company had in its possession a number of efficacy studies of Avandia that included CV adverse events. Myocardial infarction (MI, or heart attacks) was one of the most important adverse events for the study subjects (the others being mortality and congestive heart failure) and to keep the discussion focused on the issues, we confine ourselves to the MI outcome. In 2005, GSK made a meta-analysis of these studies and made another meta-analysis in 2006 with added studies; both were sent to regulatory agencies worldwide, including the FDA; both showed an excess risk of heart attacks associated with Avandia.\(^8\) A summary of the updated analysis was posted in August 2006 to the GSK Clinical Trial Register, but there was no press release or other public

---


\(^8\) Alexander Cobitz et al., A Retrospective Evaluation of Congestive Heart Failure and Myocardial Ischemia Events in 14237 Patients with Type 2 Diabetes Mellitus Enrolled in 42 Short-Term, Double-Blind, Randomized Clinical Studies with Rosiglitazone, 17 PHARMACOEPIEMIOLOGY & DRUG SAFETY 669, 770 (2008).
announcement. In June 2007, however, Steven E. Nissen and Kathy Wolski published a meta-analysis of 42 randomized controlled trials (RCTs) involving Avandia.\(^9\) Combining their data, Nissen and Wolski reported that Avandia increased the odds of MI by about 43%. Although none of the studies, considered separately, showed a statistically significant increase in MI (our calculation), the basis for Nissen and Wolski’s conclusion with regard to MI was that the meta-analytic estimate for the MI common odds ratio was 1.43 and was statistically significant ($p = 0.03$); the 95% confidence interval for the estimate extended from 1.03 to 1.98.\(^10\) In September 2007, Singh and his coauthors published a meta-analysis of four RCTs after detailed screening of 140 trials for CV events; consistent with the Nissen and Wolski meta-analysis, they reported a risk ratio of 1.42 (ci 1.06 to 1.91) for MI.\(^11\) In 2010 Nissen and Wolski published an updated meta-analysis that combined 56 studies.\(^12\) They reached the same conclusion as before: rosiglitazone therapy significantly increased the risk of MI, although the point estimate of the increase was less than before (OR 1.28; ci 1.02 to 1.63).

Nissen and Wolski’s 2007 meta-analysis touched off a firestorm. Thousands of lawsuits were filed against GSK by diabetics and others who had taken Avandia. GSK’s stock price dropped 7.8% in trading on the day their analysis was published, triggering investor lawsuits. The FDA took action. After the study appeared, the agency changed the planned agenda for meetings of two advisory committees to have them address, in a combined meeting, the CV risks of Avandia.\(^13\) The combined committees voted that Avandia increased the risk of MI events, but that the drug should remain on the market with appropriate labeling. Following that meeting, the FDA announced a requirement for a boxed warning that referenced the Nissen and Wolski study, which, it said, showed that Avandia was associated with increased risk of MI events. It added that three other studies with longer durations had neither confirmed nor excluded this risk. The warning concluded that, “[i]n their entirety, the available data on the risk of myocardial ischemia are inconclusive.”\(^14\)

---


10. Id. Hereinafter all confidence intervals (ci’s) are 95% two-sided intervals.


13. The two committees that met together were the Endocrinologic and Metabolic Drugs Advisory Committee and Drug Safety and Risk Management Advisory Committee. See CDER 2007 Meeting Documents, U.S. FOOD & DRUG ADMIN. (Sept. 11, 2007), http://www.fda.gov/ohrms/dockets/ac/cder07.htm#DrugSafetyRiskMgmt.

14. FDA Adds Boxed Warning for Heart-related Risks to Anti-Diabetes Drug Avandia, U.S. FOOD & DRUG ADMIN. (Nov. 14, 2007), http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2007/ucm109026.htm. Myocardial ischemia is a broad composite that includes MI, with less severe outcomes such as angina and chest pain.
2007 FDA did a meta-analysis and updated it in 2010 with additional trials. The 2010 analysis, combining 52 trials, found a statistically significant OR of 1.8 (ci 1.03 to 3.25). There were further events that made headlines in 2010, which we will discuss at the end of this article.

Did the Nissen and Wolski study overstate the evidence for the MI risks of Avandia? Critics attacked their methods and presented alternative analyses of the same data that produced nonsignificant findings. In exploring the issues raised by the critics, and more generally by rare-event meta-analyses, we begin with a primer on the odds ratio commonly used in these studies and the methods for combining data used in meta-analysis and then turn to the criticisms.

II. ODDS RATIOS AND METHODOLOGY

A. The 2-by-2 Table

When discrete events, such as MIs, are the subject of interest, the results of a study are usually expressed in a fourfold table, such as shown in Table 1 below. One statistic of interest is the odds ratio that is computed from the table.

<table>
<thead>
<tr>
<th></th>
<th>Event</th>
<th>No Event</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>a</td>
<td>b</td>
<td>n₁</td>
</tr>
<tr>
<td>Persons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>c</td>
<td>d</td>
<td>n₂</td>
</tr>
<tr>
<td>Persons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>m₁</td>
<td>m₂</td>
<td>t</td>
</tr>
</tbody>
</table>

The proportion of treated persons with the event is \( a/n₁ \), and this estimates the probability that a treated person would have the event. The proportion of treated persons without the event is \( b/n₁ \), which estimates the complementary probability. The ratio of the two is \( a/b \), which is the sample odds on the event for treated persons. Similarly, the sample odds on the event for control persons are \( c/d \). Hence the odds ratio on an event for treated vs. control persons is \( a/b \) divided by \( c/d \), or \( ad/bc \). Cells \( a \) and \( d \) are called the main-diagonal cells and cells \( b \) and \( c \) are referred to as the off-diagonal cells. The odds ratio is sometimes called the cross-product ratio because it is the product of the main-
diagonal cells divided by the product of the off-diagonal cells. The upper left cell (number of treated persons with events) is called the reference cell, although any other cell can be chosen as the reference cell as long as the choice is consistent across all tables. The totals for rows and columns are called the margins of the table, and $t$ is the total number of persons in the study. The odds ratio (abbreviated OR) is one of several measures of treatment effect, other common ones being the risk ratio, estimated by $\frac{(a/n_1)(c/n_2)}{(b/n_1)(d/n_2)}$, and the risk difference, estimated by $(a/n_1) - (c/n_2)$. When the event rate is low, such as it is for MI in the Avandia studies (around 0.5%), the odds ratio and the rate ratio are approximately equal. The odds ratio has some very useful statistical properties (for example, invariance) that recommend its use in epidemiologic studies.17

The basic statistical tasks of meta-analysis involve drawing inferences about the common odds ratio, that is, the assumed constant value of the true odds ratio underlying each study, based on the studies’ sample odds ratios, which of course vary from study to study.18 Three important elements of the analysis are: (1) to estimate the common odds ratio, (2) to provide ways of testing the result for statistical significance and calculating a confidence interval for the estimate, and (3) to assess whether the variation of the sample odds ratios is larger or smaller than what one would expect under the assumption of a constant true underlying odds ratio. If the assumption is found untenable, other statistical methods would need to be employed—we return to this point below.19

There are several ways of calculating a common odds ratio when confronted with multiple studies. We describe three: the inverse-variance weighting method, the conditional maximum likelihood method, and the Mantel-Haenszel method.


18. The methods we describe are still applicable if the true odds ratios do not vary substantially. If the true odds ratios are heterogeneous, other methods must be used. See the discussions of homogeneity and random effects models infra Parts III.C, III.D.

19. Even before reaching these basic statistical tasks, a hugely important phase lies in defining the criteria for which studies to select for the meta-analysis. This will include aspects of study design (for example, randomized versus observational, controlled versus uncontrolled), patient populations (for example, diabetic versus prediabetic), quality of the reporting regarding matters such as standard errors and adjustment for covariates, and so forth. Finding all relevant studies meeting the defined criteria is another important requirement. We deal with the selection of studies in the Avandia context in Part III.
B. Inverse-Variance Weighting Estimate

This method assumes that one has an unbiased (or at least consistent) point estimate of a parameter of interest together with its standard error from each of the independent studies. The parameter of interest is frequently the natural logarithm of the odds ratio, that is, the log odds ratio (LOR), because the sample LOR is more normally distributed and its variance is more stable than the odds ratio without the log. Under the assumption that the true parameter of interest is the same in each study, any weighted average of unbiased estimators is itself unbiased. To minimize the squared standard error of the weighted average the usual practice is to choose weights for each study that are inversely proportional to the variance of the respective point estimates. Thus, if the squared standard error of a point estimate is large, implying that it is not precise, the weight given to the study will be small; conversely, if the squared standard error is small, implying that the estimate is relatively more precise, the weight will be larger. In general, larger studies will have smaller standard errors and can dominate smaller studies in the averages. In the Avandia case, there were two large studies, which alone generated point estimates that were very close to the meta-analysis result for all studies.

The drawback of this method in the present case is that if a study has a zero in one or more of the cells, the logarithm of the cross-product ratio is undefined, and thus the method cannot be applied without some modification. In small studies this is a fairly common occurrence, though the usual justification of statistical validity of the inverse-variance weighting method assumes large numbers of events in each study. In the Avandia case, out of 42 studies, there were 4 studies that had double zeros (no events in either the treatment or control groups) and 24 studies that had one zero (no events in either the treatment or control groups but not both), so the problem of how to deal with tables with zero cells was extremely important. The individual study four-fold tables are listed in the Appendix. We return to this issue after we discuss the other methods.

C. Conditional Maximum Likelihood Estimate

This method picks as the estimate of the common odds ratio the value that maximizes the probability of seeing the assemblage of four-fold tables in the individual studies, regarding the marginal totals of each study as fixed at their observed values, that is, “conditioning” on the margins.20 This probability, when viewed as a function of the common odds ratio, is called the conditional likelihood of the data, and the value of the common odds ratio that maximizes

---

20. All four marginal totals are not fixed in advance by the study design, but the probability calculations enumerate all possible ways the innards of the fourfold tables might have occurred with the observed sample sizes and given the observed total number of events of interest. Regarding MIs, for example, if an inference of a significant association can be made in this conditional approach, it suffices to reject the null hypothesis for the experiment’s actual design, which allows varying number of MIs.
the conditional likelihood is called the conditional maximum likelihood estimate, or cmle, of the common odds ratio. The cmle has the characteristic that it makes the expected sum of the frequencies in the reference cells of the component studies equal to the observed sum of the reference cell frequencies. Because the expected cell frequencies depend on the common odds ratio, setting the expected sum equal to the observed sum defines the cmle only implicitly and the solution has to be found by computer iterations that close in on the cmle.\textsuperscript{21} The Nissen and Wolski study used the so-called Peto variant of this method.\textsuperscript{22} In the Peto variant, the computation begins with an initial value of the LOR equal to zero (corresponding to an odds ratio of 1) and makes only a single iteration. For this reason it is also called the one-step method. This method has shortcomings, which are discussed below.

D. Mantel-Haenszel Estimate

An alternative to the cmle of the common odds ratio is the Mantel-Haenszel (M-H) estimate.\textsuperscript{23} This is \( R/S \), where \( R \) is the sum across the tables of \( a_i d_i/t_i \), which are the products of the main-diagonal cell frequencies divided in each case by the study size \( t_i \), and \( S \) is the sum across the studies of \( b_i c_i/t_i \), which are the products of the off-diagonal cell frequencies divided in each case by the sample size \( t_i \). The M-H estimate has been shown to be a consistent estimator of the common odds ratio not only when each study has large marginal totals, but also when there are many studies each with some small marginal totals.\textsuperscript{24} It therefore has an appealingly wide applicability. It also has the advantage that a computer iteration is not required to calculate it. In large samples the cmle and the M-H estimates will produce very similar results. The cmle is somewhat more statistically efficient than the M-H, that is, it will have a smaller sampling error than the M-H.\textsuperscript{25} The M-H estimate of the common

\textsuperscript{21} Technical note: To maximize the conditional likelihood function, one takes the derivative of its logarithm with respect to the LOR as the unknown argument and sets the derivative equal to zero. The characterization of the cmle given above in the text is mathematically equivalent to setting the derivative equal to zero. Newton’s method for finding the root of an equation is then most often used to solve the equation given a starting value of the LOR, usually taken equal to zero. When the iterations converge, the cmle of the common OR is obtained by taking the antilogarithm of the maximizing LOR. Newton’s method also provides the standard error of the log cmle. This is found as the square root of the negative second derivative of the log conditional likelihood function, which in this case is equal to the square root of the reciprocal of the sum of the conditional variances of the reference cells, each evaluated at the cmle.

\textsuperscript{22} The Peto variant has been in use for meta-analysis since at least 1985, when it was proposed in Salim Yusuf et al., Beta Blockade During and After Myocardial Infarction: An Overview of the Randomized Trials, 27 PROGRESS IN CARDIOVASCULAR DISEASES 335, 338 (1985).

\textsuperscript{23} The M-H method dates to 1959, when it was proposed by Nathan Mantel & William Haenszel, Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease, 22 J. NAT’L CANCER INST. 719 (1959).

\textsuperscript{24} Norman Breslow, Odds Ratio Estimators When the Data Are Sparse, 68 BIOMETRIKA 73, 81–82 (1981).

\textsuperscript{25} A formula for the standard error of the logarithm of the M-H estimate of the common odds ratio is given in SFL, supra note 17, at 245.
E. Statistical Significance and Confidence Intervals

The computation of statistical significance and a confidence interval for the estimate of a common odds ratio is a critical part of meta-analysis. After all, one of the main purposes of combining the evidence of multiple small studies is to determine whether, in the aggregate, there is a statistically significant association between the treatment and the relevant events. The statistical significance of the combined studies may be appraised by calculating an M-H test statistic: this is the difference between the sum of the observed values of the reference cells (\(\sum_{ti} a_{i} \), say) less the sum of their expected values under the null hypothesis (that the odds ratio is 1), \(\sum m_{i} n_{i}/t_{i}\). This difference, reduced in absolute value by subtracting a continuity correction of 0.5, is squared and divided by the sum of the study variances (which are also calculated under the null hypothesis), \(\sum m_{i} m_{0i} n_{0i}/t_{i}^{2}(t_{i}–1)\). The resulting statistic has an approximate chi-squared distribution with 1 degree of freedom under the null hypothesis. If the probability of a value as large or larger than that calculated is less than 0.05, the null hypothesis is rejected and the data are found to be statistically significant. In the Avandia data the M-H chi-squared statistic with one-half continuity correction was 4.24, with p-value = 0.039, enabling an inference that the odds ratio is indeed greater than one.26

Calculation of an approximate confidence interval for the estimated common odds ratio is appropriate when a large number of study subjects are involved, either because there are a small number of studies with large marginal totals or a large number of studies with small marginal totals. In either case, the natural logarithm of the M-H odds ratio has an approximate standard error formula referred to above,27 and the 95% confidence interval is constructed by adding and subtracting the usual 1.96 standard errors around the log M-H odds ratio.28 Antilogs are then taken to retrieve the limits of the common odds ratio in the original units.

26. It has been pointed out that apparently conclusive meta-analyses may be misleading if no adjustment is made for repeated testing for significance as data accumulate with new trials. Jesper Brok et al., Apparently Conclusive Meta-Analyses May Be Inconclusive—Trial Sequential Analysis Adjustment of Random Error Risk Due to Repetitive Testing of Accumulating Data in Apparently Conclusive Neonatal Meta-Analyses, 38 INT’L J. EPIDEMIOLOGY 287, 288 (2009). It is not common for such adjustments to be made, and Nissen and Wolski did not make it in their follow-up study. For an example of adjustment for multiple comparisons in a legal context, see David H. Kaye & Joseph L. Gastwirth, Where Have All the Women Gone? The Gender Gap in Supreme Court Clerkships, 49 JURIMETRICS J. 411, 426 (2009) (analyzing hiring data for individual justices).

27. SFL, supra note 17, at 245.
28. Id. at 244–45.
When the numbers of subjects are too small, exact methods are required to compute confidence intervals. An exact, test-based 95% confidence interval is defined as follows. A possible value for the common odds ratio, say OR₀, is contained in the confidence interval if and only if it would not be rejected given the study results at level 0.05 two-tailed by an exact test if taken as a point null hypothesis.²⁹ If the value of OR₀ could be rejected, it falls outside the confidence interval. Calculation of an exact confidence interval in this way involves extensive computation, though quite routine with modern statistical software. With this introduction, we now turn to the criticisms of the Nissen and Wolski meta-analysis.

III. ISSUES RAISED BY THE NISSEN AND WOLSKI META-ANALYSIS

A. The Peto Method

Nissen and Wolski used the Peto method to calculate the common odds ratio because CV events were rare.³⁰ Diamond and his coauthors criticized this choice.³¹ The Peto method generates a fairly good approximation of the cmle (and thus a fairly good estimate of the true common odds ratio) if the cmle is close to 0 or the sample sizes are reasonably balanced and the effect size is modest.³² If these conditions do not hold, the Peto estimate can differ dramatically from the cmle. For example, in an investigation using data for three hypothetical studies, the Peto method and the cmle method were compared.³³ The studies were highly unbalanced (the unexposed subjects were between 5 and 10 times more numerous than the exposed subjects). The odds ratios for the studies were 4.50, 5.40, and 3.75 respectively. The common odds ratio estimate based on the Peto method was 11.2 while the cmle estimate was 4.35 with 95% confidence interval 1.59 to 10.8. Clearly the Peto estimate was biased and could not be consistent with any weighted average of the three odds ratios.

²⁹. To be precise, the hypothesis test envisioned here is H₀: OR=OR₀ versus H₁: OR≠OR₀. Use of exact methods is advocated in Ben Vandermeer et al., Meta-Analyses of Safety Data: A Comparison of Exact Versus Asymptotic Methods, 18 STAT. METHODS MED. RES. 421, 421 (2009).
³⁰. Nissen & Wolski, supra note 12, at 1192.
³². In Peto’s method the starting point for iteration is the null value, LOR=0, drawing a tangent to the curve of the first derivative of the log conditional likelihood function at that point, and taking as the estimate the place where the tangent crosses the LOR axis. As stated in the text, this is not a bad estimate if the cmle is close to 0 or if the numbers in the two arms of the tables are not too different and the effect size is modest. In the latter case, in particular, the curve of the first derivative is almost a straight line so the tangent to the curve from any starting point crosses the LOR axis close to where the curve crosses it. The curvature depends on the third moment of the noncentral hypergeometric distribution, which is small in this case.
The critics of Nissen and Wolski pointed out that, in fact, some of the studies were fairly unbalanced as between the treatment and control arms, that is, had allocation ratios of 3:1 or 4:1, thus creating the possibility of bias. The point would generally be well taken. However, our calculation of the \textit{cmle} estimate without limitation to a single iteration shows no difference from the Peto estimate of 1.43. The M-H estimate of the common odds ratio was also very close. But the larger point is that computers now make multiple iterations trivial, so there is no reason to use the one-step method to estimate the common odds ratio when using the \textit{cmle} or the M-H estimate could avoid serious bias.

B. The Problem of Zeros

When meta-analysis is used to combine the results of a number of small studies, some may have zero cells in one arm or the other, or both arms. Zeros in both arms means that events did not occur for either the treatment cases or the controls. As we previously noted, in the Avandia data there were four studies with zeros in both arms, which were excluded in the Nissen and Wolski analysis. Nissen and Wolski were criticized for this, but the criticism seems inappropriate given their objective, which was to estimate a common odds ratio from the studies as a step in determining whether Avandia was associated with an excess risk of MI. Double-zero studies are not informative about the comparative risk of the treatment vs. the comparator, as reflected in the odds ratio or the risk ratio, and are usually disregarded when that is the focus of the meta-analysis. In the Peto, \textit{cmle}, or M-H procedures, for example, these tables may be included or excluded from the analysis at will because either way yields identical results. Some critics complained that excluding the double-zero tables would overstate the entire level of risk, that is, the overall risk averaged over all studies, and thereby could also overstate the risk difference between arms. This is true when the statistic of interest is a risk difference, as it might be in a public health appraisal, because the risk difference multiplied by the number exposed gives the excess number of adverse events arguably attributable to the exposure. For analytic purposes, either a risk difference or an odds ratio may be used, but one generally needs to decide which measure to use in a meta-

\begin{itemize}
  \item \textsuperscript{34} Bracken, \textit{supra} note 16, at 938.
  \item \textsuperscript{35} To three decimal places, the \textit{cmle} was 1.426, the M-H estimate was 1.427 and the Peto estimate was 1.428.
  \item \textsuperscript{36} See, e.g., Diamond et al., \textit{supra} note 16, at 579.
  \item \textsuperscript{37} Technically, tables with zeros in both arms are conditionally uninformative, meaning that once one conditions on the marginal totals, there is no variation possible within the table, and thus no information about the underlying odds ratio is available. Indeed, any odds ratio at all would be compatible with a zero-zero study, because the ratio of two true odds close to zero can be arbitrarily large or small. Some statisticians believe there can be information gleaned from these tables by foregoing the conditional approach, but doing so requires making other assumptions, for example, Bayesian assumptions, which may raise more questions than they settle.
  \item \textsuperscript{38} Diamond et al., \textit{supra} note 16, at 579.
\end{itemize}
analysis because studies with a constant odds ratio will generally not have constant risk differences, and vice versa. Because odds ratios are known empirically to have greater stability across many study populations than do risk differences, a meta-analysis of the odds ratio is often preferred for etiologic research. In short, Nissen and Wolski could have used a risk difference, but cannot be faulted for using an odds ratio.

Another consideration may come into play in the context of a lawsuit. If the plaintiffs claim that a drug caused them harm, a key issue is the probability that the plaintiffs’ harm was caused by the drug as opposed to other factors. This is known as attributable risk (among the exposed) or specific causation. But to compute the probability of specific causation it is necessary to compute a risk ratio (or an odds ratio as an approximation for the risk ratio). With an odds ratio of 1.43, the probability that a random person who took Avandia and suffere...
Haldane’s correction does not eliminate all bias and the residual bias that is not corrected grows larger as the expected numbers of events in the table become smaller. In particular, we find that the residual bias becomes important when the expected number of events in any cell of a 2-by-2 table is less than 5. When the true LOR is greater than zero, the value estimated with Haldane’s correction will nevertheless be biased towards zero, that is, the residual bias is negative. Referring to the Avandia data used by Nissen and Wolski, we calculated that the residual bias in the inverse variance method may well have reduced the estimated common odds ratio by a weighted average of about 14.3%, bringing it down from 1.426 to 1.223 (this reduction made it nonsignificant). The existence of this residual bias in the small numbers case is an important reason not to use the inverse-variance method for estimating the common odds ratio. The Haldane correction does not apply to the other three methods we have discussed.

Critics of Nissen and Wolski’s meta-analysis thought their results were put in doubt because when other methods were used, they showed a smaller effect and were not statistically significant. One critic reported a risk ratio of 1.27 (ci 0.95 to 1.71) and an M-H odds ratio of 1.28 (ci 0.95 to 1.72). Another critic reported lower and nonsignificant results for odds ratios computed from the inverse-variance and M-H methods, each using a correction to accommodate zero cells. Indeed, our own calculations of the original Nissen and Wolski Avandia data show that the inverse-variance weighted average estimate of the common odds ratio, adding 0.5 to each cell to all tables, including the double-zero tables, is 1.22 (ci 0.91 to 1.65). This is substantially smaller than any of the other measures. Even if 0.5 is added only to tables with any zero cells, the result is still too small: 1.26 (ci 0.93 to 1.72). But the fact that these methods lead to lower estimates of association is not a valid basis for criticizing Nissen and Wolski’s conclusion. We have seen that the residual bias that remains after the Haldane correction accounts for the lower estimate of the common odds ratio in the inverse-variance method. And neither the cmle method, the M-H method, nor the Peto method needs such additions to keep individual odds ratios from imploding to zero or infinity. Each of those but corrects for the leading term in the approximate expression for the bias in large samples. For a discussion of this bias and its correction, see FLEISS ET AL., supra note 39, at 103, 105–06, 136–37. It appears that this practice was judicially approved in In re Gadolinium-Based Contrast Agents, No. 1:08-GD-50000, 2010 WL 1796334, at *24 (N.D. Ohio May 4, 2010) (defendant challenged an expert’s “imputation of data” in a meta-analysis; court held that imputed data was “an acceptable technique for avoiding the calculation of an infinite odds ratio.”).

43. For illustrative calculations supporting this rule of thumb see the Appendix.
44. Details of our calculation are given in the Appendix.
46. Diamond et al., supra note 16, at 579. The addition of Haldane’s bias correction for zero cells in the M-H test procedure is simply an error, whereas use of the one-half continuity correction is recommended to improve the accuracy of the chi-squared distribution as an approximation to the exact sampling distribution. This should be distinguished from the use of some correction to reduce the small bias in the M-H estimate of the common odds ratio. See infra p. 136 for further discussion.
methods involves sums of data in the cells of the fourfold tables and if one of those cells is zero it simply and appropriately contributes no evidence of risk for the corresponding treatment. For example, in one of the Avandia studies no treated persons had MIs while one untreated person did. In the M-H method recall that the numerator involves the sum of $a_d/t_i$ from the main diagonal cells across the studies and the denominator involves the sum of $b_c/t_i$ from the off-diagonal cells. In that study the $a$ cell will be 0 and the $ad/t$ component of that study will contribute nothing to the numerator, while the $c$ cell will be non-zero and the study will contribute to the denominator. The result will be an appropriate reduction in the estimated odds ratio. If 0.5 is unnecessarily added to the cells of a zero-cell table, the result will be to bias a cmle or M-H estimate of the common odds ratio downwards towards 1.47

But, if it is not necessary to add a constant to zero cells for the MH or cmle methods, is it desirable to do so to reduce bias in the estimates? The answer is that when there are sufficient data for meaningful conclusions the bias is trivial and may be disregarded. In the Nissen and Wolski initial study there were an aggregate of 86 MIs in the treatment groups and 72 in the control groups. To test how close the expected value of the M-H estimate would be to the odds ratio of 1.426 calculated for the data, we prepared a simulation study in which the observed margins were fixed and the common odds ratio was 1.426. Fixing the margins and the common odds ratio for the tables determines the probabilities of the possible outcomes within the tables. For example, the reference cell has a random number of MIs within the range permitted by the marginal totals.48 Using these probabilities, we generated 38 random tables and from these calculated the M-H log odds ratio. We did this 10,000 times and took the mean LOR. The results were as follows. Table 2 demonstrates that the bias in the three uncorrected methods is negligible.49

47. This is true for the case of equal group sizes, $n_1 = n_2$. For unequal sample sizes the bias can work in the opposite direction when the imbalance is severe. See Swee ting et al., supra note 41, at 1356–57. Swee ting also shows that adding 0.5 can reverse the direction of an observed association. Id. at 1357.

48. SFL, supra note 17, at 126 (the probabilities associated with the number of MIs in the reference cell and, as a result, the innards of the whole table, are called noncentral hypergeometric probabilities).

49. When the data are very sparse, the bias can become nontrivial and a bias correction may reduce it. For example, Swee ting et al. simulated 10 trials with an aggregate expected number of events of 9 in the treatment group and 18 in the control group. Three thousand simulations yielded a bias of -20% in the uncorrected M-H LOR. A 0.5 correction reduced this to 9% and a 0.005 correction reduced it to -1%. Swee ting et al., supra note 41, at 1367. The difference with our result is probably because of (1) the smaller number of events and (2) the fact that the simulations were not conditional on the margins, as in our case, thus generating a varying (and substantial) number of zero-zero tables. Swee ting and his coauthors also suggest corrections that have a Bayesianesque interpretation and these do seem to reduce bias by a small amount for the Mantel-Haenszel LOR estimator. The simulations made by Bradburn and his coauthors show that the uncorrected M-H LOR generally performs better than a corrected M-H LOR, and show virtually no bias in the uncorrected M-H LOR except in the extreme case where the data are essentially so sparse that no reliable conclusion can be drawn. Michael J. Bradburn et al., Much Ado About Nothing: A
Table 2. Simulated Mean LORs for the 38 Trials with at Least One MI in the Nissen and Wolski Avandia Data

<table>
<thead>
<tr>
<th>Uncorrected method</th>
<th>Mean LOR (standard error)</th>
<th>Arithmetic bias in the LOR = simulated mean minus true value (of ln 1.426 = 0.35487)</th>
<th>Geometric mean OR</th>
<th>Percentage error of the geometric mean OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mantel-Haenszel</td>
<td>0.35810 (0.00170)</td>
<td>0.00322</td>
<td>1.431</td>
<td>0.323%</td>
</tr>
<tr>
<td>CMLE</td>
<td>0.35771 (0.00169)</td>
<td>0.00284</td>
<td>1.430</td>
<td>0.284%</td>
</tr>
<tr>
<td>Peto</td>
<td>0.35838 (0.00168)</td>
<td>0.00351</td>
<td>1.431</td>
<td>0.351%</td>
</tr>
</tbody>
</table>

As these examples illustrate, the practice of adding 0.5 to sparse fourfold tables, many of which have zero cells, ought to be discontinued because it can create bias and is unnecessary given that methods are available that generally avoid the need for any such addition. One exception applies to random-effects models, for which the inverse-variance weighted average of sample log odds ratio method is often used. Here again, though, other methods are available that do not require adding corrections to cells.\(^{50}\)

C. Homogeneity

In producing an overall summary estimate it is assumed that the studies constitute repeated observations of the same true underlying odds ratio, and differences are attributable solely to sampling error. If the individual studies are not too small, one of several usual methods to test that assumption is to calculate a statistic that reflects the variability of the LORs in the individual tables around their weighted average across tables. Called the chi-squared statistic for homogeneity, it is a weighted sum of the squared differences between the separate LORs and their weighted average.\(^{51}\) The weights are the reciprocals of the variances of the individual LORs, so more precise studies have greater weight in the sum than less precise studies. Large values of chi-squared indicate that there is too much variation in individual odds ratios to have occurred by chance, assuming a constant odds ratio for all the studies. In

\(^{50}\) Comparison of the Performance of Meta-Analytical Methods with Rare Events, 26 STAT. MED. 53, 63 (2007).

\(^{51}\) For a discussion of fixed-effects versus random-effects models, see infra Part III.D.
that case, the assumption of homogeneity must be rejected. Investigators may then try stratifying studies into groups to achieve homogeneity within strata, and then analyzing strata separately. One of the most important goals of meta-analysis is to identify important sources of heterogeneity between studies. Indeed, an FDA guidance document requires this.

Nissen and Wolski reported that they tested for statistical heterogeneity using a version of the chi-squared test described above known as Cochran’s Q-test. They reported “a p-value of more than . . . 0.10 for the Q statistic, indicating a lack of heterogeneity across trials.” The Q-test is a standard method for testing homogeneity across multiple studies. However, critics objected that the Q-test had “limited ability to detect variation across studies with sparse data,” and argued “that the decision to pool all studies despite design and population heterogeneity probably led to artificial inflation and precision of the risk estimate.” They did not, however, make any calculations to support that supposition.

We did make calculations and reach a different conclusion. The chi-squared method based on sample LORs would have to use the Haldane bias correction of 0.5 to deal with zero counts. If that is done the chi-squared statistic would be 17.46 on 41 degrees of freedom and the probability of such a small chi-squared would be 0.0005. Such a result would cast doubt on the independence of the studies. However, the inverse-variance method is undesirable for the reasons stated. An exact test of homogeneity is possible because of Zelen, but it is extremely computationally intensive. A simple alternative method is as follows: (1) compute a common odds ratio by the cmle or M-H methods, (2) for each table take the squared difference between the observed value of the reference cell and its expected value given the margins of the table under the non-null distribution determined by the estimated common odds ratio, (3) compute the variance of the reference cell under the same conditions. Under the hypothesis that there is a common odds ratio, the squared difference in (2) will vary around the variance in (3) and in expectation be equal to it. Thus, (2) minus (3) will vary randomly around zero with pluses and minuses. If the odds ratios vary too much, (2) will be larger than (3) in expectation and so positive values for the differences will tend to predominate. If there are too

52. Technical note: the homogeneity statistic has an approximate chi-squared distribution with \( g - 1 \) df under the assumption of constant association, where \( g \) is the number of studies, each with sample sizes that are not too small. The choice of association measure plays a role; studies that are homogeneous with respect to, for example, the odds ratio, may appear heterogeneous with another measure, for example, the risk difference.


54. Nissen & Wolski, supra note 9, at 2459.

55. Diamond et al., supra note 16, at 578.

56. A comparable small probability results if 0.5 is added only to tables with one or two zero cells (\( p = 0.0003 \)).

57. A description of this method can be found in FLEISS ET AL., supra note 39, at 244.
many positive values the difference will be found significant and the assumption of a common odds ratio would have to be rejected. (It should be noted that no approximations or bias corrections are used here.) Applying this method to the 38 conditionally informative tables in the Avandia data showed that 31 of the 38 values were in fact negative. This is an excess number of negative values because the expected number of negative values would be 26; the excess just misses statistical significance ($p = 0.06$). Thus the odds ratios of the tables varied somewhat less around the common odds ratio than one would expect under the assumption of independent studies with a common underlying odds ratio. It seems clear that Nissen and Wolski’s failure to find heterogeneity in the component studies was not because of the low power of the chi-squared test.

**D. Fixed-Effect vs. Random-Effects Models**

Up to now we have assumed that the true value of the parameter underlying each study was substantially the same, and our effort was to estimate that common odds ratio. But, in many cases there may be variation in the true odds ratios among studies and variability in the sample odds ratios beyond what can be accounted for by sampling error within each study. In such cases interest shifts from testing the hypothesis of homogeneity as a preliminary step in the analysis, and towards specifying a model that embraces heterogeneity as a part of the uncertainty in drawing inferences. Random-effects models do this, and a random-effects meta-analysis is one that explicitly permits variation in the true parameter from study to study, and seeks to make inferences about some central feature of the distribution of the true parameters in the population, either real or hypothetical, of all such studies.

When the random-effects model is used, the cmle and M-H methods no longer provide valid standard errors for the estimated population mean parameter for the odds ratio, and even the estimated mean itself may not be optimally estimated. One must then resort to the inverse-variance weighted average method, or other, more complicated, statistical models, to estimate the population mean LOR. If the inverse-variance method is used, however, there is the problem we have noted of residual bias, which is a significant disadvan-

---

58. The method used to calculate the $p$-value was as follows. First, calculate the exact conditional probability of a negative value for the difference of terms (i) – (ii) for each table, given the margins and the conditional maximum likelihood estimate of the OR, 1.426, from the noncentral hypergeometric distribution. Then, calculate the exact distribution of the sum of 38 independent but nonidentically distributed binary outcomes (negative yes/no). The $p$-value is the one-tailed probability of at least 31 negative values in this distribution.

59. In the moot court conducted in our seminar, *Statistics for Lawyers*, at Columbia Law School in 2010, to which reference has been made, the 42 studies were the subject of testimony by Professors Bagiella and Vaughan from the Mailman School of Public Health, who acted as experts for the parties. The jury of six who heard the case was sufficiently put off by the problem of excessive homogeneity discussed by the expert critiquing the meta-analysis (who used the chi-squared method after 0.5 was added to the cells of the tables) that it voted 4-2 for the proposition that the data were not sufficient to find that Avandia increased the incidence of MIs.
With any random-effects method, though, the standard error of the final estimate (of the population mean parameter) will be larger than the standard error of the final estimate (of the assumed common parameter) in the fixed-effect model. This is because of the acknowledged presence of interstudy heterogeneity. Thus, the variance of the sample LOR in each study is the sum of the variance for that study reflecting the uncertainty of the estimated LOR for that study plus the variance across studies. This result reflects the variation of each study’s true LOR around the population mean LOR. Note that the latter is a constant across studies. This means that the confidence interval associated with a random-effects model will be wider than for a fixed-effect model, which will result in some loss of statistical power. In addition, because the interstudy variance component is the same for each study, this tends to even out the weight given to studies and hence larger studies will be less dominant in the weighting than in the case of the fixed-effect model.

The inverse-variance weighted average is not the only method available for estimating the population mean LOR. An attractive alternative for random effects models, given the assumed randomness in true odds ratios, is to employ an explicit prior distribution for the LOR with an unknown population mean and variance. One can then use ordinary maximum likelihood estimation for these two population parameters based on the marginal conditional distribution of the cell frequencies given the observed margins, where the random LORs have been integrated out using the prior distribution. A common choice of prior distribution for the LOR is the normal distribution with population mean \( M \) and variance \( D^2 \). Inference about the mean \( M \) proceeds in the usual way using the marginal maximum likelihood estimate of the mean \( M \) and its standard error, which in turn depends on the number of studies, the study sample sizes, and the estimated dispersion \( D^2 \). The same qualitative behavior obtains as described above, but no corrections for zero cells are required.

Statisticians differ in their view of the choice between fixed-effect and random-effects models. The FDA argues, in substance, that if there is substantial variation among studies beyond what would be accounted for by sampling error within studies assuming a common odds ratio, there is something causing that variability that should be explained. In that case, we believe, studies should be stratified into more homogeneous strata and the strata analyzed separately. This position is most compelling if the individual studies point in different directions as, for example, some showing a strong positive and others

60. Of the two empirical bias corrections that Sweeting reported in his study, one generally performed slightly better than the constant 0.5 bias correction while the other performed considerably worse; Sweeting et al., supra note 41, at tbl.X.


62. FDA, GUIDANCE FOR INDUSTRY, supra note 53, at 6.
Meta-Analysis of “Sparse” Data

... a strong negative effect. The average effect might be near zero, but reporting no effect would be quite misleading. On the other hand, when most or all studies point in the same direction, reporting an average effect would seem more acceptable. Some statisticians who favor random-effects models argue that studies will usually differ in various factors, such as dose, type of population, duration, end point, and so forth, and that these quite legitimately will produce variations in outcome, which should be reflected in the model assumed for the data. In effect, critics of Nissen and Wolski made this point without suggesting that they should have used a random-effects model as a solution.

Statisticians also argue that if differences among studies are not great, the random-effects model will yield the same estimates for the parameters as the fixed-effect model. That was the case for the Avandia study, where Nissen and Wolski used a fixed-effect model but the results with a random-effects model would have been the same. However, because the inverse-variance method was biased downward, its estimate, as we have seen, was lower and nonsignificant. This evident loss of power is analogous to the trade-off required in choosing a random-effects model.

The choice between fixed-effect and random-effects models may also be influenced by the intended uses of meta-analytic results. Sir Richard Peto has argued that meta-analytic estimates should be applied only to the particular studies that are its components. In this view the difference in outcomes attributable to factors of dose, population studied, endpoint, and so forth, are not random elements that need to be taken into account; thus a fixed-effect model that makes no allowance for interstudy variability because of such factors is appropriate. This is justified when the question at issue is the importance of the particular studies included in the meta-analysis. For example, if a publicly held drug company has in its possession a group of studies, not yet made public, the issue may be whether those studies indicate some statistically significant adverse effect of the subject drug such that the company would have to make the information public to comply with the federal securities laws. In such a case, a meta-analysis using a fixed-effect model would be appropriate. But in the more common situation, as when an expert is testifying that a drug causes an adverse effect and bases that conclusion on a meta-analysis, he or she is generalizing from the particular studies synthesized in the analysis. This would require a broader perspective than the restrictive Peto view and thus a random-effects model when statistical heterogeneity is in the picture. The penalty may be a loss of statistical power.

64. See Richard Peto, Why Do We Need Systematic Overviews of Randomized Trials? 6 STAT. MED. 233, 241–242 (1987) (providing a discussion between Peto and some of his colleagues).
E. Study-Level vs. Patient-Level Data

Nissen and Wolski were also criticized for using study-level data instead of patient-level data.\(^66\) They had to use study-level data because GSK had refused to make patient-level data available.\(^67\) Unquestionably, in most cases, patient-level data are superior to study-level data. For example, different studies may have different durations and this may severely bias the results unless some account is taken of this factor. A study that followed diabetics taking Avandia vs. a comparator for 24 weeks and one that followed them for 5 years might produce very different odds ratio estimates and combining them could be quite misleading. In one case, an expert for Pfizer was criticized by plaintiffs for a meta-analysis that purported to show that Celebrex and Bextra did not increase CV risks because he did not take the duration of the studies in his meta-analysis into account.\(^68\) On the other hand, when patient-level data are available, the analyst is able to use the time to each event (for example, an MI) and to compute a hazard ratio, that is, the instantaneous risk ratio at any point of time for treatment vs. comparator. For example, in a study posted in 2010, investigators used patient-level data for Medicare beneficiaries to estimate an adjusted hazard ratio for acute MI of 1.06 for rosiglitazone vs. pioglitazone (ci 0.96 to 1.18).\(^69\) This means that the risk of acute MI at any given time is increased by about 6% by taking rosiglitazone vs. pioglitazone, but the increase was not statistically significant.\(^70\)

F. Assembling the Component Studies

The process of assembling component studies to include in the meta-analysis is a huge and vital first step in the process. We include the subject here, rather than at the outset, because it was not on the front burner for Nissen and Wolski’s critics.

\(^66\) See Diamond et al., supra note 16, at 578.
\(^67\) In fact, GSK refused to make a number of its unpublished studies available to the public until it settled a lawsuit with the State of New York for failure to disclose unfavorable studies of its antidepressant drug, paroxetine. In the settlement it agreed to establish a public “clinical trial register” of all its clinical trials conducted after 2000. Nissen & Wolski, supra note 9, at 2458; Barbara Martinez, Glaxo Settles New York Suit Over Unpublished Clinical Data; Drug Maker Agrees to Make A Payment of $2.5 Million, Create Registry of All Trials, WALL ST. J., Aug. 27, 2004, at B3.
\(^68\) In re Pfizer Inc. Sec. Litig., Nos. 04 Civ. 9866 & 05 Md 1688, 2010 WL 1047618, at *7 (S.D.N.Y. Mar. 22, 2010) (testimony of Lee-Jen Wei attacked by plaintiffs). After a Rule 702 hearing, however, the court denied plaintiffs’ motion to exclude Professor Wei’s meta-analysis of study-level data, which took no account of duration, as unreliable. Professor Wei did not show that the results would be changed if duration was taken into account. Id.
\(^69\) David J. Graham et al., Risk of Acute Myocardial Infarction, Stroke, Heart Failure and Death in Elderly Medicare Patients Treated with Rosiglitazone or Pioglitazone, 304 JAMA 411, 415 (2010).
\(^70\) Id. tbl4. Another study using patient-level data by GSK investigators and others found an elevated hazard ratio of 1.30 for myocardial ischemia associated with rosiglitazone, (95% ci 1.004 to 1.69). Cobitz et al., supra note 8, at 769, 774.
We have listed protocols for meta-analysis as follows: (1) creating a pre-established research plan for including and excluding studies, which specifies criteria for the range of patients, range of diagnoses, range of study designs, and range of treatments; (2) making a thorough literature search, including an effort to find unpublished studies; (3) assembling a list of included and excluded studies, in the latter case with the reasons for their exclusion; (4) calculating a point estimate of effect, its p-value, and a confidence interval for each study (to the extent possible); (5) testing whether the studies are homogeneous, and if not making separate analyses of subgroups that are homogeneous; (6) if the studies are homogeneous, calculating a summary statistic for all of them, together with a p-value and a confidence interval for the statistic; (7) if the result is not statistically significant, calculating statistical power curves for the result against a range of alternative hypotheses; (8) calculating the robustness of the result, namely, how many negative studies would have to exist for the effect to be neutralized; (9) making a sensitivity analysis, that is, eliminating a study or studies that appear to have serious design flaws to measure the effect on the results; and (10) assessing the potential impact of measurement errors and omitted variables.\textsuperscript{71}

When the meta-analytic result is—or may be interpreted as—a narrow one relating only to the component studies, the function of the first three requirements changes to a degree. Instead of insuring the reliability of generalized inferences from the component studies, their function is to ensure the relevance of findings based on those studies. Thus, if a biased group of studies is assembled, the meta-analytic result may fairly represent them; however, the relevance of that evidence in and of itself or as a foundation for expert opinion addressing a larger issue would be in doubt. In that event, the evidence would amount to little more than a demonstration that at least one subgroup of studies can be found that shows the effect desired by the proponent. This is weak evidence unless the sample is so large that an implausible number of negative studies would have to exist to neutralize the result (the showing called for by protocol 8).

In the Avandia case, the 42 trials analyzed by Nissen and Wolski consisted of: 5 trials submitted by GSK to the FDA in 1999 in connection with proceedings that led to FDA approval of the drug; 35 studies in GSK’s clinical trial registry, of which 9 had been published and 26 remained unpublished; and 2 large then recently published trials, namely the DREAM and ADOPT trials.\textsuperscript{72} In the abstract for the article, the authors stated that they searched published literature, the FDA website, and GSK’s trial registry and identified 116 potentially relevant studies, of which 42 met the inclusion criteria. In the article, however, no details of their literature search are given and in particular

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{71} SFL, supra note 17, at 250. We note that literature searches have been facilitated to some degree by the Food and Drug Administration Amendments Act of 2007, Pub. L. No. 110–85, 121 Stat. 824 (codified as amended in scattered sections of 21 U.S.C.), which requires registration of trials of product subject to FDA regulation initiated after September 27, 2007.
\item \textsuperscript{72} Nissen & Wolski, supra note 9, at 2458.
\end{itemize}
\end{footnotesize}
there is no indication that it had included an effort to find unpublished studies (apart from those in GSK’s clinical trials registry).

In short, our first three requirements for generalizing were met only in part and requirements 8, 9 and 10 were not addressed at all.

On the other hand, this is a case where one might say that the value of the meta-analysis lies in its summary of these particular studies, even if it could not be shown that they were an unbiased sample of all similar studies. Because the included trials largely emanated from the company itself, they would not be expected, as a group, to be biased against the drug; therefore, the meta-analytic finding, even if viewed as strictly applicable only to the included studies, could nevertheless constitute important evidence. This might justify an expert in using it as a basis for more generalized testimony about causation. But the more conservative position is to confine expert opinion to an explication of the meta-analytic results for the particular studies, without permitting an opinion that reaches beyond those studies, unless there was another basis in expertise for a broader opinion.

IV. META-ANALYSIS, THE FDA, AND THE COURTS

Meta-analytic accuracy is said to be modest. A leading study of the question, by LeLorier and his coauthors, identified 12 large RCTs of treatments (the gold standard) and 19 previously published meta-analyses addressing the same questions. Because the study included primary and secondary endpoints, there were a total of 40 comparisons. They concluded that the outcomes of the 12 large trials were not predicted accurately 35% of the time by the meta-analyses and rated the agreement between the meta-analyses and the large trials as only “fair.” More specifically, the positive predictive value of the meta-analyses was 68% and the negative predictive value was 67%. Positive predictive value is probably the more important statistic. The 68% figure implies that in almost one-third of the cases in which meta-analysis found that

73. Publication bias is well recognized and its effect has been quantified. For example, a review of trials of antidepressants found that meta-analysis of only the published trials gave effect estimates 32% larger on average than when all trials sent to the FDA were analyzed. Erick H. Turner et al., Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy, 358 NEW ENG. J. MED. 252, 252 (2008). Failure to search non-English-language reports does not appear to create serious bias. See, e.g., David Moher et al., What Contributions Do Languages Other Than English Make on the Results of Meta-Analyses?, 53 J. CLINICAL EPIDEMIOLOGY 964, 964, 969 (2000).


75. Id.

76. Id. at 536 (kappa = 0.35). An earlier study, cited by the authors, reached a similarly unimpressive result. See generally J. Villar et al., Predictive Ability of Meta-Analyses of Randomized Controlled Trials, 345 LANCET 772, 775 (1995) (finding a positive predictive value of 50% to 67%).

77. LeLorier et al., supra note 74, at 541.
the treatment was beneficial and that the improvement was statistically significant, the subsequent large trial did not find that the improvement was statistically significant. There was no case in which the meta-analysis found a statistically significant positive result and the subsequent large trial found a statistically significant negative result.Treating the large trials as the gold standard, the meta-analytic results in such cases would be viewed as errors, or at least as overestimates of the evidence. However that leaves open the question whether the RCTs are accurate and adequately powered.

In comparing results from different studies there are many dimensions along which one may do so. For example, reasonable comparisons might be made (1) between nonrandomized epidemiological studies and randomized studies, (2) between small and large studies, (3) between different patient populations with different risk factors, (4) between studies included in a meta-analysis and subsequent studies, and (5) between different analytic methods for synthesizing the results for a given set of component studies. In considering what to do about Avandia, the FDA made several such comparisons.

We have noted that a number of criticisms of the Nissen and Wolski study were not well founded, but some contemporaneous and subsequent large-scale studies have raised questions over their finding. The two largest studies included in the meta-analysis, the ADOPT trial and the DREAM trial, showed elevated odds ratios for MI, but this was not significant in either case, even when combined. In addition, a subsequent large randomized study, the RECORD trial, found only a modestly elevated hazard rate for MI (hazard ratio = 1.14), which was not statistically significant (95% c.i. 0.80 to 1.63).

In a signed editorial commenting on this study in the same issue of the journal, a prominent critic wrote:

In my own review of selected meta-analyses, problems were so frequent and so serious, including bias on the part of the meta-analyst, that it was difficult to trust the overall ‘best estimates’ that the method often produces . . . [A]ny attempt to reduce the results [of disparate studies] to a single value, with confidence bounds, is likely to lead to conclusions that are wrong, perhaps seriously so. I still prefer conventional narrative reviews of the literature, a type of summary familiar to readers of the countless review articles on important medical issues.


79. These were the ADOPT (A Diabetes Outcome Progression Trial) and the DREAM (Diabetes Reduction Assessment with Ramipril and Rosiglitazone Medication) trials. See H.C. Gerstein et al., DREAM Trial Investigators, Effect of Rosiglitazone on the Frequency of Diabetes in Patients with Impaired Glucose Tolerance or Impaired Fasting Glucose: A Randomised Controlled Trial, 368 LANCET 1096, 1099 (2006); S.E. Kahn et al., Glycemic Durability of Rosiglitazone, Metformin, or Glyburide Monotherapy, 355 NEW. ENGL. J. MED. 2427, 2427 (2006). The data for those trials appear in the Appendix.

80. Philip D. Home et al., Rosiglitazone Evaluated for Cardiovascular Outcomes in Oral Agent Combination Therapy for Type 2 Diabetes (RECORD): A Multi-Centre, Randomised, Open-Label Trial, 373 LANCET 2125, 2129 (2009). The odds ratio was very similar to the hazard ratio—1.15 (ci 0.80 to 1.66). Id. Although the measures are not identical in general, in this case they are quite close and much smaller than the 43% increase in the odds found by Nissen & Wolski.
When the Nissen and Wolski study appeared, the FDA issued a safety alert, but did nothing more. Its statement explained that safety data from controlled clinical trials [evidently the Nissen and Wolski study] have shown that there is a potentially significant increase in the risk of heart attack . . . in patients taking Avandia. However, other published and unpublished data from long-term clinical trials of Avandia, including an interim analysis of data from the RECORD trial . . . and unpublished reanalyses of data from DREAM provide contradictory evidence about the risks in patients treated with Avandia . . . For these reasons, FDA is not asking GlaxoSmithKline, the drug’s sponsor, to take any specific action at this time.

The FDA’s position was later cited and accepted by the courts in dismissing lawsuits brought by investors against GSK for nondisclosure of its meta-analyses; if the data were insufficient to establish excess CV risks there was nothing to disclose.

By 2010 further inconsistencies had appeared. On the meta-analytic side, the FDA updated its 2007 meta-analysis with added studies and reported an odds ratio of 1.80 that was statistically significant for 52 component trials. This was the strongest meta-analytic finding to date. However, a large observational retrospective study of 227,571 Medicare patients aged 65 years or older, which was released in 2010, compared patients taking rosiglitazone with those taking pioglitazone, which is another thiazolidinedione that earlier studies indicated might have beneficial CV effects. The study found that the adjusted hazard ratio for MI was 1.06 and was not significant (ci 0.96 to 1.18). Here was yet another possible discrepancy between the meta-analytic findings and subsequent large-scale studies.

Thus, the FDA had to confront an apparent contradiction between meta-analytic studies, which appeared to indicate an elevated risk (albeit with controversy) and concurrent and subsequent large-scale studies that, considered separately, did not show a statistically significant increase. Should the FDA act on the former or the latter? The answer is: the FDA did both. It disclaimed reliance on the meta-analyses because of the conflicting epidemiologic and large-scale RCTs, but restricted marketing of the drug as if the risks meta-analyses had identified were probably real.

82. Id.
83. E.g., Avon Pension Fund v. GlaxoSmithKline PLC, 343 F. App’x 671, 674 (2d Cir. 2009); but see Matrixx Initiatives, Inc. v. Siracusano, 131 S. Ct. 1309, 1321 (2011) (holding that a determination of statistical significance related to adverse event reports is not required for a finding of materiality under the federal securities laws).
84. Woodcock Memorandum, supra note 15, at 6.
85. Graham et al., supra note 69, at 415. The study did find statistically significant heightened hazard ratios for stroke (1.27); for heart failure (1.25); for death (1.14); and for the composite of MI, stroke, heart failure, or death (1.18). Id.
86. Id.
tions, Janet Woodcock, Director of the FDA’s Center for Drug Evaluation and Research, noted the earlier group of related meta-analyses, and recited the findings of the FDA 2010 meta-analysis that found a “nominally significant” OR of 1.8 (c.i. 1.03 to 3.25) for nonfatal MI. She concluded that “the recent meta-analytic findings (which incorporate trials included in the prior analyses and add more) support the original concern that rosiglitazone increases the risk of heart attacks . . . .”

But she disparaged that support, writing that “It has been shown repeatedly that hypotheses generated by meta-analyses may not be verified when studied in randomized controlled trials [citing two examples] . . . . Other available evidence must be used to evaluate whether or not rosiglitazone increases the risk of heart attacks compared to placebo or standard diabetes drugs.”

She then turned to the other evidence.

Epidemiologic studies did not report any consistent increase or decrease in CV events for rosiglitazone. These results, she found, “diminish the likelihood that the meta-analytic results are correct, but do not have enough weight to dismiss them (because of the well-known limitations of epidemiologic studies).”

Observational studies compared results for patients on rosiglitazone or pioglitazone, and many of the studies found a “numerical advantage” for pioglitazone. However, a majority showed no statistically significant difference in MI between the drugs.

In her discussion of RCTs, Dr. Woodcock began by noting that “[t]he most reliable evidence about clinical outcomes comes from well-conducted randomized trials.” The DREAM and ADOPT trials had higher rates of CV events in the rosiglitazone arms, but these were not statistically significant. Neither trial had substantial numbers of CV events, which “limit[ed] how much information they contribute to the question at hand.” The RECORD trial results also showed no significant increase in MI, but its results were thrown into doubt because an FDA medical officer, who reviewed the study, found that at least 12 patients taking Avandia had suffered serious heart problems that were excluded by GSK from the tabulation. The medical officer reportedly called these “unpardonable errors” that seriously biased the trial’s conclusions.

88. Id. (observing, however, that the OR of 1.8 found by the FDA did not establish that rosiglitazone caused an 80% increase in MI compared to standard therapy because the results were based on very few events (65), were collected from trials that were not designed or executed to assess CV risk, and the confidence interval was very wide).


90. Id. at 7.

91. Id.

92. Id.

93. Id.

Dr. Woodcock concluded that the evidence pointing to a CV ischemic risk with rosiglitazone was neither “robust [n]or consistent,” but that there were “multiple signals of concern, from varied sources of data, without reliable evidence that refutes them.” According to Woodcock, the FDA permitted the drug to remain on the market only if GSK received documentation from health care providers that each patient receiving rosiglitazone was either currently receiving the drug or was unable to achieve glycemic control with other medications and who were given complete risk information. Because other medications were usually available for glycemic control, this effectively shut down sales to new patients in the United States. The FDA restricted access rather than removed rosiglitazone from the market in part because the CV safety profile was “still an open question because there are conflicting data on the existence and magnitude of the risk, and a detailed re-adjudication and analysis of data from the RECORD study needs to be conducted.”

Nevertheless, on May 18, 2011, the FDA announced new restrictions to the prescribing and use of rosiglitazone-containing medicines: health-care providers and patients must enroll in a special program to prescribe and receive such medicines. In short, the FDA gave meta-analyses significant weight even in the face of large scale RCTs that appeared to conflict with them. In an unusual coordinated announcement, the European Medical Agency went further and suspended sales entirely in Europe.

The same issue has arisen in the Avandia cases. At least one court has already confronted the question of whether meta-analysis can be a basis for reliable expert testimony when there are possibly conflicting large-scale trials. In a three-day Daubert hearing in a multidistrict litigation consolidated for pretrial in the Eastern District of Pennsylvania, one of the principal issues before Judge Rufe was whether the challenged experts could validly rely on the statistically significant findings in the meta-analyses in the face of the nonsignificant findings in the RCTs. Plaintiffs’ experts, in finding that Avandia caused MIs, relied on the meta-analyses and rejected the conflicting RCTs. Discussing the RECORD trial at length, one of plaintiffs’ experts, Dr. Eliot Brinton, criticized it because (1) of a “lack of specificity in the endpoints of interest (for example, categorizing all deaths of unknown cause as cardiovascular deaths . . .)”; (2) “prescribing doctors were permitted to measure patients’ lipid profiles and even encouraged to prescribe statins” with the

96. Id. In the same decision, GSK was required to commission an independent re-adjudication of the RECORD study. Id. at 1.
98. Harris, supra note 94.
100. Id.
result that “[s]tatın use increased 9% more in the Avandia arm than [in] the control arm” (the trial was open label); (3) “statin use (in both the treatment and control groups) also led to a lower-than-expected rate of cardiovascular disease overall in the study,” which reduced its statistical power; and (4) statistical power was also reduced “by the fact that a large percentage of study participants (in both arms) dropped out of the study.” After a lengthy discussion of Dr. Brinton’s testimony on this and other subjects, the court held that he had adequately explained why he found conflicting research “flawed and not compelling,” and concluded that “his approach to the data was scientifically reliable.”

Although Dr. Brinton apparently did not make any calculation to support his point, our calculation indicates that he was correct that the RECORD trial was underpowered against certain alternatives. First, the confidence interval for the hazard rate ratio for MI was wide (0.80, 1.63). Second, given that there were a total of 120 MIs in both groups combined, 64 in the rosiglitazone group and 56 in the control group, we calculated that the lowest detectable hazard rate ratio, with the usual 80% power at a two-tailed 0.05 level of significance, is 1.67. For a hazard rate ratio of 1.4, the value initially found by Nissen and Wolski for their odds ratio, we calculated that power is only 45%, which means that if the hazard rate ratio associated with rosiglitazone was in fact 1.4, there is only a 45% chance that the RECORD study would have found a statistically significant increased rate. Power is even lower against Nissen and Wolski’s finding of a 1.28 OR in their updated 2010 study. Thus the failure of the RECORD trial to find a statistically significantly increased hazard rate of MIs associated with rosiglitazone is not evidence that conflicts with the Nissen and Wolski finding; the RECORD trial simply did not observe enough MIs to rule out 40% or 50% elevations in risk.

Our analysis of the Avandia data in the Nissen and Wolski study shows that the objection to the Peto, or one-step, method was correct in theory, but inconsequential in fact because the same results were obtained when a multiple-step method was used. The supposition that the studies were too clini-

101. Id. at *8.
102. Id. at *9.
103. Id. at *10–12.
104. In our calculations we ignore the difference between hazard rate ratios and odds ratios, which is unlikely to be material in this application. We note that power is sufficient against the FDA’s finding in its 2010 study of an OR of 1.8. For cases discussing power in other contexts, see FED. JUDICIAL. CTR., REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 582 (3d ed. 2010) (“Thus, in Smith v. Wyeth-Ayerst Labs. Co., 278 F. Supp. 2d 684, 693 (W.D.N.C. 2003) and Cooley v. Lincoln Electric Co., 693 F. Supp. 2d 767, 773 (N.D. Ohio 2010), the courts recognized that the power of a study was critical to assessing whether the failure of the study to find a statistically significant association was exonerative of the agent or inconclusive.”).
cally heterogeneous to be combined was unsupported by calculation and, in fact, was not true; tests for homogeneity were far from statistically significant and there was some weak evidence that the event frequencies varied less about their expected values than would be permitted even by a constant OR model with independent studies. Several subsequent large-scale studies reached smaller and nonsignificant MI effects for Avandia, but these were not sufficiently powered to generate significant results for the elevated risks found by Nissen and Wolski. We have also seen that when the data are sparse, for example, when the expected value in any cell of a component study is less than five, the inverse-variance method suffers from a nontrivial residual bias that remains even after 0.5 is added to the cells of tables with zero cells. When the cmle or M-H methods are used, making such additions for tables with zero cells, as some authors have suggested and computer programs do automatically, is unnecessary, tends to inappropriately bias the principal result downward, and in close cases will mistakenly make it nonsignificant.

In the Daubert hearings that have already begun in the Avandia cases, the courts are confronting critiques of analyses and trials that become important when meta-analyses and large scale RCTs seem to reach inconsistent conclusions from the data. Although the usual mantra is that RCTs are the gold standard by which accuracy is judged, because no study is perfect, it is evident that in the future decision makers in the FDA and the judges in the adverse-event drug cases must render decisions that are more searching and nuanced than a simple rendering of preeminence to the RCTs when they seem to conflict with the meta-analyses. And, as the saga of Avandia demonstrates, those meta-analyses can be potent: In 2010 alone GSK took charges totaling $6.76 billion against earnings in dealing with the Avandia cases.
APPENDIX

We illustrate the residual bias of Haldane’s estimator of the odds ratio in the case of rare events. Suppose we observe two independent binomial samples with samples of size \( N = 1000 \) each. In the table below we consider various values of the probability of an event in a single subject in the control group, which we denote by \( P_0 \), with corresponding expected numbers of events, \( NP_0 \), as this number gets small. In each row of the table, we also determined that value of the exposed group probability, denoted by \( P_1 \), which yields an odds ratio of exactly \( OR=2 \) compared to \( P_0 \). We then calculated the exact bias of the Haldane estimator of the LOR as

\[
Bias = E \left\{ \ln \left( \frac{X + 0.5}{N - X + 0.5} \right) - \ln \left( \frac{Y + 0.5}{N - Y + 0.5} \right) \right\} - \ln 2
\]

\[
= \sum_{x=0}^{N} \ln \left( \frac{x + 0.5}{N - x + 0.5} \right) p(x) (1 - p) p(1 - p) - \sum_{y=0}^{N} \ln \left( \frac{y + 0.5}{N - y + 0.5} \right) p(y) (1 - p) p(1 - p) - \ln 2
\]

where \( X \) has a binomial distribution with sample size \( N = 1000 \) and probability parameter \( P_1 \) and \( Y \) has a binomial distribution with the same sample size and parameter \( P_0 \). As can be seen from the table, the Haldane estimate has negligible bias when the expected number of events in the control group is at least five (\( NP_0 \geq 5 \)), but becomes large for smaller expected frequencies, approaching a 50% underestimation as the expected number of events approaches zero.

<table>
<thead>
<tr>
<th>( P_0 )</th>
<th>( NP_0 ) (expected number of events per thousand)</th>
<th>( P_1 )</th>
<th>Expected value of Haldane estimate of LOR (true value is 0.69315)</th>
<th>Antilog of Haldane estimate of LOR(true OR = 2)</th>
<th>Percent underestimation error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>100</td>
<td>0.1818</td>
<td>0.69315</td>
<td>2.000006</td>
<td>-0.00%</td>
</tr>
<tr>
<td>0.01</td>
<td>10</td>
<td>0.0198</td>
<td>0.69356</td>
<td>2.000835</td>
<td>-0.04%</td>
</tr>
<tr>
<td>0.005</td>
<td>5</td>
<td>0.00995</td>
<td>0.69473</td>
<td>2.003165</td>
<td>-0.16%</td>
</tr>
<tr>
<td>0.001</td>
<td>1</td>
<td>0.001998</td>
<td>0.55041</td>
<td>1.733960</td>
<td>13.30%</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.1</td>
<td>0.00019998</td>
<td>0.10159</td>
<td>1.109635</td>
<td>44.65%</td>
</tr>
<tr>
<td>0.00001</td>
<td>0.01</td>
<td>0.0000199998</td>
<td>0.01091</td>
<td>1.010968</td>
<td>49.45%</td>
</tr>
</tbody>
</table>

We next asked if the difference in the estimate of the common odds ratio in the Avandia cases using the inverse-variance weighted average method with the Haldane one-half correction (estimated OR = 1.223) could have arisen because of the rare-events residual bias. For this hypothetical calculation, we assumed the following:

WINTER 2012 151
(a) For any study with at least one MI event in the control group, we set $P_0$ equal to the control group’s observed sample proportion.

(b) For any study with no MI events in the control group, we set $P_0 = 0.0007$. (See below for an explanation of this choice.)

(c) In either case, for the Avandia group, we set $P_1$ equal to that value which would yield a constant odds ratio of 1.426, the \textit{cmle} OR for the 42 Avandia cases.

We then calculated the exact bias of the Haldane estimate of the LOR, similarly to the illustration above, but using the actual sample sizes for each study instead of assumed equal sample sizes. After that, we calculated the weighted average of these biases using the same weights that the inverse-variance method would use to calculate the summary estimate of the LOR.

The result was that the weighted average of the Haldane biases was $-0.1538$. Exponentiating gives us the reduction factor due to bias, $\exp(-0.1538) = 0.8574$, and multiplying by 1.426 gives us the “predicted” estimate: $1.426 \times 0.8574 = 1.223$. This is just what was observed in the Avandia data using the inverse-variance weighted average of log odds ratios with Haldane’s correction, $1.223 = \exp(0.2009)$.

The value of $P_0 = 0.0007$ for tables with zero observed MIs in the control arm was arbitrarily chosen, but it appears to be reasonable. The observed pooled control group probability was $72/(72+12205) = 0.00586$. The weighted average of the selected control arm values of $P_0$ (weighted by the control arm sample sizes) is of course larger than 0.00586 (because we are replacing the zeros with something positive), but not much larger: 0.00610. On the other hand, a simple, unweighted average of the selected control arm values of $P_0$ is slightly less: 0.00478. It does not seem unreasonable therefore to assume such values for the true control arm values $P_0$.

The study-specific biases appear in the table below. The largest biases occur where the $P_0$ was assumed to be 0.0007, but there are other moderately large biases where the observed value of $P_0$ was used.
### Meta-Analysis of “Sparse” Data

<table>
<thead>
<tr>
<th>Study number</th>
<th>Avandia N</th>
<th>Avandia MIs</th>
<th>Control N</th>
<th>Control MIs</th>
<th>Haldane LOR</th>
<th>Inverse-variance weight</th>
<th>Bias in Haldane LOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>357</td>
<td>2</td>
<td>176</td>
<td>0</td>
<td>0.9092</td>
<td>0.4152</td>
<td>–0.8344</td>
</tr>
<tr>
<td>2</td>
<td>391</td>
<td>2</td>
<td>207</td>
<td>1</td>
<td>–0.1237</td>
<td>0.9310</td>
<td>–0.1618</td>
</tr>
<tr>
<td>3</td>
<td>774</td>
<td>1</td>
<td>185</td>
<td>1</td>
<td>–1.4333</td>
<td>0.7462</td>
<td>–0.1687</td>
</tr>
<tr>
<td>4</td>
<td>213</td>
<td>0</td>
<td>109</td>
<td>1</td>
<td>–1.7755</td>
<td>0.3731</td>
<td>–0.1618</td>
</tr>
<tr>
<td>5*</td>
<td>232</td>
<td>1</td>
<td>116</td>
<td>0</td>
<td>0.4119</td>
<td>0.3732</td>
<td>–0.8934</td>
</tr>
<tr>
<td>6</td>
<td>43</td>
<td>0</td>
<td>47</td>
<td>1</td>
<td>–1.0319</td>
<td>0.3688</td>
<td>–0.0712</td>
</tr>
<tr>
<td>7*</td>
<td>121</td>
<td>1</td>
<td>124</td>
<td>0</td>
<td>1.1313</td>
<td>0.3727</td>
<td>–0.2949</td>
</tr>
<tr>
<td>8</td>
<td>110</td>
<td>5</td>
<td>114</td>
<td>2</td>
<td>0.8527</td>
<td>1.6662</td>
<td>–0.0202</td>
</tr>
<tr>
<td>9*</td>
<td>382</td>
<td>1</td>
<td>384</td>
<td>0</td>
<td>1.1064</td>
<td>0.3743</td>
<td>–0.2450</td>
</tr>
<tr>
<td>10*</td>
<td>284</td>
<td>1</td>
<td>135</td>
<td>0</td>
<td>0.3604</td>
<td>0.3735</td>
<td>–0.9083</td>
</tr>
<tr>
<td>11</td>
<td>294</td>
<td>0</td>
<td>302</td>
<td>1</td>
<td>–1.0751</td>
<td>0.3741</td>
<td>–0.0870</td>
</tr>
<tr>
<td>12*</td>
<td>563</td>
<td>2</td>
<td>142</td>
<td>0</td>
<td>0.2382</td>
<td>0.4151</td>
<td>–1.3002</td>
</tr>
<tr>
<td>13</td>
<td>278</td>
<td>2</td>
<td>279</td>
<td>1</td>
<td>0.5180</td>
<td>0.9312</td>
<td>–0.0917</td>
</tr>
<tr>
<td>14*</td>
<td>418</td>
<td>2</td>
<td>212</td>
<td>0</td>
<td>0.9365</td>
<td>0.4154</td>
<td>–0.7773</td>
</tr>
<tr>
<td>15</td>
<td>395</td>
<td>2</td>
<td>198</td>
<td>1</td>
<td>–0.1785</td>
<td>0.9309</td>
<td>–0.1638</td>
</tr>
<tr>
<td>16</td>
<td>203</td>
<td>1</td>
<td>106</td>
<td>1</td>
<td>–0.6520</td>
<td>0.7420</td>
<td>–0.1660</td>
</tr>
<tr>
<td>17</td>
<td>104</td>
<td>1</td>
<td>99</td>
<td>2</td>
<td>–0.5705</td>
<td>0.9203</td>
<td>–0.0227</td>
</tr>
<tr>
<td>18*</td>
<td>212</td>
<td>2</td>
<td>107</td>
<td>0</td>
<td>0.9374</td>
<td>0.4142</td>
<td>–0.8968</td>
</tr>
<tr>
<td>19</td>
<td>138</td>
<td>3</td>
<td>139</td>
<td>1</td>
<td>0.8692</td>
<td>1.0341</td>
<td>–0.0904</td>
</tr>
<tr>
<td>20*</td>
<td>196</td>
<td>0</td>
<td>96</td>
<td>0</td>
<td>–0.7111</td>
<td>0.2490</td>
<td>–0.9340</td>
</tr>
<tr>
<td>21</td>
<td>122</td>
<td>0</td>
<td>120</td>
<td>1</td>
<td>–1.1234</td>
<td>0.3727</td>
<td>–0.0949</td>
</tr>
<tr>
<td>22</td>
<td>175</td>
<td>0</td>
<td>173</td>
<td>1</td>
<td>–1.1159</td>
<td>0.3734</td>
<td>–0.0943</td>
</tr>
<tr>
<td>23*</td>
<td>56</td>
<td>1</td>
<td>58</td>
<td>0</td>
<td>1.1513</td>
<td>0.3701</td>
<td>–0.3034</td>
</tr>
<tr>
<td>24*</td>
<td>39</td>
<td>1</td>
<td>38</td>
<td>0</td>
<td>1.0986</td>
<td>0.3678</td>
<td>–0.3669</td>
</tr>
<tr>
<td>25</td>
<td>561</td>
<td>0</td>
<td>276</td>
<td>2</td>
<td>–2.3251</td>
<td>0.4157</td>
<td>–0.0270</td>
</tr>
<tr>
<td>26</td>
<td>116</td>
<td>2</td>
<td>111</td>
<td>3</td>
<td>–0.3903</td>
<td>1.4211</td>
<td>–0.0036</td>
</tr>
<tr>
<td>27*</td>
<td>148</td>
<td>1</td>
<td>143</td>
<td>0</td>
<td>1.0711</td>
<td>0.3731</td>
<td>–0.3398</td>
</tr>
<tr>
<td>28*</td>
<td>231</td>
<td>1</td>
<td>242</td>
<td>0</td>
<td>1.1494</td>
<td>0.3738</td>
<td>–0.2477</td>
</tr>
<tr>
<td>29*</td>
<td>89</td>
<td>1</td>
<td>88</td>
<td>0</td>
<td>1.0986</td>
<td>0.3718</td>
<td>–0.3571</td>
</tr>
<tr>
<td>30*</td>
<td>168</td>
<td>1</td>
<td>172</td>
<td>0</td>
<td>1.1280</td>
<td>0.3734</td>
<td>–0.2830</td>
</tr>
<tr>
<td>31*†</td>
<td>116</td>
<td>0</td>
<td>61</td>
<td>0</td>
<td>–0.6389</td>
<td>0.2485</td>
<td>–0.9164</td>
</tr>
<tr>
<td>32*†</td>
<td>1172</td>
<td>1</td>
<td>377</td>
<td>0</td>
<td>–0.0339</td>
<td>0.3745</td>
<td>–0.7864</td>
</tr>
<tr>
<td>33*†</td>
<td>706</td>
<td>0</td>
<td>325</td>
<td>0</td>
<td>–0.7750</td>
<td>0.2497</td>
<td>–0.7157</td>
</tr>
<tr>
<td>34</td>
<td>204</td>
<td>1</td>
<td>185</td>
<td>2</td>
<td>–0.6143</td>
<td>0.9285</td>
<td>–0.0244</td>
</tr>
<tr>
<td>35*</td>
<td>288</td>
<td>1</td>
<td>280</td>
<td>0</td>
<td>1.0740</td>
<td>0.3740</td>
<td>–0.2942</td>
</tr>
<tr>
<td>36*</td>
<td>254</td>
<td>1</td>
<td>272</td>
<td>0</td>
<td>1.1709</td>
<td>0.3739</td>
<td>–0.2245</td>
</tr>
<tr>
<td>37*</td>
<td>314</td>
<td>1</td>
<td>154</td>
<td>0</td>
<td>0.3910</td>
<td>0.3736</td>
<td>–0.8630</td>
</tr>
<tr>
<td>38*†</td>
<td>162</td>
<td>0</td>
<td>160</td>
<td>0</td>
<td>–0.0124</td>
<td>0.2492</td>
<td>–0.3162</td>
</tr>
<tr>
<td>39*</td>
<td>442</td>
<td>1</td>
<td>112</td>
<td>0</td>
<td>–0.2686</td>
<td>0.3734</td>
<td>–1.3755</td>
</tr>
<tr>
<td>40*</td>
<td>394</td>
<td>1</td>
<td>124</td>
<td>0</td>
<td>–0.0522</td>
<td>0.3735</td>
<td>–1.2106</td>
</tr>
<tr>
<td>41 (DREAM)</td>
<td>2635</td>
<td>15</td>
<td>2634</td>
<td>9</td>
<td>0.4915</td>
<td>5.8637</td>
<td>0.0004</td>
</tr>
<tr>
<td>42 (ADOPT)</td>
<td>1456</td>
<td>27</td>
<td>2895</td>
<td>41</td>
<td>0.2801</td>
<td>16.258</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th>Total</th>
<th>Total</th>
<th>Total</th>
<th>Weighted Average</th>
<th>Total</th>
<th>Weighted Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>15556</td>
<td>86</td>
<td>12277</td>
<td>72</td>
<td>0.2009</td>
<td>43.274</td>
<td>–0.1538</td>
</tr>
</tbody>
</table>

* = Observed control group event frequency = 0; calculation assumed \( p_0 = 0.0007 \).
† = Observed Avandia and control group event frequencies = 0.
Inverse-variance weight = reciprocal of sum of reciprocals of observed cell frequencies to which \( \frac{1}{2} \) has been added to each cell.

WINTER 2012