A Comparison of Weighted Pooled Results from a Cochrane Meta-Analysis with Likelihood Ratio Functions to Determine Strength of Evidence

Thomas, Ronald

Wayne State University School of Medicine, Department of Pediatrics

Detroit, Michigan, USA 48178

rthomas@med.wayne.edu

Daoud, Marwan

Wayne State University School of Medicine, Department of Pediatrics

Detroit, Michigan, USA 48178

mdaoud@med.wayne.edu

CONTEXT:

Faced with small sample sizes, or inferior power of an intervention, a meta-analysis can be applied to combine "evidence" from different studies. By combining results from different studies pooled data help to optimize conclusions about the outcome of the treatment(s) investigated [1]. However, with all the potential good that a rigidly conducted meta-analysis can provide to the decision making process it still retains its critics. A major criticism is that meta-analyses include all published and unpublished data results, whether good, bad, or indifferent. The range of sample sizes from studies pooled in an analysis reveal that some can be poorly powered out, or not even at all. Pooled effects may be multivariate rather than univariate, and data summarized may not be homogeneous. Grouping different causal factors may lead to meaningless estimates of effect size. Therefore, meta-analyses may include qualitative judgments they were designed to control. This has led critics to position that meta-analysis may not be the one best method for studying the diversity of fields for which it has been used.

At the center of this statistical debate is an attempt to answer the question „What do the data state?" Two well defined statistical approaches over the years have been applied to interpret this question. Researchers know them as the frequentist and Bayesian approaches. However, both approaches cannot answer this question specifically. In a sense, both attempt to answer different questions. The frequentist asks „what should I do?" and the Bayesian asks „what do I believe?" An alternate approach to this debate has been advanced [2,3,4,5] in the literature, which meets both methodologies in the middle. This approach is referred to as The Evidential Paradigm [6]. It provides a fundamental structure for presenting and evaluating likelihood ratios (LR) as measures of statistical evidence for one hypothesis over another.

OBJECTIVE:

Our purpose for this paper was to investigate whether the likelihood ratio function could be used as a complementary tool to measure strength of evidence gathered from a rigidly conducted Cochrane meta-analysis [7], and to the degree the two techniques agreed or disagreed on their respective interpretations of evidence.

METHODOLOGY:

Law of Likelihood

The following concepts of the evidential paradigm borrow heavily from that presented in a previously published tutorial on likelihood ratios as statistical evidence [8]. The fundamental principle of statistical reasoning, that Hacking [9] named the Law of Likelihood, is as follows:

If hypothesis H_A implies that the probability that a random variable X takes the value x is P_A, while hypothesis H_B implies that the probability is P_B , then the observation X = x is evidence supporting A over B if and only if $P_A > P_B$, and the likelihood ratio, $P_A / P_B = k$, measures the strength of that evidence.

If an event is more probable under *A* than *B,* then occurrence of that event is evidence supporting *A* over *B.* A likelihood function is an expression of the conditional probabilities $P(x \mid H_1), P(x \mid H_2), \ldots$as a single

function $P(x \mid H_i)$ for $i=1,2,\ldots$. A plot of the likelihood function $(L(H_i)$ versus (H_i) reveals which hypotheses are better supported by the data because these hypotheses will have a larger $P(x \mid H_i)$ relative to other hypotheses. Statistical evidence has a different mathematical form than uncertainty. Probabilities measure uncertainty, but not strength of evidence. Likelihood ratios measure the strength of statistical evidence. The Evidential Paradigm prescribes that it is the likelihood ratio for two simple hypotheses that indicates the strength of the evidence about a parameter. One can conclude that there is strong evidence supporting one hypothesis over the other when the observed value is sufficiently large. How much larger the support must be to represent strong evidence of one hypothesis over another requires relating the likelihood ratio values to categories ranging from "weak" to "very strong". The values 8 and 32 have been suggested as benchmarks for *k.* Observations with a likelihood ratio of k=8 (or 1/8) constitute moderately strong evidence, and observations with a likelihood ratio of k=32 (or 1/32) represent strong evidence [10].

The logarithm of k has been referred to as the weight of evidence given by the observed value of x for H_A over H_B, measured in bits, nats, or bans, according to whether the logarithm is taken to base 2, base e, or base 10. A value of $k > 1$ means that the data indicate that H_A is more supported by the data under consideration than H_B. Jeffreys [11] gave a scale for interpretation of *k (weight of evidence)* in decibans (tenths of a power of 10) and bits:

k	**dB**	**bits**	**Strength of evidence**
< 1:1	< 0		Negative (supports HB)
1:1 to 3:1	0 to 5	0 to 1.6	Barely worth mentioning
3:1 to 10:1	5 to 10	1.6 to 3.3	Moderate
10:1 to 30:1	10 to 15	3.3 to 5.0	Strong
30:1 to 100:1	15 to 20	5.0 to 6.6	Very strong
>100:1	>20	>6.6	Decisive

For our re-calculation of the Cochrane data, bits were chosen to represent a value for weight of evidence:

$$\text{Weight of evidence} = \text{base 2 log } (k)$$

DATA SOURCE AND EXTRACTION

The first step in this study involved abstracting data from summative totals of the outcomes of each variables examined in the Cochrane report and compared between the two study drugs, ibuprofen and indomethacin, for the closure of the patent ductus arteriosus. The second step involved entering these data values into a relative risk conditional likelihood function syntax created using the free software program R version 2.11.1 [12] to calculate the likelihood functions for the relative risk and graphs. In the third step the values of *k* (*k* here is the likelihood ratio of the observed relative risk to the relative risk of 1) obtained were transformed to base 2 log (bits) values using SPSS software Version 18.0. The fourth step involved table reporting these values, interpreting them using the strength of evidence verbal categories, and comparing them to the weighted p-values reported in the Cochrane report to denote whether or not there was a statistically significant difference between the two study groups. How well the outcomes from these two approaches matched up were then examined.

RESULTS:

Figure 1 displays the standardized likelihood function for the primary outcome of failure to close the PDA. The y-axis indicates that only ratios of points on the likelihood function have evidential meaning. Note that the best supported value for the relative risk (the maximum likelihood estimator (MLE) of the relative risk which is the *observed* relative risk) is 0.94. This value is at the crest of the likelihood function. The usefulness of the likelihood function is that it „shows" all the likelihood ratios and provides a visual impression of the strength of evidence.

Support intervals

Inside the likelihood function are two lines drawn parallel to the x-axis, which are likelihood support intervals (SI) (Figure 1). These support intervals identify all the parameter values for the relative risk that are consistent with the data under the peak of the likelihood function. Likelihood support intervals summarize the evidence about MLE without having to report numerous likelihood ratios. A 1/*k* likelihood support interval (SI) is defined as the "set" of relative risk values where the standardized likelihood function is greater than 1/*k*. Any relative risk value falling within the 1/*k* SI is supported by the data, because the best supported hypothesis (MLE) is only better supported by a factor of *k* or less. When *k* =8 there is weak evidence supporting the obtained MLE over any other relative risk value in the interval. Fixed values of *k* =8 and 32 are employed to form moderate and strong support intervals.

In Figure 1, the 1/8 SI for the relative risk of IBU failing to close the PDA over INDO is 0.75 to 1.19 (the 1/8 SI is the line corresponding to a height of 1/8 on the y-axis). Hypotheses within the interval may be better supported over others within the interval, but the level of support is weak and less than a factor of 8. For hypothesized values outside the interval, there always exists another relative risk value (that the MLE=0.94) that is better supported by a factor greater than 8. Therefore, 1/32 SIs are judged to be „stronger" support intervals than 1/8 SIs. In the current example, hypotheses" suggesting the relative risk of IBU failing to close the PDA over INDO is between 0.70 and 1.27 (the 1/32 SI is the line corresponding to a height of 1/32 on the y-axis) and are considered consistent with the data at a weaker level.

Figure 1shows that the value of *k* (1.2), transformed to bits (0.263) , falls within the *low range* of 0 to 1.6. In other words, support of one hypothesis (IBU) over another (INDO) for failure to close the PDA was "barely worth mentioning". Figure 1 presents this result graphically. The p-value reported in the Cochrane report (0.58) supports this finding which is interpreted as a "statistically non-significant difference".

Without presenting the results from all other variables, and to show how the weight of evidence (strength) values can vary across other variables when compared between study drugs, two further examples are provided. Necrotizing enterocolitis produced a larger weight of evidence value of 3.20,

which would fall within the *high range* for "substantial" evidence. Necrotizing enterocolitis would be more likely to occur with INDO. The p-value reported by Cochrane was significant at a cut-off of 0.05 (p=0.045). Decreased urine output (<1 cc/kg/hr) produced a weight of evidence value even larger of 12.33; greater than 6.6 indicating "decisive evidence". Decreased urine output would be much more likely to occur with INDO. The p-value reported by Cochrane was also highly significant at a cut-off of ≤0.001.

CONCLUSIONS:

This study revealed that the meta-analysis performed extremely well in the interpretation of the data, contrary to the criticisms commonly raised in the application of the technique. This study also showed that the likelihood ratio function can successfully be applied for the interpretation of the strength of evidence of one hypothesis over another, and thus helps to interpret the data in a more "clinically significant" way.

REFERENCES

1. Apa. (2007, November 29). In *Wikipedia, The Free Encyclopedia*. Retrieved 14:34, January 24, 2011, from http://en.wikipedia.org/wiki/Meta_analysis.
2. Hacking I. *Logic of Statistical Inference.* Cambridge University Press: New York, 1965.
3. Royall RM. Statistical Evidence: *A Likelihood Paradigm.* Chapman and Hall: London, 1997.
4. Blume JD. *On the probability of observing misleading evidence in sequential trials*, PhD dissertation, John Hopkins University School of Public Health, 1999.
5. Edwards AWF. *Likelihood.* Cambridge University Press: London, 1972.
6. Vieland VJ, Hodge SE. Review of R. Royall (1997) Statistical Evidence: a likelihood paradigm. *Annals of Human Genetics*. 1998: 63:283-289.
7. Ohlsson A, Walia R, Shah SS. *Ibuprofen for the treatment of patent ductus arteriosus in preterm and/or low birth weight infants.* Cochrane Database Syst Rev. 2010 Apr 14: CD003481. Review.
8. Blume JD. Tutorial in Biostatistics Likelihood methods for measuring statistical evidence. *Statistics in Medicine*. 2002; 21:2563-2599.
9. Hacking I. *Logic of Statistical Inference*. Cambridge University Press. New York, 1965.
10. Royall RM. Statistical Evidence: *A Likelihood Paradigm.* Chapman and Hall: London, 1997.
11. Jeffreys, H. *The Theory of Probability*. The Oxford University Press, 1961.
12. Retrived 14:48, January 24, 2011, from http://www.r-project.org.

FIGURE 1

Ibuprofen vs. Indomethacin

Failure to Close a PDA (after single or three doses)