Combining Information from Diverse Sources

Eloise E. Kaizar
Thesis Proposal

July 6, 2005

Abstract

Research synthesis plays a central role in the process of scientific discovery, providing a formal methodology for the systematic accumulation and evaluation of scientific evidence. There are many situations in which research synthesis is required because obtaining the required information from an individual trial is not possible or practical. This thesis proposes a general method to synthesize information from multiple sources. The study of the relationship between suicide and antidepressant use in children and adolescents is one such case. To shed light on this issue, the FDA combined data from 24 randomized controlled trials using standard frequentist fixed and random effects models. However, the diversity of the trials suggested the presence of systematic effect variation that was not incorporated into these models. We applied a Bayesian hierarchical model to the data to include more appropriate variance structure and answer scientific questions regarding subsets of the data. While this technique was successful in solving the variance and subsetting issues, the collection of clinical trials share qualities that limit the generalizability of the meta-analysis. Several available administrative databases do not suffer from these same limitations of generalizability. Combining administrative databases and clinical trials in one analysis would improve the reliability and generalizability of the analysis. Several methods have been proposed to accomplish this goal, but to our knowledge, only one has been implemented and so many practical problems related to these methods (including exchangeability, bias, prior specification and model selection) have not been addressed. This thesis proposes a general method to synthesize information from diverse sources that has two of the already proposed methods as special cases and is also consistent with the third proposed method.

1 Introduction

Research synthesis plays a central role in the process of scientific discovery, providing a formal methodology for the systematic accumulation and evaluation of scientific evidence. As Cooper and Hedges note, "Research syntheses attempt to integrate empirical research for the purpose of creating generalizations. Implicit in this definition is the notion that seeking generalizations also involves seeking the limits and modifiers of generalizations" [1]. Perhaps the single most important strength of research synthesis over individual studies is the ability to investigate the robustness of a relationship and the conditions moderating its magnitude and direction. By comparing results across studies it is possible to investigate systematic variation in effect size due to patient characteristics, design features, types of treatments, providers of treatments, outcomes and time. The quantitative approach to research synthesis can go beyond the simple description and estimation of effects to address the problem of explaining why, or under what conditions, a given effect may be observed, and is an important tool for evidence-based medicine.

Given the potential for such revealing analysis inherent in research synthesis, it is disappointing that its most prominent use is the simple fixed or random effects meta-analysis of randomized controlled trials (RCTs), which do not explore the robustness of relationships or their moderators but instead focus on estimating a single effect of an intervention. These meta-analyses often combine studies that show a significant relationship with those that do not (for reason of power or lack of effect) to reconcile these results by pooling power across the studies. Meta-regression is also commonly used to combine RCTs. This technique brings the standard meta-analysis one step closer to the full research synthesis potential by including study-level covariates in the analysis.

All of these standard techniques are currently restricted to combining similar high-quality studies–most predominantly RCTs, but also occasionally observational studies. The RCT research syntheses offer one major strength. Because they are constructed from the gold standard of evidence–randomized controlled trials–their internal validity is generally reliable. Unfortunately, this type of synthesis also has a major weakness: the patient pools used in RCTs are often restricted to a small, well-defined group and so the generalizability of the meta-analyses is correspondingly limited. Observational data research syntheses are quite the opposite. Their biggest strength is the wide generalizability of their results due to their broad subject pool. Their weakness is limited internal validity, since the potential for confounding is not eliminated by randomization, as in RCTs. By combining RCTs and observational data in one carefully-constructed analysis, we may rely on the strengths of each kind of data to offset the other's weakness. Using many sources of data also allows us to fully explore the systematic variation captured both by individual patient traits and study characteristics, while providing a larger pool of data that increases the analysis' power to detect associations. This thesis proposes a general method to synthesize information from both observational and randomized studies within one unified analysis.

There are many current open questions in the medical field that would be informed by a more inclusive approach to research synthesis using different data sources. As an example, two major issues of drug safety have recently been at the forefront of US Food and Drug Administration (FDA) regulatory priorities. In both cases, there is some indication that a drug causes an increased risk of a rare adverse event. However, the individual RCTs performed by various pharmaceutical companies to test the efficacy of these drugs did not have enough power to investigate the safety of these drugs with respect to rare adverse events. The first collection of drugs to come under fire was Paxil and the other newer generation antidepressants. Through meta-analysis, the FDA was able to garner enough power to declare the existence of an association between suicidality (suicidal thinking and/or behavior) and antidepressant use in children and adolescents [2]. However, scientific questions such as who is at risk, or whether there is an association between antidepressant use and suicide remain unanswered. Hot on the heels of the antidepressants, Vioxx and other COX-2 selective and non-selective non-steroidal anti-inflammatory drugs (NSAIDs) were accused of causing an increased risk of adverse cardiovascular events. The FDA again collected many controlled and uncontrolled trials of the efficacy and safety (mainly gastro-intestinal safety) of the COX-2 drugs, but the combination of these studies was very complex and no research synthesis was presented at the FDA advisory committee meeting. The FDA found the uncombined data strong enough to ask several of the manufacturers to withdraw their products from the market and require black box warnings for the other COX-2 inhibitors [3].

Another example of a potential medical application of research synthesis of diverse data comes from the treatment of coronary artery disease. This field is the focus of much cutting-edge research and so innovation is rapid and abundant. Because of the speed of technique and technology improvement, each innovation is not the subject of a lot of study before the "next best thing" comes along. Each technique or technology has thus not been compared to a wide array of alternatives, especially for long term differences. For example, several studies have compared the long-term survival for patients undergoing coronary-artery bypass grafting (CABG) and for those undergoing percutaneous coronary intervention (PCI). But, these studies have mostly been conducted before stenting revolutionized the PCI approach. To compare stenting to CABG, we must rely on several sources of data to create a chain of evidence between CABG and stenting. This chain has been examined via two databases, but its internal validity could be strengthened by adding RCTs to the analysis [4].

These three examples are only a small sample from the vast body of medical applications that would benefit from research synthesis across diverse sources of data. For this proposal, we use the pediatric use of antidepressants and their potential association with increased risk of suicide as a case study. We have taken first steps to improving the research synthesis for this case study by taking a Bayesian hierarchical modelling approach similar to meta-regression. The power and flexibility advantages of this approach over standard frequentist methodology allowed us to reveal more of the systematic variation structure present in the FDA's collection of studies. Again, although we demonstrate the usefulness of multi-level Bayesian models in research synthesis via the antidepressant case study, the approach we outline would be useful for many medical applications. In our combination of information from several RCTs, we are confident in the internal validity of our findings. However, RCTs exclude a variety of patients, and so we are concerned about the generalizability of our results to patients that are not like the patients included in the trials. To accomplish this, we propose to build on currently proposed ideas to develop methodology to combine information from RCTs with data that could be more representative of the general patient population (including observational studies, administrative databases, and surveys). Currently, three methods have been proposed to combine data from two or more types of sources. These are the Confidence Profile Method (CPM), response surface methodology (RSM), and cross design synthesis (CDS). Only the CPM has been fully applied to real problems, and so the applicability of RSM and CDS in the medical arena has not been established and several pragmatic issues in the application of all three methods remain open questions. This thesis proposes a general method to synthesize information from both observational and randomized studies within one unified analysis. The CPM and RSM are special cases of our proposed method, which is also consistent with CDS.

The remainder of this proposal is organized as follows. In Section 2, we review the current state of the analysis of our case study–the association between suicide and antidepressant use in children and adolescents; this section includes a summary of the data, the FDA analysis, the Bayesian hierarchical model we used to analyze the data, and a discussion of the limitations of the current Bayesian hierarchical model for the case study and general application. In Section 3, we discuss the three currently proposed approaches to combining randomized and observational data (RSM, CPM, and CDS). Finally, in Section 4, we describe how we intend to use the ideas in each of these approaches to combine data, using simulation to demonstrate our developed methodology in the antidepressant case study.

2 Current State of Our Antidepressant Randomized Controlled Trial Analysis

In October 2004, the FDA issued a letter to the manufacturers of newer generation antidepressants informing them of the need to "caution practitioners, patients, family members or caregivers about an increased risk of suicidal thinking and behavior (suicidality) in children and adolescents...who are taking antidepressant medications." The FDA ordered the manufacturers to include a "black box" warning on the labels of all their antidepressants. The warning is expected to cause a sharp decline in the numbers of children and adolescents who take antidepressants.

Leslie, Newman, Chesney and Perrin review the evidence presented to, deliberations of, and recommendations from the FDA scientific advisory committees that led to the black box warning [5]. The primary source of evidence that informed the advisory committees' recommendation was based on an FDA meta-analysis of 24 randomized placebo-controlled trials of the efficacy of antidepressants in children and adolescents with several different psychiatric diagnoses. The FDA examined studies of both Selective Seratonin Reuptake Inhibitors (SSRIs) and atypical antidepressants, but did not consider older classes of antidepressants such as tricyclics and Monoamine Oxidase Inhibitors (MAOIs). Among 4,582 patients studied, there were *no* completed suicides. However, the overall rate of suicidality (suicidal behavior and/or ideation), the FDA's primary outcome, was 1.9%.

The database consists of the 24 randomized placebo-controlled trials of antidepressants previously meta-analyzed by the FDA and presented to an FDA expert advisory panel in September 2004 [6, 7]. Among the 24 placebo-controlled trials, 16 studied efficacy for Major Depressive Disorder (MDD), 4 studied efficacy for Obsessive Compulsive Disorder (OCD), 3 studied efficacy for Social or General Anxiety Disorders, and one studied efficacy for ADHD. The drug formulations investigated included 5 SSRIs (Celexa, Luvox, Paxil, Prozac and Zoloft) and 4 atypical antidepressants (Effexor, Remeron, Serzone and Wellbutrin). The outcome of interest is the number of subjects per treatment group (within each study) for whom suicidal behavior and/or ideation was reported. Our primary measure of association between suicidal behavior and/or ideation with treatment is the log odds ratio, which we also refer to as the 'drug effect'. In the remainder of this section, we review the frequentist fixed and random effects models used by the FDA, then present our Bayesian hierarchical model, the results from the analysis based on this model, and finally discuss the limitations of this approach and opportunities to enhance it. A full report of these analyses appears in Kaizar, Greenhouse, Seltman, Kelleher (2005) [8].

2.1 Fixed and Random Effects Models

The FDA used two types of standard models to combine the 24 randomized controlled trials to quantify an overall drug effect. Both methods are a weighted average of the change in risk within each study. They differ in the method used to estimate the weights and to estimate the variance for the overall drug effect point estimate. The **Cochrane-Mantel-Haenszel** [9] estimate of the log odds ratio is a heuristically-derived fixed effects method, while the **DerSimonian-Laird** [10] estimate is a random effects method that weights each study according to the inverse of the sum of the within- and between-study variances.

Figure 1 displays a forest plot of the odds ratio (OR) of suicidality for each study (weighted according the Mantel-Haenszel method), as well as the fixed and random effects estimates of the overall OR and its confidence interval. In effect, this display summarizes the results of the FDA analysis.

In both the Mantel-Haenszel and DerSimonian-Laird methods, the variance estimates are plugged into the model as if they were known thus underestimating (i.e., estimating to be zero) the extra variance introduced by estimating this parameter. Further, the FDA considered the different drug formulations and psychiatric diagnoses to be equivalent in their effect on suicidal behavior/ideation. Both of these sources of underestimated variance may have artificially inflated the FDA's confidence in the presence of an association between suicidality and antidepressant use in children. To incorporate more sources of variance, we constructed a fully Bayesian hierarchical model that explicitly models the subject-level data and includes levels for each potential extra variance component. While the form of the model is not novel, we are not aware of its previous use in clinical trial meta-analyses.

2.2 Bayesian Hierarchical Model

We specify a four-level hierarchical model for the drug effect. Level 1 models the number of patients who had suicidal behaviors and/or ideations in each treatment group as Binomial distributions [11]; Level 2 models the study drug effect as a Normal distribution; Level 3 models the effect of study-level variables (such as drug formulation or psychiatric diagnosis) as a Normal distribution; and Level 4 models the overall drug effect as a Normal distribution. The structure of this model also allows us to include treatment arms with no events, whereas frequentist models require some form of adjustment for zero counts. In fact, the 4 studies in which there were no events in either arm (comprising 16% of the sample) were excluded from the FDA meta-analysis; for studies in which only one arm had no events the FDA added a correction factor of 0.5 events to all study arms.

To fix ideas we specify the model for the case where the study-level (Level 3) variable is drug formulation. Figure 2 is a diagram of this model. We let r_{ij}^C denote the observed number of subjects with suicidal behaviors/ideations in the control group of study $i = 1 \dots I$ of drug $j = 1 \dots J$ and n_{ij}^C denote the number of subjects in the control group of study i of drug j. We let π_{ij}^C denote the probability of a subject in the control group of study i of drug j having suicidal behavior/ideation. Corresponding values for the treatment groups are r_{ij}^T, n_{ij}^T, and π_{ij}^T. We let μ_{ij} denote the true log OR for study i of drug j, λ_j denote the true log OR for drug j, and θ denote the true overall log OR for all antidepressants together. The complete specification for our four-level hierarchical model in which the drug effect is measured by the log OR is given by Equations 1 through 6:

$$\begin{aligned}
r_{ij}^C &\sim \mathrm{Bin}(n_{ij}^C, \pi_{ij}^C) && (1)\\
r_{ij}^T &\sim \mathrm{Bin}(n_{ij}^T, \pi_{ij}^T) && (2)\\
\mathrm{logit}(\pi_{ij}^T) &= \mathrm{logit}(\pi_{ij}^C) + \mu_{ij} && (3)\\
\mu_{ij} &\sim N(\lambda_j, \tau_j^2) && (4)\\
\lambda_j &\sim N(\theta, \omega^2) && (5)\\
\theta &\sim N(\nu, \phi^2) && (6)
\end{aligned}$$

The variance component for the second level, τ_j^2 (Equation 4), denotes the within-drug formulation variance, a measure of the variability of individual study log ORs *within* drug formulation j. For example, τ_{Celexa}^2 is a measure of the variability of the study-level log ORs within the collection of Celexa studies. The variance component for the third level, ω^2 (eq. 5), denotes the between-drug formulation variance, a measure of the variability of the formulation-level log ORs among the different drug formulations. Finally, ϕ^2 denotes the prior variance of the overall log OR, which quantifies our *a priori* uncertainty of the value of the overall log OR.

We use the Deviance Information Criterion (DIC) to evaluate the fit of our models [12]. DIC calculations require the specification of a single likelihood and single prior distribution, collapsing the problem to a two-level hierarchy. Thus, we must specify the level of the hierarchy that is of interest (i.e., the "focus"). For example, for comparisons with fixed or random effects models, we chose to focus on Level 2 (π^C and μ) and for all other comparisons to focus on Level 3 (π^C and λ). The interpretation of DIC is similar to that of AIC. If two models have DIC values within two points of each other, there is little evidence that either has better fit [12]. Ordinarily, one might choose the most parsimonious of those with the best fit as the 'best' model. In this case, the more complex models allow us to assess different questions of scientific interest, and so the less parsimonious models have some value. We therefore choose to analyze these models in addition to the simplest ones.

Initially, we specify a prior distribution for the overall drug effect θ using a diffuse ($\phi^2 = 9$) Normal distribution centered at no effect ($\nu = 0$). By centering this diffuse prior at "no effect" we are reflecting the lack of consensus among the clinical community as to the risks and benefits of antidepressant use in children prior to the FDA's issuance of the black box warning. The choice of prior variance on θ with prior standard deviation $\phi = 3$ assigns probability $\frac{2}{3}$ on the odds ratio scale (e^θ) to $\{\frac{1}{20} < e^\theta < 20\}$. We also consider in our sensitivity analysis tighter distributions on the odds ratio corresponding to $\phi = 1.6$ which assigns probability $\frac{2}{3}$ to $\{\frac{1}{5} < e^\theta < 5\}$ and $\phi = 0.7$ which assigns probability $\frac{2}{3}$ to $\{.5 < e^\theta < 2\}$. For prior distributions on the variance parameters τ_j^2 and ω^2, we specify conjugate Inverse Gamma distributions with shape and rate 0.001 (Gamma parameterization: mean=shape/rate). These distributions approximate a flat prior on the variance parameters.

Our analytic strategy is as follows:

1. We replicate the FDA's fixed- and random-effects models, in which we consider all the studies to be exchangeable, by collapsing the model described in Equations 1 through 6. For the random-effects model, this means setting $\lambda_j = \theta$ and $\tau_j = 0$ for all j. For the fixed-effects, we further collapse the model by setting $\mu_{i,j} = \theta$ for all i, j and $\omega = 0$.

2. We evaluate various sources of study-level variation using our 4-level model. We take **drug formulation**, **drug class**, **study length**, and **psychiatric diagnosis**, in turn, as our level 3 study factor, as illustrated for drug formulation in Figure 2. In each of these models, we consider studies of the same study-level group (e.g., drug formulation) to be exchangeable.

We explore the sensitivity of our results to the specification of prior distributions for θ and for the variance parameters τ^2 and ω^2, as well as to the influence of individual studies and drug formulations. To investigate the sensitivity of our results to the specification of the original skeptical prior distribution for θ we reanalyze the data by setting the prior mean for θ at $\nu = 2$, indicating an *a priori* belief in a large increase in risk of suicidal behavior/ideation due to the use of antidepressants, and by setting the prior standard deviation for θ at ϕ=0.7, 1.6 and 3.0. To assess sensitivity to prior variance specification, we fix the between Level 3 variance (ω^2) at different plausible values and examine how our results change as a result of changing the variance. We repeat this analysis for the *within*-Level 3 variance (τ^2). Finally, to assess sensitivity to single study influence, we leave each study out of our model one at a time, allowing us to examine how our results change as a result of the influence of each study.

2.3 Results

Recall that Figure 1 displays the results of the FDA's fixed effects Mantel-Haenszel and random effects DerSimonian-Laird models.

Fixed Effects Models As a preliminary analysis, we first consider a Bayesian version of the fixed effects analysis used by the FDA by fixing $\tau_j = 0$ for all j and fixing $\omega = 0$. In words, all the studies are treated as exchangeable and pooled together. Table 2 displays the log OR, credible interval and probability that the log OR is greater than zero (where zero is no effect) under the fixed effects model. The probability that the log OR of suicidality is greater than zero is nearly one, indicating strong evidence for a link between antidepressant use and suicidality.

Random Effects Models A Bayesian version of the FDA's random effects analysis is obtained by fixing $\omega = 0$. This model allows for a between-study component of variance but treats the different study effects as exchangeable, regardless of their study-level characteristics (e.g., drug formulation). Table 2 displays the results from this model. The probability the log OR is greater than zero (no effect) is very close to one and indicates strong evidence for a link between antidepressant use and suicidality. We use Level 2 focused DIC (displayed in Table 2) to compare the fixed- and random-effects models. Because the DIC for these models are within two points of each other (157.4 vs. 159.3), we can conclude that the two models fit the data equally well.

Four Level Hierarchical Model We next evaluate various sources of study-level variation (including drug formulation, drug class, study length and diagnosis) using our four-level hierarchical model (Equations 1-6).

Table 2 displays the results for the overall antidepressant effect for each of the four stratified models. We see that the 95% credible interval for the overall effect, θ, does not include zero when drug formulation is the stratification variable ([0.156, 1.281]), but the interval *does* include zero for each of the three other stratification schemes.

We next examined individual strata results (i.e. posterior distributions for the λs). For the four-level model where drug formulation is the Level 3 variable, the probability that θ is greater than 0 for each of the individual drug formulations are each no less than 0.96. However, many of the 95% credible intervals did include zero, and so there was not strong support for a link between each drug and an increased risk of suicidality. Specifically, the credible intervals for Zoloft, Remeron, Serzone, and Wellbutrin included zero.

We present the results for the other three stratified models in Table 3. For the drug class variable, we see support for an association between SSRI class antidepressants and suicidality (95% credible interval [0.266, 1.267]), whereas we do not see similar support for an association involving the atypical antidepressants (95% credible interval [-1.041, 1.295]). Similarly, we see support for an antidepressant-suicidality link among longer studies, but not shorter ones (95% credible interval for ≤ 8 weeks and ≥ 9 weeks are [-0.191, 1.217] and [0.274, 1.334], respectively). Lastly, the results of our model support an antidepressant-suicidality link among studies of MDD, but not among studies of other diagnoses (95% credible intervals are [0.255, 1.292] and [-0.579, 1.218], respectively).

The posterior distribution of ω^2, the Level 3 variance, shows that this variance may be quite small, with each of the 95% credible intervals having a lower bound of just 0.001 (see Table 2). Thus, this model is potentially not very different from the random and fixed effects models where this component of variance is set to $\omega^2 = 0$. However, the DIC values indicate that the extra component of variance does not overfit the data as compared to the fixed and random effects models. In fact, the DIC values for the drug formulation model are consistently lower than the corresponding fixed and random effects models for all four subsets of the data. The DIC values for the other categorization models are not lower than the fixed effects model DIC, but are all within or nearly within the two-point equivalency range (see Table 2). Overall, the DIC values suggest that the four-level models do not diminish, and often improve, the fit of the model to the data as compared to the fixed and random effects models.

We also compare the fit of the four-level models with each other using Level 3 focused DIC. We did not find any one model fitting the data better than any other since the DICs all fall within a two point range from 153 to 155.

Sensitivity Analysis We found the model robust to changes in the prior mean and variance of the overall effect θ. Moving the mean from zero (no effect) to 2 (strong effect) and moving the prior standard deviation from 3 to 0.7 changed the posterior probabilities of an effect only slightly. The conclusions of our analysis do not change even when we use a strong prior (variance $= 0.49$) centered at a large effect (log OR $= 2$, OR ≈ 7). For example, in Figure 3, we display 95% credible intervals for the overall effect θ and group effects λ_{MDD} and $\lambda_{\mathsf{OtherDiagnosis}}$ under six different prior means and variances for θ. The overall credible interval only excludes zero (no effect) when the prior has mean 2 and variance 0.49. The MDD credible interval does not include zero and the Other Diagnoses credible interval does include zero for all of the prior distributions we examined.

We conclude from the variance sensitivity analysis that while the mean effects for the groups become more different as ω^2 increases, the significance of an association between suicidality and antidepressant use remains strong at all values of ω^2 for those groups identified as at significantly higher risk when the variance was unrestricted. Similar results were obtained for changes in the within-group variances τ_j^2.

Lastly, our exploration of the sensitivity of our results to single study influence via leave-one-out analyses showed that the exclusion of any single study did not change the conclusions of our analysis.

Conclusions The results of the analysis using Bayesian hierarchical models shows strong support for the FDA conclusion that significant increased risk of suicidality is associated with antidepressant use among children being treated for depression or being treated with an SSRI. However, the analysis does not support a similar conclusion for children being treated for other psychiatric diagnoses or being treated with an atypical antidepressant, a result that has not been reported previously.

2.4 Discussion

The four-level model presented here overcomes many of the limitations of the standard fixed and random effects models applied by the FDA. First, this model follows methodology outlined by Warn, et al. to directly incorporate the binary data into a model rather than model the log OR derived from the data [13]. This is an improvement because it eliminates the "plug-in" estimator for the log OR variance, which artificially deflates the variance for the overall estimate. Second, the random effects model assumed that the log OR for each study was exchangeable regardless of study properties. The hierarchical model relaxes this assumption by only modelling similar studies as exchangeable. Third, the hierarchical model approach addresses more complex and scientifically-relevant inquiries. For the question of association between suicidality and antidepressant use, the fixed and random effects models could answer questions only about overall effect for all antidepressant formulations for all psychiatric diagnoses. The FDA attempted to answer the question of association within particular groups by applying the fixed and random effects models to subsets of the data. While this would have been appropriate if an adequate number studies had been available, the limited quantity of our data makes it difficult to distinguish between lack of association and lack of power. The hierarchical model shares information across groups to help increase the effective sample size and increase the power.

In addition, the Bayesian approach gives the hierarchical model several advantages over the frequentist methods used by the FDA. First, it allows exploration of the balance between the data and prior distributions to better understand the sample size/power issues. It also easily includes trials with no events in either arm, thus incorporating 16% of the data that the FDA omitted from their analyses. Finally, the models presented here provide posterior probabilities of associations, which are more clinically meaningful than frequentist confidence intervals.

Generally speaking, the Bayesian hierarchical model approach demonstrated in this case study is appropriate for many medical meta-analyses–particularly those with binary outcomes. The advantages outlined for the antidepressant case study are broadly appealing. Specifically, the Bayesian hierarchical model presented here:

1. directly models binary outcomes
2. relaxes the assumption that all studies included in the analysis are exchangeable
3. addresses more complex and scientifically-relevant questions by modelling study-level effects
4. allows exploration of power issues via exploration of robustness to prior distribution specification
5. incorporates trials with no events in either arm without adjustment provides posterior probabilities of effects, which are more clinically meaningful than frequentist confidence intervals.

Unfortunately, even with all its advantages, the Bayesian hierarchical model presented here can not overcome limitations to the generalizability of the results to the population at large due to restrictive trial eligibility or limited scope of the trial design. The specific reasons for limited generalizability are almost as numerous as the randomized studies themselves. For our case study, there are three main reasons the results can not be generalized to the pediatric population at large:

1. This analysis addresses the question of an association between antidepressant use and *suicidality*, not *suicide*. The link between suicidality and suicide is one that has not been studied extensively, and so we must be cautious in concluding that antidepressants relate to suicide in the same way they may relate to suicidality.
2. All of the studies included in this analysis lasted sixteen weeks or less. The risk/benefit balance for antidepressant use in children may tip the other direction to show a decreased risk of suicide associated with antidepressant use, if the drugs were given more time to work. That is, even if antidepressants are associated with an increase in risk of suicidality in the first sixteen weeks, they may also be associated with a larger decrease in risk of suicidality after 16 weeks. In this situation, the risk/benefit analysis may change dramatically.
3. Most of the studies analyzed here excluded children deemed to have a high baseline risk of suicide. These children may react differently to antidepressants than those with a baseline risk of suicide, and so the results of this analysis can not be readily generalized to this higher-risk group.

Limitations to generalizability can be overcome with more data, specifically data that informs us about the portions of the population or specifics about the intervention that are missing from our RCT meta-analysis. Again, the specific data that would be useful for a particular analysis varies uniquely with each application, and could include prospective or retrospective observational studies, administrative databases, or surveys. For our antidepressant case study, the additional data must include enough subjects to use suicide (rather than suicidality) as the outcome event, or it must somehow relate suicidality to completed suicide. It must include data collected over long periods of antidepressant use, and it must include *all* groups of children and adolescents who may benefit from antidepressant use (including those at baseline risk for suicide). These three qualities would all be present in large administrative databases. The British national health database [14] and a claims database for a collection of HMOs in the United States [15] have each been analyzed in attempts to shed more light on the nature of any relationship between suicide and antidepressant use in children and adolescents.

The United States claims database is a record of all paid insurance claims for both health care services and prescriptions garnered from several HMOs (including HMOs involved in Medicaid managed care) located in all four major regions of the United States. This database provides information about roughly 25,000 depressed adolescent patients from all walks of life and with a wide range of depression severity and suicide risk. Current research using this database indicates there is no relationship between antidepressant use and suicidal behavior in adolescents with MDD [15].

The United Kingdom General Practice Research Database contains information on about nine million patients of UK primary care practices, including the diagnoses, treatments and mortality of each individual. Again, the patients in this database cover the spectrum of possible patient characteristics. Current research using this database examines the rates of suicide among children and adolescents who filled prescriptions for three newer generation antidepressant or a tricyclic antidepressant, and does not find any significant difference between the tricyclic and newer generation antidepressants [14].

These two administrative databases offer long-term data on a potentially more representative sample of the US and UK populations, and may even be large enough to use completed suicide as an outcome variable. By combining these two administrative databases (as well as any other available data) with the randomized controlled trial data in a single analysis, we may be able to obtain a better understanding of the relationship between suicide and antidepressant use in children and adolescents. If we are not able to obtain access to the raw data from these two databases, sufficient summaries of the each have been published for us to simulate possible data sets that will be similar enough to the true data to allow development of methodology to synthesize the database and RCT data. The remainder of this proposal describes the extensions to the current Bayesian hierarchical model that would be needed to appropriately synthesize all the data. Although we are inspired by the suicide case study, the current model and proposed extensions could be used in a range of applications.

3 Current Approaches to Evidence Synthesis

A formal synthesis of the evidence that is available from *all* data sources can be valuable for investigating the effects of treatments when implemented across large and diverse populations, and in general service settings (i.e., outside of the controlled clinical setting). In addition to RCTs, appropriate sources of data could include open label (uncontrolled) trials, prospective or retrospective observational studies, administrative databases and surveys. Regardless of source, the individual subject data may be available or the data may only have been released in aggregate. Perhaps the most valuable achievement obtained by combining studies of various designs is the capturing of all the strengths of the various designs, while at the same time mitigating their limitations. In our suicide example, observational studies allow for greater generalizability, while the RCTs contribute greater internal validity. When taken alone, RCTs provide powerful evidence for causality; the degree to which this causality evidence is eroded by adding observational studies to the analysis is not clear. Instead of focusing on the causal nature of research synthesis, we limit ourselves to associative conclusions.

Despite the interpretational advantages of using a variety of data sources, evidence synthesis has traditionally concentrated on information derived from RCTs (via meta-analysis). Current alternatives to RCT syntheses include: (1) qualitative reviews of empirical studies, (2) descriptions of patients treated outside of research protocols such as in open label trials or epidemiological studies, (3) reliance on expert clinical opinion, and (4) analysis of large administrative databases such as health insurance claim files. All of these only utilize *similar* sources of data, but combining data from *different* sources would reveal a more complete picture of the evidence. Very little work has focused on determining appropriate methods to combine data from dissimilar sources. Three main approaches have been proposed: the Confidence Profile Method, the response surface methodology, and cross-design synthesis.

3.1 The Confidence Profile Method

In the late 1980s, Eddy introduced a general method to reduce biases in studies, which he named the confidence profile method (CPM) [16, 17, 18]. He proposed a variety of applications for his method, including combining different types of data into a single analysis. The CPM was conceived for problems in which a researcher would like to estimate a parameter θ, but has data that can be used only to estimate this parameter with bias.

For a more concrete example, consider combining two types of studies attempting to measure the same parameter θ. The first type of study is a controlled trial performed in a clinical setting; we believe this trial measures the parameter without bias. The second type of study is performed in the field; we believe this trial to have some bias, which could destroy the exchangeability of this study with the first type. We model the parameter actually measured in the field (θ^*) as a function of the parameter we want to measure (θ) and some other parameters (α and β), which may or may not be known. For simplicity, we assume a linear relationship:

$$\theta^* = g(\theta) = \alpha + \beta\theta.$$

Assume that the measurements from the clinic studies (x_i) come from a normal distribution with mean θ and standard deviation σ_x:

$$x_i \sim \text{Normal}\left(\theta, \sigma_x^2\right) \quad , \; i = 1 \ldots n$$

The CPM gives a similar distribution for the measurements from the field studies (y_j):

$$y_j \sim \text{Normal}\left(\theta^*, \sigma_y^2\right) = \text{Normal}\left(g(\theta), \sigma_y^2\right) = \text{Normal}\left(\alpha + \beta\theta, \sigma_y^2\right) \quad , \; j = 1 \ldots m$$

In eliminating the bias of the second type of study, the CPM induces partial exchangeability between the first (unbiased) type of study and adjusted versions of the second (biased) type of study so they can be combined into a single analysis or research synthesis. This adjustment to the likelihood is a simple calibration example; Eddy has formulated adjustments for a variety of biases, a sample of which are listed in Table 1.

In Eddy's proposal of the CPM, he focuses on cases where the bias-correcting parameters (here α and β) are considered be fixed and known, but also included examples of how to incorporate uncertainty about these parameters via Bayesian methods [17]. More recently, Wolpert has revived the mostly overlooked CPM in a full Bayesian construction under the name *adjusted likelihoods*, a term originally coined by Eddy [19].

3.2 Response Surface Methodology

In the late 1980s and early 1990s Rubin proposed applying RSM to research synthesis [20, 21]. RSM has been most widely been applied in experimental situations. Its goal is to adequately describe the response of an experiment as the experimental conditions are varied by approximating the response surface as an n^{th} order polynomial of the experimental conditions [22].

In the usual RSM context, the experimenter is interested in discovering some optimal point that maximizes or minimizes the experimental response over certain ranges of experimental conditions. To find such a point, she uses RSM to determine the optimal positions in the experimental condition

Bias	**Parameter Adjustment**	
Dilution and Contamination	$\theta^* = (1-\alpha)\theta + \alpha\theta^u$	$\alpha \in [0,1]$ = the proportion of people who are in the treatment group but do not get treatment θ^u = outcome for those who do not get the treatment
Measurement Error of a Dichotomous Outcome	$\theta^* = (1-\alpha)\theta + \beta(1-\theta)$	α = P(record failure | true success) β = P(record success | true failure)
Loss to Follow-up	$\theta^* = \frac{(\theta - \lambda\theta^L)}{1-\lambda}$	λ = proportion of subjects lost to follow-up θ^L = outcome in the lost group
Measurement Error of a Continuous Outcome (Calibration), or Population Bias	$\theta^* = \alpha + \beta\theta$	
Intensity Bias, dichotomous	$\theta = \frac{(\theta^*)^{1/\tau}}{(\theta^*)^{1/\tau} + (1-\theta^*)^{1/\tau}}$	τ = the effect with the actual intensity divided by the effect with the intended intensity
Intensity Bias, continuous	$\theta^* = \tau\theta$	τ = the effect with the actual intensity divided by the effect with the intended intensity

Table 1: Examples of Adjustments for the Confidence Profile Method [18]

domain to most efficiently describe the response surface. She will also often perform several studies in succession, using the previous experiments to determine the direction of steepest descent or ascent to determine the best domain points for her next set of experiments. In each of these stages, the experimenter would traditionally use least squares estimation or ridge regression to estimate the local surface, but work has been done to apply Bayesian methods to the problem [22]. Once the location of the optimal point has been found, we use the polynomial models used to estimate (predict) the response at that point.

In the context of meta-analysis, we do not have the luxury of choosing our experimental points, but we can still use the idea of a response surface to predict response values at pre-specified points, and use RSM to make suggestions as to which future studies would contribute the most to our knowledge of the response surface. Vanhonacker takes a classic RSM approach to meta-analysis, using standard least squares to estimate the effect of explanatory variables of interest on the effectiveness of advertising when other nuisance variables are set to a combination that had not been present in any individual study [23]. This estimation is at the heart of the goal of combining data from different types of studies, since we can then specify an "ideal" study design and find the treatment effect if that design was possible. Vanhonacker's model is:

$$Y_i = Z_i\beta_1 + X_i\beta_2 + \epsilon_i \tag{7}$$

$$Y_i = Z_i\gamma_1 + X^*\gamma_2 + \xi_i \tag{8}$$

where Z denotes the vector of variables of interest, X is the vector of nuisance variables that capture the study design set to the values that were actually measured and X^* is the vector of nuisance variables set to the point of interest. Thus, the model in Equation 7 describes the situation we actually see in our experiments and the model in Equation 8 describes what we would have wanted to measure, with the overall goal being to accurately estimate γ_1. Vanhonacker made the usual assumption that the errors ($\{\epsilon_i\}$ and $\{\xi_i\}$) are each identically and independently distributed with a Normal distribution. He relates the two models to find an unbiased estimate of γ_1 as a function of β_1 and γ_2; he then substitutes the least squares estimate of β_2 for γ_2 to calculate an estimate of γ_1:

$$\hat{\gamma}_1 = \hat{\beta}_1 + (Z'Z)^{-1}Z'(X - X^*)\hat{\beta}_2$$

Although this seems to be a reasonable approach, Vanhonacker has not explored the effect of estimating γ_2 with β_2; it is possible that the mean squared error (MSE) of his method is larger than the MSE of simply using β_1 to estimate γ_1.

Vanhonacker has also made two very strong assumptions that are not likely to hold in practice for meta-analysis. First, he assumed linear bias based on the design variables. Second, he assumed a very restricted variance structure in which there is no between-study variance and homogeneous within-study variance. Methods have been developed to relax first assumption. The most common method of relaxing the second assumption is to assume a hierarchical model that allows the structure of the within- and between-study variance to be very complex.

3.3 Cross Design Synthesis

In 1992, the GAO proposed an approach to research synthesis that they termed cross-design synthesis (CDS). CDS differs from CPM and RSM in that it is a broad approach to research synthesis, rather than a specific technical method to combine a collection of datasets. (In fact, CDS includes both CPM and RSM as possible steps in an analysis.) The GAO assumed access to only aggregate data (and not individual patient-level data) and focused on the specific problem of combining randomized controlled studies and administrative databases. As such, CDS is composed of four steps:

1. **Examination of random studies for external validity.** The GAO suggested two approaches to completing this task: assessment of the patient selection process (including the definition of the inclusion criteria, methods for recruitment, exclusion criteria, and self-selection of eligible patients), and empirical assessment of the resulting patient pool (including comparison of the trial pool to the general population on both demographics and outcome levels in a method that is independent of treatment assignment in the trial).

2. **Examination of administrative databases for internal validity.** The GAO focused on the "balance" of comparison groups for this task. They suggested assessment of the methods used by the data base analysts to identify comparison groups (either "natural cohorts" or treatment/control groups) and to balance these (using, for example, matching, propensity scores, poststratification, or regression). Further, they propose exploring evidence that such a balance was actually achieved, using methods that include "no difference tests, sensitivity analyses, and goodness-of-fit tests". The GAO report also promotes comparison of the administrative database conclusions with the conclusions of other studies (both observational and random) to identify (if the studies do not match) or rule out (if they do match) imbalances. Finally, CDS includes secondary analyses to test for imbalances if the original data is available.

3. **Adjustment of data to improve validity.** Although Eddy's approach was noted by the GAO in their report describing CDS, they advocate the use of weighting procedures to improve the generalizability of the RCTs and to improve the internal validity of the databases. For the databases, the GAO also suggests the use of *a priori* adjustment of effect (via the CPM) and

computing the effect from a treatment/control comparison if a natural cohort comparison had originally been used, and vice versa.

4. **Combination of information within and between study types.** This step is further subdivided into three substeps:
   (a) *Create a framework for the research synthesis* based on two stratifications. The first stratification is based on study type (randomized vs. observational). The second is based on the patient population included in the trials so that the patients in the RCTs are all in one group, while the observational study patients are divided according to how similar they are to the RCT patient pool.
   (b) *Combine studies within each study design,* while accounting for study "certainty", differences in design, and reliability. This may be accomplished through the use of quality weights, projection (via RSM), models (via CPM) or stratification.
   (c) *Synthesize information across study designs,* either by using RSM to project to the strata with no observations (which are included by design in step 4a), by simply presenting the stratified results, or by using quality weights, projection (via RSM), models (via CPM) or stratification.

4 Proposed Work

We propose to use the CDS approach as a framework to develop more complete methodology to synthesize RCT and observational data. We will use the antidepressant case study to demonstrate the use of our methodology; although we will most likely not have access to administrative data within the timeframe of this dissertation, we will use published summaries of administrative data to simulate reasonable approximations to these databases and other potential sources of data (whether or not they actually exist for this topic).

We will follow the CDS to complete the first three tasks of evaluating the RCTs for external validity, evaluating the administrative databases for internal validity, and adjusting the data as appropriate to improve validity. However, the fourth task (combining information within and between study types) is the most challenging is the area we focus on for clarification of ideas and methodological improvement. We discuss each of this task's three substeps separately.

4.1 Create a framework for the research synthesis.

We agree that this framework is necessary, but also believe the GAO-suggested two-strata framework is too simple. In our case study, the patient characteristics differ across the 24 RCTs, and therefore should not be lumped into a single stratum. Instead, we plan to use RSM to model the effect size as a function of patient and trial characteristics. We will still be mindful of the "holes" in our data where we do not have any study whose patients are characterized by certain combinations of individual and study level characteristics.

4.2 Combine studies within each study design.

We have already constructed a Bayesian hierarchical model to combine RCT results. The model we constructed may be too simple, since we did not consider interaction of multiple study characteristics. We will explore expanded models to see if any of these fit the data better. We must also construct a similar model to combine data across the simulated administrative databases. For both of these, we must consider using weights, propensity scores or the CPM to be sure the populations being combined are adjusted to be comparable where appropriate.

When constructing these models, we must keep in mind two general concerns for all Bayesian hierarchical models: exchangeability and prior distributions.

4.2.1 Exchangeability.

Mathematically speaking, exchangeability of random variables $\lambda_1 \ldots \lambda_n$ occurs when the permutation of the indices $1 \ldots n$ does not change the joint probability function of the random variables. As aptly stated by Spiegelhalter, et. al. the assumption of exchangeability is "mathematically equivalent to assuming the [λs] are drawn at random from some population distribution, just as in a traditional random-effects model." [24]

Practically speaking, exchangeability is usually assumed when there is lack of information that differentiates data points, or when we have no reason to believe that subsets of the random variables are systematically different from each other. For example, measurements of a sample of people are usually assumed to be independently and identically distributed Normally with the same mean and standard deviation, since the researcher has no information about the people in the sample that would lead her to believe she should differentiate them into subpopulations with different means or standard deviations. Bias also plays a large role in exchangeability.

With the need to induce exchangeability in mind, the usual method of creating a hierarchical model for meta-analysis is to *a priori* classify each study according to its relationship to the other studies, then classify how these groups are related to each other, and continue until all the studies fit into an exchangeability structure. Malec and Sedransk have taken a different approach. Instead of creating the structure for the hierarchical model a priori and then applying this set model to the data, they explore a method to let the data instead of investigator intuition guide the classification of studies as exchangeable. In essence, they allow the classification of a study also be a random variable that is included in the hierarchy. [25]

We will explore the use of Malec and Sedransk's methodologies in conjunction with the CPM to combine information from studies with similar designs.

4.2.2 Prior Specification.

While prior distributions for means may be easily elicited, specification of priors for variances is more challenging. In addition, a "noninformative" prior would be useful for interpreting the influence of the prior specification on the model results. Daniels, DuMouchel and Spiegelhalter have all proposed similar "noninformative" prior distributions for between-study variances (τ^2 in Equation 4) [26, 27, 28]. However, their results are restricted to a two-level model, and do not address how they may be adapted to suit multi-level models. For example, in the model discussed in Section 2, these priors are not readily adapted for the between-study variance (ω^2).

To illustrate the type of "noninformative" priors currently in use, we describe Daniels' approach [26]. He notes that the estimator of a mean via a Bayesian hierarchical approach with conjugate priors amounts to a shrinkage estimator, where a function of the variances in the levels of the hierarchy constitute the shrinkage parameter. Because the shrinkage parameter is bounded between zero and one, a uniform prior on the shrinkage parameter may be uninformative for the shrinkage, and is thus in this sense uninformative for the variances. Daniels did not examine extensions of this idea to incorporate the shrinkage from several tiers of a hierarchy.

It seems straightforward to consider shrinkage parameters at several levels of a hierarchy. As a simple example, consider the model for our case study (Equations 1 through 6), where I=J=1, and consider two levels: (1) the overall mean, θ, and (2) the group mean λ. The posterior mean for each of these is a shrinkage estimator:

$$\begin{array}{ll} \hat{\theta} = \left(\frac{\phi^2}{\omega^2+\phi^2}\right)\lambda + \left(\frac{\omega^2}{\omega^2+\phi^2}\right)\nu \quad , & \text{shrinkage parameter} = \frac{\omega^2}{\omega^2+\phi^2} \\ \hat{\lambda} = \left(\frac{\omega^2}{\tau^2+\omega^2}\right)\mu + \left(\frac{\tau^2}{\tau^2+\omega^2}\right)\theta \quad , & \text{shrinkage parameter} = \frac{\tau^2}{\tau^2+\omega^2} \end{array} \tag{9}$$

The second line of Equation 9 represents the model for which Daniels and others have proposed a uniform prior on the shrinkage parameters. Specifically, Daniels proposed using a uniform prior for the shrinkage parameter ($\frac{\tau^2}{\tau^2+\omega^2}$) and an independent log-uniform prior for ω^2. We propose adding a second layer to accommodate the second layer of our model, giving us a uniform prior for the shrinkage parameter $\frac{\tau^2}{\tau^2+\omega^2}$ conditional on ω^2, a uniform prior for the shrinkage parameter $\frac{\omega^2}{\omega^2+\phi^2}$ conditional on ϕ^2, and an independent log-uniform prior for ϕ^2. The implications of structuring the prior distributions are not apparent. What prior distributions does this scheme induce on each of the variance components? How do these distributions compare to the usually-employed conjugate Inverse Gamma distributions?

Of course, we will also devise methods to assess the sensitivity of our results to prior density specifications.

4.3 Synthesize information across study designs.

Again, we will use RSM to synthesize the available RCT and simulated administrative information. Generally speaking, we will add the type of study (RCT vs. administrative) to the model as another level of the hierarchy. Of course, we will also need to account for the interaction of the study type with other study variables. In this overall synthesis, we will also explore exchangeability and prior specification (as discussed in the previous subsection). We will additionally be concerned with projection of our estimates to parts of the variable space that have little or no information, as these regions of the variable space may represent a sort of “ideal” study design.

4.4 Summary

The limited internal validity of observational studies has long been recognized as a caveat when interpreting results of analyses of this type of data. This weakness caused the medical community to rely heavily on RCTs because their randomization reduced the possible impact of confounding. However, the community is just now really recognizing the limits of RCTs in terms of generalizability and power to detect associations with rare events. As such, there has been an increasing call for methodology to combine RCTs with observational data to obtain a better understanding of relationships between interventions and outcomes. To date, there have been three serious proposals as to how to accomplish this synthesis: CPM, RSM, and CDS. All three approaches have desirable properties, but none has been fully implemented in a real practical problem. For this dissertation, we propose to build on the framework provided by these three approaches to develop and refine a practical methodology to synthesize data from diverse sources. We propose to apply our developed methodology to a case study of the association between antidepressant use and suicide in children and adolescents; for this application we will use all available data, supplemented where necessary or desirable by simulated data.

References

[1] Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. Russell Sage Foundation; 1994.

[2] FDA. FDA launches a multi-pronged strategy to strengthen safeguards for children treated with antidepressant medications. `http://www.fda.gov/bbs/topics/news/2004/NEW01124.html`; 2004.

[3] FDA. FDA announces series of changes to the class of marketed non-steroidal anti-inflammatory drugs (NSAIDs). `http://www.fda.gov/bbs/topics/news/2005/NEW01171.html`; 2005.

[4] Hannen EL, Racz MJ, Walford G. Long-term outcomes of coronary-artery bypass grafting versus stent implantation. N Engl J Med 2005;352:2174–83.

[5] Leslie, Newman, Chesney, Perrin. The Food and Drug Administration's deliberations on antidepressant use in pediatric patients. Pediatrics 2005;to appear.

[6] Hammad TA. Review and evaluaton of clinical data. `http://www.fda.gov/ohrms/dockets/ac/04/briefing/2004-4065b1-10-TAB08-Hammads-Review.pdf`; 2004.

[7] Hammad TA. Results of the analysis of suicidality in pediatric trials of newer antidepressants. `http://www.fda.gov/ohrms/dockets/ac/04/slides/2004-4065S1\_08\_FDA-Hammad.ppt`; 2004.

[8] Kaizar EE, Greenhouse J, Seltman H, Kelleher K. Do antidepressants cause suicidality in children: A bayesian meta-analysis. Clinical Trials 2005;Submitted.

[9] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–748.

[10] DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986;7:77–188.

[11] Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: A comparative study. Statistics in Medicine 1995;14:2685–2699.

[12] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (Pkg: P583-639). Journal of the Royal Statistical Society, Series B: Statistical Methodology 2002;64(4):583–616.

[13] Warn DE, Thompson SG, Spiegelhalter DJ. Bayesian random effects meta-analysis of trials with binary outcomes: Methods for the absolute risk difference and relative risk scales. Statistics in Medicine 2002;21(11):1601–1623.

[14] Jick H, Kaye JA, Jick SS. Antidepressants and the risk of suicidal behaviors. JAMA 2004; 292(3):338–343.

[15] Valuck RJ, Libby AM, Sills MR, Geise AA, Allen RR. Antidepressant treatment and risk of suicide attempt by adolescents with major depressive disorder - a propensity-adjusted retrospective cohort study. CNS Drugs 2004;18(15):1119–1132.

[16] Eddy DM. The use of confidence profiles to assess tissue-type plasminogen activator. In: Wagner GS, Califf R, editors. Acute Coronary Care 1987. Martinus Nihjoff Publishing Company; 1986. .

[17] Eddy DM. The Confidence Profile Method: A Bayesian method for assessing health technologies. Operations Research 1989;37:210–228.

[18] Eddy DM, Hasselblad V, Shachter R. Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence. Statistical Modelling and Decision Science. Harcourte Brace Jovanovich; 1992.

[19] Wolpert RL, Mengersen KL. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: Effects of environmental tobacco smoke. Statistical Science 2004; 19(3):450–471.

[20] Rubin DB. A new perspective on meta-analysis. In: Straff KWWachter&ML, editor. The Future of Meta-Analysis. Russel Sage Foundation; 1990. p. 155–165.

[21] Rubin DB. Meta-analysis: Literature synthesis or effect-size surface estimation? Journal of Educational Statistics 1992;17:363–374.

[22] Box GEP, Draper NR. Empirical model-building and response surfaces. John Wiley & Sons; 1987.

[23] Vanhonacker WR. Meta-analysis and response surface extrapolation: A least squares approach. The American Statistician 1996;50:294–299.

[24] Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Wiley; 2004.

[25] Malec D, Sedransk J. Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. Biometrika 1992;79:593–601.

[26] Daniels MJ. A prior for the variance in hierarchical models. Canadian Journal of Statistics 1999; 27:567–578.

[27] DuMouchel W. Hierarchical bayes linear models for meta-analysis. Tech. Rep. 27; 1994.

[28] Spiegelhalter DavidJ. Bayesian methods for cluster randomized trials with continuous responses. Statistics in Medicine 2001;20(3):435–452.

	Model					
	Fixed Effects	Random Effects	Drug Formulation Grouping	Drug Class Grouping	Study Length Grouping	Psychiatric Diagnosis Grouping
Overall Drug Effect						
log OR	0.758	0.729	0.736	0.597	0.671	0.616
(95% Credible Interval)	(0.318, 1.220)	(0.253, 1.209)	(0.156, 1.281)	(-1.283, 1.869)	(-0.688, 1.652)	(-0.787, 1.754)
Pr(log OR > 0)	0.999	0.998	0.988	0.901	0.931	0.897
Between-Group Variance						
Mean	0	0	0.081	5.0	3.1	4.5
(95% Credible Interval)	(0, 0)	(0, 0)	(0.001, 0.658)	(0, 20.4)	(0, 11.3)	(0, 16.2)
DIC						
Level 2-focused	157.4	159.3	156.8	159.8	159.4	158.3
Level 3-focused	–	–	155.0	153.4	153.7	154.1

Table 2: Posterior summary measures for θ and ω^2 across models

Level 3 Variable	Overall or Subgroup	Mean log OR (95% Credible Interval)	Pr(log OR > 0)
Drug Class	SSRI	**0.762** **(0.266, 1.267)**	0.999
	Atypical	0.455 (-1.041, 1.295)	0.850
Study Length	$\leq$ 8 Weeks	0.581 (-0.191, 1.217)	0.940
	$\geq$ 9 Weeks	**0.811** **(0.274, 1.334)**	0.999
Psychiatric Diagnosis	MDD	**0.783** **(0.255, 1.292)**	0.996
	Other	0.496 (-0.579, 1.218)	0.874

Table 3: Posterior summary measures for study-level models

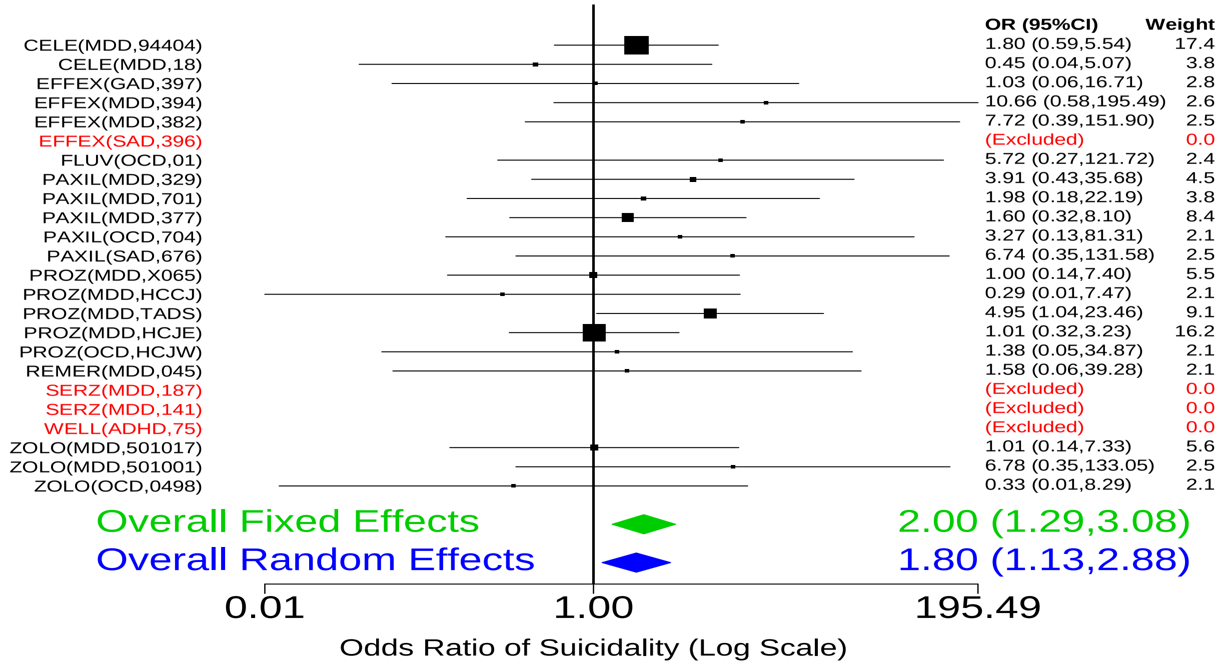

Figure 1: Forest plot of studies included in the meta-analysis; studies in red had no events in either the control or treatment arms.

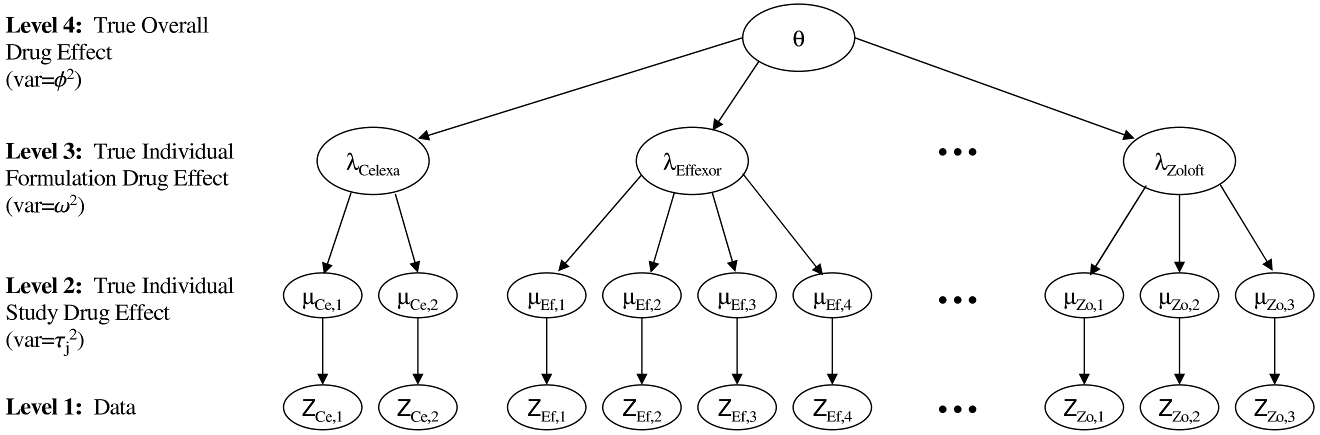

Figure 2: Bayesian Hierarchical Model for studies of SSRIs in treatment of MDD

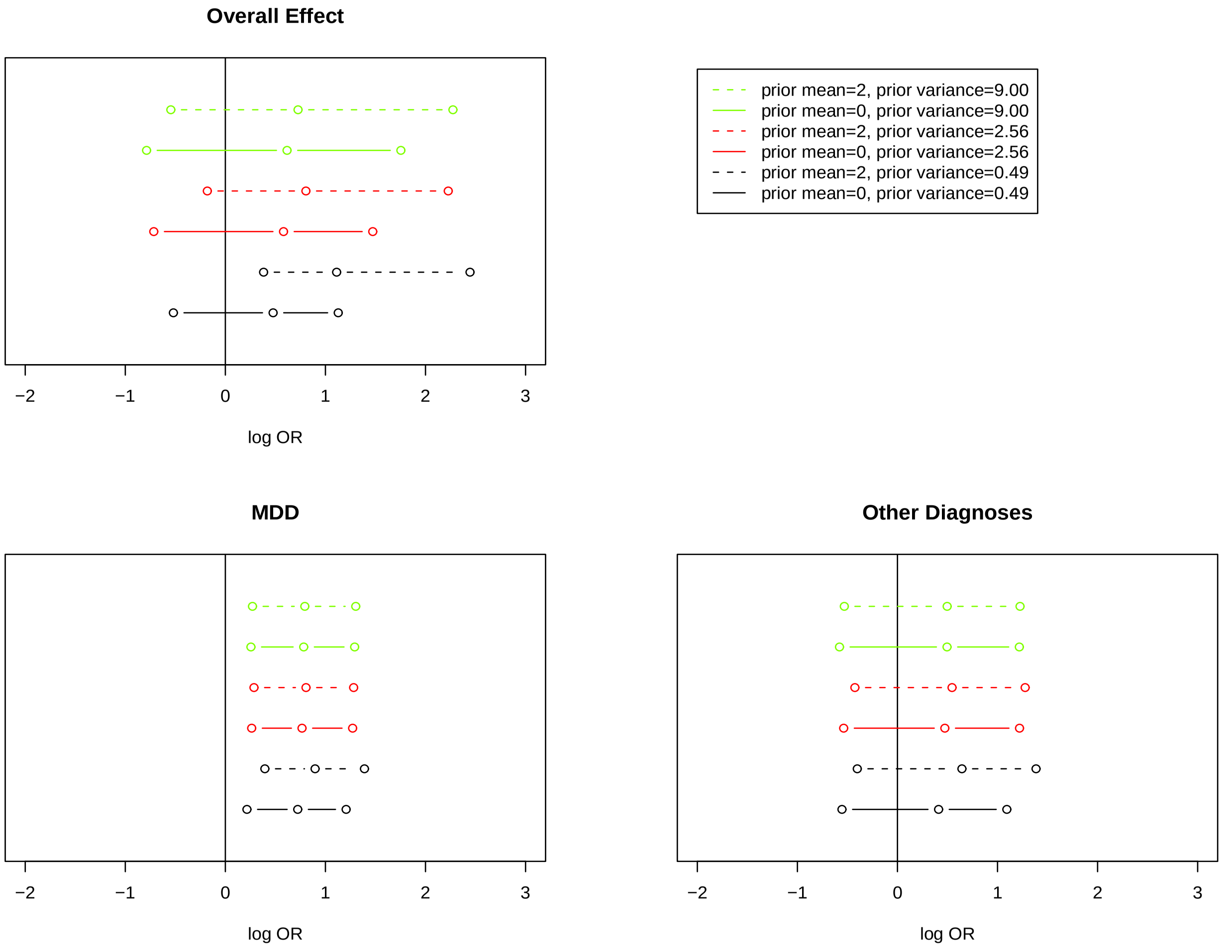

Figure 3: Sensitivity analysis for the prior mean and variance on the overall effect (θ) for the 4-level model in which the Level 3 variable is psychiatric diagnosis.