Final Report

Meta-Analysis: Synthesis and Appraisal of Multiple Sources of Empirical Evidence

With the increasing concern in science and medicine for issues such as more complete use of all sources of evidence and reproducibility for single statistical studies, multiple studies and meta-analysis are becoming central to scientific advancement. The Statistical and Applied Mathematical Sciences Institute (SAMSI) addressed this topic through a research program from June 2-13, 2008. It brought together leading statisticians and scientists with interests in meta-analysis, to assess the existing methodology, and develop needed new methodology, and explore pedagogy for bringing the methodology to the broader scientific community.

1 Scientific Overview

1.1 General Background on Meta-Analysis

Seldom is there only a single empirical research study or source of evidence relevant to a question of scientific interest. However, both experimental and observational studies have traditionally been analyzed in isolation, without regard for previous similar or other closely related studies. A new research area has arisen to address the location, appraisal, reconstruction, quantification, contrast and possible combination of similar sources of evidence. Variously called meta-analysis, systematic reviewing, research synthesis or evidence synthesis, this new field is gaining popularity in diverse fields including medicine, psychology, epidemiology, education, genetics, ecology and criminology.

Statistical methods for combining results across independent studies have long existed, but require renewed consideration, development and wider dissemination by inclusion in the mainstream statistics curriculum. The possibility that the due consideration of all relevant evidence should be accepted as standard practice in statistical analyses deserves investigation. The combination of results from similar studies is often known simply as 'meta-analysis'. Common examples are combining results of randomized controlled trials of the same intervention in evidence-based medicine; of correlation coefficients for a pair of constructs measured similarly across studies in social science; or of odds ratios measuring association between an exposure and an outcome in epidemiology. More complex syntheses of multiple sources of evidence have developed recently, including combined analyses of clinical trials of different interventions, and combined analysis of data from multiple microarray experiments (sometimes called cross study analysis). For straightforward meta-analyses,

general least-squares methods may be used, but for complex meta-analyses, the technical statistical approach is not so obvious. Often likelihood and Bayesian approaches provide very different perspectives; and in practice the possible benefits of more complex approaches may be hard to discern as many meta-analyses are compromised by limited or biased availability of data from studies as well as by varying methodological limitations of the studies themselves.

The presence of multiple sources of evidence has long been a recognized challenge in the development and appraisal of statistical methods - from Laplace and Gauss to Fisher and Lindley. In the 1980s Richard Peto argued that a combined analysis would be more important than the individual analyses, a view taken still further by Greenland and O'Rourke who have suggested that that individual study publications should not attempt to draw conclusions at all, but should instead only describe and report results, so that a later meta-analysis can more appropriately assess the study's evidence fully informed by other study designs and results. Will combined analyses actually replace individual analyses (or at least decrease their impact)? If so, it is time to reexamine the perennial problems of statistical inference in this context.

The concept of multiple sources of evidence itself needs to be generalized and applied more generally and creatively through many areas of statistical research. Multiple sources should not just be taken as separate studies or even the possible simple regrouping of subsets of observations within studies but the bringing to bear of seemingly distinct information sources on given question and even the "creation" of multiple sources as in Bayesian Additive Regression Trees (BART) where differing regression trees are purposefully grown to be later advantageously combined. In some fields, terms like data fusion and data integration are being used for this more general sense of utilizing multiple sources of evidence. A single strategy of no pooling, complete pooling or partial pooling of separate studies perhaps needs to give way to adaptive strategies where the degree of pooling is individually chosen for each and possibly every parameter in the joint probability models used to represent all the relevant sources of evidence.

1.2 Specific motivation for the focus of this program

This program comprised two weeks of research, mixing tutorials, research presentations and working group activities on the subject. The goal of this program were three-fold: 1) to bring the area to the attention of statistical researchers, whose expertise is critical to substantiate and clarify the necessary statistical theory and methodology; 2) to nurture the necessary interdisciplinary collaboration and communication between statistical researchers and statisticians who currently work or plan to work with basic and applied science researchers and

3) to provide an entry point into the field to interested students and faculty, and to allow researchers already specialized in the domain to exchange recent results and information.

2 Program Structure

2.1 Leaders

The program was initiated by Keith O'Rourke. The tutorials in the first week were organized jointly by Vanja Dukic, Ken Rice and Keith O'Rourke. Ingram Olkin opened the program with the lead tutorial followed by Keith O'Rourke, Ken Rice, Vanja Dukic and Julian Higgins. The data analysis sessions were given by Keith O'Rourke and Ken Rice. The working group leaders chosen for the second week of workshops were Dalene Stangl, Ken Rice, Vanja Dukic, Julian Higgins, Keith O'Rourke and David Dunson.

2.2 Program Attendance

The program attracted 44 participants in the first week, which featured tutorials and data analysis workshops (see below). A total of 31 participants either continued from the first week or joined the program in the second week; the second week involved working groups and a final summary session (see below). All in all, 55 participants attended some portion of the program.

2.3 Tutorials and Opening Workshop

The introductory overview was given by Ingram Olkin, Stanford University. It provided participants with an elementary but thorough introduction to the challenges and opportunities of dealing with multiple studies in the context of biomedical and social science research. Many real application examples were covered to illustrate basic and advanced methods and highlighted the numerous scientific issues and challenges that inevitably arise.

An introductory overview on the likelihood basis for multiple data sources was then given by Keith O'Rourke, Duke University. This provided participants with an elementary introduction to working directly with likelihoods to contrast and combine data based information. This was done both for individual observations - where individual observation likelihoods were contrasted and combined to obtain the usual study estimates - and for studies where study level likelihoods were contrasted and combined. A general meta-analysis approach was then presented in terms of the contrast and combination of likelihoods. Various problems that arise with likelihoods in meta-analysis were then discussed. These problems are largely due to fact that although likelihood concentrates for common parameters (of interest) it

expands in dimension for arbitrary (nuisance) parameters and unfortunately there usually are many of these arbitrary (nuisance) parameters.

The advanced tutorials were then started the next day with Keith O'Rourke more thoroughly reviewing the likelihood approach and underlining issues of sparseness with the historical Neyman-Scott examples. Vanja Dukic then covered the integrated likelihood approach as well as some preliminaries for a Bayesian approach. Ken Rice then covered the conditional likelihood approach both from a classical and Bayesian perspective as well as providing material on exchangeability and other more general aspects of meta-analysis. Following this, Vanja Dukic and Ken Rice more fully covered Bayesian approaches to meta-analysis. The tutorial sessions ended with Julian Higgins giving a thorough overview of current (practical) challenges in undertaking meta-analyses in clinical research.

Inter-dispersed with these tutorials, Keith O'Rourke gave a "Data Analysis Session" on likelihood calculations in R and Ken Rice gave one on implementing Bayesian meta-analyses in WinBUGs.

3 Working Groups

In the second week, working groups were formed based on the participant research interests. There were six working groups formed comprised of 31 participants, with group sizes ranging from 7 to 17. The Working Groups were

1. Decision theory
2. The role of priors for bias and random effects
3. Bias modeling and information from observational studies
4. ROC and survival analysis
5. Networking, multiple treatments and multivariate
6. Genetics

Here is a summary of their activities.

3.1 Decision analysis group

During the week this group explored two questions:

1. In reporting estimates of treatment effect and heterogeneity, is there a loss function for which usual estimates reported are optimal?

2. In non-inferiority trials, how does one choose the delta by which a new treatment is considered "good-enough" relative to the standard treatment and placebo [this question was motivated by an FDA inquiry to the program].

The group studied what is currently done in non-inferiority trials, discussed the difference between random and fixed effects, raised and discussed a concern about only looking at inter-study variability in average treatment effect rather than also being concerned with within study treatment effect variance in choosing between drugs and discussed how "ideally" one would like to address the problems versus how one can take what is currently done and make an improvement that has a chance of being implemented. At the end there seemed to be a consensus that it was necessary to be clear about what the "real" question was and for exactly what "population" so that a full and complete modeling of the decision and its relevant consequences could be undertaken.

As a result of the discussions, some members of the group have written a technical report that has been submitted for publication but is still under review. The reference of the paper is: E. Moreno, F. J. Giron, F.J. Vazquez-Polo and M.A. Negrin (2008). Optimal decisions in cost-benefit analysis. Tech. Report. Dpt. Statistics, University of Granada.

3.2 Role of priors for bias and random effects group

This group focused on priors for random effects and bias - two areas in meta-analysis where there is usually a small amount of sample information and hence the choice of priors can be critical.

The discussions around random effects necessarily started with the choice of the parametric distribution of random effects – meant to represent the physical variation in effects from study to study – and then priors for the parameters in these distributions and then non-parametric approaches to random effects. In this group, roughly as in the Decision Analysis group, it was found necessary to be clear about what the "real" question was, what the random effect distribution was meant to represent and exactly what parameters were of inferential interest. A quick review of some current choices for priors for random effects seen in the meta-analysis literature was also undertaken.

The discussion of priors for biases, such as may vary with varying assessed study quality, largely revolved around the possibilities of obtaining empirically motivated priors from the empirical literature. A possibly relevant data set of clinical research studies with various methods of appraising their quality was acquired and a method for investigating quality effects identified in a paper by Greenland and O'Rourke. This likely will become a student project in the near future.

Currently, the group leader, Ken Rice is working on a paper with Keith Abrams entitled Estimating population-averaged contrasts under exchangeability; the role and influence of random-effects distributions.

3.3 Bias modeling group

The bias modeling group undertook the challenge of issues and methods for the contrast and combination of biased and confounded sources of information. There ended up being a focus on two main topics - 1) propensity score issues and methods in multiple observational studies and 2) investigations of bias modeling using both RCTs and Non-Rcts together.

The propensity score focus ended up involving two projects, Project A where there was individual-level data available from multiple observational studies and Project B where was only study level data available. Project A , was lead by Elizabeth Stuart of John Hopkins University and B was lead by Robert Platt of McGill University. The motivating question for A was with regard to how propensity scores should be estimated in this setting and the motivating question for B was with regard to whether or not and if so – how propensity score-based subclass estimates from the multiple studies should be combined to get an overall estimate of the effect. Elizabeth Stuart and Robert Platt have since been collaborating on these projects and anticipate involving students in the future.

The investigations of bias modeling using both RCTs and Non-Rcts was lead by Dan Jackson of MRC Cambridge and involved the adaption/extension of methods developed by Steyerberg and the motivating question was with regard to explicating the necessary assumptions and critically assessing their appropriateness. Dan Jackson is continuing to work on the adaption of the Steyerberg method and its extensions and in related work with the Fibrinogen Studies Collaboration [published in Statistics in Medicine (2009) – Systematically missing confounders in individual participant data meta-analysis of observational cohort studies]; he has found the discussions at SAMSI useful in his thinking further about applying Steyerberg type methods.

Also of note, Elizabeth Stuart – partly as a result of the SAMSI meeting – is planning to organize a 2010 JSM Invited Session on methods for assessing generalizability.

3.4 ROC and survival analysis group

This working group addressed the issues of synthesizing evidence from independent studies about diagnostic test accuracy or survival times – both of which entail individual study and pooled curves or distributions. Both parametric and non-parametric approaches were of interest.

They currently have one paper in preparation, with Jean-Francois Plante of University of Toronto, Vanja Dukic of Chicago University, David Dunson of Duke University and possibly Dalene Stangl of Duke University on Bayesian non-parametric meta analysis of ROC curves. The abstract is as follows: Most standard meta-analytic methods combine information on single parameter, such as treatment effect. For meta-analysis of diagnostic test accuracy, measures of both sensitivity and specificity from different trials are of meta-analytic interest, summarized as a bivariate measure of accuracy, or possibly as a receiver operating characteristic (ROC) curve. Motivated by an analysis of serum progesterone tests for diagnosing non-viable pregnancy, we develop simple fixed-effects and random-effects summary ROC curve estimators, based on a flexible density estimation technique. We compare the performance of the new estimator to the simpler bivariate normal summary ROC estimator.

3.5 Network meta-analysis group

Network meta-analysis refers to the situation in which studies brought together for synthesis have compared different subsets from a finite collection of treatments. By exploiting 'chains' of evidence, such as making inference on treatment A vs treatment B by contrasting studies of A vs C with studies of B vs C, a network of interrelationships among the studies is created. These meta-analyses are often, and perhaps more appropriately, called multiple treatments meta-analyses (MTM), or mixed treatment comparisons (MTC) meta-analyses.

The working group tackled a variety of problems associated with network meta-analysis. Particular progress was made on methods for illustrating the network graphically. If every study makes a pair-wise comparison – i.e. includes exactly two treatments – then simple graphs with nodes for treatments and lines for comparisons are sufficient to represent the dataset. However, if some studies include three or more treatments, as is typically the case, then such representations do not adequately illustrate the important difference between within-study (direct) comparisons and across-study (indirect) comparisons. In this case, comparisons that come from the data are not independent. The group proposed a diagram in which the distinction is made by using separate lines or shapes for different study designs. Since the workshop, some progress has been made in using graph theory to examine 'loops' of evidence in the network.

The importance of separating direct from indirect evidence is largely in order to investigate whether the network of evidence is coherent. Coherence is defined informally as mismatch between direct and indirect sources of evidence, or between two different indirect sources of evidence, on any particular comparison. It is a special kind of heterogeneity between studies that focuses on between-design differences rather than between-study differences. Two statistical methods for tackling incoherence have been proposed, by T. Lumley

(Network meta-analysis for indirect treatment comparisons, Stat Med 2002; 21: 2313-2324) and Lu and Ades (Assessing evidence inconsistency in mixed treatment comparisons, JASA 2006; 101: 447-459). The former adds a random effect across all studied pair-wise comparisons and tests whether the variance of this random effect is zero. The Lu and Ades approach adds a random effect across each independent evidence cycle, and tests whether the variance of this random effect is zero. There are fewer independent evidence cycles than there are comparisons. However, counting the number of independent evidence cycles is not trivial when there are multi-arm studies. The group discussed other approaches, such as fitting a model than assumes coherences and comparing deviances with a 'free' model that makes no assumptions about chains of evidence. Three of the workgroup members (Dan Jackson, Jessica Barrett, Julian Higgins; working with Ian White) have a paper in preparation about some of these ideas.

The working group also discussed technical issues about making inferences in network meta-analyses. Restricted maximum likelihood is often used; Lumley uses the function lme in R with a slightly unusual construction for random-effects variances. Inference is less straightforward with multi-arm trials or logistic models. We explored profile likelihood, and inverting the observed information matrix. Plans were made to investigate the use of conditional likelihood, integrated likelihood, and inverting expected information matrix.

3.6 Meta genetics group

This working group address issues of multiple sources of evidence for genetics, focusing on Gene Expression Meta Analysis, Meta Analysis for Genetic Association Studies and Accounting for Dependence in High-Dimensional Predictors. Since this summer's program, the meta-genetics working group has been quite productive. The active core of this group consists of David Dunson at Duke University , Fei Zou at UNC Biostatistics and Fei Liu at the University of Missouri Columbia. They have submitted the following paper to Biometrics: Liu, F., Dunson, D.B. and Zou, F. (2008). High-dimensional variable selection in meta analysis for censored data. Biometrics, submitted. In addition, they have another paper under way: Liu, F., Dunson, D.B. and Zou, F. (2009). Annotated relevance vector machine with application to polymorphism selection. In preparation.

Their following summary highlights some of the work undertaken to date which represents the most exciting research happening in the program and provides a nice example of both a generalized concept of multiple sources of evidence and the replacement of a single strategy of no pooling, complete pooling or partial pooling of studies with an adaptive strategy where the degree of pooling is individually chosen for different coefficients.

In large scale genetic epidemiology studies that collect massive numbers of single nu-

cleotide polymorphisms (SNPs) or gene expression measurements, it is extremely challenging to identify genes that are predictive of disease phenotypes given the modest sample size of most studies relative to the number of genes. Due to concern about false positive rates, it is crucial to replicate findings about disease genes in multiple studies. Standard approaches take multistage testing approaches in which one tests if genes identified in initial studies are significant in follow-up studies. This strategy is shown to have major disadvantages in terms of power and type I error rates compared with an innovative approach developed in the SAMSI meta-genetics working group based on simultaneous selection through a multi-task relevance vector machine (MT-RVM) procedure. This approach, which is related to methods used in signal processing, borrows information across studies in the degree of shrinkage of gene-specific coefficients towards zero. The method is scalable to large numbers of genes, can accommodate censored data commonly collected in disease recurrence studies, and clearly outperforms common competitors, such as Lasso. In addition, the meta-genetics group is currently pursuing a new procedure that allows information on gene function annotation to be incorporated, while automatically learning how predictive each annotation source is. The annotated relevance vector machine (aRVM) procedure should be very widely useful in machine learning and other applications beyond genetics, as it allows an adaptive targeted search for important predictors enabling an effective reduction in dimensionality and mechanism for borrowing information across disparate studies.

4 Post Program Activities

1. At the Eastern North American Region 2009 meeting of the International Biometric Society most of the working group leaders and some of the participants presented their research. In particular, a session "Advances in Meta-Analysis" was organized, based on the program, with presentations by Eloise Kaizar, Robert Platt, Vanja Dukic,and Dalene Stangl.

2. Professional Courses:

   - Keith O'Rourke gave a two day course on meta-analysis for Statisticians and Students at the University of Alberta in July 2008.
   - Keith O'Rourke gave an Advanced Meta-analysis Short Course at the University of Alberta.