Just the History from The combining of information: Investigating and synthesizing what is possibly common in clinical observations or studies via likelihood.

Keith O'Rourke

Department of Statistics,

University of Oxford,

April, 2007

Abstract

The combining of information: Investigating and synthesizing what is possibly common in clinical observations or studies via likelihood.

A thesis submitted by Keith O'Rourke of Worcester College towards a D.Phil. degree in the Department of Statistics, University of Oxford, Trinity Term, 2003 and passed in Trinity Term, 2007.

The thesis is to develop an analytical framework for a flexible but rigorous model based investigation and synthesis of randomized clinical trials - regardless of outcome measure, probability model assumed or published summary available. This involves the identification of relevant statistical theory, the development and adaptation of necessary techniques and the application of these to a number of examples.

A new strategy for the investigation and synthesis of RCTs regardless of outcome measure, probability model assumed or published summary available was developed to accomplish this. No such general strategy has been explicitly set out before. It provides a quite general method, and with adequate sample information, results in arguably correct and adequate techniques for the assumptions made.

A new method of numerical integration was developed to incorporate flexible random effects models; an importance sampling approach was developed to obtain the needed observed

summary likelihoods and a Monte Carlo based diagnostic to assess the adequacy of sample information was produced but remains to be further researched.

Contents

1 Short history of likelihood for meta-analysis

1.1 Pre-Fisher

Meta-analysis, or at least the combination of observations "not necessarily made by the same astronomer and or under different conditions", figured prominently in the initial development of statistical theory in the 18th and 19th centuries[29][14][21]. There was a belief (or at least a hope) that something could be gained from combining the observations however, exactly how the observations should be combined to achieve exactly what gain was far from obvious. This spurred the development of "Bayesian-like" methods that utilized the likelihood as a means of combining observations and offered justifications of this method of combination as providing the most "probable" true value (though originally conceived somewhat less directly as the most probable error of measurement made in the observations). The justifications though, were not as well formalized and understood as current Bayesian justifications of, for instance, the most probable value of an unknown parameter, given an explicitly known prior probability distribution for that unknown parameter[14]. The confusion in the justifications in fact, was wide-spread before and some time after Fisher's thesis of 1912[8], when according to Hald, Fisher vaguely drew attention to some of the difficulties. An exception may have been Keynes' 1911 paper[18] which will be mentioned below where the role of various assumptions was very clearly delineated.

Often, justifications of intuitively reasonable combinations that involved either the mean or various weighted means of multiple observations were argued about and sought. In fact, the

early attempts to justify combinations based on "likelihood" were largely abandoned when the mathematical analysis under the "primitive" probability models assumed for the observations was found intractable at the time[14]. One early attempt was by Daniel Bernoulli in a 1778 paper entitled "The most probable choice between several discrepant observations and the formation therefrom of the most likely induction." which was reprinted in Biometrika 1961[17].

In the English translation, most references to what was being combined in the paper were to observations, but the term observers was also used - suggesting that observations were not highly distinguished as coming from the same or different investigators/studies. In this he enunciated a principle of maximum likelihood "of all the innumerable ways of dealing with errors of observations one should choose the one that has the highest degree of probability for the complex of observations as a whole." He believed that the mean was a poor combination using intuitive arguments that observations further from the "centre" should be given less weight, except in the case where the observations were believed to be Uniformly distributed - where he incorrectly (from a likelihood combination perspective) believed all observations should be equally weighted.

He assumed a semicircular distribution and derived the likelihood function as

$$L = \prod_{i=1}^{n} \sqrt{a^2 - (y_i - \mu)^2}$$

and tried to maximize L^2 with respect to μ but was unable to do this for more than 2 observations, as it lead to an equation of the fifth degree. For just 2 observations, it was maximized by the mean. For some numerical examples with 3 observations he noted that L^2 was maximized by weighted means. The idea of using a probability model to determine the best combination was definitely there, and he did realize that the probability of individual independent observations multiplied to provide the joint probability of the complex of observations. Interestingly, he actually used the smallest observation as the origin i.e.

$$y_i - \mu = \left(y_i - y_{(1)}\right) - \left(\mu - y_{(1)}\right)$$

which emphasizes the correction that needs to be added to $y_{(1)}$ - $u - y_{(1)}$ as the unknown. Unfor-

tunately for him, if instead of $\partial L^2/\partial u = 0$ he had used $\partial \log L/\partial u = 0$ he would have found

$$\begin{aligned} \sum \frac{y_i - \mu}{a^2 - (y_i - \mu)^2} &= 0 \\ \sum \frac{y_i}{a^2 - (y_i - \mu)^2} - \mu \sum \frac{1}{a^2 - (y_i - \mu)^2} &= 0 \end{aligned}$$

$$\begin{aligned} \mu \sum \frac{1}{a^2 - (y_i - \mu)^2} &= \sum \frac{y_i}{a^2 - (y_i - \mu)^2} \\ \mu &= \sum \frac{y_i}{a^2 - (y_i - \mu)^2} / \sum \frac{1}{a^2 - (y_i - \mu)^2}. \end{aligned}$$

This shows that $\hat{\mu}$ is the weighted average of the observations, the weights being the reciprocal of the squared density, that is, increasing with the distance from μ. It is also unfortunate that he did not consider multiplying individual observation likelihoods assuming $Uniform(u - h, u + h)$ with h known, as the mathematics is simple and the best combination involves only the most extreme observations on each side of the centre – about the most different combination from the one he intuitively thought best in this case (the equally weighted mean which puts equal weight on all observations).

Somewhat later, circa 1800, Laplace and Gauss fully investigated the multiplying of probabilities of individual observations as the means of combination of observations given a common parameter[21]. Laplace initially concentrated on the probable errors, and often specifically the most probable error, given all the observations (and a more or less explicit assumption of a prior uniform distribution on the possible errors). Gauss moved towards concentrating on the probable values rather than errors and specifically the most probable value given all the observations (and a very explicit assumption of a prior uniform distribution on the possible values). Gauss was also perhaps the first with some real practical success. He reversed the reasoning that Bernoulli had used earlier – rather than trying to establish that the mean is the best combination for some "motivated by first principles" distribution, and he found the distribution for which "likelihood multiplication" would determine that the best combination was the mean. According to Hald, he did not check the distribution empirically[14].

In 1839 Bienayme had remarked that the relative frequency of repeated samples of binary outcomes often show larger variation than indicated by a single underlying proportion and proposed a full probability-based random effects model (suggested earlier by Poisson) to account for this.

Here, the concept of a common underlying proportion was replaced by a common distribution of underlying proportions. It is interesting that a random effects model where what is common in observations is not a parameter, but a distribution of a parameter, followed so soon after the development of likelihood methods for combination under the assumption of just a common parameter.

The 1911 paper of Keynes mentioned above, acknowledged and revisited Gauss's derivation of the *Normal* distribution as the only symmetric distribution whose best combination was the mean and also investigated this for the median and the mode. Here, simply for interest in itself, the result of the *Normal* distribution as the only symmetric distribution whose best combination is the mean, is presented in modern form but following that given in Keynes' paper.

The assumption of a symmetric distribution obviously does not imply that

$$f(y_i;\mu) = Be^{\theta(\mu-y_i)^2}.$$

It is required to show that

$$\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\mu} f(y_i,\mu)}{f(y_i;\mu)} = 0 \text{ being equivalent to } \sum_{i=1}^{n} (\mu - y_i) = 0$$

along with symmetry does imply this.

The most general form of $\sum_{i=1}^{n} (\mu - y_i) = 0$ is $\sum_{i=1}^{n} g(\mu)(\mu - y_i) = 0$, where g is an arbitrary function of μ. Assuming $g(\mu)$ to be twice differentiable, without loss of generality, one may write $g(\mu) = \varphi''(\mu)$. Since y_i is arbitrary, the equivalence requires

$$\frac{\frac{\partial}{\partial\mu} f(y_i,\mu)}{f(y_i;\mu)} = \varphi''(\mu)(\mu - y_i)$$

or

$$\log f(y_i;\mu) = \int \varphi''(\mu)(\mu - y_i)\, d\mu + \psi(y_i),$$

where $\psi(y_i)$ is an arbitrary function of y_i. Integration by parts gives

$$\log f(y_i;\mu) = \varphi'(\mu)(\mu - y_i) - \varphi(\mu) + \psi(y_i).$$

Now, it is required that $f(y_i;\mu)$ be symmetric about μ, *i.e.* invariant under $y \to 2\mu - y$. Thus

$$\varphi'(\mu)(\mu - y_i) - \varphi(\mu) + \psi(y_i) = \varphi'(\mu)(y_i - \mu) - \varphi(\mu) + \psi(2\mu - y_i)$$

or

$$2\varphi'(\mu)(\mu - y_i) = \psi(2\mu - y_i) - \psi(y_i).$$

Taylor expand $\psi(2\mu - y_i)$ about $\mu = y_i$.

$$\psi(2\mu - y_i) = \psi(y_i) + 2\psi'(y_i)(\mu - y_i) + 2\psi''(y_i)(\mu - y_i)^2 + \sum_{j=3}^{\infty} \frac{2^j}{j!}\psi^{(j)}(y_i)(\mu - y_i)^j,$$

which results in

$$\varphi'(\mu)(\mu - y_i) = \psi'(y_i)(\mu - y_i) + \psi''(y_i)(\mu - y_i)^2 + \sum_{j=3}^{\infty} \frac{2^{j-1}}{j!}\psi^{(j)}(y_i)(\mu - y_i)^j,$$

which simplifies to

$$\varphi'(\mu) = \psi'(y_i) + \psi''(y_i)(\mu - y_i) + \sum_{j=3}^{\infty} \frac{2^{j-1}}{j!}\psi^{(j)}(y_i)(\mu - y_i)^{j-1}.$$

But $\varphi'(\mu)$ is a function of μ alone for arbitrary y_i, which implies that $\psi''(y_i) =$ a constant along with $\psi'(y_i) - y_i\psi''(y_i) = 0$, which implies that $\psi(y_i) = ky_i^2$. Then

$$\varphi'(\mu) = 2k\mu \quad \Rightarrow \quad \varphi(\mu) = k\mu^2 + C.$$

Substituting in the equation for $\log f((y_i;\mu))$,

$$\begin{aligned} \log f(y_i;\mu) &= 2k\mu(\mu - y_i) - k\mu^2 - C + ky_i^2 \\ &= k(\mu - y_i)^2 - C, \end{aligned}$$

or

$$f(y_i;\mu) = Ae^{k(\mu - y_i)^2}.$$

Note that

$$\int f(y_i;\mu)\, dy_i = 1 \quad \Rightarrow \quad k < 0.$$

Keynes' paper provides a good indication of the central role played by the combination of observations in statistics prior to Fisher. Apparently though, only Keynes, Gauss, Laplace and perhaps a few others were fully aware of the need for, and arbitrariness of, a prior distribution for the probability justifications for the combination, and both Gauss and Laplace became at some point uncomfortable with this and turned to sampling distribution-based justifications instead[14]. In particular, Gauss developed optimal combinations based on a restriction to unbiased linear combinations of unbiased estimates (i.e. least squares or inverse variance weighted combinations). This approach allowed for varying but known differences in the variances of the estimates and implicitly assumed the estimates and variance were uncorrelated so that weighted averages of unbiased estimates would give unbiased combinations (which is, of course, trivially true for known variances).

Somewhat later, based on Laplace's expositions of his own and Gauss's work, Airy made an extension for the estimation of unknown variances in 1861[1]. He also made a related extension to Bienayme's "random effects model" by developing methods based on within day and between day variances of observations to allow for imperfect but partial replication of independent estimates. Consideration of the within day sampling errors had shown that in some applications, observations on different days were not in fact replicating in the usual sense – they had larger variations than would be expected from the within day sampling errors. It was conceptualized that there was some unknown day error and that some allowance should be made for this.

The two-stage summary approach to meta-analysis used today is close to this approach, but where the implicit assumption is often violated as, for instance, with effect measures which are slightly correlated with their variances.[15] Fisher though, as we will see below, returned to likelihood (separated from the prior) and again provided arguments for likelihood multiplication as the "best" basis for combining observations in the early 1900s. Pearson wrote an editorial on Airy's book[21] and Fisher, as a graduate student, either studied Airy's book or related ones on the combination of observations[14]. Pearson meta-analysed medical examples in the early 1900s, drawing attention to opportunities suggested by the heterogeneous study outcomes. Fisher and Cochran meta-analysed agricultural trials in the 1930s[11][3]. Fisher drew attention to the need to carefully consider the reasons for less than perfect replications between trials (i.e. whether in fact it was a treatment interaction with place and time or differing measurement errors) and various ways of dealing with it for different inferential purposes. It apparently is one of Fisher's few publications on

random effects models (private conversation with D. Sprott and J. A. Nelder). Cochran explicated the full $Normal - Normal$ random effects model with a likelihood-based meta-analysis in 1937. Further details are given in O'Rourke[21].

1.2 Early Fisher

In some ways, perhaps most interesting of all, Fisher in his 1925[9] and 1934[10] papers in which he mainly developed his theory of statistics, thought through the issues of multiple experiments when addressing the loss of information when summarizing data. In the 1925 paper, he points out that if there is no loss of information in a summary (i.e. when there are sufficient statistics) then the summary of two combined samples from the same population must be some function of the two summaries of the individual samples without recourse to the individual observations from either sample. He then concludes the paper with a section on ancillary statistics whose purpose was defined as providing a true, rather than approximate, weight for combining the multiple individual sample summaries.

In the case of a [small] number of large samples, he shows that the likelihood from all the individual observations collected from all the samples can be recovered from the MLEs of the multiple individual samples via a weighted average of those MLEs with weights equal to the observed information (second derivative of the log-likelihood evaluated at the MLE) of each individual sample. Essentially this is because, for large samples, the log-likelihoods are approximate quadratic polynomials and their addition only involves their maximums (MLEs) and curvatures (observed informations evaluated at the MLE essentially estimated without error and taken as known).

Following Hald[14] and using modern notation

$$l = \log L(\theta, y_{all}) = \sum_k \log L(\theta, y_k) = \sum_k l_k$$

and therefore

$$l^{'}(\widehat{\theta}) = \sum_k l_k^{'}(\widehat{\theta}).$$

By Taylor series expansion about $\widehat{\theta}_k$

$$l_k^{'}(\widehat{\theta}) = l_k^{'}(\widehat{\theta}_k) + (\widehat{\theta} - \widehat{\theta}_k) l_k^{''}(\widehat{\theta}_k) + ...$$

$$\sum_k l_k^{'}(\widehat{\theta}) = \sum_k l_k^{'}(\widehat{\theta}_k) + \sum_k (\widehat{\theta} - \widehat{\theta}_k) l_k^{''}(\widehat{\theta}_k) + ...$$

$$\sum_k l_k^{'}(\widehat{\theta}) = \sum_k 0 + \sum_k (\widehat{\theta} - \widehat{\theta}_k) l_k^{''}(\widehat{\theta}_k) + ...$$

$$\sum_k l_k^{'}(\widehat{\theta}) = \sum_k (\widehat{\theta} - \widehat{\theta}_k) l_k^{''}(\widehat{\theta}_k) + ...$$

but $\sum_k l_k^{'}(\widehat{\theta}) = 0$ so

$$\sum_k (\widehat{\theta} - \widehat{\theta}_k) l_k^{''}(\widehat{\theta}_k) \approx 0$$

$$\widehat{\theta} \sum_k l_k^{''}(\widehat{\theta}_k) - \sum_k \widehat{\theta}_k l_k^{''}(\widehat{\theta}_k) \approx 0$$

$$\widehat{\theta} \sum_k l_k^{''}(\widehat{\theta}_k) \approx \sum_k \widehat{\theta}_k l_k^{''}(\widehat{\theta}_k)$$

$$\widehat{\theta} \approx \frac{\sum_k \widehat{\theta}_k l_k^{''}(\widehat{\theta}_k)}{\sum_k l_k^{''}(\widehat{\theta}_k)}$$

Since each of the estimates $\widehat{\theta}_k$ is asymptotically $Normal(\theta, 1/nI)$, the combination based simply on the unweighted average $\overline{\theta} = \sum_k \widehat{\theta}_k/m$ would have variance $1/mnI$. Note, however, that the above combination recovers the likelihood from the full data and the $\widehat{\theta}$ from this is asymptotically $Normal(\theta, 1/l^{''}(\widehat{\theta}))$.

The advantage is perhaps more easily seen in terms of variances from the finite sample version given by Rao[24] -

> "Suppose that we have two independent samples X and Y, giving information on the same parameter θ, from which estimates $T_1(x)$ and $T_2(y)$ obtained are such that
>
> $$\begin{aligned} E[T_1(X)] &= E[T_2(Y)] = \theta, \\ V[T_1(X)] &= v_{1,} V[T_2(Y)] = v_{2,} \end{aligned}$$
>
> where v_1 and v_2 are independent of θ. Further, suppose that there exist statistics $A_1(X)$ and $A_1(Y)$ such that
>
> $$\begin{aligned} E[T_1|A_1(X) &= A_1(x)] = \theta, \\ E[T_2|A_2(Y) &= A_2(y)] = \theta, \end{aligned}$$

$$V[T_1|A_1(X) = A_1(x)] = v_1(x),$$
$$V[T_2|A_2(Y) = A_2(y)] = v_2(y),$$

where x and y are observed values of X and Y, respectively, and $v_1(x)$ and $v_2(y)$ are independent of θ. Then, we might consider the conditional distributions of T_1 and T_2 given A_1 and A_2 at the observed values and report the variances of T_1 and T_2 as $v_1(x)$ and $v_2(y)$, respectively, as an alternative to v_1 and v_2. What is the right thing to do?

Now, consider the problem of combining the estimates T_1 and T_2 using the reciprocals of v_1, v_2 and $v_1(x)$, $v_2(y)$ as alternative sets of weights:

$$t_1 = (\frac{T_1}{v_1} + \frac{T_2}{v_2})/(\frac{1}{v_1} + \frac{1}{v_2}),$$

$$t_2 = (\frac{T_1}{v_1(x)} + \frac{T_2}{v_2(y)})/(\frac{1}{v_1(x)} + \frac{1}{v_2(y)}).$$

It is easy to see that the unconditional variances of t_1 and t_2 satisfy the relation

$$V(t_1) \geq V(t_2)$$

[by application of the Guass-Markov Theorem, conditional on x and y]."

In the 1934 paper, he addressed the same question for small samples (where the log-likelihoods can be of any form) and concluded that, in general, single estimates will not suffice but that the entire course of the likelihood function would be needed. He then defined the necessary ancillary statistics in addition to the MLE in this case as the second and higher differential coefficients at the MLE (given that these are defined). These would allow one to recover the individual sample log-likelihood functions (although he did not state the conditions under which the Taylor series approximation at a given point recovers the full function - see Bressoud[2]) and with their addition, the log-likelihood from the combined individual observations from all the samples.

The concept of ancillary statistics has changed somewhat since - in fact very soon afterwards, as a year later Fisher treated "ancillary" as a broader term of art not specifically wedded to local behavior of the likelihood function[30]. This was its original conceptualization though – how to "correctly" (without loss of information) combine results from separate sample summaries, given a choice of what the separate sample summaries should be but no access to the individual observations in the separate samples. Here, "correctly" is defined as getting some multiple of the likelihood function from all the observations but with access only to the collection of summaries.

It is perhaps tempting to suggest that Fisher's key ideas in his theory of statistics (the breadth of which is for instance reflected in Efron's claim that modern statistical theory has added only one concept, that of invariance, which is not well accepted[7]) arose from his thinking of statistics

as the combination of estimates. Fortunately for us, Fisher as much said so in a 1935 paper read at the Royal Statistical Society[12]. In discussing overcoming the preliminary difficulty of multiple criteria for judging estimates – better for what? – he argued

> "Whatever other purpose our estimate may be wanted for, we may require at least that it shall be fit to use, in conjunction with the results drawn from other samples of a like kind, as a basis for making an improved estimate. On this basis, in fact, our enquiry becomes self contained, and capable of developing its own appropriate criteria, without reference to extraneous or ulterior considerations."

And later in the next paragraph –

> " . . . , where the *real* problem of finite samples is considered, the requirement that our estimates from these samples may be wanted as materials for a subsequent process of estimation [combined somehow with results drawn from samples of a like kind?] is found to supply the unequivocal criteria required." [italics in the original]

1.3 Late Fisher

Fisher continued to exhibit numerous references to multiple estimates or studies in his 1956 book Statistical Methods and Scientific Inference[13]. For instance, on page 75, he states

> "It is usually convenient to tabulate its [the likelihoods] logarithm, since for independent bodies of data such as might be obtained by different investigators, the "combination of observations" requires only that the log-likelihoods be added."

On page 163 he further notes

> "In practical terms, if from samples of 10 two or more different estimates can be calculated, we may compare their values by considering the precision of a large sample of such estimates each derived from a sample of only 10, and calculate for preference that estimate which would at this second stage [meta-analysis stage] give the highest precision."

Finally on page 165 he concludes

> ". . . it is the Likelihood function that must supply all the material for estimation, and that the ancillary statistics obtained by differentiating this function are inadequate only because they do not specify the function fully."

Given this, it is suggested that Fisher considered the theory of estimation as validly based on the idea of retaining "all" of the likelihood in the estimates "summarized" from studies so that the

overall likelihood-based on the individual observations from similar studies could be re-constituted by just using the studies' estimates. This metaphor or model of estimation was continually referred to through many of his publications - though perhaps even few familiar with Fisher's work have noticed that (AWF Edwards, private communication). Fisher was even cited as being the main impetus for one of the earliest papers on p-value censorship[28]. There is some note of it given in Savage [26], which suggested to the author that Fisher's papers should be reviewed for this, and also in Rao[24].

In conclusion, the early development of statistics in the context of combination of observations and Fisher's numerous and continued references to multiple estimates or summaries in his statistical writing suggests that statistical theory should be easily relatable to meta-analysis as some of the roots and elaborations of statistical theory were based on meta-analytical considerations.

1.4 Post-Fisher

The history chapter in this thesis started with the combination of observations made by different astronomers and geodesists in the late 1700's and early 1800's and then concluded with some excerpts from Fisher's 1956 book. Unfortunately, the quantitative combination of estimates from randomized clinical trials was quite rare before about 1980 so there is a need to bridge the gap. Meta-analysis for psychological and educational research started somewhat earlier, and by 1976 Glass highlighted the desirability of the tradition of combining estimates from different studies and apparently first coined the term meta-analysis. Some authors argue that meta-analysis methods for clinical research were initially based on this activity in psychological and educational research. In educational and psychological research however, studies would very often use different outcomes or scales, and to this end, Glass proposed the use of an index of effect magnitude that did not depend on the arbitrary scaling of the outcomes so that combining in some sense, made sense. Presumably, in response to this, Hedges and Olkin wrote a book[15] in 1985 directed (as the authors indicated) at providing different statistical methods from those of Fisher & Cochran that were designed to specifically deal with this new and different kind of meta-analysis - that of combining different outcomes using an index of magnitude. In 1990, Olkin[20], quoting Fisher, again highlighted this arguably different class of meta-analyses (which apparently are more common in psychology and education than clinical research) of determining the combined significance of independent tests on outcomes "that may be of very different kinds" (by combining their p_values.)[21].

Hedges and Olkin's book, although a substantial and now classic book for combining different outcomes using an index of magnitude, is somewhat out of place for the more usual situation encountered in clinical research where a series of randomized clinical trials have identical or very similar outcomes. Here Fisher and Cochran's methods would be arguably more appropriate. (With recent changes in clinical research, specifically the inclusion of Quality of Life measures which are comprised of various scales, this may be less the case for those outcomes.)

DerSimonian and Laird[6], published in 1986 what was perhaps one of the first "modern" papers on statistics for meta-analysis for randomized clinical trails. It drew on and referenced a 1981 paper[25] that W. G. Cochran was the senior author on (published posthumously) that was comprised of simulation studies of various estimators of combined estimates from Cochran's 1937 $Normal - Normal$ random effects model[3]. DerSimonian and Laird chose to adopt one the closed form non-iterative formulas from this paper and adapted it for binary outcomes. Two more methodological as well as statistical papers appeared in the next year - Sacks et al[16] and L'Abbe, Detsky and O'Rourke[19] (the author of this thesis). The authors of these three papers had been loosely collaborating since 1985. In particular, Chalmers had provided a draft of his quality scoring system and DerSimonian and Laird had provided their draft paper to the author when the L'Abbe group were developing their ideas and paper. There it was suggested that logistic regression be used for conducting meta-analyses of randomized two group experiments with binary outcomes as it provided a likelihood-based approach (the author was the statistician on the paper and wrote the statistical appendix for it). First, the logistic regression is set up to include an indicator term for study, a term for treatment group, and an interaction term (treatment by study). The indicator term for study allows a separate baseline estimate for each study so that each study's treatment effect estimate contribution is relative to its own control group. The treatment group term allows for a common treatment effect estimate and the interaction term allows for a separate treatment effect estimate for each individual study (the same as one would get using each study's data alone). The consistency of study results is then quantitatively analyzed by investigating the variation in the individual study treatment effect estimates and their confidence intervals and, less preferably, the statistical significance of omitting the interaction term in the logistic regression. A warning about the low power of this test was given along with a suggestion that clinical judgement was preferable. With the omission of the interaction term, a common "pooled" treatment effect is constructed along with estimates and likelihood ratio-based confidence intervals and tests. The likelihood for

the confidence intervals for the common treatment parameter τ is obtained by profiling out the within study baseline parameters c_i

$$L(y_1, ..., y_n; \tau, \hat{c}_1, ..., \hat{c}_n)$$

which is of course equal to

$$\prod_i L(y_i; \tau, \hat{c}_i)$$

as the $\hat{c}_i, s$ are mutually independent. Thus it was equivalent to the approach in this thesis, but with the marginal likelihood being immediately given by sufficiency and random effects neglected.

Random effects were later allowed for in a technical report[22] using a method from Cox and Snell[5] that Venables and Ripley claim was first suggested by Finney in 1971[31] and is now often referred to as quasi-likelihood - where the scale parameter, rather than being set equal to one, is estimated by the deviance or Pearson Chi-square statistic divided by the residual degrees of freedom. Quasi-likelihood though, is a much more general approach, not tied to specific estimates of scale. To second order, this scale estimate has the effect of simply increasing the standard error of the MLE as the MLE itself is unaffected. As reviewed in appendix E, Tjur gave reasons for preferring that the MLE be unaffected, which McCullagh was then easily able to set aside. Many authors though, simply reject this allowance for random effects by scale estimation as being *ad hoc*. Stafford's adjustment[27] was adopted earlier in this thesis, as it provides an asymptotic rationale for the allowance for random effects which may overcome such objections to its use. But, unless the likelihoods are essentially quadratic, as is usually the case with binary outcomes, it is unlikely to modify the fixed effect likelihoods to adequately approximate possibly true level 2 likelihoods.

In Statistics in Medicine in 1986[23], Richard Peto provided an explanation for a statistical method he had used in earlier applications. For ruling out the null hypothesis of no effect, he had used a test based on the unweighted sum of observed minus expecteds Oi – Ei, and for combined estimation of an odds ratio, he had used a weighted sum of Oi – Ei with the weights being the inverse variance of Oi – Ei. These quantities could be directly motivated as being quadratic approximations to maximum likelihood estimation under a conditional logistic regression model, as for instance was shown in Cox[4] and referenced by Peto[32] in 1985. Of course there is always more than one way to motivate a quantity – it is just suggesting this is one possible way.

In 1986, Peto emphasized entirely different justification of the use of Oi – Ei by starting with the

question “But why use observed minus expecteds rather than some logistic model?” His answer had two parts – one was that observed minus expecteds would be readily understandable to physicians and that it provided a typical estimate of odds ratios that did not depend on assumptions of the sort needed for logistic regression (although this does follow from assuming a conditional logistic regression model and approximating the conditional MLE by the score statistic – see O’Rourke[21]). Unfortunately, he did not define what he meant by “typical” nor the “depend[ence] on assumptions". Perhaps most strikingly, he dismissed the use of random effects models using very similar arguments that Fisher had used for the certain cases where Fisher thought random effects specifically should not be considered – see O’Rourke[21]. It is perhaps more tenuous to relate this Oi – Ei approach back to Fisher and Cochran than the approach of DerSimonian and Laird and L’Abbe, Destky and O’Rourke but more or less indirectly the methods of Fisher and Cochrane became central for the meta-analyses of randomized clinical trials.

The pressure for clinical researchers to actually carry out meta-analysis of randomized controlled trials in their various fields had been building perhaps soon after Archie Cochrane published an essay in 1979, in which he suggested that "It is surely a great criticism of our profession that we have not organized a critical summary, by speciality or subspecialty, adapted periodically, of all relevant randomized controlled trials" . In 1985, an international collaboration to prepare systematic reviews of controlled trials in the field of pregnancy and childbirth, resulting in the publication in 1989 of: "Effective Care in Pregnancy and Childbirth (ECPC): A Guide to Effective Care in Pregnancy and Childbirth (GECPC)", and "The Oxford Database of Perinatal Trials (ODPT)". Encouraged by the reception given to the systematic reviews of care during pregnancy and childbirth, Michael Peckham, first Director of Research & Development in the British National Health Service, approved funding for "a Cochrane Centre" to facilitate the preparation of systematic reviews of randomized controlled trials of health care, in 1992. Later that year, "The Cochrane Centre" opened in Oxford, UK. In 1993, an international and comprehensive concept of the Cochrane Collaboration was presented at a conference ("Doing more Good than Harm") organized by Kenneth Warren and Frederic Mosteller at the New York Academy of Sciences, and in June of that year the development of Cochrane Collaboration’s Handbook as a tangible means to facilitate the preparation of systematic reviews of randomized controlled trials of health care began with the arrival of the 1st Cochrane Visiting Fellow at the UK Cochrane Centre.

In 1993, a Cochrane Collaboration Workshop on statistical methods for data synthesis was

conducted and a report drafted. The list of participants included D. Altman, P. Armitage, C. Baigent, J. Berlin, M. Bracken, R. Collins, K. Dickersin, D. Elbourne, R. Gray, K. McPherson, A. Oxman, M. Palmer, R. Peto, S. Pocock, K. Schulz and S. Thompson, all of whom were statisticians, epidemiologists or physicians with expertise in statistical methods for data synthesis. The workshop was convened to develop guidelines on statistical methods for data synthesis for the Cochrane Collaboration's eventual handbook and to identify useful research topics in that area.

In the report, the deliberations are outlined and a set of implications for the Cochrane Collaboration are given. It was assumed that only published summary statistics would be available for the foreseeable future, although the preferability of having individual participant data was indicated. Issues of inclusion criteria for systematic reviews were not considered except for those having to do with methodological quality. There was a major discussion on effect measures with greatest emphasis on binary outcomes where the relative merits of odds ratio versus relative risk were discussed at length. Here the odds ratio was favoured as a default, but it was stated that the relative risk and risk difference should not be ruled out as options. Some felt the choice of effect measure should depend in some part on a test of heterogeneity, while others disagreed. Several participants felt it would be preferable to use different measures for presentation than were used for analyzing the data. Continuous outcome measures received much less attention with weighted mean differences being suggested as appropriate, along with the possible consideration of standardizing by the control group standard deviation (to get an "effect size"). Most felt the area merited deeper study - difficulties being anticipated about choice of effect measure, the issue of data distribution, use of medians rather than means, handling of before and after measurements, weighing of studies and missing data. Further research on these was recommended. The issue of binary and continuous data also arose with some suggestion of automatic transformation of continuous outcome to binary, but further study was recommended. Here some of the issues now resolved by this thesis were being identified and highlighted 20 years ago.

As for approaches to aggregation, many but not all, recommended the use of a test of heterogeneity with the issue of low power being identified as a concern along with a suggested Type I error level of .10 rather than the customary .05. As for aggregation, given the determination that "substantial" heterogeneity is not present, after some discussion and a suggestion that results would be similar, a fixed effect models was decide upon as the default approach. Considerable disagreement ensued, however, when the discussion turned to the preferred approach under condi-

tions of statistically demonstrable heterogeneity. Both random and fixed effect models had strong proponents. The report cautions that characterizing in a few words the differences between fixed and random effects proponents would be challenging.

Some claimed the fixed effect approach was "assumption free" and is not [should not be] directly influenced by heterogeneity while others claimed that it would produce an artificially narrow confidence interval, as it does not reflect between-trial variance. They suggested random effects did not make as stringent an assumption as there being no differences between the underlying true treatment effects in the individual trials and hence was preferable. Common ground under these widely contrasting views was then summarized : the analyst should attempt to explore the reasons for the heterogeneity and explain it, especially with regard to varying methodological quality, that the ruling out of an overall null hypothesis of no effect in all trials need not distinguish the alternative to be fixed or random but "at least one of the trials" has an effect; that whether heterogeneity was present or not, the fixed effect estimate is an informative average measure of treatment effect; and, finally, that as random effects methods have rather amorphous assumptions, it was an area requiring more research into the importance of the assumptions and robustness to them. Here, the pragmatic concern arose regarding random effects methods giving relatively more weight to smaller studies when these often are of poorer quality and more subject to publication bias.

The entire discussion regarding appropriate approaches for aggregation under conditions of heterogeneity pertained to binary data. The same general principles, however, were thought to apply to continuous data and it was mentioned that the same discussion about fixed versus random effects models had occurred many years ago, relative to continuous data. They felt they should acknowledge that various fixed and random effects approaches were available and that future research should compare DerSimonian and Laird's approach to those based on maximum likelihood methods.

This thesis (passed in 2007) provides a general approach for both discrete and continuous data, regardless of reported summaries, based on the observed summary likelihood. Additionally, DerSimonian and Laird's approach can be compared to likelihood methods using numerous assumed distributions for random effects. It is a bit surprising that it has taken 20 years for this to be undertaken.

References

[1] Airy, G. *On the algebraical and numerical theory of errors of observations and the combination of observations.* MacMillan and Co., Cambridge, 1861.

[2] Bressoud, D. *A radical approach to real analysis.* The Mathematical Association of America, Washington, 1994.

[3] Cochran, W. G. Problems arising in the analysis of a series of similar experiments. *Journal of Royal Statistical Society Supplement 4*, 1 (1937), 102–118.

[4] Cox, D. *The analysis of binary data.* Methuen, London, 1970.

[5] Cox, D. R., and Snell, E. J. *Analysis of binary data.* Chapman & Hall Ltd, 1989.

[6] DerSimonian, R., and Laird, N. Meta-analysis in clinical trials. *Controlled Clinical Trials 7*, 3 (1986), 177–88.

[7] Efron, B. 1996 R. A. Fisher lecture. *Statistical Science 13*, 2 (1998), 95–122.

[8] Fisher, R. On an absolute criterion for fitting frequency curves. *Messeng. Math 41* (1912), 155–160.

[9] Fisher, R. Theory of statistical estimation. *Phil. Trans., A 222* (1925), 309–368.

[10] Fisher, R. Two new properties of mathematical likelihood. *Journal of the Royal Statistical Society, Series A, General 144* (1934), 285–307.

[11] Fisher, R. *The Design of Experiments.* Oliver and Boyd, Edinburgh, 1935.

[12] Fisher, R. The logic of inductive inference. *Journal of the Royal Statistical Society 98* (1935), 39–54.

[13] Fisher, R. *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh, 1959.

[14] Hald, A. *A history of mathematical statistics from 1750 to 1930.* John Wiley & Sons, 1998.

[15] Hedges, L. V., and Olkin, I. *Statistical methods for meta-analysis.* Academic Press, Orlando, FL, 1985.

[16] H.S., S., Berrier, J., Reitman, D., Ancona-Berk, V., and Chalmers, T. Meta-analyses of randomized controlled trials. *N Engl J Med. 316* (1987), 450–455.

[17] Kendall, M., Bernoulli, D., Allen, C., and Euler, L. Studies in the history of probability and statistics: XI. Daniel Bernoulli on maximum likelihood. *Biometrika 48* (1961), 1–18.

[18] Keynes, J. The principal averages and the laws of error which lead to them. *Journal of the Royal Statistical Society 74* (1911), 322–331.

[19] L'Abbe, K. A., Detsky, A. S., and O'Rourke, K. Meta-analysis in clinical research. *Ann.Intern.Med. 107*, 2 (1987), 224–233.

[20] Olkin, I. History and goals. In *The future of meta-analysis*, Wachter and Straf, Eds. The Belknap Press of Harvard University Press, Cambridge, Massachusettes, 1990.

[21] O'Rourke, K. Meta-analytical themes in the history of statistics: 1700 to 1938. *Pakistan Journal of Statistics [Split into Series A and B, 1986-1994] 18*, 2 (2002), 285–299.

[22] O'Rourke, K., and et al. Incorporating quality appraisals into meta-analyses of randomized clinical trials. Tech. rep., University of Toronto. Dept. of Statistics, 1991.

[23] Peto, R. Discussion. *Statistics in Medicine 6* (1987), 242.

[24] Rao, C. R. A. Fisher: The founder of modern statistics. *Statistical Science 7*, 1 (1992), 34–48.

[25] Rao, P., Kaplan, J., and Cochran, W. Estimators for the one-way random effects model with unequal error variances. *Journal of the American Statistical Association 76* (1981), 89–97.

[26] Savage, L. On rereading Fisher. *Annals of Statistics 4* (1976), 441–500.

[27] Stafford, J. E. A robust adjustment of the profile likelihood. *The Annals of Statistics 24* (1996), 336–352.

[28] Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association 54*, 285 (1959), 30–34.

[29] Stigler, S. M. *The history of statistics: The measurement of uncertainty before 1900.* Harvard University Press, 1986.

[30] Stigler, S. M. Ancillary history. In *State of the Art in Probability and Statistics* (2001), pp. 555–567.

[31] Venables, W. N., and Ripley, B. *Modern Applied Statistics With S (4th ed.)*. Springer-Verlag, New York, 2002.

[32] Yusuf, S., Peto, R., and Lewis, J. Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases 27*, 5 (1985), 335–371.