A Bayesian Meta-Analysis of Drug Court Cost-Effectiveness

P. Mitchell Downey
John K. Roman

DCPI District of Columbia Crime Policy Institute
Urban Institute | Brookings

Dec 2010

The District of Columbia Crime Policy Institute (DCPI) was established at the Urban Institute in collaboration with the Brookings Institution, through the jointly administered Partnership for Greater Washington Research with funding from the Justice Grants Administration in the Executive Office of the Mayor. DCPI is a nonpartisan, public policy research organization focused on crime and justice policy in Washington, DC. DCPI's mission is to support improvements in the administration of justice policy through evidence-based research. For more information on DCPI, see http://www.dccrmepolicy.org

This report was funded by the Justice Grants Administration, Executive Office of the District of Columbia Mayor under sub-grant: 2009-JAGR-1114.

Contents

1

Introduction

In recent years, meta-analysis has been widely adopted as a means of informing sound policy decisions. A particularly successful application of meta-analysis has been in the evaluation of the effectiveness of new medical interventions, particularly the results of new drug trials. The goal of this type of study is to draw a general conclusion from the results of many clinical trials of the same drug which often produce very different results to determine whether the drug is safe enough and effective enough to be made available to the public. Over the past two decades, meta-analysis has become increasingly common in the study of programs and policies designed to reduce crime.

More recently, researchers have begun to combine meta-analyses with cost-benefit analysis. Meta-analysis answers the question: does the intervention produce positive outcomes? Cost-benefit analysis moves beyond this question to ask whether an effective intervention creates enough of a benefit to justify the cost. Thus, the combined meta-analysis and cost-benefit analysis attempt to answer the question: What is the most cost-effective way to increase public safety?

DCPI has developed an empirical model that combines meta-analysis and cost-benefit analysis to help answer this question. The DCPI model uses Bayesian methods to test whether the expected outcomes of implementing a policy or combination of policies in Washington, D.C., is worth the investment. The DCPI model will incorporate both the costs of delivering services in Washington, D.C., and the benefits of those services to District citizens. In particular, the DCPI model will incorporate benefits to D.C. citizens from reductions in the risk of becoming a victim of crime.

The goal of this paper is to demonstrate how Bayesian statistics can be used in conjunction with meta cost-benefit analysis. To do this, we use data from 86 drug court evaluations previously coded for meta-analysis (Shaffer 2009). We then follow Drake, Aos and Miller (2009) in the development of estimates of the Washington, DC specific costs, recidivism rates and criminal justice system resource utilization. The calculation of the price of crime to victims is drawn from Roman (2009).

In this paper, we first describe a brief history of the use of a combined cost-benefit and meta-analytic model. We then discuss the advantages of using Bayesian statistics, rather than traditional methods, for applied policy research. Next, we apply this method to a practical policy problem: should the District of Columbia implement a drug court? We chose drug courts as our beta test of the model because there have been several prior meta-analyses of drug courts which allows us to assess the reliability of our effect size estimates. We discuss the Bayesian methods used in this analysis and our preliminary findings. We conclude with a discussion of how differences between our results and prior results should be interpreted.

We note that the goal of this research is to validate the model, in particular to compare effect sizes generated here with effect sizes generated by others using the same or similar data. In this model iteration, only the prices of treating District of Columbia residents are drawn from DC data. Thus, the findings in this

paper should not be interpreted as an evaluation of the Superior Court Drug Intervention Program (SCDIP). Future research will re-populate these models using DC-specific data which will allow us to estimate the optimal size and composition of the current drug court (SCDIP). Ultimately, the model can be used to make evidence-based decisions when policymakers are confronted with difficult choices between successful programs when the resources to fund those programs are limited.

1.1. PRIOR RESEARCH

In 1999, researchers at the Washington State Institute of Public Policy (WSIPP) created what is to our knowledge the first cost-benefit model that built upon meta analytic data. The goal at WSIPP was straightforward: WSIPP sought to take research results about what programs have been shown to prevent crime in other places and times, estimate the cost of implementing those programs in Washington State, and estimate the net costs and benefits of running those programs in Washington. The WSIPP cost-benefit model is widely acclaimed and represents a substantial step toward evidence-based governance.

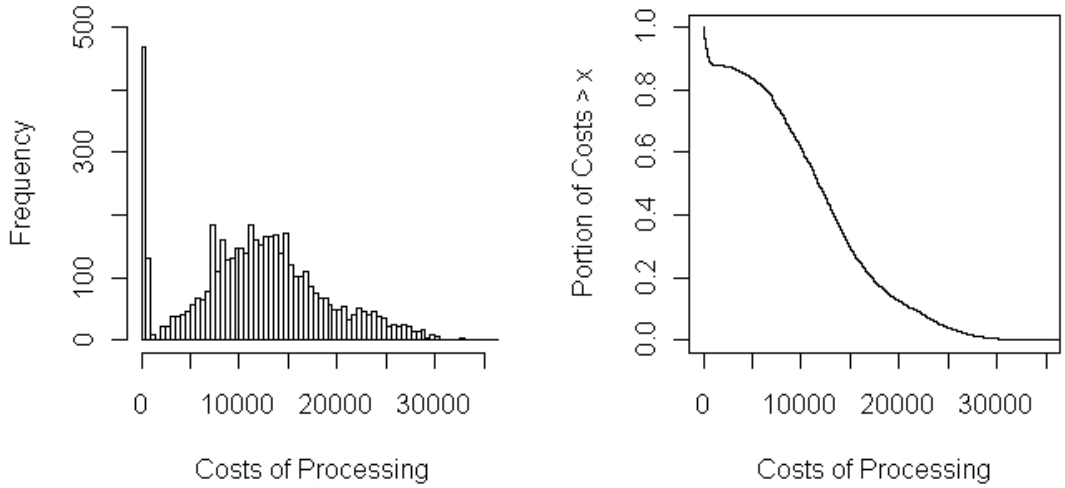
The WSIPP model uses frequentist statistical methods to estimate the costs and benefits of innovative policies and programs. Frequentist statistics, however, have several notable limitations in this application. First, there is a general problem in meta-analysis that if the underlying studies are unreliable, any estimate of the aggregated average effect is also unreliable. Second, the frequentist models generally do not account for uncertainty in each step of the estimation process, focusing on variation around the meta-analytic effect size, but not on the variation around costs and benefits. While researchers using frequentist statistics could estimate variation in each stage of estimation, combining that uncertainty across multiple estimates (of effect, costs and benefits) is a much more difficult problem for frequentists.

Before discussing the statistical advantages of a Bayesian model, it is important to point out some practical advantages of the approach. Most importantly, these models solve an ages old problem with frequentist statistics: frequentist statistics produce output that is extremely difficult for non-statisticians to interpret. For instance, frequentist analysis is commonly used to generate measures of central tendency (most often the mean) and other summary statistics which are simply convenient measures of distance from the mean of a sample. In the case of a normal distribution, a standard deviation, for example, identifies how far from the mean you would find two-thirds of all possible means of a sample. Two standard deviations from the mean include 95 percent of all possible means from a sample. This information is conventionally used to determine whether there is a real effect of an intervention.

For instance, a test of whether a drug court decreases recidivism might compare new arrests among a number of drug court clients and compare those outcomes to new arrests in a group of eligible non-participants. That test would produce a mean effect, for example a reduction from an average of 5 new arrests to 2 new arrests. However, if there was a lot of uncertainty around that estimate, for example, because there were a few people with many arrests than other tests would be needed to determine whether the difference could plausibly have been caused by chance. A typical analysis would test whether the plausible range of the true mean reduction in new arrests includes 0. Frequentist statistics produced in this analysis would include the standard deviation described above, and most notably, would produce an estimate of whether the probability that new arrests would have dropped from 5 to 2 purely by chance is less than 5 percent. If the researcher the probability of such a drop occuring purely by chance, in the absense of a true effect, is less than 5 percent, that finding is commonly labeled as a statistically significant effect. The choice of 5 percent is based on convention and is completely arbitrary.

In applied policy analysis, that way of presenting information is extremely cumbersome. It is equivalent to forecasting tomorrow's weather by stating that the mean expected rainfall is 0.40 inches with a standard deviation of 0.25 inches and thus there is a non-significant chance of rain. A much more intuitive forecast is to simply state that there is an 80 percent chance of rain with rainfall between one-quarter and one-half an

Figure 1.1: Examples of Cost Data



inch. In effect, this is precisely the output generated in a typical Bayesian analysis. Consider the following example.

Figure 1.1 describes the results of a simulation estimating the costs of processing typical drug court clients. The figure on the left shows the frequency with which each possible cost was observed in the data, with the most frequent observation being about 0 (suggesting the defendant had minimal involvement with the court, possibly only a single hearing) and then a large number of normally distributed costs averaging close to 13,000. The second figure presents the probability than any single court will have each level of costs. It shows, for example, that one in eight courts will have more than double the typical drug court costs. While this difference in non-significant, it is critical information to a cost conscious court system considering whether to implement a drug court. While frequentist statisticians could produce a similar finding, it is outside that tradition to do so. And, if those data in Figure 1.1 were produced by combining several distributions (such as the distribution of residential treatment, outpatient treatment, etc.) it would be almost impossible to reproduce Figure 1.1 using frequentist statistics.

1.2. AN ALTERNATIVE APPROACH

We are proposing using Bayesian statistics to address these problems. The Bayesian model explicitly introduces uncertainty into each step of the process. Rather than generating point estimates in each step that are then used as inputs into the next step without accounting for the uncertainty of the point estimate, we obtain a distribution for each of the parameters in our model. This is preferable to simply incorporating point estimates in subsequent steps.

In short, we believe that the key benefit of Bayeisan methods is realized in multistage models, where the results from one step are fed as input into the next step. In models like these, it is always possible to use a single point estimate as teh input into future stages. However, doing so implicitly assumes that this is exactly the right value, estimated with no error. This assumption disregards all common knowledge about statistics and produces spurious results that do not accurately reflect reality. Modern Bayesian methods, which permit using a range of plausible values as inputs into future steps, relax this unrealistic assumption and thus better describe the true reality.

1.3. DRUG COURTS

Drug Courts typically involve court-supervised treatment for non-violent, drug-involved offenders who meet legal eligibility guidelines (Harrell 1999). Typically, offenders are offered a reduction in penalties upon successful completion of the program and demonstrating that they are drug free. Participation is almost universally voluntary. Once clients are determined to meet the courts criminal justice and clinical eligibility criteria, the defendant is given the option of entering the court or continuing their case in a more traditional manner. While programs differ in their components, treatment participation is usually mandatory, a program of court supervision is imposed, illicit drug and alcohol use is monitored closely with frequent drug tests, and a schedule of requirements for program advancement and graduation (or failure) is defined upon entry, with a system of graduated sanctions/incentives in place to ensure accountability (Harrell, Cavanagh, Roman 1998).

Defendants targeted for Drug Court are generally non-violent offenders whose current involvement with the criminal justice system is due, primarily, to their substance addiction. Defendants eligible for the drug court are identified as soon as possible after arrest and, if accepted, they are offered an immediate referral to a multi-phased treatment program entailing multiple weekly (often daily) contacts with the treatment provider for counseling, therapy, and education; frequent urinalysis (usually at least weekly); frequent status hearings before the drug court judge (bi-weekly or more often at first); and a rehabilitation program entailing vocational, educational, family, medical, and other support services. Unlike traditional treatment programs, becoming clean and sober is only the first step toward drug court graduation. Almost all drug courts require participants to obtain a GED, maintain employment, be current in all financial obligations (which often includes drug court fees) and child support payments, if applicable, and to have a sponsor in the community. Many programs also require participants to perform community service hours to make restitution to the community they have harmed.

However, adult drug courts vary enormously in their practices. Despite the general approach shared among drug courts, screening and eligibility criteria, program completion requirements, treatment, sanctions, and termination criteria vary widely (Government Accountability Office [GAO], 2005). Drug courts handle cases before a defendant is sentenced through pre-trial diversion, post-plea (but pre-sentencing) intervention, or a combination of pre- and post-plea approaches. Some courts exclude individuals who have prior convictions or are only users of marijuana or alcohol, whereas others consider such users eligible to participate. Treatment, sanctions, and drug testing represent key components of drug court programs; however, referral to additional services and support (e.g., Alcoholics Anonymous), use of graduated versus case-by-case sanctions, and the nature and frequency of drug testing are not consistent across all programs.

Although most drug courts serve non-violent offenders, participants in different drug courts may display a range of demographic and socioeconomic characteristics, criminal histories, and substance abuse profiles. While a 2005 GAO review found that participants were generally in their early 30s, male, and unemployed upon program entry, these characteristics were not uniform across the programs reviewed. For example, the average age at program entry ranged from 24 to 36 years, the percentage of respondents who were male ranged from 46 percent to 88 percent, and the percent of participants who were employed at program entry ranged from 16 percent to 82 percent. Further, among the evaluations included in the GAO (2005) review, adult drug court completion rates ranged from 27 to 66 percent.

The vast majority of adult drug court evaluations have found that drug courts are associated with reduced recidivism. From the late 1990s through the mid 2000s, a series of narrative literature reviews agreed that most studies show reductions in recidivism (e.g., Belenko 1998, 1999, 2001; GAO 2005; Roman and DeStefano 2004). More recently, three reviews have emerged that employed formal meta-analytic techniques, enabling quantitative generalizations. The first, conducted by the Washington State Institute for Public Policy, considered evaluations of 57 adult drug courts (Aos, Miller, and Drake 2006). The review found that the average adult drug court reduces recidivism by 8 percentage points. The analysis determiend that although drug courts were, on balance, cost-effective, 7 of the 11 other adult criminal justice programs

considered were even more cost-effective. The next considered results from 55 sites, including 49 adult and six juvenile drug courts (Wilson, Mitchell, and MacKenzie 2006). In 48 of 55 sites, drug court participants had lower rearrest or re-conviction rates than their comparison groups and the sites averaged an estimated 13 percentage point reduction in recidivism. Shaffer (2006) employed comparable meta-analytic techniques with an overlapping, but slightly larger group of 61 adult and 21 juvenile drug court evaluations. This analysis reported an average recidivism reduction of 10 percentage points for adult drug courts and five points for juvenile drug courts. This analysis added that drug court programs designed to last from 8 to 16 months were more effective in reducing recidivism than those designed to last for either shorter or longer timeframes.

A third review conducted by the GAO (2005) omitted evaluations whose designs were seriously compromised: for example, by comparing only successful participants (graduates) to the comparison group or by making no effort to control for baseline differences between participants and comparison offenders. Previous reviewers had all drawn attention to the low scientific quality of much of the literature, and the Schaffer analysis detected a noticeably smaller effect size for studies implemented with a higher quality methodology. However, the GAO results continued to be positive. Drug courts significantly reduced the rearrest rate in 10 of 13 sites, and significantly reduced the re-conviction rate in 10 of 12 sites.

1.4. OVERVIEW OF THE PAPER

The cost-benefit analyses based on meta-analytic impact estimates discussed here can be usefully divided into three conceptual domains: estimated impacts, costs, and benefits. We discuss each in turn.

We first focus on the estimated impacts i.e., the programs effects. These are based on meta-analytic results. The first step in any meta-analytic model is to collect the data. We have, to this point, used data collected and coded for other meta-analyses. One critical dimension of the coding, particularly salient in social science program evaluations, is the quality of the evaluation. Fortunately, all social science meta-analyses carefully code this information, making any available meta-data suitable for our purposes. After coding, we model measured program effects according to a three-level hierarchical model (Smith, et al., 1995). This modeling uses a Bayesian framework with a diffuse prior, estimated using Markov chain Monte Carlo methods based on extensions of the hierarchical models presented in Gelman, et al. (2006) and Gelman and Hill (2005). The results of the quantitative meta-analytic model are distributions of the effect size. These include the estimated mean and variance of program effects, enabling the construction of a predictive interval of effectiveness. This concludes the meta-analytic impact estimates.

We next turn to the costs of program operation. Currently, no strong framework has been developed for the consideration of implementation costs in criminal justice programs. For this reason, we focus primarily on operating costs. From past cost-benefit analyses of the programs under study, we collect information about the resources (e.g., staff time) that went into program operations. That is, instead of collecting estimated costs of program participation, we collect the quantities of resources used. We then use extant data sources and expert perspectives from the agencies who would be operating the program to obtain DC-specific prices for these resources. In this way, we account for the fact that drug treatment or probation officers might be more costly in the District than in places where the program has been implemented before, but our estimated costs will draw heavily on past empirical evaluations of the program. Where possible, we collect a range of program costs to account for the fact that not all program participants require the same attention or involvement.

Finally, we must estimate the benefits of the program. Benefits can be conceptualized as the product of a quantity of resources saved and the price of those resources. Prices are based on budgets of involved agencies, which allow us to estimate, for instance, the cost of one year of probation. The prices of crime (the costs to society of criminal victimization) are based on jury-award data (Roman, 2009).

The quantities of resources saved are more difficult to estimate. The impact analysis provides the estimated range of arrests that will be prevented by program participation. Using various data sources from DC agencies, we are able to estimate what types of crimes these arrests most likely would have been. We are also able to estimate the probability that these arrests would have led to pretrial detention or supervision; criminal court trials; conviction; jail, prison, or probation; etc. Thus, from the estimated range of arrests prevented, we can extrapolate to the estimated range of the number of each of these costs that would be prevented. Our simulation-based Bayesian methods account for uncertain outcomes throughout the process and provide a final range of the costs avoided to society. These can be combined with the program costs to provide final estimates (including reasonable ranges and probabilities) of the net benefits of program operations.

1.5. BAYESIAN METHODS

Bayesian methods follow a different tradition than most statistics, referred to here as frequentist or classical.[1] This section overviews the features of Bayesian statistics which we consider relevant for our purposes, and interested readers are referred to Gelman, et al (2006) for a more comprehensive discussion of the comparative strengths and weaknesses of frequentist and Bayesian methods.

Bayesian methods are drawn from a well-accepted law of elementary probability:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{1.1}$$

where $p(\theta|y)$ is typically referred to as the posterior probability of θ, a vector of model parameters, conditioned on the data. That is, the output of a Bayesian analysis is a distribution defining the probability of each possible value of the parameter θ, given the data that was observed. Since $p(y)$ depends on neither the model nor the parameters, it is the same for all parameters and models. Thus, it contributes nothing to estimating particular parameters, and only serves to ensure that $p(\theta|y)$ integrates to 1, as all probabilities must. As such, it is usually referred to as a normalizing constant, and dropped from consideration, since any estimated distribution can simply be scaled by some c to ensure that it sums/integrates to 1. Therefore, Bayes' Theorem is typically reduced to:

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{1.2}$$

In the above equation, $p(\theta)$ is the prior distribution of θ and represents the analyst's prior beliefs about the parameters. The controversy over Bayesian methods most often centers on this. It is often argued that $p(\theta)$ makes the analysis entirely subjective. However, in most models, it is possible to choose some prior distribution that does not influence or has minimal influence over the final results. This is the approach taken here, where the prior distribution has little or no impact, negating most of the criticisms of Bayesian methods. In fact, if $p(\theta) \propto 1$, or we have no beliefs that any one value of θ is any more likely than another, then $p(\theta|y)$ is simply $p(y|\theta)$, which is simply the likelihood function. This is an important result.

Most frequentist analyses are based on the methods of maximum likelihood estimation. This means that given the model used, the estimated parameters are those which maximize the likelihood of the observed data. Intuitively, this makes sense, as the parameters are selected to maximize the probability that the true, observed data would be generated from the model, or maximize the likelihood that what *did* happen *would* happen. However, with no prior beliefs, the posterior distribution from a Bayesian analysis is identical to the liklihood function. Thus, while maximum likelihood methods select the peak of this distribution, Bayesian methods provide the full distribution. In this context, frequentist results are a subset of Bayesian

[1] The term classical is often used, although Bayesian statistics are typically considered older than "classical" statistics.

results. For this reason, Bayesian methods provide a much more natural way of dealing with complex models, where parameter values are contingent on the values of other parameters, and multi-stage models, where the output of one estimation is used as the input to another. In frequentist models, it is most often assumed that the estimated parameter is the true value of the parameter, ignoring the fact that it is estimated with uncertainty. Complex and multi-stage models simply take estimated values as true values.

Since both complex and multi-stage models use parameters to generate new parameters, it is vital that uncertainty in every stage of the process be carried on to the next. This is best done through Bayesian analysis. For this reason, we believe that Bayesian methods are the most appropriate and accurate for the models used here. Particularly given modern Bayesian methods, especially Markov chain Monte Carlo simulation, used in all analyses in this paper, a Bayesian approach provides more flexibility to estimate complicated models than any alternative. For example, frequentist analogs to some of the models presented later operate as Bayesian models except with additional assumptions that certain parameters are known, when these parameters are not known. We believe that these pragmatic concerns drastically outweigh any philosophical debate, and make a Bayesian approach most appealing and useful for practical policy analysis.

2

Data

2.1. META-ANALYTIC DATA

The data come from Shaffer (2009). Shaffer collected meta-analytic data through the traditional means of large literature reviews and careful coding of information provided in the published studies. Concerned that data aggregated in this way was insufficient to determine how drug courts work and what types are most effective, Shaffer constructed a detailed survey about procedures and programmatic elements and administered the survey by phone and mail to 86 drug courts. This approach allows a richer understanding of variation in drug courts, to provide more meaningful policy analysis than simply answering the question, "Do drug courts work?" In this analysis, we use the Shaffer data to estimate the adult drug court effect size. We do not exploit the additional information provided through surveys in the analyses conducted here.

2.2. COST DATA

To estimate the costs of operation for a drug court in the District of Columbia, we interviewed the Director of Treatment for the Federal Pretrial Services Agency, who heads the current drug court in the District (SCDIP). We asked a series of questions derived from personal experience in drug courts, past drug court research, and familiarity with the DC criminal justice system. We asked about costs (both financial and non-financial resources) of treatment, supervision and case management, sanctions, hearings, and administration. Throughout the interview, we put as much emphasis as possible on obtaining a reasonable range of plausible costs and outcomes, rather than simple point estimates. This was done in order to fully capture variation in drug court processing costs, which has been shown to be substantial (Downey and Roman, 2011).

Finally in order to estimate the costs associated with business as usual case processing, we use data from the Multi-Site Adult Drug Court Evaluation (MADCE), funded by the National Institute of Justice (Rossman, 2011). That study collected data on drug court participants in 23 jurisdictions, and data on drug court eligible arrestees in six jurisdictions. Since we are primarily interested in the new costs associated with drug court, we use the MADCE data to estimate the resources that would have been consumed by drug court participants had they not received drug court, and subtract those costs to obtain the new, additional (marginal) costs specific to drug court.

2.3. BENEFIT DATA

In this analysis, we only estimate benefits that accrue to the criminal justice system, and to victims of crime. While it is relatively easy to observe the costs to the criminal justice system, estimating harms to victims is much more complex. Robust estimates of the price of criminal victimization, measured as the costs of crime to victims, inform a wide range of policy analysis and are widely applied. However, the most commonly cited studies are constrained by limited data and cannot directly estimate prices and thus the studies cannot correct for sampling bias and do not report estimated variance in prices. A recent study (Roman, 2009) combines individual and aggregate data and analyzes individual-level data from two sources: jury award and injury data from the RAND Institute of Civil Justice and crime and injury data from the National Incident-Based Reporting System. Propensity score weights were developed to account for heterogeneity in jury awards with respect to legal claims. Data from the jury awards are interpolated onto the NIBRS data using combinations of all attributes observable in both data sets. From the combined data, estimates are developed of the price of crime to victims for thirty-one crime categories, and these prices of crime are used in this research.

3

Methods

3.1. META-ANALYTIC METHODS

Throughout this report, we denote the estimated effect from the ith study as e_i. Specifically, in all analyses, the log odds ratio of recidivism was used. In addition to being approximately normally distributed, the log odds ratio has a number of desirable properties for analyzing binary outcomes such as recidivism (Lispey and Wilson, 2006). The log odds ratio of recidivism for the treatment (T) and control (C) groups is defined as follows

$$e = Ln\left[\frac{p_T/(1-p_T)}{p_C/(1-p_C)}\right]. \tag{3.1}$$

A negative e_i indicates that drug court reduced the likelihood of recidivism and a positive e_i indicates drug court increases the likelihood. The maximum likelihood estimate of the log odds ratio of recidivism for the ith study can be simplified to

$$e_i = Ln\left[\frac{n_{iT}^{+}n_{iC}^{-}}{n_{iT}^{-}n_{iC}^{+}}\right] \tag{3.2}$$

which, as laid out in Wolpert and Mengerson (2004), has approximate variance

$$\sigma_i^2 = 1/n_{iT}^{+} + 1/n_{iT}^{-} + 1/n_{iC}^{+} + 1/n_{iC}^{-} \tag{3.3}$$

where n refers to the number of study participants, a superscript of + indicates recidivism, a superscript of − indicates non-recidivism.

3.1.1. Fixed and Random Effects Models

Meta-analyses take a variety of different forms, however most can be classified as either fixed or random effects. A fixed effects analysis assumes that all studies are measuring the same effect. In this case, the maximum likelihood estimate of the mean is

$$\mu = \frac{\sum_i^I e_i \sigma_i^{-2}}{\sum_i^I \sigma_i^{-2}} \tag{3.4}$$

and that for the variance is

$$\sigma^2 = 1/\sum_i^I \sigma_i^{-2}. \tag{3.5}$$

This formulation is the familiar aggregate mean, where each study's estimate is weighted by the inverse of its variance. The intuition is that each study's estimate is estimating the same thing: a single mean impact. Some studies provide a very precise estimate of this impact, while others estimate it with less precision. Thus, to identify the true mean impact, one should consider all studies, but base estimation more on those with more precise estimates.

The random effects formulation, however, does not assume that all studies are estimating the same true mean impact. Rather, it assumes that each study is estimating its own true impact. Those impacts are related, but not identical. More formally, it assumes that the true impact estimated by each study is drawn from a common population. For each study, there is a measured treatment effect (e_i), which is estimated with some uncertainty (the squared standard error of the estimate, σ_i^2). Following the convention in the literature (Higgins, Thompson, and Spiegelhalter, 2009; Sutton and Abrams, 2001; Turner, et al, 2009; Welton, et al, 2009; Wolpert and Mengerson, 2004) we will assume that the squared standard error of the estimate is the true variance of the effect, as Welton, et al (2009) has shown that this rarely affects analytic results. Thus, we have

$$e_i \sim N(\theta_i, \sigma_i^2) \tag{3.6}$$

where θ_i is the true effectiveness of the ith court. The second level of the hierarchical model assumes that the ith true treatment effectiveness is from a larger population of treatment effectiveness:

$$\theta_i \sim N(\mu, \tau^2). \tag{3.7}$$

The random effects model has a second source of variation. Within each study, there is σ^2. This can be thought of as the uncertainty of the study's estimated effect (e_i) or as the variance in esmtimated impact if one were to replicate the same analysis[1]. This is captured by σ_i^2. However, there is also variation in effectiveness among the various studies, or drug courts under study, captured by τ^2.

Frequentist or classical statistics also have random effects models. The difference between these models and their Bayesian analogs is small but important. The frequentist analysis estimates a single value τ^2 which is carried throughout the remainder of the analysis as though it is the true value. The Bayesian analysis, on the other hand, treats both parameters as being uncertain by estimating a full distribution for each. The implications of this are three-fold. First, the Bayesian approach allows the analyst to specify a prior distribution for each parameter, thus synthesizing the information in the data and any outside source of relevant information (Gelman, et al, 2006). Second, by assuming that the estimate of τ^2 is the true estimate, classical analyses do not generate a range of possible values for τ^2. Thus, though the analyst can be certain that the estimate of τ^2 is the most likely of possible values, there is little information about how far off one could potentially be from the true value.

[1]Though the first is a Bayesian way of thinking and the second is fundamentally frequentist, for our purposes, both can be used interchangably.

Finally, and relatedly, by assuming that the estimated value of τ^2 is the true value, a frequentist analysis greatly overstate the confidence in the final estimate of μ, which is often the primary parameter of interest in the model because it represents the overall average effect of the treatment. Sutton and Abrams (2001) showed that the variance of the estimate of τ^2 (a measure of uncertainty about the estimate) is often quite large and that including this uncertainty dramatically reduces the precision of the estimate of μ. Since τ^2 is, in fact, an unknown quantity, treating it as such and transferring that uncertainty throughout each stage of the analysis is an important and realistic step. By assuming that the maximum likelihood estimate of τ^2 is the true value, a frequentist analysis will artificially inflate confidence in the estimated aggregate impact of the program.

We use Bayesian random effects meta-analyses instead of Bayesian fixed-effects models because they have been shown to be most theoretically and practically appropriate in almost all circumstances (Wolpert and Mengerson, 2004). They are particularly important in our context for two reasons. First, though the mean effect size is relatively unchanged between fixed and random effects analyses, the variance of that estimated effect, which represents the uncertainty surrounding the estimate of the mean effect size, is significantly higher with random effects (Higgins, Thompson, and Spiegelhalter, 2009). Random effects seem appropriate given that our motivation for employing Bayesian methods is largely derived from attempts to fully represent uncertainty inherent in *any* predictive meta-analysis and to present policy-makers with the most realistic estimate of the full range of effects that can be reasonably expected.

Further, and more importantly, random effects estimation is most appropriate when the underlying studies are estimating different effects because the actual treatment itself is heterogenous, whereas fixed effects are appropriate when each study is estimating the same true underlying effect (Sutton and Abrams, 2001). Since each study is estimating the effectiveness of a different drug court and drug courts are notoriously heterogenous (Zweig, Rossman and Roman, 2010), assuming that each study is estimating the same effect (which is to say that each drug court studied in our sample is equally effective), seems highly inappropriate. As the saying goes, "If you've seen one drug court, you've seen one drug court." Thus, random effects models allow us to determine the aggregate, average effectiveness of drug courts without assuming that each drug court is equally effective (as fixed effects would do).

Finally, it is worth noting that fixed effects models are a special case of random effects models (Wolpert and Mengerson, 2004). If $\tau^2 = 0$, then there is no between-study variance, so all studies are estimating the same true treatment effect. In this case, the random effects model reduces to the fixed effects model. Thus, if our analyses reflect that much of the probability mass of τ^2 is near zero, fixed effects models may be more appropriate and are worth investigation. For these reasons, all analyses in this report are conducted using random effects models of the form laid out in equations (8) and (9).

3.1.2. Study Quality

It would be inappropriate, however, to analyze all studies collectively, because the 86 studies vary substantially in their quality. Thus, even when allowing the measured treatment effect to vary across the different studies, it seems inappropriate to allow poorly identified studies, most suceptible to bias, to have the same influence on the estimate of the aggregate mean as well-designed, well-executed studies. Meta-analysis, particularly in social science, has long recognized the importance of this and has adopted a number of approaches to address the problem. Examples abound in social science, criminal justice (Aos, et al, 2006), and even drug court research (Wilson, Mitchell and Mackenzie, 2006). We adopt many of the methods employed by those analysts.

Before accounting for study quality, we must define some metric by which to measure it. The primary metric employed in criminal justice research is the Maryland Scientific Methods Scale (SMS) for program evaluations. The scale defines five levels of scientific rigor, which are, in order of most to least rigorous (Sherman, et al, 1998):

1. Random assignment to treatment and control groups.
2. A comparison between multiple treatment and multiple control groups.
3. A comparison between the treatment group and a comparable control group.
4. Either a relationship between outcomes before and after a program, or the comparison of individuals in treatment and control groups with no efforts to ensure comparability.
5. Simple correlation between an outcome and some risk factor at a single point in time.

Studies which are more rigorous are expected to be less subject to bias and are more likely to reliably estimate the true treatment effect of interest. We see strong advantages to the use of the Maryland SMS, and Aos, et al (2001) have demonstrated that it is widely applicable to a number of different policy considerations. Perhaps its greatest strength is that it is not dependent on the context. Since it is purely methodological, it can be applied to any issue, and used for a variety of puposes.

However, in particular contexts it may not represent the optimal means by which to define study quality. We see two arguments. First, a well-designed study is not always that which is least subject to bias. For instance, two of the four random assignment studies of drug courts in our sample of 86 studies include very small sample sizes, and one has substantial attrition as well. To further illustrate this point, consider a well-designed quasi-experiment where the comparison group was drawn from a sample of eligible drug court participants excluded from the program because it had reached capacity. While the first study has the highest value according to the SMS and the latter study is assigned a middle ranking, there is no reason, *a priori*, that it should be assumed that the first is better simply because of the study design.

Second, in a particular context, the Maryland Scientific Methods Scale may overlook valuable information which can be used to gauge the quality and reliability of the research. Thus, though generality is its most attractive feature, it can also be a drawback. For instance, consider two studies of the effectiveness of drug court. Both have a treatment and control group, and both use propensity scores to balance both groups on observable characteristics such as age, race, gender, number of prior arrests, and substance dependency. Both studies would receive identical scores on the Maryland Scale. However, suppose that one study's comparison group was drawn from those who were rejected from the program by the prosecutor, while the other's was again drawn from indivduals excluded due to capacity restraints. Clearly, the studies are of different qualities with different levels of reliability and susceptability to bias. Though the samples may be balanced on observable characteristics, unobservable characteristics are often of greatest concern to social science researchers.

With these considerations, we define our own scale of rigor to be applied specifically to drug court research and specifically for use in meta-analysis. Our scale, from most rigorous to least rigorous, is as follows:

1. Random assignment to treatment and control groups.
2. Studies that successfully match individuals from each group or balance the groups on observable characteristics, where the comparison group was eligible drug offenders who were not referred, the regular probation or diversion population, or a historical comparison group.
3. Studies which used multivariate statistical analysis or unsuccessfully attempted matching, where the comparison group was made up of eligible drug offenders who were not referred, the regular probation or diversion population, or a historical population.
4. Studies where the comparison population was made up of individuals who declined participation, were rejected from the program, had a non-eligible drug offense, or where there was substantial (differential) attrition or all of the above, regardless of the study design.

While the Maryland Scale is discussed in depth elsewhere and is largely adapted from a long scientific literature (see the Cochrane Collaborative for more information), our scale warrants some brief discussion. The highest level of rigor, as in the Maryland Scale, random assignment, is based on study design. Beyond

that, however, our scale was designed to envelop two different dimensions: composition of the comparison group and analytic methods. The next level includes studies where we feel that the comparison group is most comparable to the treatment group and least likely to be subject to unobserved bias.

The second level of our scale requires matching studies. We do not feel that matching studies are better at identifying program effects than well-executed studies using multivariate controls. However, the log odds ratio, our measure of the effect of the program, does not utilize the information from the statistical controls. That is, the log odds ratio is based solely off of the number of indiduals in each group who recidivated. Thus, while the studies themselves may account for all confounding factors, our measure of the effect size does not. This shortcoming builds bias into the log odds ratio defined for studies which use multivariate analysis. However, as long as the comparison group is reasonable, while we are not as confident in the log odds ratio as we are for studies where the groups were carefully matched, we do feel that the bias is limited. Thus, we rank these studies the second lowest.[2]

Finally, the last group includes studies with poorly selected comparison groups. Those who would not be eligible and those who either declined or were rejected would be expected to naturally do worse than eligible individuals who were accepted, regardless of the intervention. Thus, the final category includes studies which have bias built in.

Throughout the analyses presented in this report, we use both the study quality defined by the Maryland Scale (which is far more generally applicable) and that defined above (which is specific to drug courts and meta-analysis) and run each model twice.

3.1.3. Independent Analyses

Once we have appropriately labeled poor studies, the natural next step is to estimate the effectiveness of drug court while controlling for study quality information. The most basic way of doing so is to meta-analyze the highest quality evidence separately. This way, the information derived from weak, potentially biased studies has no impact on the information derived from higher quality studies. Random effects analyses allow for variability between studies of the same methodological rigor, and results provide estimates of the average treatment effect within each study type, the variability between estimates of the same type, and a mean effect for each study which is adjusted to draw power from studies of the same rigor.[3]

3.1.4. Weighted Analysis

Independent analyses, however, are not ideal. Applied policy research has little interest in the variation in estimated effects across different research designs. The goal of these analyses is to inform policy-makers about the true effects of a particular program to inform implementation decisions. Thus, we are only interested in the "true" effects of the program and the heterogeneity of true effects across different drug courts. Because random assignment is typically considered the ideal research design, we would most likely use the results from the analysis only of randomized studies to estimate the effects (and a predictive interval for those effects) of drug court. Within the dataset, there are only four randomized trials[4], one of which had attrition of nearly 50% and another which had a sample size of less than 80. Neither of these studies seems appropriate for determining the true effectiveness of drug court. By restricting our inference to only the randomized trials, we ignore 78 studies which potentially contain a wealth of useful information.

[2]One potential solution would be to use the estimated coefficient measuring the treatment effect from the multivariate analysis instead of the simple number of individauls from each group who recidivated (see Aos, et al, 2006). This method has methodological implications too complex to be discussed here. We encourage the interested reader to see Lipsey and Wilson (2006), especially pages 67-71.

[3]This is an important results in Bayesian hierarchical models often referred to as "shrinkage". For more, see Gelman, et al (2006) or Raudenbush and Byrk (2002).

[4]Deschenes, et al (1995); Breckenridge, et al (2000); Dickie (2001); Gottfredson, et al (2003)

Thus, an alternative approach is to include all studies in the analysis, but to assign weights such that studies with better designs have greater influence on the results. This approach has been used in a number of contexts (Begg and Pilote, 1991; Li and Begg, 1994; Larose and Dey, 1997). Aos, et al, weight studies according to the Maryland Scientific Methods Scale such that studies in the last two categories are excluded from the analysis, studies in the third level are weighted by 0.5, those in the fourth are weighted by 0.75, and randomized trials (the first category in the Scale) are not downweighted at all. The intuition behind these weights is that studies with worse designs are less likely to reliably identify the true treatment effect. The weights reduce their influence on the overall estimated mean.

We conduct a random effects meta-analysis of the effectiveness of drug courts using the weighting scheme designed by Aos, et al, and the analog based on our categorization (studies in the first level weighted by 1, those in the second weighted by 0.75, those in the third weighted by 0.5, and the rest excluded). It is important to note that the specific values of the weights selected are arbitrary (Aos, et al, 2006). Conceptually, more rigorous studies are given greater weight, but the selection of 1, 0.75, and 0.5 has little empiricial or theoretical foundation. We adopt the same values, with the same limitation.

3.1.5. Regression Analysis

Weights, however, are not ideal either, particularly in the presence of bias. Weighting an estimate does not change its estimate of the treatment effect. A weight is often misinterpreted as pulling the estimate towards zero. This is not correct. Rather, weights simply reduce the influence that a particular estimate has on the aggregate mean estimate. In this way, downweighting an observation has the same effect as increasing its variance. Thus, in short, a weighting approach assumes that the mean estimated effect of a subgroup of studies does not depend on the quality of those studies (Greenland and O'Rourke, 2001). This is tantamount to assuming that poorly designed studies still provide unbiased estimates of the true treatment effect, but estimate it less precisely (greater variance). Thus, while weights account for underestimated variance, any directional bias in the estimate is unaddressed.

Regression methods are often used to identify these systematic biases (e.g., Lipsey, 2009). Each study's estimated effect can be regressed on a discrete set of quality variables. This way, coefficients on study quality variables pick up any systematic deviation from the omitted category, most naturally random assignment.[5] This method can be used to detect bias resulting from study design.[6]

Sutton and Abrams (2001) showed that a random effects meta-analysis model can easily be extended to a meta-regression by including study-level variables at the second level. Thus, as in the typical random effects formulation,

$$e_i \sim N(\theta_i, \sigma_i^2). \tag{3.8}$$

However, now

$$\theta_i = \mu + \sum_{j=1}^{J} \beta_j x_{ij} + \tau z_i \tag{3.9}$$

where μ is still the grand mean and τ still captures between study heterogeneity because

[5] It should be noted here that this method, as explicated here, only detects additive bias (e.g., when the effect estimated from a weak study is x units higher or lower than the true effect). For multiplicative biases (where the estimated effect is x times larger than the true effect), a different formulation is required.

[6] Technically, it is not appropriate to assume causality in this relationship. As Wilson (1995) has argued, it is possible that study design is correlated with program quality so that weak study designs produce worse estimated effects because they are evaluating weaker programs.

$$z_i \sim N(0,1). \tag{3.10}$$

The variable x_{ij} is a binary variable indicating whether or not the ith observation falls in the jth quality classification. Thus, β_j captures the systematic bias in studies of type j.

Because this is still a Bayesian analysis, the results will provide full (posterior) distributions for μ, τ^2, and the various β_j's. Thus, the β_j's provide a distribution of the potential bias of each study type. If any of these β_j's appear to be significantly different from zero, the corresponding study type produces estimated effects which are persistently different from those derived from random assignment (which we select to be the omitted category). Thus, systematic bias is present which is unaccounted for by the weighting strategy (Greenland and O'Rourke, 2001).

Through a Bayesian random effects regression, we can identify whether any types of studies introduce persistent bias. We conduct such a regression using study quality definitions from both the Maryland Scale and our own definitions. If our model fits the data better than the Maryland Scale, then our categories more accurately identify sources of heterogenity among estimated drug court impacts. Two Bayesian models can be compared through Bayes' Factors (Kass and Raferty, 1995).

Bayes' Factors (sometimes called BFs) are derived directly from Bayes' Theorem. They are useful for comparing the probability of one model being the true model, relative to another specified model (Gelman, et al, 2006). The probability of a model (M_i), given the existing data (D) (also known as the posterior probability of the model), can be evaluated directly using Bayes' Theorem as

$$p(M_i \mid D) = \frac{p(D \mid M_i)p(M_i)}{p(D)}. \tag{3.11}$$

If we wish to compare two models, we can calculate the ratio of their posterior probabilities as

$$\frac{p(M_1 \mid D)}{p(M_2 \mid D)} = \frac{p(D \mid M_1)p(M_1)}{p(D)} \times \frac{p(D)}{p(D \mid M_2)p(M_2)} = \frac{p(D \mid M_1)p(M_1)}{p(D \mid M_2)p(M_2)} \tag{3.12}$$

where $p(D)$ cancels itself out because the same data is used in both models. The $p(M_i)$'s allow the researcher to specify prior probabilities for the models. That is, the researcher can say, *a priori*, how much more likely one model is than the other. If the researcher considers both models equally likely, then $\frac{p(M_1)}{p(M_2)}$ reduces to 1 and the Bayes Factor is simply

$$\frac{p(D \mid M_1)}{p(D \mid M_2)}. \tag{3.13}$$

The functional form of $p(D \mid M_i)$ is simply that of the likelihood function, in which case the Bayes' Factor is simply a likelihood ratio test. However, the important difference is that in a Bayes' Factor, the true parameter values are assumed unknown, whereas a likelihood ratio test assumes that the estimated values are the true values . Thus, to incorporate all uncertainty about parameter values, $p(D \mid M_i)$ must be integrated over all possible values of the parameters in M_i so that

$$p(D \mid M_i) = \int_{-\infty}^{\infty} p(D \mid M_i, \Phi_i)p(\Phi_i \mid M_i)\, \partial\Phi_i \tag{3.14}$$

where Φ_i is the vector of model parameters in model i and $p(\Phi_i \mid M_i)$ is the prior density of those parameters conditioned on the model (Sutton and Abrams, 2001). We employ Bayes Factors using the estimated

posterior distributions as $p(\Phi_i \mid M_i)$ to compare the random effects regression model fit to the Maryland Scale's definitions of study qualtiy and that fit to our own categories to determine which classification better explains the heterogenity in estimated effects, which may be the result of biases inherent in research design.

3.1.6. Three-Level Hierarchical Model

However, a random effects regression of the form used here,

$$e_i \sim N(\theta_i, \sigma_i^2) \tag{3.15}$$

$$\theta_i = \mu + \sum_{j=1}^{J} x_j + \tau z_i \tag{3.16}$$

has the undesirable property of most regressions: it constrains the variance of each of the k sub-groups to be identical. That is, the between-study variance (τ^2) does not depend on the type of study. Therefore, this method allows the mean of estimated effects to depend on the quality of the study, addressing biases resulting from design, but without letting the variance vary across study type. Thus, the model implies that studies of all types are equally reliable and consistent in estimating the class-specific mean. This is the opposite problem from the weighting approach, which assumed that study quality led to no bias, but simply reduced the consistency and reliability of the estimates.

An ideal model would allow the mean of the estimated effect to vary depending on study quality, in addition to allowing studies of different quality to estimate this effect with different levels of reliability. That is, an ideal model would also have a variance term to explain between-study, within-type variance. As illustrated in Sutton and Abrams (2001), this can be achieved with a three-level hierarchical model (see Smith, Abrams and Jones, 1995; Prevost, Abrams and Jones, 2000). The model is now of the form:

$$e_{ij} \sim N(\theta_{ij}, \sigma_i^2) \tag{3.17}$$

where

$$\theta_{ij} = \eta_j + \upsilon_j \epsilon_i \tag{3.18}$$

and

$$\eta_j = \mu + \tau z_j \tag{3.19}$$

where z and ϵ are standard normally distributed random variables (with mean 0 and variance 1). Here, there is a class specific mean (η_j), which varies from the aggregate mean (μ). Now, instead of representing the variance between studies, τ represents the variance between different types of studies' estimated effect. Within each study type, υ_j represents the heterogeneity among studies of that type. As in the previous analyses, each study has its own true mean effect (θ_{ij}), which is assumed to be estimated with known variance (σ_i^2) and will be shrunk towards its class-specific mean (η_j), and μ, τ, and υ are assumed to be random variables themselves.

Thus, the hierarchical model proposed by Sutton and Abrams (2001) allows both the mean and the variance of the estimated effectiveness of drug court to vary by study design, incorporating the attractive advantages of both the weight-based and regression-based approaches.

Several of the parameters in the model are of particular interest for inference. The global mean, μ is of obvious potential interest, since it is the average effect across all studies (accounting for the fact that some types have different variances and biases). However, μ does not take into account the ordinal nature of the quality categories. That is, each quality group's influence on μ is a function only of 1) the difference between effects estimated from that group and those estimated from other groups and 2) the heterogeneity (variance) across studies within that group. Thus, the influence does not depend on the relative rigor of the method or how confident we are in results generated by those studies. So μ does not know that the quality groups are ranked.[7]

Thus, another parameter which is of primary interest is the η_j corresponding to the mean of the random assignment studies. A property of Bayesian hierarchical models is that they "shrink" the class-specific means to be more like the global mean. Thus, while the η_j is still dominated by the information from the random assignment studies (which we, again, consider to be the most rigorous designs), it is also influenced by the estimates from other types of studies. The degree of influence from other studies depends on the precision of the random assignment estimates relative to the precision of estimates from other study types.

Additionally, the υ_j's have important policy implications. These parameters represent the heterogeneity among studies of the same type. There are two main sources of heterogeneity among studies within the same design group. The first is artificial variation due to the design and execution of the studies themselves, which is rarely of interest. The second is variation in program effectiveness. This is of obvious interest to policymakers considering implementing drug courts, and our use of Bayesian methods was largely driven by an ambition to capture and appropriately represent this variation.

Unfortunately, it is difficult to parse out the former from the later when considering the various υ_j's. One approach would be to only consider the random assignment studies' υ_j's because they would largely be driven by cross-court variation rather than variation resulting from the study design. This estimate, however, is unlikely to be precise due to the small number of random assignment studies. More generally, the analyst could consider any (or all) study design(s) where additional study-resultant variation was unlikely. This does not require that the studies provide unbiased estimates of the target impact, but that all studies within the same group are biased by (roughly) the same amount[8] so that the cross-study variation would still be driven by cross-court heterogeneity. Finally, the analyst could consider the υ_j's as a whole. If they are all fairly similar, then it is unlikely that the differences in study design play a large role in the size of υ_j and it is likely that they are measuring true heterogeneity in effectiveness.

Finally, from a research standpoint, which can be useful in designing future studies to inform policy, it is important to consider τ^2, which represents the variation across different types of studies. That is, τ^2 is a measure of the influence that the study design has on estimated outcomes. If τ^2 is very small, study design is somewhat less important.[9] To consider an extreme example, if $\tau^2 = 0$, all studies estimate the same mean effect. There is no clustering of estimated effects by study type and a weighting procedure and simple random effects model would be sufficient. On the other hand, a large τ^2 indicates the opposite: that study design leads to substantially different esimated effects.

The various considerations available in this method highlight its flexibility and practicality. Inference can be based on general notions, such as comparing υ_j's or η_j's for study types of varying quality, rather than strict definitions of the relative quality of various methods (as is required for a weighting scheme). Likewise, the flexible method models all relevant information, rather than making the draconian assumption that study design has no impact on the consistency of the estimated effect (as in the regression-based analysis). This technique allows a range of important inferences to be drawn without placing too high of demands on the data, which is often questionable in a social science meta-analysis.

[7] This can be taken into account with weights, however some arbitrary decision must still be made for weight values.

[8] Again, this is assuming an additive bias.

[9] The size of τ^2 can be considered relative to the size of heterogeneity between studies of the same type (υ_j) or differences between the mean effectiveness estimated by studies of different types ($\eta_j - \eta_k$ for $j \neq k$).

3.2. COST METHODS

In line with the methods used throughout this analysis (and modern Bayesian methods, more generally), all final program costs were simulated. That is, for each iteration (N) we simulate a number of individuals who participate in the program (n) and for each, we simulated all aspects of the drug court process (such as the number of months they would participate, whether or not they would graduate, how much treatment they would use, how many sanctions they would incur, etc.) based on the probabilities and confidence intervals estimated in the interview. Random simulation is the most efficient and effective way to fully represent heterogeneity in process costs.

For each simulated individual, simulated resource use (such as days in residential treatment) are conditioned on simulated values for other individual characteristics to accurately reflect the correlation between resource use and avoid overestimating variation in process costs. For example, each individual is estimated to graduate or not, given a graduation probability of 0.4 (corresponding to the true 40% program graduation rate). The length of time an individual who *does not* graduate stays in the program is estimated according to the reported distribution of time spent in the program among those who *do not* graduate. That individual's drug treatment experience, then, is conditioned on how long he/she is simulated to be in the program.

Through the interviews, we obtained all necessary unit prices, probabilities, and confidence intervals to simulate the program costs. However, some of these costs would have been incurred even without the program. Thus, we need to calculate marginal costs of the program by comparing program costs with the costs of the alternative. To do so, we relied on MADCE data (Rossman, 2011). For each of n program participants, we sampled resource use (supervision officer's time, drug tests received, drug treatment, etc.) from one of the 440 comparison group participants from that study, and assumed that this is the level of resources that the individual would have used had he/she not been in the program. These resources were valued using the resource prices specific to DC that were collected from the interview with the Director of Treatment. Thus, by subtracting valued resource use from simulated program costs, for each individual we can develop an individual level measure of marginal program costs.

3.3. BENEFITS METHODS

Because this is a meta-analysis, we can only measure, estimate, and value impacts that were measured and estimated in past studies. Thus, we are largely limited to valuing reductions in rearrest as the sole benefit of the program, though many authors have suggested that there are likely many other important social benefits derived from drug court (WSIPP, 2003).[10] Thus, our focus is on the benefits of drug court which manifest in terms of reductions in arrest, incarceration, and criminal activity.

Roman (2009) developed estimates of the victim costs of various crime types and provided a number of percentiles for each crime type. We prefer this data to the more commonly used Miller and Cohen (1996) estimates. However, those estimates present expected costs and benefits as point estimates with no variation. Not every victimization incident is equally costly: some assaults result in little damage, while some result in massive and costly injury. Thus, we prefer victims prices that include uncertainty in estimating social costs of crime. To incorporate this variation, we used a linear approximation to the full distribution of crime costs.

First, we used the comparison group from the above mentioned multi-site drug court study and obtained the empirical distribution of crimes for which they were arrested during the 18 month follow-up period. The comparison group's arrests could be divided into 20 major crime types. Thus, for any given arrest averted by the program, we were able to simulate which of those 20 crime types the arrest would have been.

[10] Recent findings suggest that many of the benefits that were expected and hoped for may not actually exist. Impacts such as reductions in government support, medical care received, and use of other public services, which many authors suggested likely add up to substantial social benefits from drug court, do not appear to come to fruition (Downey and Roman, 2011).

We constructed a linear approximation of the distribution of the costs of each type of crime by obtaining the 10th, 25th, 50th, 75th, and 90th percentiles of each category and then assumed linearity between each of these points. Then for each crime, we simulated a uniformly distributed random number along the interval from 0 to 100. We matched the randomly distributed number with the linear approximation so that, for instance, if the random number was 20, we assumed that the cost of that crime were the approximation of the 20th percentile of costs. This non-parametric method allows us to sample from all parts of the distribution of crime costs without having to specify or estimate the form of the distribution.[11]

3.4. FULL COST-BENEFIT METHODS

To summarize, our approach relied heavily on simulations to capture the true variation in every step of the analysis. By simulating, we can estimate the extent of aggregate variation in outcomes. First, we developed posterior distributions of the parameters of interest in the meta-analysis. Then, from these posterior distributions, which are variable themselves, since neither parameters of the distribution are known with certainty, we simulated 2,000 observations of the log odds ratio expected from a new drug court. All of the following steps were taken for each of the 2,000 simulations of the log odds ratio so that full distributions of all outcome measures could be estimated.

Assuming a 43% base rate (the mean of comparison group recidivism rates in our sample), we converted this log odds ratio into a treatment group recidivism rate. We then specified how many participants the program would serve in a year. Arbitrarily, based on past experiences with drug courts and past research, we selected 150 participants. We then simulated how many of those individuals would have been rearrested given the comparison group recidivism rate (43%) and given the treatment group recidivism rate (calculated from the log odds ratio). The difference between these is the estimated number of arrests prevented by drug court.

For each arrest prevented, we simulated what type of criminal event led to that arrest, given the probabilities estimated from arrest records of the comparison group of the MADCE study. However, not every criminal incident results in arrest. If each criminal incident is an independent event, and the clearance rate is the probability that any particular criminal incident will result in arrest, then the number of criminal incidents that will occur before an arrest follows a negative binomial distribution (parameterized by p = the clearance rate). Thus, for each crime category in which there was at least one arrest prevented, we simulated a random negative binomially distributed random variable to estimate how many crimes lie behind this arrest. We also simulated whether an arrest would result in incarceration given the proportion of arrests that resulted in an incarceration in the MADCE comparison group. To estimate how long the resultant incarceration would last, we used the statistical tables from Durose, Farole, and Rosenmerkle (2009) to calculate the probabilities of various sentences for each crime type. Since DC has truth-in-sentencing laws, at least 80% of a sentence will be served. For each incarceration prevented, we simulated a uniformly distributed random variable on the interval (0.8,1) to simulate what portion of that sentence would have been served.

Finally, for each crime prevented, we simulated a random number uniformly distributed on the interval (1,100), as described in the previous section, to estimate the victimization costs of that crime. The final program benefits, for each of the 2,000 iterations, are the total costs of victimization avoided, the costs of arrests avoided, and the costs of incarceration avoided. These can be compared to the total simulated marginal costs of program participation for the 150 simulated participants to determine the net benefits of the program. This was done for 2,000 iterations to provide the full distribution of expected net benefits, after accounting for variation and uncertainty in each step of the process.

[11] Estimating the forms of the distributions of crime costs is non-trivial, since all are highly skewed and abnormal.

4

Results

4.1. STUDY QUALITY

4.1.1. Maryland Scale

The Maryland Scientific Methods Scale is defined as follows (Sherman, et al, 1998):

Level 5. Random assignment and analysis of comparable units to program and comparison groups.

Level 4. Comparison between multiple units with and without the program, controlling for other factors, or using comparison units that evidence only minor differences.

Level 3. A comparison between two or more comparable units of analysis, one with and one without the program.

Level 2. Temporal sequence between the program and the crime or risk outcome clearly observed, or the presence of a comparison group without demonstrated comparability to the treatment group.

Level 1. Correlation between a crime prevention program and a measure of crime or crime risk factors at a single point in time.

The meta-analytic data used in this analysis captured many of these important methodological considerations in three variables denoting research design, composition of the comparison group, and similarity of treatment and control groups on key variables (such as demographics and criminal history). From these variables, we attempted to classify our 86 studies along the Maryland Scale.

Level 5. Reserch design was random assignment. (n = 4)

Level 4. Either statistical matching with no differences between treatment and control group on key variables *or* multivariate analysis. (n = 26)

Level 3. Statistical matching studies that did have differences between the treatment and control groups. (n = 31)

Level 2. All remaining studies. (n = 25)

Table 4.1: Classification of Study Quality

Categories	Our Level 5	Our Level 4	Our Level 3	Our Level 2
MD Level 5	4	0	0	0
MD Level 4	0	8	5	13
MD Level 3	0	9	19	3
MD Level 2	0	0	12	13

Level 1. The data were restricted to studies which used an explicitly defined treatment and control group, so there were no Level 1 studies.

4.1.2. Our categories

The following classification scheme was created to address the main identification concerns specific to drug court research and application in meta-analysis. We refer to Levels 2-5 to maintain rough conceptual comparability with the Maryland Scale, although we do not define a Level 1. [1]

Level 5. Research design was random assignment. (n = 4)

Level 4. Comparison group was probationers, eligible but not referred offenders, or historical. Research design was matching. No differences or minor difference between treatment and comparison group. (n = 17)

Level 3. All other studies where comparison group was probationers, eligible but not referred offenders, or historical (this includes matching studies with major differences between treatment and comparison group *and* studies that used multivariate analysis). (n = 36)

Level 2. Any study with a comparison group which we consider very weak (individuals who declined participations, were rejected, or had noneligible offenses), an ambiguous comparison group, or those that didn't report the composition of the comparison group. (n = 29)

4.1.3. Comparing study quality definitions

Overall, most studies (51%) were in the same categories under both schemes. Table 4.1 displays how often the two categorizations matched and differed.

The largest two categories of different classifications were those that scored a 2 on the Maryland Scale but a 3 on our scale (these are primarily matching studies which reported substantial differences between the treatment and control groups, but for which the control group was at least well chosen) and studies that Maryland coded as 4 that we coded as 2. All of the multivariate analyses are in this category, which may be good studies (unknown) but are bad for our methods of meta-analysis because effect sizes are coded using the log odds ratio, which does not depend on the coefficients estimated in the multivariate analysis.

4.2. META-ANALYTIC RESULTS

4.2.1. Model 1: Independent Analyses of subgroups

After acknowledging that different studies are of different qualities, the next natural step is to consider only the results from the studies of the highest quality: random assignment. However, with only 4 random

[1] This to is to maintain rough comparability, since our studies do not represent Level 1 on the Maryland Scale.

assignment studies, the results, unsurprisingly, are of little use.

Interest in this model centers on two parameters: μ and τ. μ represents the mean effectiveness of drug courts, while τ is the heterogenteity in this effect. Thus, τ is the variation around μ. However, in addition to capturing cross-court variation around μ, Bayesian methods also capture the variation in the estimate of μ itself, by estimating a full distribution for the parameter. Thus, while τ captures the variation around μ, the posterior distribution of μ captures the variance in the estimate of μ. As Higgins, et al (2009) put it, Bayesian random effects models capture heterogeneity of program effects (τ^2) and uncertainty of estimated program effects (Var(μ)).

Figure 4.1 displays the posterior densities estimated for both μ and τ. Both are very wide distributions, indicating significant uncertainty in the estimates. For instance, the distribution of μ indicates that, given a base comparison group recidivism rate of 43%, there is a 95% chance that the *mean* recidivism rate of drug court participants is between 3.5% and 81%. After including the inter-court heterogeneity (τ), the 95% predictive interval for any particular court's recidivism rate goes from 0% to 100%, conveying virtually no information. This underscores the primary flaw in conducting independent analyses of subgroups of studies.

Figure 4.1: Model 1 Posterior Distributions

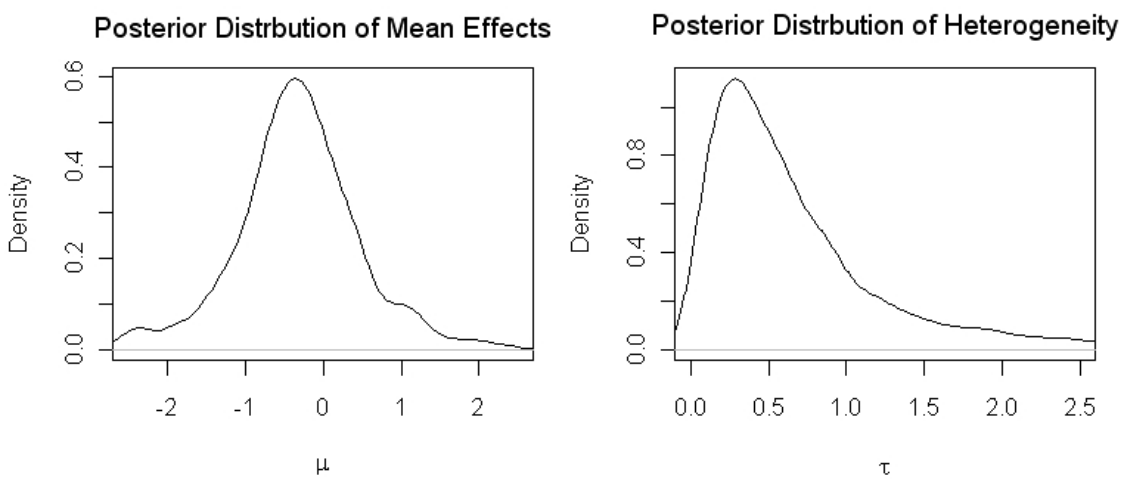


Table 4.2 displays the results from the analyses of all subgroups, as defined by the Maryland Scale and our scale. To contextualize the numbers, which are all on the log odds ratio scale, it is helpful to have some frame of reference. The mean recidivism rate among control groups in our studies was 43%. Thus, we present all numbers as recidivism rates of drug court participants, assuming a base rate of 43% for the comparison group. Given this, a log odds ratio of 0 corresponds to a treatment group recidivism rate of 43%. The log odds ratios of -0.5, -0.8, and -1.2, correspond to recidivism rates of 31%, 25%, and 18%, respectively, while 0.5, 0.8, and 1.2 correspond to 55%, 63%, and 71%, respectively. Log odds ratios outside the interval (-2, 2) seem implausible in this analysis, as -2 and 2 correspond to treatment group recidivism rates of 10% and 85%.

There are several noteworthy observations in Table 4.2. First, the mean and median estimate of μ is below zero (indicating drug court reduces recidivism) for every category. The mean of μ tends to be between -0.4 and -0.6. Assuming a base recidvism rate of 43%, this translates to a treatment group recidivism rate from 29% to 33%, roughly a 10 to 15 percentage point reduction in recidivism. However, all credibility intervals span zero, meaning that there is a substantial probability that the mean drug court effect is an increase in recidivism. It is also important to address τ. We remind the reader that individual drug court effects are normally distributed around the mean, and thus, 95% of drug courts' effectiveness is expected to lie between $\mu - 2\tau$ and $\mu + 2\tau$. Thus, the large standard deviations indicate that the above analyses include virtually all

Table 4.2: Results of Independent Models

Quality Level		Maryland Classes Mean	Maryland Classes Median	Our Classes Mean	Our Classes Median
		(95% Cr. Int.)		(95% Cr. Int.)	
Level 5	μ:	-0.365	-0.345		
		(-3.03, 1.71)			
	τ:	0.735	0.525		
		(0.05, 2.92)			
Level 4	μ:	-0.551	-0.567	-0.646	-0.621
		(-2.10, 1.06)		(-1.96, 0.50)	
	τ:	0.707	0.700	0.509	0.500
		(0.48, 1.00)		(0.28, 1.85)	
Level 3	μ:	-0.558	-0.567	-0.447	-0.453
		(-1.57, 0.53)		(-2.36, 1.36)	
	τ:	0.376	0.375	0.819	0.800
		(0.18, 0.63)		(0.60, 1.10)	
Level 2	μ:	-0.342	-0.339	-0.392	-0.391
		(-2.41, 1.68)		(-1.30, 0.54)	
	τ:	0.947	0.925	0.312	0.300
		(0.60, 1.40)		(0.10, 0.60)	

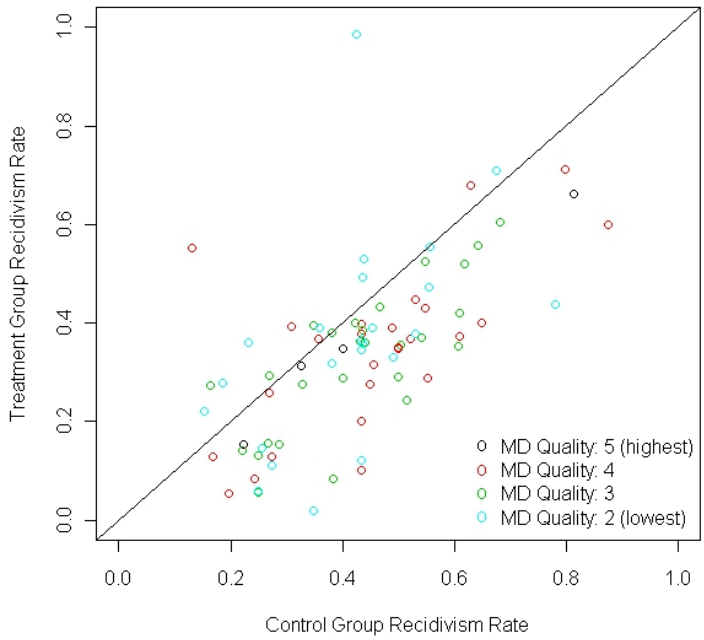
possible outcomes within the range of reasonably expected values. That is, independent analyses provide almost no information.

One further observation from the table is worth mentioning. It is often hypothesized that weak studies are less able to identify causal impacts. Unobserved characteristics conflate the estimates so that the comparison group does not accurately reflect what would have happened without the program. Most often, it is argued that the treatment group was likely more motivated to succeed than the comparison group, that some unknown portion of the observed outcome difference is the result of that motivation, that the estimated program impact overstates the true impact, and that the results should be discounted for this reason. However, the results in Table 4.2 demonstrate that this may not be the case. The weakest studies (by both definitions) estimated effect sizes that were smaller or comparable to the strongest studies. Higher quality non-random studies, regardless of definition, estimated larger effect sizes than the weakest studies, and according to our quality classifications, Level 4 estimates were even higher than Level 3. Thus, the link between weak study designs and magnitude of estimates is not as clear as is often assumed, at least within drug court research.

Figure 4.2 presents the raw data, with no statistical adjustments. On the x-axis is the recidivism rate of the control group. On the y-axis is the recidivism rate of the treatmetn group. The diagonal line represents where the rates are equal and thus there is neither a decrease nor an increase as a result of the program. Thus, studies above the diagonal line saw higher recidivism than studies below the diagonal line. The color of each dot corresponds to the quality of the study (according to the Maryland Scale).

Figure 4.2 confirms the results discussed above presented Table 4.2. The data presents little evidence that weaker studies tend to overestimate the effectiveness of drug court. Light blue studies, the weakest appear to be more likely than others to fall above the diagonal line. In fact, none of the Level 5 report that drug court did not work and only 15% and 10% of the Level 4 and 3 studies did, respectively, while 33% of Level 2 studies found that drug court did not work. Taken together, these results imply that the weakest studies are far more likely to find that drug court did *not* work, rather than the especially large effects weak studies are often assumed to produce.

Figure 4.2: Raw Recidivism Rates by Study Quality

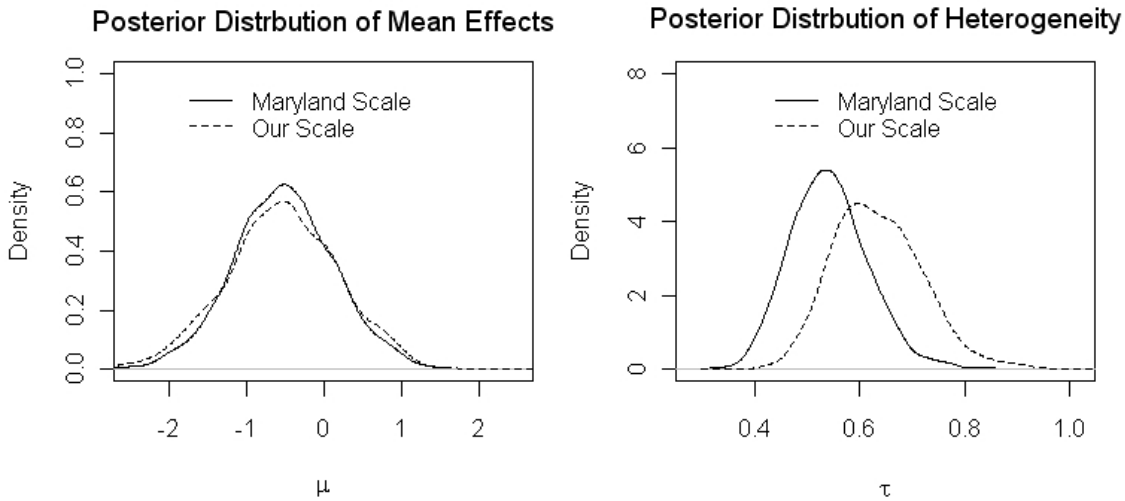


4-18 2.jpeg

4.2.2. Model 2: Weighted Synthesis of Different Study Types

We next analyzed all studies together, but weighted those of inferior quality to have less influence over the resultant estimates. As discussed in Section 3.1.4, we weight Level 5 studies by 1, Level 4 studies by 0.75, Level 3 studies by 0.5, and Level 2 studies are excluded from the analysis (weighted by 0). The results of our analysis are displayed here in graphical and tabular form.

Figure 4.3: Model 2 Posterior Distributions



The first important observation is that the precision of the estimates appears to have improved over Table

Table 4.3: Results of Weighted Analysis

	Maryland Classes		**Our Classes**	
	Mean	Median	Mean	Median
	(95% Cr. Int.)		(95% Cr. Int.)	
μ:	-0.512	-0.522	-0.536	-0.524
	(-1.80, 0.81)		(-1.99, 0.87)	
τ:	0.542	0.550	0.635	0.625
	(0.40, 0.70)		(0.48, 0.83)	

4.2. The credibility intervals are much smaller than those in Table 4.2. On the one hand, this is to be expected as more data is used in this analysis than those. On the other hand, since diverse estimates are being included together, it is plausible that the credibility interval would have increased. The narrowing of the credibility interval may suggest that the estimates themselves are not all that diverse. The credibility interval of τ narrowed more than that of μ, indicating that the model is significantly better at accurately estimating the degree of cross-court heterogeneity.

We also highlight the consistency of the estimated mean with the subgroup analyses. The mean of the estimated mean, which still has considerable uncertainty, indicates that roughly 31% of the treatment group can be expected to be re-arrested, a 12 percentage point drop relative to the comparison group. However, this is only the mean effect, and the mean value of τ indicates that any particular drug court's experience may reasonably be as high as 60% ($\mu + 2\tau$) or as low as 12% ($\mu - 2\tau$). In short, these estimates still lack ideal precision, though they represent a significant improvement over the independent analyses of subgroups.

Finally, we wish to contrast our categorization with the Maryland scale. The estimates of μ and τ are roughly comparable between the two classifications, however both are more precisely estimated with the Maryland Scale. Further, estimated heterogeneity is lower with the Maryland Scale than our drug court specific scale, indicating that many of the heterogeneous studies excluded by the Maryland Scale may have been included (in weak study categories) with our scheme. These results seem to suggest that the Maryland Scale may more successfully identify weak and potentially biased studies.

4.2.3. Model 3: Regression Analysis

In the presence of bias, weighting studies does not adequately adjust for study quality. As discussed, weights do not move the location of the estimate to account for bias, only diminish its influence on the final results. Thus, we used a random effects regression to test the possibility that different study types have different mean estimates.[2] Our results are displayed here, graphically and in tabular form. A key advantage of this method, which represents its most significant gain over the independent analyses of subgroups, is that estimates of variance around the mean and of the mean both borrow from other studies. Therefore, although each study type has a totally independent mean, the precision of that estimated mean is markedly improved over the independent analyses. To highlight this benefit, the first two figures below (in the top two panels) depict the posterior densities of the mean of each study category, while the second two (the bottom two panels) depict the matching densities that were estimated with independent models. Different study types' mean effects can be compared within each figure; different classification schemes can be compared horizontally; and the merits of regression analysis relative to separate analyses can be vertically compared.

The advantages of the regression approach, combining all the data, are clear. The distributions are estimated with much greater precision. We note that the plots are on different scales and so the lower two distributions in the figure are more than twice as wide as the upper two distributions. Further, for all subclasses except random assignment, the 95% credibility interval for the mean now excludes zero. Though the random assignment-specific mean is estimated with far more precision, it still has a wide credibility interval,

[2] It is most helpful to incorporate group estimated coefficients with the grand mean by defining $\mu_j = \mu + \beta_j$ and that is the approach taken here.

Figure 4.4: Model 3 Posterior Distributions

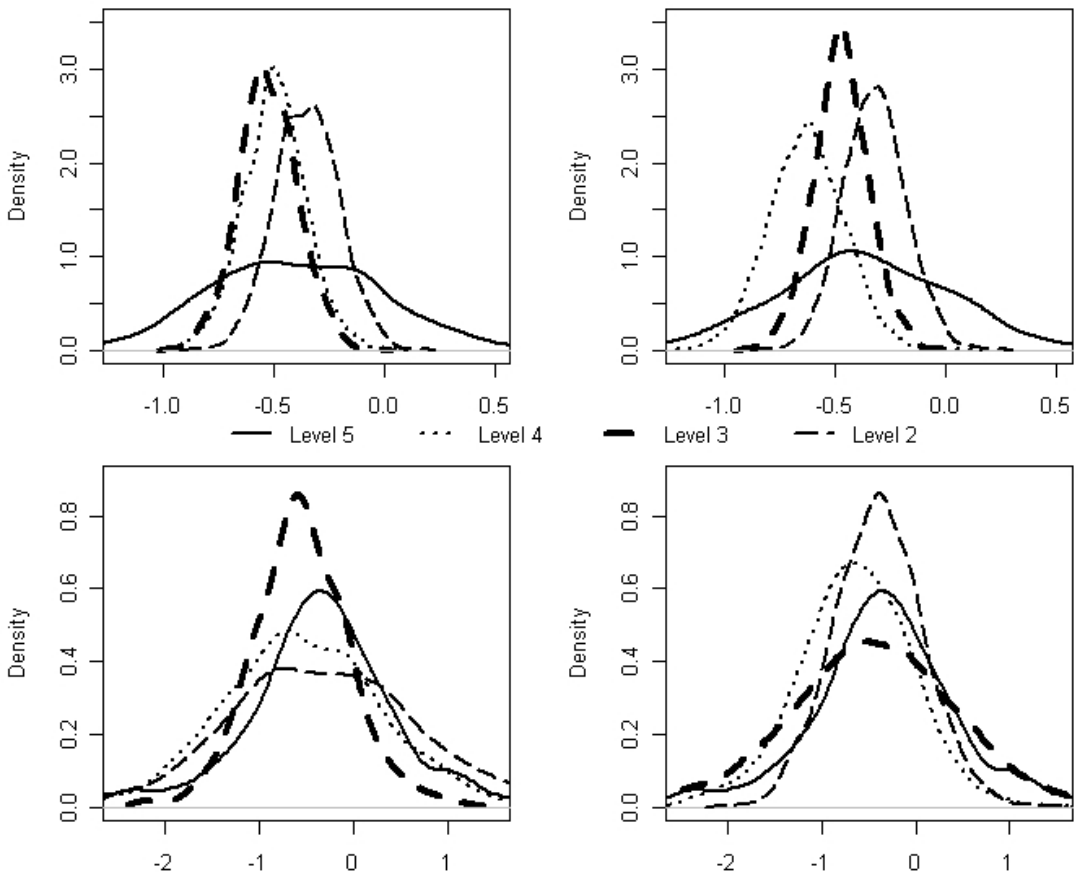


Table 4.4: Results of Regression Analysis

	Maryland Classes		**Our Classes**	
	Mean	Median	Mean	Median
	(95% Cr. Int.)		(95% Cr. Int.)	
μ_1:	-0.401	-0.408	-0.379	-0.382
	(-1.15, 0.36)		(-1.16, 0.39)	
μ_2:	-0.493	-0.492	-0.631	-0.630
	(-0.78, -0.21)		(-0.95, -0.30)	
μ_3:	-0.523	-0.528	-0.473	-0.474
	(-0.79, -0.26)		(-0.72, -0.22)	
μ_4:	-0.348	-0.348	-0.321	-0.321
	(-0.63, -0.06)		(-0.59, -0.04)	
τ:	0.616	0.600	0.614	0.600
	(0.48, 0.78)		(0.48, 0.78)	

with substantial probability mass on both sides of zero. Importantly, the studies showing the smallest estimated effects are those of the worst quality, confirming that presuming that weak studies automatically overestimate impacts appears to be inappropriate.

Also notable in the table, regardless of the classification system used, τ appears to be unchanged, suggesting that we may be near the true value. With a mean of 0.6, there is substantial heterogeneity among courts.

If we take the true mean to be the rough median of estimated means, -0.5, and the mean of τ, 0.6, to be its true value, then a 95% predictive interval for the effectiveness of a randomly selected drug court would be (-1.7, 0.7), which assuming a 43% base rate suggests a treatment group recidivism rate between 12% and 60%. This underscores the importance of presenting a policymaker with accurate information about the uncertainty and variation in program impacts, not just the mean program effectiveness with a confidence interval indicating plausible values of that mean. The results above suggest that there is more than a 98% probability that drug courts, on average, reduce recidivism, but that effect varies significantly across courts, so that any given court only has an 80% chance of reducing recidivism. Simply presenting the estimate of the mean effect overstates the probability of success by a factor of 10.

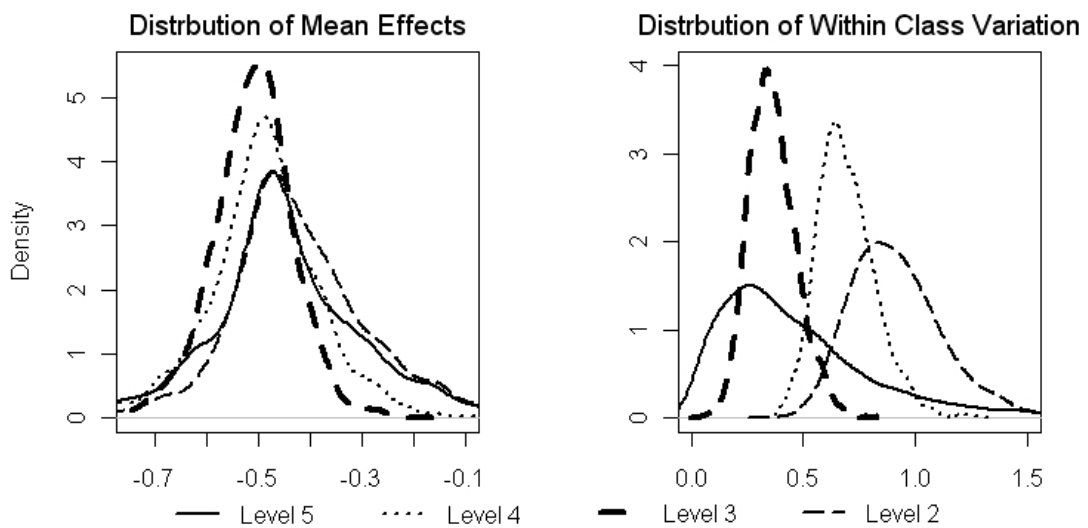
4.2.4. Model Selection

We used Bayes' Factors to compare model fit for the two definitions of study quality. For each estimated set of parameters, we calculated the probability of the observed data. We then summed this for all estimated values of the parameters. This is a numerical approximation of integrating across all model parameters. The results indicated that the odds that data were generated according to the Maryland categories as opposed to our own are 1.6:1. This suggests that the Maryland classification scheme more acurately fits the data, and as such, this is the only classification scheme used in the final anlaysis.

4.2.5. Model 4: Three-Level Hierarchical Model

However, across all study types, the above model constrains between-study variation to be equal. This assumption seems overly restrictive. Thus, we employ a three-level hierarchical model which allows both the mean and heterogeneity to vary across different types of studies. The results, including estimated means and variances for each group, are displayed here in graphical and tabular form.

Figure 4.5: Model 4 Posterior Distributions



The final results are, as discussed in the methods section, rich with information, both of practical significance for policymakers as well as importance for the research literature. It is natural to start inference with the mean effects, and it is noteworthy that each study class has virtually the same mean. The mean estimates and 95% credibility intervals overlap significantly. Also noteworthy, none of the 95% mean intervals include zero, indicating that for any type of study (including the least precisely estimated: random assignment), the mean effect of drug court is a reduction in recidivism. Among the groups analyzed, the weakest studies

Table 4.5: Results from 3-Level Hierarchical Model

Quality Level	Mean(η_j)	Median(η_j)	Mean(v_j)	Median(v_j)
	(95% Cr. Int.)		(95% Cr. Int.)	
Level 5	-0.447	-0.461	0.487	0.400
	(-0.80, -0.08)		(0.05, 1.53)	
Level 4	-0.492	-0.490	0.694	0.675
	(-0.74, -0.27)		(0.48, 0.98)	
Level 3	-0.510	-0.507	0.363	0.350
	(-0.68, -0.36)		(0.18, 0.60)	
Level 2	-0.417	-0.434	0.918	0.900
	(-0.68, -0.11)		(0.58, 1.39)	
Overall	-0.484	-0.485	0.183	0.175
	(-0.63, -0.32)		(0.03, 0.43)	

(Level 2) had the smallest estimated effect sizes, consistent with the findings throughout the analysis but contrary to popular intuition. Finally, we note the precision of the final estimate of μ. The 95% credibility interval (-0.51, -0.45) is consistent with the estimates throughout the various models and specifications employed, as is its mean of -0.48.

The findings suggest that the 95% interval for the mean treatment recidivism rate is 29-35%, indicating that the average drug court effect is a 8-14 percentage point reduction in recidivism (consistent with findings from other drug court meta-analyses).

However, there is considerable variation around that mean effect. Presenting such a simplified estimate to policymakers would lead them to be overly confident in the expected outcome of implementing a drug court. In this model, variance around the mean is captured in two separate terms (each which is estimated with uncertainty): τ represents variation across different study types and the v_j's capture variation across studies of the same type. It is immediately evident that variation within study type differs significantly across different types. The Level 3 studies appear to be the most consistent, while the Level 2 studies (the weakest) are the most variable. In other words, not all weak studies are created alike. Generally, though, across all study types, it appears that most of the variation between studies is variation within-type (v_j) rather than variation across type (τ). In fact, τ is surprisingly small, with a mean only 40% of the mean of v_5 (the random assignments), 25% the mean of v_4, 50% the mean of v_3, and only 20% the mean of v_2. It appears that most study quality impacts manifest themselves as decreasing the reliability of estimates, rather than introducing systematic bias, supporting the weighting approach demonstrated earlier.

4.2.6. Selecting the Final Meta-Results

Finally, we must select which model, and which estimates from that model, to carry forward through the full cost benefit analysis.[3] As discussed in the methods section, we prefer Model 4 because it captures many of the features we believe to be present in drug court research. However, we are simply interested in understanding the true effects of drug court, and Model 4 provides 5 estimated means and 4 possible estimates of the between court variation around these means.

We choose to use the estimated distribution of μ as the distribution of the true mean drug court effect. Although μ is influenced by all study types (including the weakest), because the means of each study type are so similar, we do not feel that this introduces bias. Using η_5, the mean of random assignment studies, would be an obvious alternative, since they are the least likely to be subject to bias, however, because there

[3]Technically, we are not required to select a single model. The Bayesian literature includes well-developed methods of "model averaging," where multiple distinct models are combined, weighted by the relative probability that each model represents the true data generating process (Hoetling, et al, 1999; Raferty and Zheng, 2003). Though we acknowledge the usefulness of these techniques, we do not employ them here.

are so few random assignment studies, η_5 is estimated imprecisely. The next natural choice, then, is η_4, the mean of the next highest quality studies. This would also be a strong choice. We choose to use μ instead because it has a (slightly) smaller mean estimated effect. Thus, we make the more conservative assumption here.

We must also choose some parameter to represent variation between courts, given that not all courts lead to the mean effect. We choose to use v_3. The argument for doing so builds off that made in the methods section. Each v is made up of two components: variation due to true court outcomes and practices, and variation due to research design and execution. From a policy perspective we are only interested in the former, but the two components are indistinguishable from one another. However, we argue that the effects of research design is a strictly positive component. That is, research design will never lower the variation in estimated outcomes below the variation in court performance, it will only increase it. Thus, if we believe that research design is independent of true court performance,[4] then courts within each subgroup have equal variation in performance. We can most closely estimate variation in true performance by taking the minimum observed between-study variance (again, assuming that research design can only increase not decrease between study variance). Thus, we use v_3 to measure heterogeneity in court impacts.

4.3. COST RESULTS

Drug court operations are expensive. The mean marginal costs of drug court participation were $ 10,190 per participant.[5] Consistent with expectations and past research (Downey and Roman, 2011), this mean conceals significant volatility. To demonstrate this heterogeneity, we simulated 5,000 program participants, enough to roughly demonstrate the asymptotic properties of the distribution. The 95% interval of participant costs is ($ 116, $ 26,456), with a mean cost of $ 11,853 per participant. The following two figures display this heterogeneity in different ways. The first is a simple histogram, while the second is the plot of the inverse cumulative distribution function. For any particular value along the x-axis, the y-axis displays the probability that a participant's costs will exceed this value.

The histogram clearly shows clustering around zero. This is because the interview indicated that 20% of those who fail are totally non-compliant. They attend no hearings and no treatment, and fail from the program within one month. Clearly, these individuals incur virtually no cost. The remainder of the distribution appears to be roughly normally distributed, although the positive tail is rather large. This is likely due to rare but expensive long-term residential treatment. Roughly 40% of participants receive residential treatment, which can last up to 60 days and cost anywhere between $ 2,000 and $ 4,000 per month. The second plot demonstrates that the probability of very costly individuals is non-trivial. Nearly one in eight participants is expected to exceed double the mean program costs.

However, displayed in Figure 4.6 are total program costs, not marginal program costs. Some of the resources used by participants would have been used even in the absence of the program. However, estimated comparison group costs tend to be quite low. The mean cost of a comparison individual is only $ 1,720, with a 95% interval of ($ 0, $ 8,723). The histograms in Figure 4.7 display the distribution of comparison group costs and the marginal costs of program participation among the 5,000 participants simulated.

Unsurprisingly, the costs among non-participants also tend to cluster at zero, and so the marginal program costs exhibit the same property. There is also a non-trivial chance that drug court participation will be less

[4] We reiterate the concern of Wilson that this is not the case.

[5] Harrell, Cavanaugh, and Roman (1998) did a rigorous random assignment cost-benefit analysis of the Superior Court Drug Initiative Program. Earlier versions of this analysis used the estimates presented there for estimated program costs, after adjusting for inflation. We chose to replace these estimates by conducting interviews for two reasons: 1) those estimates were based on processing and costs from 1995, and 2) because they were top down estimates, there was no way to construct confidence intervals from those data, so heterogeneity among indivduals' costs was unaccounted for. However, estimates based on interviews with the Director of Treatment at SCDIP provided mean participant costs that were within several hundred dollars (per participant) of the inflation adjusted costs from Harrell, et al, (1998).

Figure 4.6: Properties of the Distribution of Processing Costs

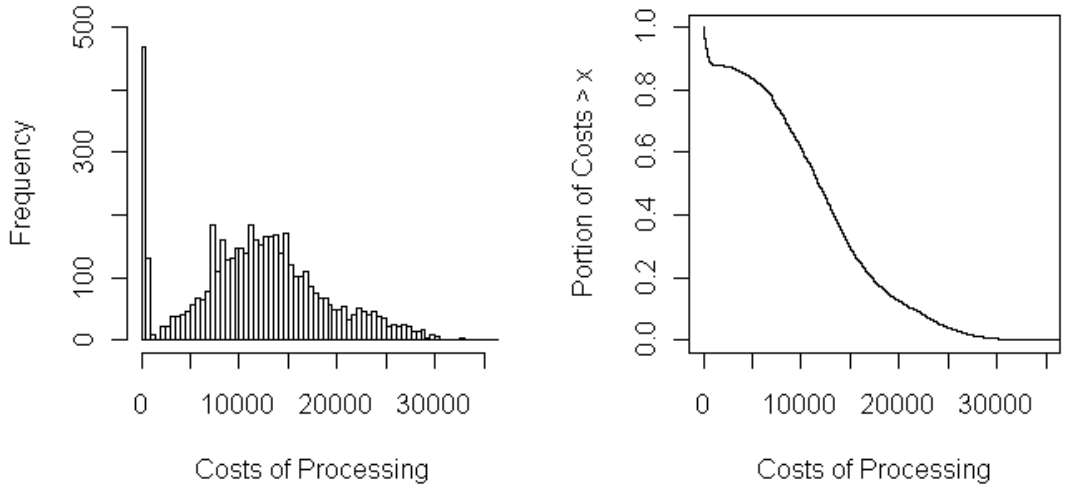
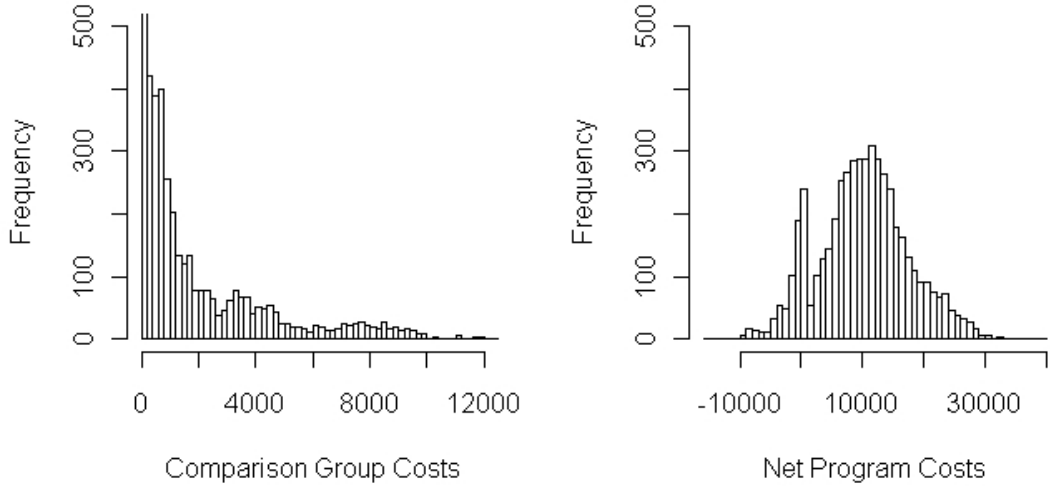


Figure 4.7: Comparison Costs and Marginal Costs



expensive than the alternative.[6] In fact, for 10% of the simulated sample, drug court participation is the less costly option. The final mean marginal costs of drug court participation are $ 10,133, with a 95 % interval of (-$ 3,511, $ 25,429). Clearly these results indicate that drug court participation tends to be expensive. To decompose these costs, Table 4.6 displays resources used by the participants and the expected resources used were they not participating in drug court.

Table 4.6 validates the cost estimates displayed in the figures above. All levels of resource usage are reasonable. Furthermore, it is interesting to note that for virtually all categories, the upper bound of the 95% credibility interval is higher for the comparison group than the treatment group, although the treatment group mean is considerably higher. This suggests that a subset of offenders are highly involved, have substantial needs, and are very costly. These individuals would incur high costs regardless of the program and program involvement might actually lower their costs. Most additional costs of drug court are from typical offenders, who would otherwise use little or none of the resources which are often prescribed by drug court. This finding, though driven by simulated data, has important theoretical implications for

[6]This could happen, for instance, if regular hearings and less severe drug treatment were able to deter the need for more serious treatment such as residential treatment.

Table 4.6: Resource Use Among Program Participants and Comparison Group

Resource Used	Treatment Mean	Treatment Median	Treatment (95% Cr. Int.)	Comparison Mean	Comparison Median	Comparison (95% Cr. Int.)
Outpatient Treatment (in hours)	120	111	(0, 312)	18	0	(0, 115)
Residential Treatment (in days)	16	0	(0, 58)	8	0	(0, 60)
Supervision Officer's Time (in hours)	8.4	8.6	(0, 17)	6.4	3.5	(0, 34)
Hearings (count)	8	9	(0, 14)	2	0	(0, 17)
Drug Tests (count)	31	30	(1, 65)	15	5	(0, 85)
Administrative Costs (in dollars)	850	881	(33, 1617)			

cost-benefit analyses of drug court: most of the costs are incurred in dealing with typical offenders. Given that other research indicates that most of the benefits accrue from dealing with high risk offenders, rather than the typical ones, these findings bolster the suggestion that drug courts are best for high level, high risk offenders (Marlowe et al, 2003).

4.4. BENEFIT RESULTS

The meta-analytic results indicate that drug courts almost certainly reduce crime. The estimated mean is considerably below zero, and there is an 87% chance that any randomly selected court reduces recidivism. Assuming a base rearrest rate of 43% and 150 participants, the Figure 4.8 shows the expected number of arrests prevented. At times, this number can be quite large. There is a 55% chance that more than 15 arrests will be prevented (or 1 per 10 participants) and a 17% chance that more than 30 will be prevented (1 per 5 participants). The 95% predictive interval is (-15,43). Thus, the final results indicate that there is little doubt that the average drug court effect is a reduction in recidivism, but that reduction tends not to be overwhelmingly large, and that there is no guarantee that any particular drug court is effective (although they are on average).

However, most of the crimes prevented are insignificant in nature. Collectively, theft, drug offenses, tresspassing, and traffic offenses make up over 60% of the comparison group's arrests. These crimes are nearly costless to society, and so in terms of benefits of reduced victimization, preventing these crimes contributes little. In fact, this is typical of the crimes that the comparison group committed. Of the 20 considered crimes,[7] only three had median costs above $ 1,000. Table 4.7 displays the distribution of crimes committed, and the distribution of their costs.

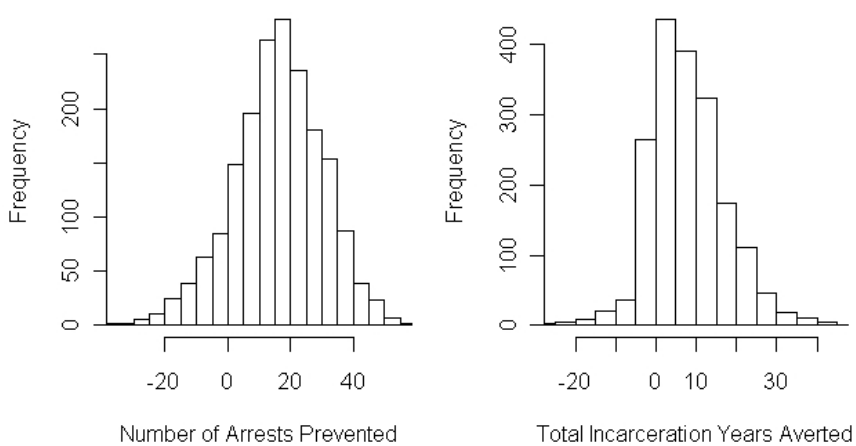
Table 4.7 shows that a given arrest has only a 9.1% chance of being one of these costly crime types. Unsurprisingly, though a significant number of arrests are averted by drug court, the benefits of those averted arrests are quite small. Though there is an 85% chance that the benefits will be positive (resulting from reduced arrests), there is only a 67% chance that they will exceed $ 1,000 per participant, and only a 28% chance that they will exceed $ 5,000 per participant. Figure 4.8 displays histograms representing the most important impacts of the program on the city's criminal justice system: how many arrests are prevented and how many total years of incarceration are saved.

[7] If no individual in the MADCE comparison group was arrested for a particular crime, that crime type was not considered in this analysis (that is, its probability was considered zero).

Table 4.7: Distribution of Crime Types

Crime Type	Probability	Arrest Cost	Costs of Crime (percentiles) 10th	25th	50th	75th	90th
Assault	0.073	$ 354	$ 611	$ 13,285	$ 66,644	$ 155,270	$ 334,515
Gambling	0.011	20	0	0	0	0	0
Prostitution	0.016	20	0	0	0	0	0
Weapons	0.011	20	0	0	0	0	0
Disorderly Conduct	0.014	57	162	162	162	162	162
DUI	0.048	54	0	0	170	1,700	7,700
Family	0.002	57	0	0	0	0	0
Trespassing	0.11	57	0	0	0	0	0
Robbery	0.011	1,003	18,908	68,326	88,915	334,515	605,225
Burglary	0.069	747	12	222	782	2,210	5,279
Theft	0.155	264	192	192	192	192	192
Motor Vehicle Theft	0.007	756	41	2,550	6,800	17,000	39,100
Forgery	0.011	264	0	0	34	459	1,870
Fraud	0.021	264	0	2	170	1,771	7,990
Stolen Property	0.037	20	0	0	0	0	0
Damage to Property	0.002	57	0	0	0	0	0
Bad Checks	0.002	264	0	0	34	459	1,870
Drug/Narcotics	0.231	25	0	0	0	0	0
Traffic Offenses	0.126	57	0	0	0	0	0
Other Offenses	0.043	0	0	0	0	0	0

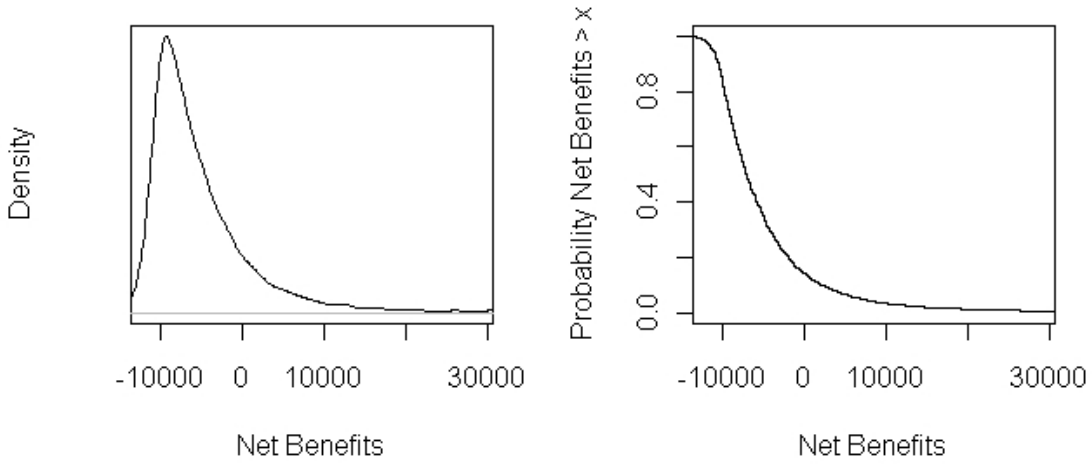
Figure 4.8: Histograms of Primary Criminal Justice Impacts



4.5. FULL RESULTS

To determine the final results, we must calculate the net benefits of program participation. For each iteration, we calculate the total marginal costs of program participation across all 150 simulated participants. We then calculate the total program benefits across all arrests, crimes, and incarcerations averted. We simply difference the two to obtain the net benefits of drug court participation. We divide by 150 to present net benefits per participant. Figure 4.9 displays the density of net benefits and the inverse cumulative distribution function, which represents, for any level of net benefits on the x-axis, the probability that the experienced net benefits will exceed that.

Figure 4.9: Distribution of Net Benefits



The results indicate that drug court benefits rarely exceed costs. Only 14% of the time are net benefits positive. When they are positive, however, they have the potential to be quite large. The 99th percentile of net benefits is nearly $ 23,000 *per participant*. The costs of drug court, per participant, however, can also be quite large. Though they are never expected to exceed -$ 15,000, there is a 25% chance that net costs per participant will be greater than $ 10,000 per participant. As discussed in the benefits section, these bleak results are largely driven by the relatively benign nature of the crimes that most drug involved offenders mostly commit. The kinds of violent crime that carry large social costs are rarely averted because they rarely occur in the first place.

It is critical here to highlight a key limitation of this study. The impacts of drug court are based on evaluations of various drug courts around the country. The benefits of drug court are largely based on the MADCE, the largest evaluation of drug courts ever conducted. The costs of drug court are largely, but only partly, based on the Superior Court Drug Intervention Program (SCDIP). Thus, the benefits and costs of drug court do not use the same court. If the effects of SCDIP are greater than the average drug court, these methods will present an unfair view of drug court. As such, ***it is not a responsible use of this research to conclude that SCDIP is not cost-effective.*** We see good reason, though we cannot definitively prove it, that SCDIP is more effective than the average drug court. As such, it is possible that this research matches cost estimates of a particularly effective drug court to benefits of only average drug courts. We see considerable value to the analyses conducted here, but emphasize that **they are not a substitute for an evaluation of SCDIP.** They simply illustrate that most drug courts implemented in DC would not be cost-effective. We reiterate that these results do suggest that there is a considerable possibility that a drug court implemented in DC could be very cost-effective.

5

Conclusions

5.1. METHODOLOGICAL IMPLICATIONS

We have attempted to demonstrate the advantages, in terms of flexibility, thoroughness, communicability, and accuracy, of Bayesian methods and inference which considers variability around the mean as much as it does the true mean effect. We believe these analyses demonstrate the importance of presenting this information for responsible decision making. We argued that the mean drug court effect is certainly a reduction in arrests and that there is a 99% chance that this mean reduction is greater than 7 percentage points. However, there is a 29% chance that any new drug court will not experience a reduction as large as 7 percentage points, and an 17% chance that a new drug court will not reduce the rearrest rate at all.

We have also attempted to highlight the implicit assumptions of some of the most commonly used methods for adjusting for differences in study quality. Namely, weight-based modeling assumes that weaker studies are less reliable, but does not take into account any systematic direction of bias. On the other hand, regression-based modeling assumes that study quality affects the mean estimate, but not the reliability or consistency with which that mean is estimated. We have demonstrated that three level hierarchical models, as advocated by Smith, Abrams, and Jones (1995) have the attractive feature of relaxing both assumptions of constant mean and constant variance, and shown that such models can still produce estimates with equal, or near-equal, means for all subgroups if the data are best described that way.

Throughout, we have used two different definitions of study quality, each with its own advantages and disadvantages. The results show that the more general classification scheme better fits the data, but we hope to have highlighted that there are alternatives and initiatied an important discussion about how best to operationalize study quality. Little work has been done in this area, despite acknowledgement of its importance from virtually every meta-analyst.

5.2. POLICY IMPLICATIONS

We have attempted to answer several questions with this analysis.

5.2.1. Do drug courts work?

Yes. It is virtually certain that the average drug court effect is a reduction in recidivism. This finding holds for studies of all levels of rigor.

5.2.2. How consistently do drug courts work?

There is considerable variation in drug court effectiveness. Overall, there is an 87% chance that a new drug court will effectively reduce recidivism.

5.2.3. How well do drug courts work?

Drug courts at the top 5% can be expected to reduce recidivism by 23 percentage points, while the weakest 5% see an increase in recidivism up to 3 percentage points. In aggregate, the number of arrests averted is typically not large. 95% of drug courts with 150 people will prevent less than 39 arrests.

5.2.4. What types of arrests are prevented?

Most of the time, arrests of drug offenders are for relatively minor offenses. The most common offenses are drug offenses, thefts, traffic offenses, and trespassing. Together, these offenses make up nearly 2/3 of crimes averted. It is worth noting that none of these offenses is particularly costly to victims. As a result, the social benefits of prevented crime tend to be small.

5.2.5. What are the costs of drug court?

The average drug court participant costs roughly $ 10,000 more than he/she would without the program. This masks significant variability across participants, which often carry virtually no costs, but could cost up to $ 30,000.

5.2.6. Do the benefits of drug court outweight its costs?

This research suggests probably not. On average, drug court will cost $ 5,000 more per participant than is yielded in benefits, and there is only an 14% chance that benefits will exceed costs. That said, the benefits are potentially large. There is a 1% chance that the benefits could be as high as $ 23,000 per participant, which across 150 participants is an aggregate social gain of $ 3.4 million.

5.2.7. What does this imply about DC's Superior Court Drug Intervention Program?

Nothing. The Superior Court Drug Intervention Program (SCDIP) was cooperative in helping the research team conduct the project. However, this is NOT an evaluation of SCDIP. The costs of drug court operations were based on SCDIP, but the costs of business as usual (which forms the marginal costs of drug court) were not. Likewise, the impact of drug courts presented here is not based on SCDIP. The information here does not say that SCDIP is a more expensive than normal drug court. Even if it did, it is entirely possible that SCDIP is a more successful than normal drug court, justifying the additional costs. The results presented here say nothing about SCDIP operations or effectiveness and cannot responsibly be interpreted as such.

5.2.8. Final Recommendations

This study is not designed to determine whether the District of Columbia should continue or expand drug courts, since critical DC data–including patterns of criminal offending for those eligible for a DC drug court are not included in this version of the model. Rather, we have attempted describe what type of

information we will have available for policymakers, including the expected impact of drug court on arrests and the variability in expected outcomes. We also will present the expected costs of court processing and the estimated social benefits of the averted crime. This does not represent the sum total of information that policymakers should consider. For instance, concerns about fair treatment of an often disadvantaged population, pressures on overcrowding of the corrections system, and broader social impacts of heightened incarceration of drug involved offenders are all important considerations which cannot be incorporated in a cost-benefit analysis.

Further, this analysis is limited to only estimating program effects on rearrest rates in the near term following program enrollment. Because a meta-analysis can only estimate impacts which have been estimated by existing studies, impacts such as employment, social service reliance, and medical system use, to name a few of the potential impacts of drug court, cannot be considered here. Additionally, the long term benefits of drug court cannot be estimated, because all studies included relatively short follow-up periods (typically a few years or less), though theory and intuition suggest that reducing substance dependence among a disadvantaged population could have large long-term effects.

Finally, this analysis estimated the impacts of implementing a generic drug court to deal with a generic drug court population. It provided no information on what types of offenders should be targeted, nor which policies can make drug courts most effective. Past research has shown that drug courts have different effects on different types of people. More information about for whom drug court can work best and the characteristics of the population in question should be considered before implementation decisions are made, and this model has the flexibility to incorporate that information. Programs targeted at the populations for which they are most effective can improve the chances of receiving the high social benefits demonstrated here to be possible. Likewise, we found that there are considerable differences between drug courts. Further research on what types of courts are most effective can help reform drug courts to be more effective, as this research has demonstrated that there is capacity for effectiveness.

Bibliography

[1] S. Aos, M. G. Miller, and E. K. Drake. Evidence-based public policy options to reduce future prison construction, criminal justice costs, and crime rates. Technical report, Washington State Institute for Public Policy, 2006.

[2] Steve Aos, Polly Phipps, Robert Barnoski, and Roxanne Lieb. The comparative costs and benefits of programs to reduce crime. Technical report, Washington State Institute for Public Policy, 2001.

[3] Colin B. Begg and Louise Pilote. A model for incorporating historical controls into a meta-analysis. *Biometrics*, 47(3):899–906, 1991.

[4] S. Belenko. 1999. *Research on Drug Courts: A Critical Review 1999 Update*, 2(2):1–58, 1999.

[5] S. Belenko. Research on drug courts: A critical review 2001 update. Technical report, The National Center on Addiction and Substance Abuse at Columbia University, 2001.

[6] Steven Belenko. Research on drug courts: A critical review. *National Drug Court Institute Review*, 1(1):1–42, 1998.

[7] E.P. Deschenes, S. Turner, and P.W. Greenwood. Drug court or probation: An experimental evaluation of maricopa county drug court. *Justice System Journal*, 18:55–73, 1995.

[8] J.L. Dickie. Summit county juvenile court drug court (evaluation report: July 1, 1999-june 30, 2000). Technical report, The Institute for Health and Social Policy, University of Akron, 2000.

[9] P. Mitchell Downey and John K. Roman. *Vol. 4: Impact of Drug Courts*, chapter Results from the MADCE Cost Benefit Analysis. Urban Institute Press, 2010.

[10] Elizabeth K. Drake, Steve Aos, and Marna G. Miller. Evidence-based public policy options to reduce crime and criminal justice costs: Implications in washington state. *Victims and Offenders*, 4:170–196, 2009.

[11] Matthew R. Durose, Jr. Donald J. Farole, and Sean P. Rosenmerkel. Felony sentences in state courts, 2006. Technical report, Bureau of Justice Statistics, 2009.

[12] GAO. Adult drug courts: Evidence indicates recidivism reductions and mixed results for other outcomes. Technical report, U.S. General Accounting Office, 2005.

[13] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2004.

[14] Andrew Gelman and Jennifer Hall. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[15] D.C. Gottfredson, S.S. Najaka, and B. Kearley. Efffectiveness of drug treatment courts: Evidence from a randomized trial. *Criminology and Public Policy*, 2:171–196, 2003.

[16] S. Greenland and K. O'Rourke. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics*, 2:463–471, 2001.

[17] A. Harrell and J.K. Roman. Reducing drug use and crime among offenders: The impact of graduated sanctions. *Journal of Drug Issues*, 31(1):207–232, 2001.

[18] A.V. Harrell, S. Cavanagh, and J.K. Roman. Findings from the evaluation of the dc superior court drug intervention program. Technical report, The Urban Institute, 1998.

[19] A.V. Harrell, J.K. Roman, and E. Sack. Drug court services for female offenders, 1996-1999: Evaluation of the brooklyn treatment court. Technical report, The Urban Institute, 2001.

[20] Julian P.T. Higgins, Simon G. Thompson, and David J. Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society*, 172(1):137–159, 2009.

[21] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

[22] Jr. James F. Breckenridge, L. Thomas Winfree, James R. Maupin, and Dennis L. Clason. Drunk drivers, dwi "drug court" treatment, and recidivism: Who fails? *Justice Research and Policy*, 2(1):87–106, 2008.

[23] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[24] Daniel T. Larose and Dipak K. Key. Weighted distributions viewed in the context of model selection: A bayesian perspective. *Mathematics and Statistics*, 5(1):227–246, 1996.

[25] Zhaohai Li and Colin B. Begg. Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *Journal of the American Statistical Association*, 89(428):1523–1527, 1994.

[26] Mark W. Lipsey. The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims and Offenders*, 4:124–147, 2009.

[27] Mark W. Lipsey and David B. Wilson. *Practical Meta-Analysis*. Sage Publications, 2001.

[28] D. B. Marlowe, D.S. Festinger, K.L. Dugosh, and P.A. Lee. Are judicial status hearings a "key component" of drug court? six and twelve month outcomes. *Drug and Alcohol Dependence*, 79(2):145–155, 2005.

[29] Ted R. Miller, Mark A. Cohen, and Brian Wiersema. Victim costs and consequences: A new look. Technical report, National Institute of Justice, 1996.

[30] Teresa C. Prevost, Keith R. Abrams, and David R. Jones. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics In Medicine*, 19:3359–3376, 2000.

[31] Adrian E. Raftery and Yingye Zheng. Discussion: Performance of bayesian model averaging. *Journal of the American Statistical Association*, 98(464):931–938, 2003.

[32] J.K. Roman and C. DeStafano. *Juvenile Drug Courts and Teen Substance Abuse*, chapter Drug Court Effects and the Quality of Existing Evidence, pages 107–135. Urban Institute Press, 2004.

[33] John K. Roman. *What is the Price of Crime? New Estimates of the Cost of Criminal Victimization*. PhD thesis, Universtiy of Maryland - College Park, 2009.

[34] Shelli Rossman. *Vol. 1: The Multi-Site Adult Drug Court Study Overview and Design*, chapter Introduction: Study Context and Objectives. Urban Institute Press, 2010.

[35] D.K Schaffer. *Reconsidering Drug Court Effectiveness: A Meta-Analytic Review*. PhD thesis, University of Cincinnati, 2006.

[36] Lawrence W. Sherman, Denise C. Gottfredson, Doris L. MacKenzie, John Eck, Peter Reuter, and Shawn D. Bushway. Preventing crime: What works, what doesn't, what's promising. Technical report, National Institute of Justice, 1998.

[37] Teresa C. Smith, Keith R. Abrams, and David R. Jones. Hierarchical models in generalised synthesis of evidence: An example based on studies of breast cancer screening. Technical report, International Society for Biostatistics Conference, 1995.

[38] Alex J. Sutton and Keith R. Abrams. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, 10:277–303, 2001.

[39] Rebecca M. Turner, David J. Spiegelhalter, Gordon C.S. Smith, and Simon G. Thompson. Bias modelling in evidence synthesis. *Journal of the Roya*, 172(1):21–47, 2009.

[40] N.J. Welton, A.E. Ades, J.B. Carlin, D.G. Altman, and J.A.C. Sterne. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society*, 172(1):119–136, 2009.

[41] David B. Wilson. *The role of method in treatment effect estimates: evidence from psychological, behavioral, and educational treatment intervention meta-analyses.* PhD thesis, Claremont Graduate School, 1995.

[42] David B. Wilson, Ojmarrh Mitchell, and Doris L. MacKenzie. A systematic review of drug court effects on recidivism. *Journal of Experimental Criminology*, 2:459–487, 2006.

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

[44] WSIPP. Washington state's drug courts for adult defendants: Outcome evaluation and cost-benefit analysis. Technical report, Washington State Institute for Public Policy, 2003.

[45] J. Zweig, S. Rossman, and J. Roman. *Vol. 2: What's Happening with Drug Courts? A National Portrait of Adult Drug Courts*. Urban Institute Press, 2010.

[0]

DCPI

DISTRICT OF COLUMBIA CRIME POLICY INSTITUTE

URBAN INSTITUTE | BROOKINGS

September 2010

URBAN INSTITUTE
Justice Policy Center
2100 M St NW
Washington, DC 20037
http://www.urban.org