

# BAYESIAN ASSESSMENT OF PUBLICATION BIAS IN META-ANALYSES OF CERVICAL CANCER AND ORAL CONTRACEPTIVES

Bonnie LaFleur, Sue Taylor, University of Colorado  
D. D. Smith and R. L. Tweedie, Colorado State University  
Sue Taylor, UC Health Sciences Center, 4200 E. Ninth Ave, Denver CO 80262

**KEYWORDS:** Cervical cancer, oral contraceptives, publication bias, meta-analysis, Bayesian meta-analysis, file-drawer problem

## Abstract

Meta-analysis is well known to be susceptible to publication bias (PB), caused by lack of publication of all works on the area in question. This paper applies a recently developed Bayesian method for assessing PB to a collection of studies of the possible effects of oral contraceptive use on incidence of cervical cancer. We build on a meta-analysis developed in Delgado-Rodriguez *et al.* [2]. We apply a more formal approach to evaluating and adjusting for PB than the ad hoc funnel plot procedure used in that paper. We conclude that there is support for the existence of publication bias in the main data set analyzed in [2], and we show it is probably caused by the explicit disregard of 'low quality' studies in [2]. Overall we conclude that there is rather weak support for a positive association between oral contraceptive use and incidence of cervical cancer.

## 1 Introduction

Meta-analysis is a widely applied technique for statistically combining analyses from individual studies into a single analysis (Hedges and Olkin [7]). Concerns about the way in which such data should be combined have led to a great deal of literature discussing benefits and problems of this technique (for a more thorough discussion of the issues concerning meta-analysis see for example Mosteller and Chalmers [10]; Hedges and Olkin [7]; NRC Report [11]; Light and Pillemer [8]).

One important issue in meta-analysis is the need to gather all relevant results, in order to fully describe relationships of interest without bias, and when only a subset of the complete set of results is used, 'publication bias' (PB) may result.

It is a common belief [1, 6] that statistically significant results are published differentially in scientific journals, since some scientists (for example, students with PhD and Masters' theses) may not submit non-significant results, and editors may not publish them if submitted. If there is a non-representative number of statistically significant results in the literature, then this will obviously bias a literature-based meta-analysis toward statistically significant conclusions.

In this paper we consider a published meta-analysis of studies of the association between the use of oral contraceptives and cervical cancer, and apply new techniques developed in Givens, Smith, and Tweedie [5] to the problem of assessing and adjusting for PB in this meta-analysis.

The results we consider were collected and analyzed by Delgado-Rodriguez *et al.* [2], who evaluated 62 published relative risks, from 51 papers. They consider the results in two groups: Group I, comprising the complete set of 62 results, and Group II, comprising 26 'methodologically acceptable' results which also contained sufficient information to carry out a meta-analysis. In this paper we accept the groupings in [2] rather than critically reviewing them, but we note that, although in [2] the process of selecting for quality seems detailed and adequate, the criteria used for selection are not clearly stated.

Delgado-Rodriguez *et al.* [2] use a graphical ('funnel plot') method suggested by Vandembroucke [13] to examine potential PB in Group II. In this paper our main aim is to present a more formal method of evaluation of potential PB to supplement this ad-hoc

Table 1  
Analysis of Relative Risk of Cervical Cancer Associated with Oral Contraceptives

Type of Analysis	Group I	Group II
Dysplasia [2]	1.31 (1.24, 1.38)	1.52 (1.27, 1.82)
Carcinoma <i>in situ</i> [2]	1.29 (1.18, 1.41)	1.52 (1.31, 1.76)
Invasive cancer [2]	1.13 (0.99, 1.27)	1.21 (1.06, 1.37)
Overall Fixed Effects	1.30 (1.24, 1.35)	1.37 (1.26, 1.49)
Overall Random Effects	1.15 (1.10, 1.30)	1.38 (1.17, 1.63)
Overall Bayesian Model	1.13 (0.95, 1.34)	1.46 (1.08, 1.94)

graphical method.

Our method is based on a detailed Bayesian model for the probabilities with which publications may be ‘suppressed’, as developed in [5], and we show that it indicates that there may indeed be a suppression of several results. The method uses a data augmentation technique which not only assesses the existence of PB, but also allows us to adjust the Bayesian posterior distribution of the relative risk for such bias. We find that the overall excess risk for cervical cancer in this situation may be overstated by a factor of almost 2 if we allow for PB in this way.

In this cervical cancer context, the main meta-analysis is carried out on the restricted Group II results, so that we know a number of ‘low-quality’ results have been suppressed. By comparing the results of the PB adjustment method with the original larger Group I, we further show that our method seems to a large extent to be identifying the possible effect of this suppression on the grounds of quality, which has implications on the overall conclusions about this relationship.

## 2 Bayesian and other methods of meta-analysis

The overall measure of association used in [2] is the relative risk, and confidence interval, for each of three separate degrees of outcome (dysplasia, carcinoma *in situ*, and invasive cancer). Delgado-Rodriguez *et al.* [2] appear to have carried out fixed effects meta-analyses of each of these three outcomes despite finding [2, p 371] that the study populations were heterogeneous. We summarize their results in the top half of Table 1.

When the results are broken up by outcome there are too few relative risks in each subgroup for our methods of PB to be applicable with any degree of confidence, and so we further analyzed the complete

sets of results in Groups I and II using both random effects models and Bayesian hierarchical methods, and the results are in the bottom part of Table 1. Grouping these outcomes may be unacceptable, and consequently we advise caution in interpreting the results below, even though the random effects and Bayesian methods are designed to allow for some variation between such outcomes [9].

We used a simple hierarchical model for both the random effects and the Bayesian model. If the log relative risk in study  $i$  is defined by  $Y_i$ , and the variance of  $Y_i$  is given by  $\sigma_i^2$ , then this model is

$$Y_i \sim N(\mu_i, \sigma_i^2),$$

$$\mu_i \sim N(\Delta, \tau^2)$$

so that the overall mean is  $\Delta$  and the inter-study variability is described by  $\tau^2$ . In the random effects model we estimate  $\tau^2$  using the method of DerSimonian and Laird [3, 9]; in the Bayesian model we use the method of DuMouchel [4, 12], with the  $\sigma_i^2$  taken as known and the priors on both  $\Delta$  and  $\tau^2$  being uninformative.

Our tests of both Group I and Group II show significant heterogeneity, in accord with the assessment in [2]. For results with this degree of heterogeneity both the random effects and Bayesian hierarchical methods are preferable to the fixed effects model [11], although we report the latter in Table 1 for comparison with the subgroup results of [2].

For the ‘good quality’ results that make up Group II, the random effects estimate of relative risk is not greatly different to the fixed effects estimate, but in taking account of the heterogeneity we get a result which is substantially less significant than that of the fixed effects model. The Bayesian model gives much more weight again to the effect of heterogeneity (the posterior mean of  $\tau^2$  is 0.6, whereas the DerSimonian-Laird frequentist estimate of  $\tau^2$  is 0.1), and this changes the posterior mean estimate somewhat as well as widening the CI even further.

Most striking is the change in using random rather than fixed effects in Group I. Here the random effects and the Bayesian methods both show a halving of the excess risk, and the latter gives an insignificant overall excess risk, in contrast to the inference from the fixed effects model.

### 3 Assessing Publication Bias

Using a funnel plot approach, Delgado-Rodriguez *et al.* [2] claim to find symmetry, which suggests no PB in these studies. However, they curiously construct a funnel plot using relative risk on a logarithmic axis and do not invert standard error, which is an important measure of study variability. We find it difficult to make symmetry judgements using such axes. In contrast, using the more standard form of funnel plot from [8], we do see an indication of PB in Group II at least. In Figures 1 and 2 we give funnel plots of the data in Groups I and II. We judge that the funnel plot for Group II indicates potential PB, with the lack of results in the lower left corner being a classic indication [8] of suppression of low relative risk high variability results.

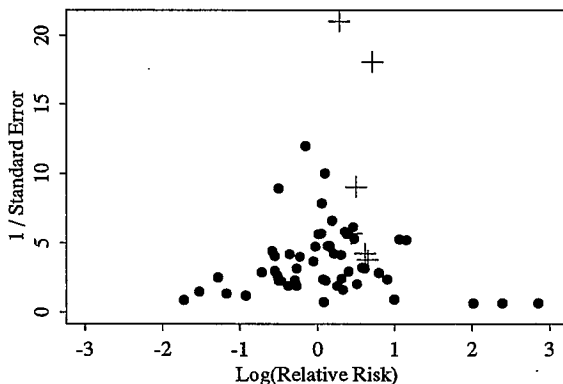


Figure 1: Funnel plot for the 62 cervical cancer results in Group I. The six results omitted for the analyses in Table 3 are designated by crosses. One cross is obscured.

Following the method of [5], which is a Bayesian approach related to the frequentist ideas in [1] and [6], we *augment* the data set by ‘filling in’ the missing results on the assumption that suppression has taken place according to the  $p$ -value of the study. Specifically, we define the three  $p$ -value intervals

$$I_1 = [0, 0.01], \quad I_2 = (0.01, 0.05], \quad I_3 = (0.05, 1]$$

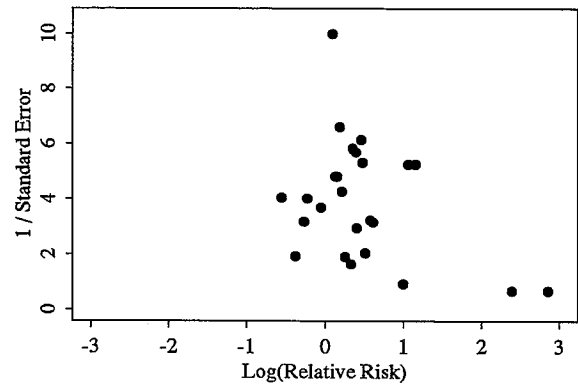


Figure 2: Funnel plot for the 26 cervical cancer results in Group II

and assume there is a probability  $w_j$  that a study with  $p$ -value in  $I_j$  was actually published. If  $m_j$  is the number of missing results with  $p$ -value in  $I_j$ , and if  $n_j$  is the number of observed results with  $p$ -value in  $I_j$ , then this implies that

$$m_j \sim \text{NegBin}(n_j, w_j), \quad j = 1, 2, 3.$$

In particular we note then that the expected number of missing results with  $p$ -value in  $I_j$  is  $n_j[1 - w_j]/w_j$  under this model, so that we will always expect it to predict *some* missing results unless  $w_j$  is one. As developed in [5], the method then augments the data set with these results, placed in a manner consistent with the data, the prior distributions, and the underlying model, and we can estimate an overall relative risk by essentially combining the real data and the estimated augmentations.

In the Bayesian framework we place the prior distributions indicated in Section 2 on  $\Delta$ ,  $\sigma_i^2$  and  $\tau^2$ . We also assume identical empirical priors for the unknown (augmenting) variances  $\sigma^2$ , based on the observed  $\sigma_i^2$ . We assumed uniform priors on the  $w_j$  over the ranges indicated in each case. The model is fitted by Gibbs sampling: for details see [5].

Table 2 contains details of several analyses of the Group II results. We give the posterior mean and 95% credible interval in each case, and also the posterior mean of the total number of results estimated to be present in each of the  $p$ -value intervals (i.e. the observed results  $n_j$  plus the posterior mean of the missing number  $m_j$ ).

The Bayes analysis is taken from Table 1. In Monotone Analysis A we assume that  $w_1 = 1.0$  (i.e. that

Table 2:  
Analysis of Publication Bias in Group II

Type of Analysis	Posterior RR	95% CI	$n_1 + m_1$	$n_2 + m_2$	$n_3 + m_3$
Bayes (no PB)	1.46	(1.08, 1.94)	4	7	15
Monotone Analysis A	1.29	(1.07, 1.53)	4	8.76	23.81
Monotone Analysis B	1.22	(0.99, 1.61)	4	10.45	39.40
Structured Priors	1.28	(1.06, 1.56)	4.13	7.65	22.32

all highly significant results are published), and that the priors on both  $w_2$  and on  $w_3$  are uniform on  $[0.5, 1.0]$  (so that the mean of each is 0.75, and thus the prior means of the associated negative binomial variables are  $E(m_2) = n_2(1 - w_2)/w_2 = 7/3$  and  $E(m_3) = n_3(1 - w_3)/w_3 = 5$ ). The monotonicity being enforced in this case is that the values of  $w_j$  must satisfy  $1 = w_1 \geq w_2 \geq w_3$ , so that even though the priors allow for both  $w_2, w_3$  to range over  $[0.5, 1.0]$  they can only do so subject to this constraint: see [5] for details.

We note that in this analysis the posterior means are  $m_2 = 1.76$  and  $m_3 = 8.81$ , so that the data have moved the posterior mean in the second interval down and that in the third interval up considerably from their prior values. This indicates that indeed there may be more missing results corresponding to  $I_3$  than was predicted by the prior.

The overall effect of the data augmentation is to almost halve the posterior mean excess risk, from 0.46 to 0.29. Figure 3 illustrates the change in the posterior distributions for the un-augmented and the augmented data sets, which is striking even though the 95% posterior credibility intervals still indicate significance in both cases.

In Monotone Analysis B we assume that the priors on both  $w_2$  and on  $w_3$  are uniform on  $[0.2, 1.0]$ , so that now the prior mean of each is 0.6, with the means of the associated negative binomial variables at this value being  $E(m_2) = 14/3$  and  $E(m_3) = 10$ .

In this case the data again drive the posterior mean of  $m_3$  much higher, to 24.40, although it lowers the posterior mean of  $m_2$  slightly to 3.45 from its prior. The net effect is to lower the estimate of relative risk slightly further again, indicating that the numerous missing results are being assigned low values. However, they are results of low precision, and so they are not heavily weighted in the overall estimate.

Thirdly, in the Structured Priors model, we dropped the monotonicity restriction but sought to provide the same type of ordering by making the priors in the first two intervals rather more restrictive.

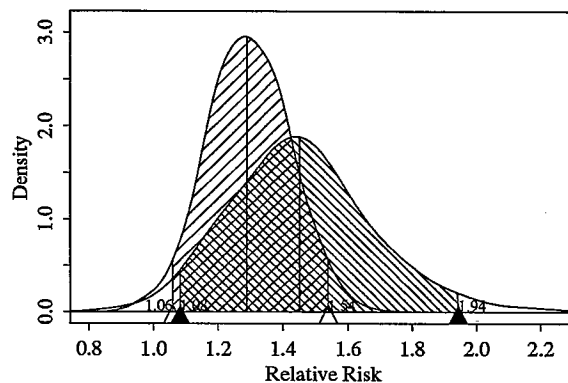


Figure 3: Relative risk posteriors for the 26 cervical cancer results in Group II. The left posterior was calculated using data augmentation, and the rightmost one assumes no publication bias.

We assume the priors on both  $w_1$  and on  $w_2$  are uniform on  $[0.85, 1.0]$ , and that on  $w_3$  is (as in Monotone Analysis A) uniform on  $[0.5, 1.0]$ . As can be seen, the posterior means on the relative risk and on the missing numbers in the three intervals are very close to those of Monotone Analysis A, as we might expect intuitively.

## 4 Suppressed results and missing results

We conclude from these sensitivity analyses that in Group II there are relatively few missing results with  $p$ -values in  $I_1$  and  $I_2$  but that there may be a substantial number in  $I_3$ , corresponding to the set of nominally insignificant results: since the posterior mean of  $m_3$  is considerably higher than the prior mean we conclude that this is not just an artifact of the model.

As discussed in the Introduction, the analysis we have just carried out is on a subset of the data, as-

Table 3  
Analysis of Publication Bias in Group I restricted to 56 Studies

Type of Analysis	Posterior RR	95% CI	$n_1 + m_1$	$n_2 + m_2$	$n_3 + m_3$
Monotone Analysis B (Table 2)	1.22	(0.99, 1.61)	4	10.45	39.40
Bayes (no PB)	1.07	(0.89, 1.26)	4	10	42
Monotone Analysis C	1.08	(0.88, 1.28)	4	11.20	73.08
Unstructured Priors	1.17	(0.83, 1.58)	12.24	10.64	71.10
Complete Data Set	1.13	(0.95, 1.34)	10	10	42

sessed in [2] to be of ‘higher quality’. Thus one explanation for the number of missing results might well be that they can be equated with the *known* results which are in Group I but have been excluded from Group II.

To evaluate this possibility we considered a set of 56 results from all results in Group I, where we suppress only those six results with  $p$ -values in  $I_1$  which were not in Group II. We do this since our monotone PB algorithm does not allow for any missing results from  $I_1$ , and it is clear that these would not be reproduced by the algorithm, although those in the other two intervals may be.

We see that in these 56 results, the distribution over the  $p$ -value intervals is startlingly close to those of Monotone Analysis B in Table 2: in forming Group II there were 3 results actually suppressed in  $I_2$  against a mean posterior estimate of 3.45, and 27 actually suppressed in  $I_3$  against a mean posterior estimate of 24.4.

For the 56 results we then ran Bayesian meta-analyses both allowing and not allowing for PB. The results are in Table 3.

The Bayesian analysis gives a mean posterior estimate of overall relative risk which is considerably lower than that calculated following the augmentation of the Group II data by the PB algorithm. Hence we deduce that the augmented results have typically been placed with a distribution reasonably consistent with that of the initial 26, whereas the real suppressed group in this set of 56 has a rather lower mean than either the original or the augmented group.

We also assess whether there is support for further publication bias in the 56 results. Here Monotone Analysis C was run on the same basis as Monotone Analysis B (i.e. priors for  $w_2$  and  $w_3$  uniform on  $[0.2, 1]$ ), and the Unstructured Priors analysis was not monotone, and allowed priors for  $w_1, w_2$  and  $w_3$  which were all uniform on  $[0.2, 1]$ . When we allow the data to decide on these further missing results,

of course it estimates that some do exist, since our priors on  $w_j$  tend to force this.

However, in Monotone Analysis C, although a mean of 32 extra results were estimated to exist, the posterior distribution of the relative risk is virtually unchanged by these augmentations. We conclude that the augmentations essentially replicate the data, and are probably an artifact of the priors used: that is, there is little to indicate that more results are missing.

Most interestingly, in the Unstructured Priors analysis, we virtually exactly replace the 6 results omitted in the 56 results: as noted in Table 3, we have 8.24 extra results to replace these 6, and although we have too many extra results in  $I_3$ , these have very little weight and make rather little difference to the estimate of the relative risk.

## 5 Conclusions

Overall, we conclude that in the Group II data set, the Bayesian augmentation method has identified substantial publication bias in  $I_2$  and  $I_3$ . These missing results seem to be rather similar in structure to those actively suppressed by Delgado-Rodriguez *et al.* [2], and there is little evidence that there are substantial further missing results: that is, their literature search has been successful in constructing the Group I set. Our method is sensitive to the priors used for the probabilities  $w_j$  of suppression, and it appears that these should be wide in this case (based on reproducing the Group I results): however, the use of a sensitivity analysis such as that in Table 2 seems advisable.

When carrying out a meta-analysis, the suppression of so many results may be a real danger: we have seen that the suppressed results in this context are reasonably consistent with the ‘gaps’ in the distribution of the unsuppressed results, so it is perhaps arguable that they are providing valid results despite any perceived lack of quality. Moreover, their

suppression leads to an overall estimate of the effect of oral contraceptives that is stronger than that obtained if all of these results are incorporated.

Of considerable practical interest is the suppression of the 6 highly significant results in  $I_1$ . These results will not (and almost certainly should not) be detected by an algorithm for publication bias, since they are different in kind from the observed results in Group II, so that internal data in Group II cannot predict their presence. It should be noted that when we allow missing results in  $I_1$  in the 56 results of Group I, we do get noticeable agreement with the known suppressed results, which may be taken as an indication that these large results do fill 'gaps' in the study distribution.

Large studies typically have an influence of considerable magnitude on meta-analyses. It is a matter of epidemiological and medical rather than purely statistical judgment whether to include or exclude them, but we feel that only a strong justification should be used to exclude such influential (in a statistical sense) results.

Epidemiologically, the message here is twofold. Firstly, the use of fixed effects models seems quite inadequate, independently of any PB considerations; and secondly, there is an indication that the effect of excluding the 'low quality' results has substantially biased the results. The use of all the Group I results (or the Group II results after allowing for missing ones) shows virtually no support for an association of oral contraceptive use with incidence of cervical cancer, in contrast to the positive if rather weak support in the Group II results. Further research is clearly needed to warrant the conclusion that there is a real association here.

## References

- [1] K.B.G. Dear and C.B. Begg. An approach for assessing publication bias prior to performing a meta-analysis. *Stat. Science*, 7:237-245, 1992.
- [2] M. Delgado-Rodriguez, M. Sillero-Arenas, J.M. Martin-Moreno, and R. Galvez-Vargas. Oral contraceptives and cancer of the cervix uteri. *Acta Obstet. Gynecol. Scand.*, 71:368-376, 1992.
- [3] R. DerSimonian and N.M. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177-188, 1986.
- [4] W. DuMouchel. Bayesian meta-analysis. In *Statistical Methods for Pharmacology*, pp. 509-529. Marcel Dekker, New York, 1990. ed. D. Berry.
- [5] G.H. Givens, D.D. Smith, and R.L. Tweedie. Estimating and adjusting for publication bias using data augmentation in Bayesian meta-analysis. Technical Report 95/31, Department of Statistics, Colorado State University, 1995.
- [6] L.V. Hedges. Modeling publication selection effects in meta-analysis. *Stat. Science*, 7:246-255, 1992.
- [7] L.V. Hedges and I. Olkin. *Statistical Methods for Meta-analysis*. Academic Press, 1985.
- [8] R.J. Light and D.B. Pillemer. *Summing Up: the Science of Reviewing Research*. Harvard Univ. Press, 1984.
- [9] K.L. Mengersen, R.L. Tweedie, and B.J. Biggerstaff. The impact of method choice in meta-analysis. *Australian J. Statistics*, 37:19-44, 1995.
- [10] F. Mosteller and T. Chalmers. Some progress and problems in meta-analysis of clinical trials. *Stat. Science*, 7:227-236, 1992.
- [11] NRC Committee on Applied and Theoretical Statistics. *Combining Information: Statistical Issues and Opportunities for Research*. National Academy Press, Washington, 1992.
- [12] R.L. Tweedie, D.S. Scott, B.J. Biggerstaff, and K.L. Mengersen. Bayesian meta-analysis, with application to studies of ETS and lung cancer. *Lung Cancer*, 14:S171-S194, 1996. Suppl. 1.
- [13] J.P. Vandenbroucke. Passive smoking and lung cancer: a publication bias? *Br. Med. J.*, 296:391-392, 1988.