Meta-analysis of incidence of rare events

Peter W Lane

Stat Methods Med Res 2013 22: 117 originally published online 4 January 2012
DOI: 10.1177/0962280211432218

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
22(2) 117–132

DOI: 10.1177/0962280211432218
smm.sagepub.com
SAGE

Meta-analysis of incidence of rare events

Peter W Lane

Abstract

This is a review of methods for the meta-analysis of incidence of rare events using summary-level data. It is motivated and illustrated by the dataset used in a published analysis of cardiovascular safety in rosiglitazone trials. This review compares available methods for binary data, considering risk-difference, relative-risk and odds-ratio scales, fixed-effect and random-effects models, and frequentist and Bayesian approaches. Particular issues in this dataset include low incidence rates, the occurrence of studies with no events under one or all treatments, and discrepancy among results achieved using different statistical methodologies. The common method of adding a correction factor to handle zeroes may introduce bias where the incidence of events is small, as in this case. Alternative analyses on the log-odds scale are shown to give similar results, but the choice between them is less important than the potential sources of bias in any meta-analysis arising from limitations in the underlying dataset. It is important to present results carefully, including numerical and graphical summaries on the natural scale of risk when the analysis is on a statistically appropriate scale such as log-odds: the incidence rates should accompany an estimated ratio (of odds or risk) to put the analysis into the proper context. Beyond the statistical methodologies which are the focus of this paper, this dataset highlights the importance of understanding the limitations of the data being combined. Because the rosiglitazone dataset contains clinically heterogeneous trials with low event rates that were not designed or intended to assess cardiovascular outcomes, the findings of any meta-analysis of such trials should be considered hypothesis-generating.

Keywords

Adverse event, continuity correction, prediction, rosiglitazone, safety, zero count

1 Introduction

Standard methods for the meta-analysis (MA) of binary data run into problems when proportions are small. This is a particular issue in the MA of adverse events associated with medical treatments, which is becoming more common as observational information grows in large medical databases, and clinical trial results are published on pharmaceutical company websites. Difficulties arise when the analysis is done either at the patient level using individual patient data, or at the study level using just summary counts from each trial. I concentrate here on MA of study-level summaries, which is

Quantitative Sciences, GlaxoSmithKline R&D, Stevenage, UK

Corresponding author:
Peter W Lane, Statistical Consulting Group, Quantitative Sciences, GlaxoSmithKline R&D, Stevenage, STV 2F137, UK.
Email: peterwlane@gmail.com

far more common in the assessment of adverse events, though patient-level analyses are to be preferred when the data are available.

Apart from the division into patient-level and study-level approaches, there are two other main categorisations of MA. Fixed-effect methods summarise the evidence actually provided by the available studies, based on an underlying assumption that the effects of treatment are common to all studies, whereas random-effects methods consider the studies as representative of a population across which the treatment effects vary. Both these approaches share problems when events are rare. The other contrast is between Bayesian and frequentist methods: the first requires prior information on the model parameters, and treats these parameters as random variables, whereas the second avoids the need for, and the possible controversy resulting from, specification of the priors in distributional form, but cannot provide probability statements about the parameters. Recent meta-analyses of the cardiovascular safety of Avandia (rosiglitazone maleate) illustrate these issues, which is explored in this article.

One crucial issue of MA in the area of drug safety, or of drug efficacy assessed with a binary outcome, is the approach to presentation of results. An uninformed belief that analysis should be done on the scale on which results need to be presented for the purpose of interpretation, can lead to analysis being carried out without regard for the appropriateness of the underlying model. So I consider here in detail the reporting of combined estimates on a scale different from that on which analysis was performed, and also introduce some modifications to standard graphical presentations. It must be remembered that, regardless of whether patient- or summary-level data are used, MA of trials not intended to assess the outcome of interest are considered exploratory, not definitive.

2 Example

A study-level analysis by Nissen and Wolski,[1] referred to as N&W from now on, used a dataset of clinical trials to evaluate the potential effect of Avandia on myocardial infarction (MI) and cardiovascular death (CVD). N&W found 48 trials satisfying their stated criteria (Phase II to IV, 24 weeks or longer duration, randomised comparator group, and similar duration of treatment in all groups). Table 1 lists the trials, as given by Tian et al.,[2] (N&W omitted six trials with no events from their list).

Before applying any analytical methods to this collection of data, it is important to consider just what they represent, as this has implications on any interpretation of the results. First, note that the data, and hence the results, are dependent on the searching methods used to find the trials, as there is always potential for publication bias in any systematic review. For example, a follow-up analysis by Friedrich et al.[3] reported finding six more trials with no events on the Register. The impact of this difference would depend upon the chosen statistical method among those to be described.

There are several other major issues to take into account, and some minor ones, which could impact on all of the analytical methods to be described.

(1) Most of the trials were not designed to study cardiovascular problems, and so events such as MI (myocardial infarction) were not prospectively recorded and adjudicated as they would be in studies designed for this purpose. Interpretation of meta-analyses of trials not intended to assess cardiovascular outcomes is inherently limited by the underlying data. Hence, the results of such meta-analyses are considered hypothesis-generating.
(2) The comparator groups in each trial vary widely: some received Placebo, while others received alternative medication such as Metformin (MET), Sulfonylurea (SU) or Insulin.

Table 1. Trials used in meta-analysis of Avandia, including those with no events

	Rosiglitazone			Comparator			Duration
Trial	N	MI	CVD	N	MI	CVD	(weeks)
011	357	2	1	176	0	0	24
020	391	2	0	207	1	0	52
024	774	1	0	185	1	0	26
093	213	0	0	109	1	0	26
094	232	1	1	116	0	0	26
684	43	0	0	47	1	0	52
143	121	1	0	124	0	0	24
211	110	5	3	114	2	2	52
284	382	1	0	384	0	0	24
008	284	1	0	135	0	0	48
264	294	0	2	302	1	1	52
185	563	2	0	142	0	0	32
334	278	2	0	279	1	1	52
347	418	2	0	212	0	0	24
015	395	2	2	198	1	0	24
079	203	1	1	106	1	1	26
080	104	1	0	99	2	0	156
082	212	2	1	107	0	0	26
085	138	3	1	139	1	0	26
095	196	0	1	96	0	0	26
097	122	0	0	120	1	0	156
125	175	0	0	173	1	0	26
127	56	1	0	58	0	0	26
128	39	1	0	38	0	0	28
134	561	0	1	276	2	0	28
135	116	2	2	111	3	1	104
136	148	1	2	143	0	0	26
145	231	1	1	242	0	0	26
147	89	1	0	88	0	0	26
162	168	1	1	172	0	0	26
234	116	0	0	61	0	0	26
330	1172	1	1	377	0	0	52
331	706	0	1	325	0	0	52
137	204	1	0	185	2	1	32
002	288	1	1	280	0	0	24
003	254	1	0	272	0	0	32
007	314	1	0	154	0	0	32
009	162	0	0	160	0	0	24
132	442	1	1	112	0	0	24
193	394	1	1	124	0	0	24
dream	2635	15	12	2634	9	10	156
adopt	1456	27	2	2895	41	5	208
282	70	0	0	75	0	0	24
369	25	0	0	24	0	0	26

(continued)

Table 1. Continued

	Rosiglitazone			Comparator			
Trial	N	MI	CVD	N	MI	CVD	Duration (weeks)
096	232	0	0	115	0	0	26
044	101	0	0	51	0	0	26
325	196	0	0	195	0	0	24
004	676	0	0	225	0	0	24

(3) In some trials, rosiglitazone was tested as an adjunct therapy with MET, SU or Insulin, whereas in others it was a monotherapy.
(4) The populations also varied widely: most trials treated patients with diabetes, but some looked at psoriasis or Alzheimer's disease; some trials targeted mild forms of diabetes, while others looked at seriously ill patients already requiring insulin treatment.
(5) The duration of the trials was very variable, ranging from 24 to 208 weeks.
(6) Different doses of rosiglitazone were used, from 2 up to 8 mg.
(7) Events were rare: for MI, less than 2% p.a. (per annum) in all but five treatment groups, and no incidence at all in nearly half the treatment groups; for CVD, less than 0.5% p.a. in all but five, and no incidence in 70% of the groups. Of the 48 trials, 10 have no reported MI events at all, and 26 more have none under one or other treatment, leaving only 12 trials with events in both groups. The problem is even more pronounced for the CVD data, for which 25 trials have no events, and 17 have none under one treatment, leaving only six.

All of these considerations cast doubt on the results of any analysis of these data, regardless of method, and make interpretation of any analysis of this dataset unclear. In this paper, I am using the dataset simply for illustration of the statistical techniques and not attempting to draw clinical conclusions.

3 Risk difference

Before undertaking a stratified analysis, careful thought must be given to the outcome measure to be combined across trials. Common choices are the risk difference, relative risk and odds ratio. A fourth is the hazard ratio, relevant to time-to-event analysis, which I do not consider here; but note that when the event-rate is low, and follow-up is reasonably balanced across groups within a study, then a study-stratified odds-ratio analysis would well approximate a stratified time-to-event analysis.

An analysis using risk difference has limitations. Firstly, the duration of the trials varies widely (24–208 weeks), and secondly the incidence rates also vary very widely; for MI, apart from the zeroes, the range is from 0.06% in Study 330 (on psoriasis patients) to 3.1% in Study 211 (on patients with both diabetes and congestive heart failure). The same studies exhibit the extremes of CVD, 0.06% and 2.2%. The potential for an increased risk under treatment of an event like MI or CVD seems very likely to be related to duration: the alternative would be to hypothesise a one-off increase in risk due to starting a treatment, with no additional risk thereafter, however long the treatment was applied. This may be a suitable model for some interventions and associated risks, but does not seem reasonable here.

Table 2. Notation

	Event	No event	Total
Rosiglitazone	s_{Ri}	f_{Ri}	n_{Ri}
Comparator	s_{Ci}	f_{Ci},	n_{Ci}
			n_i

Similarly, it seems more sensible scientifically to hypothesise that any effect of treatment would be larger if the background risk were larger. Consider two people, one who has a history of congestive heart failure and has a 3% background risk of MI, and one who is in the early stages of diabetes but otherwise healthy and has a 0.5% risk. If they are given a drug with a side-effect of deterioration in cardiovascular health, this is likely to increase the risk of MI for both of them. But the side-effect is almost certain to be more pronounced for the person who is more ill, whose risk of MI would therefore increase by more than that of the person who is less ill. In other situations, this might not be the case, and a model of constant additional risk may be more appropriate.

I will summarise three available methods of MA of the risk difference. The most popular method for any outcome measure is the inverse-variance method.[4] This requires the individual risk differences (d_i) from each trial together with their SEs (Standard Errors, s_i), and produces a combined estimate by weighting by the squared inverse of SE ($w_i = 1/s_i^2$).

$$\text{Combined estimate } = \Sigma w_i d_i / \Sigma w_i$$

$$\text{SE} = \surd(1/\Sigma w_i)$$

Here, d_i is the observed difference in incidence, $s_{Ri}/n_{Ri} - s_{Ci}/n_{Ci}$ and the usual estimator of s_i is

$$\surd\left(s_{Ri} f_{Ri}/n_{Ri}^3 + s_{Ci} f_{Ci} n_{Ci}^3\right)$$

where s_{Ri} and s_{Ci} are the numbers of events for Rosi and Comp, f_{Ri} and f_{Ci} the numbers of non-events, $n_{Ri} = s_{Ri} + f_{Ri}, n_{Ci} = s_{Ci} + f_{Ci},$ *and* $n_i = n_{Ri} + n_{Ci}$ the total number of subjects in Trial i (Table 2).

With rare events, a difficulty immediately arises: what do we do with the result $s_i = 0$ in the trials with no events (where $s_{Ri} = s_{Ci} = 0$)? One approach is to exclude these trials (referred to as 0-0 trials from now on), as their contribution to the combined estimate is undefined ($d_i = 0$ and $w_i = \infty$). This seems contrary to the inverse-variance approach, as we would be excluding the trials with smallest variance (i.e. variance estimated as zero): the 'true' inverse-variance combined estimate of the risk difference should be zero, because these trials with a zero risk difference have infinite weight and those with a non-zero difference have finite weight. Nonetheless, carrying through the exclusion of these trials gives a combined estimate of a 0.18% greater risk of MI on rosiglitazone, with 95% CI [0.07, 0.28] and a p-value of 0.001, and a 0.13% greater risk of CVD, [0.03, 0.22], $p = 0.009$. A second approach is to add a "correction factor" before analysis, but I leave this to a later section on odds ratios, where the objections to this approach are the same.

One of the issues of retrospective MA is the interpretation of p-values, such as 0.009 above. They do not have the same precise frequentist meaning as p-values from a prospective analysis of a clinical trial. This is because the decision to focus on a particular outcome (here MI and CVD) has been informed before planning the MA, by awareness of the results of individual trials that will be included in the MA. This multiplicity issue is complex and difficult to adjust for, but clearly implies that inference based on these p-values is not robust. I will continue to quote p-values for the various methods illustrated, but they should all be interpreted with this important issue in mind.

Apart from the problem of interpretation of the inverse-variance estimate, given the varying duration of the trials and the range of incidence, there is another important issue. The SE from the formula decreases as the observed proportion decreases, and so a large weight is derived for any trial that has a small observed proportion. In this example, the largest weight is for Trial 330 (one of the psoriasis trials), in which the observed proportions for both MI and CVD were 1/1172 under Rosi (Rosiglitazone) and 0/377 under Comp (Comparator). The values of d_i and s_i are both 0.00085, giving a weight of 1.4×10^6 which is 40% of the total for MI and 33% for CVD. By comparison, the large trials, DREAM and ADOPT, contribute only 9% and 2% to the weight for MI and 8% and 15% for CVD. It does not seem reasonable to give Trial 330 so much weight just because its incidence rate is so low. Without this trial, the combined estimate is 0.14%, [0.03, 0.25], $p = 0.009$ for MI and 0.10%, [0.003, 0.19], $p = 0.04$ for CVD, so it is fortuitous that this trial gives individual estimates not too far from the means of the other trials.

A second approach to analysing risk differences was given by Greenland and Robins,[5] and this is the one usually recommended for risk differences of rare events because of the problems noted above. It is an adaptation of the Mantel-Haenszel method, originally derived for odds-ratio estimates, and takes into account the underlying model of a constant risk difference across trials.

$$\text{Combined estimate } = \Sigma w_{MHi} d_i / \Sigma w_{MHi}, \quad \text{where} \quad w_{MHi} = n_{Ci} n_{Ri} / n_i$$

$$\text{SE} = \surd(\Sigma\{(s_{Ri} f_{Ri} n_{Ci}{}^3 + s_{Ci} f_{Ci} n_{Ri}{}^3)/(n_{Ci} n_{Ri} n_i{}^2)\})/\Sigma w_{MHi}$$

This gives positive weight even to 0-0 trials. Unscaled by duration, the estimate is 0.19% [0.01, 0.37], $p = 0.034$, for MI, and 0.11% [0.00, 0.21], $p = 0.048$, for CVD. The larger p-value reflects the different weighting used in this method, though again it happens not to affect the estimates very much. Scaling by duration, the estimate is 0.12% p.a. [0.01, 0.23], $p = 0.034$, for MI, and 0.07% p.a. [0.00, 0.13], $p = 0.049$, for CVD.

More recently, Tian et al.[2] suggested a third way to analyse the N&W data on the risk difference scale, including information from the 0-0 trials. This method is similar to the method of combining p-values, but based on consideration of confidence intervals with varying coverage. Analysing risk differences, their estimate is 0.18% [–0.08, 0.38], $p = 0.27$, for MI, and 0.06% [–0.13, 0.23], $p = 0.83$, for CVD. Their method demonstrates that statistical significance of the risk difference depends on the methodology employed, so that the results by any of the methods cannot be considered robust, or firm conclusions drawn.

Random-effects and Bayesian methods can also be used, but I will leave the illustration of these approaches to analysis on the more appropriate odds-ratio scale.

4 Relative risk or odds

Both the major problems with the risk-difference scale can be overcome by using a risk-ratio scale: the issue with duration disappears because one incidence count is divided by the other (as long as the assumption of constant risk over time is accepted, and also the assumption that follow-up across treatment groups is reasonably balanced within study), and the model is designed for looking at multiplicative effects. There is virtually no difference here between the risk-ratio and odds-ratio scales because the risks are so small. I shall use the odds-ratio scale since it is conceptually more appropriate for modelling risks bounded in the range [0, 1], and is also the more common and the one that N&W used.

The inverse-variance method using the individual log odds-ratios is problematic because for MI only 12 trials have finite values of this statistic (whichever treatment group has zero incidence, the log odds-ratio is infinite); for CVD, there are only six. It does not seem reasonable to analyse just these few trials.

A more representative analysis is provided by logistic regression, fitting additive effects of treatment and study to the observed proportions on the logistic scale. Most implementations of logistic regression exhibit some minor technical problems, but the generalised-linear-model algorithm converges to give an estimate of 1.427 [1.030, 1.977], $p = 0.0327$ for MI and 1.664 [0.975, 2.840], $p = 0.0619$ for CVD. The minor problems with this method are caused by the 0-0 trials: the maximum-likelihood estimates of the study effects for these trials are all infinite on the log-odds scale, and most software will converge with some largish negative value on the log-odds scale, such as –10 (corresponding to odds of 0.00005). But this does not affect the fit of the model (as long as the algorithm converges), and some software such as LogXact handles the infinite values intelligently. To avoid any difficulty with logistic regression, the 0-0 trials can just be excluded from the analysis. However, excluding such trials raises other issues, which I discuss below.

A major part of any meta-analysis should be the investigation and interpretation of heterogeneity. There are generally two types of heterogeneity: clinical and statistical. I discussed some aspects of clinical heterogeneity in Section 2, which limit the interpretability of the rosiglitazone MAs. Statistical heterogeneity can result from aspects of clinical heterogeneity or from other causes. Statistical heterogeneity can be assessed by the Cochrane Q-statistic, though its use as a test for heterogeneity is generally accepted as being inappropriate because the result is driven more by the number of trials than by the size of the heterogeneity.[6] Another approach is to use the I^2 statistic, which is a measure of the percentage of the total variance that is attributable to difference between trials. In this example, both the Q and I^2 statistics are zero. This is a likely occurrence in the meta-analysis of rare events because of the sparseness of the data; see, for another example, Box 4 in Sutton et al.[7] A consequence is that the commonly used DerSimonian–Laird random-effects method gives exactly the same result as the corresponding fixed-effect inverse-variance approach.

Cai et al.[8] proposed a method based on a Poisson model, with fixed- or random-effects, and also find little evidence of statistical heterogeneity, giving an estimate of relative risk for MI with either method of 1.33 [0.96, 1.84], $p = 0.087$. They state that this method is more powerful than standard methods when there is a high level of between-study variation. Shuster et al.[9] proposed a random-effects approach that effectively treats each trial as having equal weight, and show very different results compared to the fixed-effects estimates for both MI (1.51 [0.91, 2.48], $p = 0.11$) and CVD (2.37 [1.38, 4.07], $p = 0.002$). Both Cai and Shuster demonstrate that results from analyses of the N&W dataset depend on the choice of statistical method and illustrate the underlying fragility of the data.

There are at least four other fixed-effect non-Bayesian methods to analyse odds ratios.[10] One is a simpler approach to estimation that has been widely used in meta-analysis, using the idea of 'efficient scores' and Fisher's information. This can be seen as a one-step approximation to the likelihood method; but the approximation is often good (unless the data are extreme with respect to the distributional assumptions) and the calculations are certainly easier – though the power of current computers now removes the original force of this argument.

A second method is to condition on the observed margins: in this case, the total number of events in each study. This margin is a measure of the average event rate in each study, and can be seen to be ancillary to the effect of interest – the way the rate differs between treatments within each study. Rather than using the binomial distribution to model the observed counts, the hypergeometric distribution is appropriate when the margin is treated as fixed. The method is called conditional logistic regression, and in this context is an extension of Fisher's exact test. It is often recommended for models in which the number of parameters is a substantial proportion of the number of observations. Here, the number of observations is the number of patients, not the number of trial arms, so there should not be the same issue as with matched binary data (Agresti,[11] Section 10.2.3). In my experience of meta-analysis, the results are little different from those from unconditional logistic regression, and that is certainly the case in this example. Cai et al.[8] investigate a conditional method with a random-effects model based on the beta-binomial distribution; however, they report that the unconditional approach is more efficient than the conditional when there is a high level of between-study variation, and gives similar performance when there is little between-study variation.

The Peto method[12] is a simpler version of the conditional approach. This uses the scoring method, and is virtually the same as the scoring method for the unconditional approach (the only difference in calculation is that the variance of the contributing estimate from each trial is larger by a factor of $n/(n–1)$, where n is the total number of patients in the trial). The calculations can be done as for the inverse-variance method using Peto estimates of log odds-ratio from each study:

$$\log \text{OR} = (o_i - e_i)/v_i$$
$$\text{SE}(\log \text{OR}) = \sqrt{\{1/v_i\}}$$

Where

$o_i = s_{Ri}$ is the observed number of events in the Treatment arm),
$e_i = (s_{Ri} + s_{Ci})n_{Ri}/n_i$ is the expected number
$v_i = (s_{Ri} + s_{Ci})(n_i - (s_{Ri} + s_{Ci}))n_{Ri}n_{Ci}/(n_i{}^2(n_i - 1))$

N&W chose the Peto method. This is generally considered a reasonable method for analysing rare events when the treatment effect is not large.[13] It has been reported as behaving poorly when randomisation ratios are far from 1, i.e. 8:1 or greater.[14,15] In this set, there were three trials with a 4 : 1 ratio, and three with 3 : 1; in addition, one of the two very large and long trials (ADOPT, containing 43% of the events) had a 1 : 2 ratio. The behaviour of the method can be tested by detailed simulation, and this shows that the combined estimate has little bias here and the properties of the significance test are satisfactory. The main reason for choosing the Peto method, as opposed to the inverse-variance method with conventional log odds-ratios, is to allow inclusion of the trials with events in only one arm.

The Mantel-Haenszel method (MH) was originally designed for odds ratios.[16] It combines estimates on the odds-ratio scale rather than the log-odds-ratio scale:

$$\text{Combined estimator } = (\Sigma s_{Ri} f_{Ci}/n_i)/(\Sigma s_{Ci} f_{Ri}/n_i)$$
$$\text{SE} = \sqrt{(\Sigma\{A_iC_i/C^2 + (A_iD_i + B_iC_i)/CD + B_iD_i/D^2\}/2)}$$

where $A_i = (s_{Ri} + f_{Ci})/n_i, B_i = (s_{Ci} + f_{Ri})/n_i, C_i = s_{Ri}f_{Ci}/n_i, D_i = s_{Ci}f_{Ri}/n_i, C = \Sigma C_i, D = \Sigma D_i$.

So 0-0 trials still have no contribution, but trials with 0 only in the Rosi arm (zero odds-ratio) increase the denominator, and trials with 0 only in the Comp arm (infinite odds-ratio) increase the numerator. Note that some packages, such as RevMan,[17] implement the MH method using a different formula, with the denominator expressed as a sum of odds ratios $(s_{Ri}/f_{Ci})/(s_{Ci}/f_{Ri})$ multiplied by weights $s_{Ci}f_{Ri}/n_i$. This is identical when counts are non-zero, but is undefined for studies with no events in the Comparator arm. In practice, the alternative formula is used in

Table 3. Results of seven methods of analysing the odds or risk ratios

	MI			CVD		
	Est	95% CI	*p*-value	Est	95% CI	*p*-value
Logistic	1.427	1.030, 1.977	0.0327	1.664	0.975, 2.840	0.0619
Cai	1.33	0.96, 1.84	0.087	–	–	–
Shuster	1.51	0.91, 2.48	0.11	2.37	1.38, 4.07	0.0017
Scoring	1.429	1.031, 1.980	0.0320	1.641	0.980, 2.748	0.0595
Conditional	1.426	1.029, 1.975	0.0328	1.663	0.975, 2.837	0.0621
Peto	1.428	1.031, 1.979	0.0321	1.640	0.980, 2.744	0.0597
MH	1.427	1.054, 1.932	0.0215	1.698	1.042, 2.767	0.0337

conjunction with correction factors, but the original formula should be used when correction factors are not used.

The variance of the estimator is not useful for constructing a confidence limit because of the skewness of the distribution on this scale (Whitehead,[10] p. 219). Instead, the formula given by Robbins et al.[18] can be used, derived from the log-odds-ratio scale.

Table 3 summarises the methods in this section.

These methods show similar estimates, though different conventional inference if the *p*-values are taken at face value (see Section 4). However, the analysis is still bedevilled by the trials with no events, 10 for MI and 25 for CVD: they make no contribution at all to the combined estimates in any of these methods. Intuitively, there seems to be some information about the relative sizes of the risks in these trials from the fact that an equal number of events (i.e. none) were observed on the two treatments. But this feeling is really based on prior expectation: we expect that the risk is not actually zero, and that therefore there is an underlying relative risk. Unless we can quantify this intuition, there is in fact no statistical information about relative risk because the observed risks are zero.

There are three ways to address this problem, discussed below.

4.1 Bayesian approach

One way is to adopt a Bayesian approach, and be explicit about prior expectations. The whole point of meta-analysis is to include all relevant information in the dataset being analysed, so there is rarely any prospect of a prior that is generated from actual data exactly corresponding to that in the MA, though there may be other contextual information, e.g. from non-randomised studies or from other drugs in the same class that could be used to derive weakly informative priors. Priors can also be based on elicited information from experts; however, different people (e.g. industry clinicians and regulators) will naturally not agree on the prior. As a result, Bayesian methods for meta-analysis are usually based on 'non-informative' priors (Cochrane Collaboration,[19] Section 16.8.1).

A fixed-effect analysis can be carried out using the logistic regression approach, adopting Normal priors with mean 0 and variance 10,000 (say) on the log-odds scale for the treatment and study effects. (This can be done, for example, using WinBUGS, or the MCMC Procedure in SAS 9.2.) As with the non-Bayesian methods, the 0-0 trials do not contribute, and an estimate for MI using SAS (see Appendix 1.1) with 100,000 simulations after a start-up of 10,000, and thinning by a factor of 2, is 1.45 with 95% credible interval [1.03, 1.98]. This matches the result reported by Friedrich et al.,[3] and is similar to the non-Bayesian methods described above, as is to be expected in the absence of prior information. Friedrich et al. applied several methods to account for the 0-0 trials, some of which resulted in non-significant findings, again indicating that formal inference is sensitive to the method chosen to analyse this dataset. Similarly, Diamond et al.[20] performed a Bayesian analysis and reported an estimate of 1.26 [0.93, 1.72], but they included a 'standard' continuity correction (see next section).

A Bayesian random-effects analysis gives slightly different results. Using WinBUGS (for variety, see Appendix 1.2) with 190,000 simulations after a start-up of 10,000, and no thinning, with a 'non-informative' Uniform[0, 10] prior for the SD of the log odds-ratio between the treatment across trials, the estimate is 1.47 (median value of OR posterior distribution, Monte-Carlo error 0.007) with 95% credible interval [0.97, 2.36]. The small increase in the estimate can be ascribed to a small amount of information from the 0-0 trials, and the increase in the interval to a non-zero estimate of heterogeneity: the SD is estimated as 0.25 (median of posterior, with MC error 0.008) and 95% interval [0.005, 1.102]). Cai et al.[8] investigate several Bayesian approaches, and report an estimate of 1.26 with 95% credible interval of [0.80, 1.82] from a fully Bayesian random-effect model for relative risk.

4.2 Continuity correction

A second approach is to adjust the data. A small number can be added to the zeroes to make them non-zero, on the basis that this is a more 'likely' result to have observed (based on prior belief) apart from the fact that the sample size or trial duration was too small. This is referred to as a 'continuity correction' because it attempts to redress the fact that observation of small counts is necessarily discrete; but it is usually applied only to zero observations, and counts of 1 or 2 are not similarly adjusted. The number is usually added to all four counts from a trial in which one is zero (i.e. numbers of events and non-events for both arms), in an attempt to reduce bias. Some software does this adjustment automatically, using the value 0.5, and some research has indicated that this produces less biased results than ignoring the zero results.[21,22] However, that research looked at underlying risks no smaller than 10%: here we have less than 1%, and the impact of all the added 0.5 s on these small counts can potentially swamp any real effects.[7] Whereas we might expect the adjustment to correct bias when a zero count is unlikely (i.e. when risks are not very small), it is likely to create bias when a zero is likely, as here. Sweeting et al.[13] provided a detailed study of correction factors in the analysis of odds ratios of sparse events, with an informative erratum[23] showing in particular that the Mantel-Haenszel method performs well with no correction factors, for balanced data.

Using the standard 0.5 adjustment (adding to both counts in trials where either or both counts are zero), the combined estimate for all the odds-ratio methods drops to 1.24 [0.92, 1.65], $p = 0.16$ for MI, and 1.14 [0.76, 1.70], $p = 0.53$ for CVD. In an attempt to reduce the swamping effect, a value of 0.1 for the adjustment raises the estimate back up to 1.38 [1.00, 1.90], $p = 0.048$, and 1.48 [0.90, 2.43], $p = 0.12$, respectively, while 0.01 returns us almost to the results omitting the zero trials. An alternative adjustment (which raises the same objections as above) has been suggested[23] called 'Treatment-arm correction', which adjusts for imbalance in numbers of subjects in the two arms. The usual implementation is to add $n_{Ci}/(n_{Ci} + n_{Ri})$ to both counts on the control arm and $n_{Ri}/(n_{Ci} + n_{Ri})$ on the treatment arm. This gives 1.34 [1.00, 1.08], $p = 0.053$ for MI and 1.36 [0.90, 2.04], $p = 0.14$ for CVD. Diamond et al.[20] and Friedrich et al.[3] report results from the Mantel-Haenszel method with various alternative corrections.

Table 4. Matching of trials with no events in the Avandia meta-analysis

Trials with no MI events				Matched trials with some MI events			
Trial	Dur.	Treatments	Ratio	Trial	Dur.	Treatments	Ratio
095	26	Rosi + Ins vs Ins	2 : 1	082	26	Rosi + Ins vs Ins	2 : 1
234	26	Rosi + SU vs SU	2 : 1	079	26	Rosi ± SU vs SU	2 : 1
331	52	Rosi vs Plac	2 : 1	330	52	Rosi vs Plac	3 : 1
009	24	Rosi + Met + Ins vs Ins	1 : 1	347	24	Rosi + Ins vs Ins	2 : 1
282	24	Rosi + Met vs SU + Met	1 : 1	284	24	Rosi + Met vs Met	1 : 1
369	26	Rosi vs SU	1 : 1	162	26	Rosi + SU vs SU	1 : 1
096	26	Rosi + SU vs SU	2 : 1	079	26	Rosi ± SU vs SU	2 : 1
044	26	Rosi + Met vs Met	2 : 1	094	26	Rosi + Met vs Met	2 : 1
325	24	Rosi vs SU	1 : 1	143	24	Rosi + SU vs SU	1 : 1
004	24	Rosi ± SU vs SU	3 : 1	132	24	Rosi + SU vs SU	4 : 1

4.3 Matched pooling of trials

The third way is to combine trials to avoid having any components of the analysis that have no events. This is standard procedure in the analysis of sparse contingency tables, and here (as there) there is the danger that the pooling will introduce the type of bias seen with Simpson's Paradox.[24] It is clearly best to match the trials as closely as possible, to minimise the potential for this type of bias. A simple pooling of all small trials by randomisation ratio was carried out by Nissen and Wolski[25] in a follow-up paper summarising 56 trials; for example, 31 small trials with 1 : 1 randomisation were pooled. To avoid the potential for bias as far as possible, it is better to pool as little as possible. Table 4 shows a matching that pools pairs of trials with more or less the same duration, treatment regimens and randomisation ratio.

The effect of including this extra information in a pooled form is minimal: the Peto estimate become 1.433 [1.035, 1.985] $p = 0.0303$. It is at first sight surprising that the inclusion of more patients on Rosi than on Comp, all with no further events, actually increases the combined estimate slightly. In fact what happens is that pooling slightly reduces the underlying heterogeneity, and this slightly increases the effect on the marginal scale.

Whatever method is used for the MA, there are limitations: sparse data, clinical heterogeneity, lack of prospective planning to assess CV events, and the difficulty of interpreting p-values from retrospective analyses. Nonetheless, there is a suggestion in some analyses that there may be slightly more heart attacks experienced by the patients in the trials who received rosiglitazone. In those analyses that suggest a potential elevated risk, we can quantify the potential risk with the estimated odds ratio 1.43, which we can also interpret as a risk ratio because the underlying risk is low. So we could state that some analyses suggest a potential 43% higher risk of MI. This sounds large, until you realise that it is a percentage of a percentage. As far as patients and physicians are concerned, what is much more relevant is the risk difference. This natural requirement leads some people to conclude that the analysis should be done on the risk scale, regardless of the problems outlined above. But this is not necessary. Instead, we can simply use the model we have fitted on the log-odds scale to estimate, or more formally to "predict", the risk difference. This could be done for any patient for whom we know the risk of a heart attack when not on a rosiglitazone treatment (though remember that the analysis has done some peculiar averaging over very different comparator treatments). So if the patient is mildly diabetic and has a risk of 0.5%, say, the model predicts that the risk rises on average to 0.7% – a risk difference of 0.2% or two in a thousand. In fact, of course, any real patient will know what other treatment they are receiving already, so this average effect over all treatments derived from this collection of trials is not actually meaningful to that patient: this is one of the major difficulties with the decision to combine all the information from different comparator treatments in a single MA.

5 Reporting results

Having fitted the model, it is useful to produce a summary on the risk scale that averages over the studies, adjusting for the differences between them. This can be called a 'prediction'[26] as it predicts what the average risk might be under adjusted conditions: in this case, with equal numbers of patients on each treatment within a trial. This is the sort of statistic that is calculated by the LSMEANS statement in the GLM procedure in SAS, but is unfortunately not available in the GENMOD procedure usually used for logistic regression. It is calculated by the PREDICT statement in GenStat (Appendix 1.3).

This gives the results:
Comparator 0.44% (s.e. 0.053)
Rosiglitazone 0.63% (s.e. 0.069)
Difference 0.19% (s.e. 0.088)

Note how close this turns out to be to the result of the risk-difference analysis at the start of this section. However, the patients over whom the risk has been averaged were exposed to the treatments for varying periods, so the average risks also average over the pattern of exposure times across the trials. A more useful summary, if we believe the risk to be proportional to exposure in any trial, can be obtained by adjusting for exposure. This is done simply by including the log of the duration (in years) as an 'offset' variable in the logistic regression and further adjusting the prediction to give the values relevant for one year's exposure (Appendix 1.4).

Introducing the offset does not change the fit of the model, because the exposure was assumed equal for both treatments in a trial. The resulting values can then be interpreted as average risks per annum.

Comparator 0.34% (s.e. 0.050)
Rosiglitazone 0.48% (s.e. 0.059)
Difference 0.14% (s.e. 0.066)

The combined estimate in this analysis is, nominally, just significantly different from zero: we would expect an estimate of that size or more by chance in every 30 such analyses if there were actually no effect at all, and clinical trial reports list dozens of types of adverse event. The number of events, even in this large collection of trials is low: the odds-ratio would not have been significant at the 5% level if any two events in the rosiglitazone group of the large ADOPT study had not occurred, or any three events in a smaller trial. If about a dozen events had occurred in the comparator group rather than the rosiglitazone group, the odds-ratio would actually be less than 1.

It is instructive to note that these predictions on the risk scale do depend on the 0-0 trials, unlike the estimates on the log-odds scale. This is because the averaging of the effects is done over all the

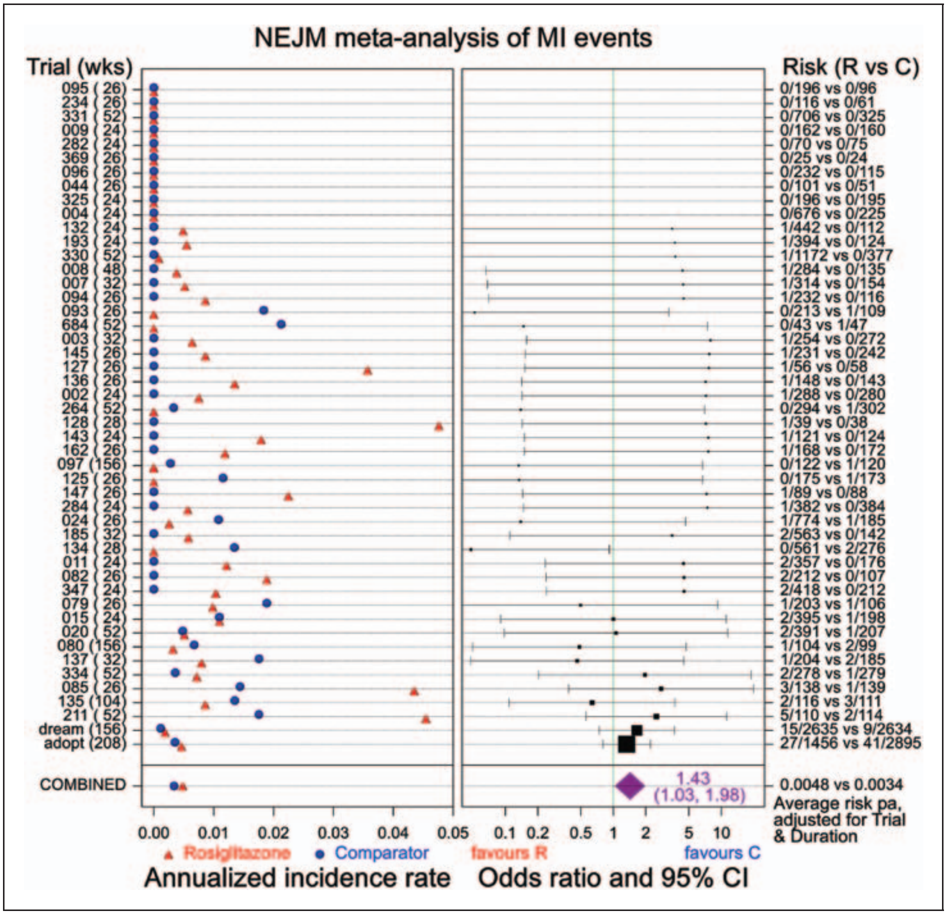

Figure 1. Enhanced forest plot of the Peto meta-analysis of all 48 trials.

trials. If we had left out all ten 0-0 trials in the analysis of MI, the results above would have been as follows.

Comparator 0.39% (s.e. 0.058)
Rosiglitazone 0.55% (s.e. 0.068)
Difference 0.14% (s.e. 0.076)

The results of meta-analysis are usually presented in a forest plot. This display has the advantage of giving a clear visual representation of the combined effect together with those from the contributing studies. However, when analysing on a relative scale, the display can be enhanced by adding a second panel to show the incidence rates from which the relative statistics have been described (as described by Amit et al.,[27]). This can be further enhanced by including the predictions above, as in Figure 1.

Following N&W, Dahabreh[28] reanalysed the N&W dataset adding new results from the interim analysis of a large randomised clinical trial (RECORD), and updated event counts in the other large studies (ADOPT and DREAM), which had completed processes of event adjudication. The headline estimate for MI of 1.43 reduced to somewhere between 1.23 and 1.33, and that for CVD of 1.64 reduced to 1.01–1.13, depending on the method of handling the trials with no events. The follow-up paper by Nissen and Wolski,[25] already mentioned, included further trials (including RECORD) in another Peto analysis that gave an estimate of 1.28 for MI and 1.03 for CVD. An investigation by Mannucci et al.[29] found 164 trials with duration >4 weeks, and analysed the results using the DerSimonian-Laird random-effects model, giving estimates of 1.14 [0.90, 1.45] for MI and 0.94 [0.68, 1.29] for CVD. However, they used the program Comprehensive Meta-Analysis, which automatically makes the 0.5 correction to zero counts. Most recently, Kaul and Diamond[30] reanalysed the data in Nissen and Wolski[25] with various methods using corrections, again showing the fragility of the *p*-values for analysis of MI, though all analyses of CVD showed odds ratios around 1.0 with no statistical significance.

6 Conclusions

There is a broad range of meta-analytic methods that can be used to combine summary-level data from trials with low event rates. However, each of the available methods has limitations. Risk difference is easily interpretable and can include zero-event trials, but is not the best choice when duration and incidence rates vary among trials. Odds-ratio-based methods are commonly used, but do not incorporate zero-event trials, thereby excluding available trials and data from the combined estimate. Continuity corrections that allow for inclusion of zero-event trials can produce misleading results when incidence rates are low. Bayesian analyses also can include zero-event trials, but specification of priors can be a source of disagreement. Matched pooling of similar trials can lead to a Simpson's Paradox type of bias. Thus, there is no perfect meta-analysis technique for rare events, and investigators should consider sensitivity analyses employing numerous methodologies. The N&W example illustrates how meta-analytic results for sparse data can be discordant depending on the statistical technique employed to analyse the data, with some findings achieving statistical significance while others do not. Other investigators have applied different statistical techniques to the N&W dataset and found results that were no longer statistically significant. This discrepancy is an indication that the underlying data are not robust. Where findings from a meta-analysis of rare events in trials not intended to assess the outcome of interest are of borderline significance, it is helpful to consider and present various sensitivity analyses. When results are sensitive to the choice of statistical methodology, the results should be considered hypothesis-generating.

Where the statistical significance of the results depends on the choice of methodology, the findings must be interpreted cautiously. However, the choice of statistical method used to examine the data is less important than the choices that have to be made in selecting and evaluating the studies, and the inherent limitations of combining data never intended to assess the outcome of interest. There is a large literature on the potential biases associated with these choices, and attention to this aspect of a meta-analysis is given prominence in the extensive advice provided by the Cochrane Collaboration.[19] In this example, the data are extremely sparse and were not collected to address the question of cardiovascular safety. Thus, regardless of the methodology employed, the results should be considered exploratory and hypothesis generating.

More attention also needs to be paid to the presentation of the results of meta-analysis of binary data. It is not acceptable to report simply a combined estimate of a statistical ratio, without presenting also the associated incidence rates to put the analysis in context. Statistical methods are available to take results from analyses done on one scale in order to provide meaningful summaries on another scale for ease of interpretation. Improved graphical displays, as illustrated here, can be used to summarise all the information accessibly.

Acknowledgements

I am indebted to my colleague James Roger in GSK's former Research Statistics Unit for providing the Bayesian analyses, and to him and Valerii Fedorov, Head of the RSU, for many discussions.

References

1. Nissen SE and Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med* 2007; **356**: 2457–2471.
2. Tian L, Cai T, Pfeffer MA, Piankov N, Cremieux P-Y and Wei LJ. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 x 2 tables with all available data but without artificial continuity correction. *Biostatistics* 2009; **10**: 275–281.
3. Friedrich JO, Beyene J and Adhikari NKJ. Rosiglitazone: can meta-analysis accurately estimate excess cardiovascular risk given the available data? Re-analysis of randomized trials using various methodologic approaches. *BMC Res Notes* 2009; **2**: 5, doi: 10.1186/1756-0500-2-5.
4. Birge RT. The calculation of errors by the method of least squares. *Phys Rev* 1932; **40**: 207–227.
5. Greenland S and Robins JM. Estimation of common effect parameter from sparse follow up data. *Biometrics* 1985; **41**: 55–68.
6. Higgins JP and Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002; **21**: 1539–1558.
7. Sutton AJ, Cooper NJ, Lambert PC, Jones DR, Abrams KR and Sweeting MJ. Meta-analysis of rare and adverse event data. *Expert Rev Pharmacoecon Outcomes Res* 2002; **2**: 367–379.
8. Cai T, Parast L and Ryan L. Meta-analysis for rare events. *Stat Med* 2010; **29**: 2078–2089.
9. Shuster JJ, Jones LS and Salmon DA. Fixed vs random effects meta-analysis in rare event studies: the rosiglitazone link with myocardial infarction and cardiac death. *Stat Med* 2007; **26**: 4375–4385.
10. Whitehead A. *Meta-analysis of controlled clinical trials*. Chichester: Wiley, 2002.
11. Agresti A. *Categorical data analysis*. Hoboken. NJ: Wiley, 2002.
12. Yusuf S, Peto R, Lewis J, Collins R and Sleight P. Beta-blockade during and after myocardial infarction: an overview of the randomized trials. *Progr Cardiovasc Dis* 1985; **27**: 335–371.
13. Sweeting MJ, Sutton AJ and Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med* 2004; **23**: 1251–1375.
14. Greenland S and Salvan A. Bias in the one-step method for pooling study results. *Stat Med* 1990; **9**: 247–252.
15. Bradburn MJ, Deeks JJ, Berlin JA and Localio AR. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med* 2007; **26**: 53–77.
16. Mantel N and Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Instit* 1959; **22**: 719–748.
17. Deeks JJ and Higgins JPT. Statistical Algorithms in Review Manager 5, http://www.cochrane.org/sites/default/files/uploads/handbook/Statistical_Methods_in_RevMan5-1.pdf (accessed 7 September 2011).
18. Robbins J, Greenland S and Breslow N. A general estimator for the variance of the Mantel-Haenszel odds ratio. *Am J Epidemiol* 1986; **124**: 719–723.
19. Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Interventions, http://www.cochrane.org/training/cochrane-handbook (accessed 7 September 2011).
20. Diamond GA, Bax L and Kaul S. Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. *Ann Intern Med* 2007; **147**: 578–581.
21. Gart JJ and Zweifel JR. On the bias of various estimators of the logit and its variance with application to quantal assay. *Biometrika* 1967; **54**: 181–187.
22. Sankey SS, Weissfeld LA, Fine MJ and Kapoor W. An assessment of the use of the continuity correction for sparse data in meta-analysis. *Commun Stat – Simul Comput* 1996; **25**: 1031–1056.
23. Sweeting MJ, Sutton AJ and Lambert PC. Correction. *Stat Med* 2006; **25**: 2700.
24. Blyth CR. On Simpson's paradox and the sure-thing principle. *J Am Stat Assoc* 1972; **67**: 364–366.
25. Nissen SE and Wolski K. Rosiglitazone revisited. *Arch Intern Med* 2010; **170**: 1191–1201.
26. Lane PW and Nelder JA. Analysis of covariance and standardization as instances of prediction. *Biometrics* 1980; **38**: 613–621.
27. Amit O, Heiberger R and Lane PW. Graphical approaches to the analysis of safety data from clinical trials. *Pharmaceut Stat* 2008; **7**: 20–35.
28. Dahabreh IJ. Meta-analysis of rare events: an update and sensitivity analysis of cardiovascular events in randomized trials of rosiglitazone. *Clin Trials* 2008; **5**: 116–120.
29. Mannucci E, Monami M, Di Bari M, Lamanna C, Gori F, Gensini GF, et al.. Cardiac safety profile of rosiglitazone: a comprehensive meta-analysis of randomized clinical trials. *Int J Cardiol* 2010; **143**: 135–140.
30. Kaul S and Diamond GA. Rosiglitazone and cardiovascular risk: Nissen and Wolski 2010 updated meta-analysis revisited. *J Am College Cardiol* 2011; **57**(14, Supp 1): E1205.

Appendix: Software code

1.1 SAS code for Bayesian fixed-effect analysis

```
proc mcmc data=MI48 ntu=10000 nmc=100000 nthin=2 propcov=quanew
       diag=(mcse ess) outpost=beetleout seed=246810;
 parms alpha beta;
 parms _study1-_study&nstudy. ;
 prior alpha beta _study1-_study&nstudy. ~ normal(0, var=10000);
 array _study[&nStudy];
 p = logistic(alpha + beta*type + _study[studyid]);
 model mi ~ binomial(n,p);
run;
```

1.2 Winbugs Code For Bayesian Random-Effects Analysis

```
 {
 for( i in 1 : Num ) {
  rc[i] ~ dbin(pc[i], nc[i])
  rt[i] ~ dbin(pt[i], nt[i])
  pc[i] <- 1 - 1/(1+exp(mu[i]))
  pt[i] <- 1 - 1/(1+exp(mu[i]+delta[i]))
  mu[i] ~ dnorm(0.0,1.0E-4)
  delta[i] ~ dnorm(d, tau)
          }
 d ~ dnorm(0.0,1.0E-6)
 tau <- 1/(sigma*sigma)
 sigma ~ dunif(0,10)
 }
```

1.3 GenStat code for predictions

```
model [distribution = binomial] mi; nbinomial = subjects
 fit treatment, study
 predict treatment
```

1.4 GenStat code for predictions adjusting for duration

```
model [distribution = binomial; offset = log(duration)] mi; nbinomial = subjects
 fit treatment,study
 predict [offset = 0] treatment
```