Bayesian Graphical Models, Intention-to-Treat, and the Rubin Causal Model

David Madigan
Box 354322, Department of Statistics
University of Washington
Seattle, WA 98195
madigan@stat.washington.edu

Abstract

In clinical trials with significant noncompliance the standard intention-to-treat analyses sometimes mislead. Rubin's causal model provides an alternative method of analysis that can shed extra light on clinical trial data. Formulating the Rubin Causal Model as a graphical model facilitates model communication and computation.

1 INTRODUCTION

The clinical trials literature distinguishes between two types of objectives. "Use-effectiveness" or "pragmatic" trials seek to provide valid estimates and tests for the effect on outcome of assignment to therapy. "Method-effectiveness" or "explanatory" trials on the other hand seek to assess the effect of the actually administered therapy (Fisher et al., 1990). Intention-to-treat (ITT) is the standard analytic technique for estimating use-effectiveness. This approach compares the outcomes for subjects on the basis of assigned treatment and ignores the actual treatment received. ITT has a key and useful role in clinical trial analysis. However, in many cases, ITT analysis may produce erroneous inferences about the effectiveness of treatment. In such situations, we argue that an alternative analytic procedure should supplement the ITT analysis.

We begin with a motivating example. Cytomegalovirus (CMV) retinitis is a major cause of morbidity in patients with AIDS. Autopsy and clinical data indicate that up to 40% of AIDS patients experience sight-threatening or life-threatening CMV disease (Drew, 1992). The drug ganciclovir is the standard treatment approach for CMV and the oral form of the drug received FDA approval in 1994 for CMV treatment (Fisher and Barton, 1996).

In September 1995, the National Institute of Allergy and Infectious Diseases' Community Programs for Clinical Research on AIDS (CPCRA) issued the results of a study that considered the use of oral ganciclovir as a prophylactic intervention. This was a double-blind, placebo-controlled randomized trial involving 994 patients. A news article in *The Lancet* reported the disappointing news that "Oral ganciclovir fails to prevent CMV in HIV trial" (September 30, 1995, p.895.). The article went on to state that "oral ganciclovir did not prevent symptomatic CMV disease to a clinically or statistically significant degree" (McCarthy, 1995).

The CPCRA data analysis used an "intention-to-treat" method; that is, the analysis ignored the actual drugs used by the study participants and instead compared the outcomes of the subjects assigned to placebo with the outcomes of the subjects assigned to oral ganciclovir irrespective of the actual treatment received. However, after the CPCRA study began, the results of a different study involving 725 subjects became available. This study showed a 49% decrease in the number of clinical CMV infections in the group that received prophylactic oral ganciclovir (Drew et al., 1995, Spector et al., 1996). Consequently, for ethical reasons the CPCRA allowed the subjects in its placebo arm to take oral ganciclovir. The intention-to-treat analysis ignored this fact, and as a result was biased in favor of the no-treatment effect hypothesis (the degree of bias is unclear because subject exposure to ganciclovir in the placebo arm averaged only 2.1 months as compared with 9.3 months in the treatment arm). *The Lancet* report did not mention this problem.

Despite many examples like the preceding one (see also Pocock and Abdalla, 1998), intention-to-treat (ITT) analyses have become a mainstay of the clinical trialist dealing with non-compliance, often to the exclusion of other analyses. The FDA guideline (FDA, 1988), the European Union's Guidance from the Commission for Proprietary Medical Products, and a number of similar documents, as well as numerous publications in the medical literature, support ITT analysis. Some authors go so far as to issue broad recommendations that "the primary analysis of a randomized clinical trial should compare patients in their randomly assigned treatment groups" and that the validity of statistical analyses that

consider the actual treatment received "will be undermined" (Lee et al., 1991). Other authors have called for a more cautious attitude towards ITT - see for example Feinstein (1991, p.361), Jones et al. (1996), Lewis (1995), Lewis and Machin (1993), Salsburg (1994), and Sheiner and Rubin (1995) - but their advice often goes unheeded.

The "Rubin Causal Model" (Rubin, 1974, Holland, 1986) as applied to randomized trials with non-compliance by Imbens and Rubin (1997) provides an alternative to ITT. This approach combines Bayesian analysis with the counterfactual perspective introduced to statistics by Neyman (1923). We emphasize that we are not suggesting that Imbens-Rubin Causal (IRC) models replace ITT analyses, but rather that they be used in a supplemental manner to shed extra light on trial data. We agree that with the suggestion of Spiegelhalter et al. (1994) that clinical trialists should present the results of a Bayesian analysis separately from the conventional "results" section in an additional formal section on "interpretation."

2 INTENTION-TO-TREAT

Sheiner and Rubin (1995) argue that method-effectiveness is more relevant to medical decision making than use-effectiveness, and it is clear that ITT analysis can be highly inappropriate for estimating and testing method-effectiveness (e.g., the ganciclovir trial described above). Sheiner and Rubin (1995) also argue that, faced with substantial non-compliance, ITT analyses can be misleading even for use-effectiveness trials since compliance patterns in the clinical trial context may be quite different from compliance patterns in normal practice. Recent results on the effectiveness of protease inhibitors in HIV/AIDS care provide a troubling example of this problem. Early clinical trials showed that upwards of 90% of HIV/AIDS patients responded to treatment with multiple protease inhibitors with viral loads dropping to undetectable levels. However, data presented by Deeks at a September 1997 infectious disease conference (ICAAC-97) suggested that response rates in routine care settings may be much lower. Of 136 HIV-infected people using protease inhibitors and reviewed by Deeks and colleagues, 53% had detectable levels of the virus.

Peduzzi et al. (1993) and Sheiner and Rubin (1995) provide a detailed critique of other simplistic forms of analyses such as "as-treated", "per-protocol", "censored method", and "transition method" which result in biased estimates of causal effects. Unlike ITT analyses, "as-treated" and "per-protocol" biases tend to be in the anti-conservative direction (that is, tend to support the alternative hypothesis. The direction of the conservative-ness is reversed in equivalence trials).

So, if the standard methods of analysis fail to adequately address method-effectiveness, do more satisfactory methods exist? The next section describes the IRC model, which, we argue yields valid estimates of method-effectiveness even in the face of confounding non-compliance. The IRC approach does require that the trialist adopt a relatively elaborate model - simple data summaries will not suffice - and, compared to ITT, this involves increased subjectivity and uncertainty. However, the price that ITT pays for greater objectivity and certainty is a failure to estimate method-effectiveness (Sheiner and Rubin, 1995). The IRC model, at the very least, enables trialists to explore the potential usefulness and impact of more accurate estimates of treatment effects.

3 THE IRC MODEL FOR NON-COMPLIANCE IN RANDOMIZED TRIALS

3.1 INTRODUCTION AND NOTATION

A statistical study for causal effects compares the results of two or more treatments on a population of units (e.g., plots of land, animals, people), each of which in principle could be exposed to any of the treatments (Rubin, 1990). In what follows we shall refer to the units as "subjects" and we shall assume that the trial comprises two treatments which we label "E" or "1" for an experimental new therapy and "C" or "0" for an existing or placebo therapy ("C" for Control). The trial follows N subjects for a specified time period (e.g. one year) and measures some health outcome Y (e.g. survival) at the end of that period. Our goal is to estimate the causal effect of E relative to C. Intuitively, this causal effect for a particular subject is the difference between the result if the subject had been exposed to E and the result if, instead, the subject had been exposed to C (Rubin, 1978).

Let $Y_i(j)$ be the health outcome (e.g. survival) for subject i if *all* subjects were assigned to treatment j (i=1,…,N, j=0,1). We define the ITT causal effect of assignment for subject i to be $Y_i(1)$-$Y_i(0)$. This definition does not make much sense without the Stable-Unit-Treatment-Value-Assumption (SUTVA). This assumption says that $Y_i(j)$ is *stable* in the sense that it would take the same value for all other treatment allocations such that subject i receives treatment j. This assumption is not innocuous - the health outcome for Subject A could depend on Subject B's treatment assignment if, for example, A and B were in the same household and the treatment had a psychological component. However, SUTVA is generally not contentious in the sorts of randomized studies considered here. With SUTVA we can consider $Y_i(j)$ to be the outcome for subject i if subject i were assigned to treatment j. We note also that other causal effect

definitions are possible, e.g. $Y_i(1)/Y_i(0)$, but we do not pursue this further here.

Population-level causal effects are usually of more interest than subject-level effects and we adopt the common approach of simply averaging the subject-level causal effects. In what follows we will be especially interested in sub-population average causal effects, such as: ave($Y_i(1)$-$Y_i(0)$ | i-th subject is male).

Similar to the definition of $Y_i(j)$, we define $D_i(j)$ to be an indicator for the treatment that subject i would receive given the assignment j, j=0,1 (the "treatment status.") For now, we shall assume that $D_i(j)$ is binary. Thus we can now define a 4-vector of "semi-latent" variables for i-th subject:

$$(D_i(0), D_i(1), Y_i(0), Y_i(1)).$$

These variables are semi-latent in the sense that for any one subject, we will generally observe at most two of the four variables, i.e., either $D_i(0)$ and $Y_i(0)$, or $D_i(1)$ and $Y_i(1)$. For any particular subject, either or both potentially observable variables may be missing.

For each subject, $\underline{D_i}=(D_i(0), D_i(1))$ describes the compliance behavior. Imbens and Rubin distinguish four categories of subjects. Subject i is a:

Complier,	if $D_i(0)=0$ and $D_i(1)=1$,
Never-taker,	if $D_i(0)=0$ and $D_i(1)=0$,
Always-taker,	if $D_i(0)=1$ and $D_i(1)=1$,
Defier,	if $D_i(0)=1$ and $D_i(1)=0$.

(In Section 4.2 we will extend this framework to include partial compliance.) Now we are ready to define some sub-population causal effects of interest. The complier average causal effect (CACE) is given by:

$$\text{CACE} = \text{ave}(Y_i(1)-Y_i(0) \mid D_i(0)=0 \text{ and } D_i(1)=1).$$

Similarly we can define the defier average causal effect (DACE), the always-taker causal effect (AACE), and the never-taker causal effect (NACE). Of the four sub-population causal effects, two, AACE and NACE, do not address causal effects of the receipt of treatment since the former compares outcomes both with treatment, and the latter compares outcomes both without treatment. For compliers, assignment to treatment agrees with receipt of treatment and CACE compares outcomes with drug to outcomes without drug. For such complier subjects, following Imbens and Rubin (1997), we will attribute the effect on Y of assignment to treatment to the effect of receipt of treatment. This attribution is what trialists typically do in randomized trials with full compliance. The DACE is also of some interest although in what follows, we will focus on the CACE as the primary estimand of interest.

3.2 BAYESIAN INFERENCE WITHOUT COVARIATES

Recent developments in Bayesian computation render the estimation of the CACE straightforward, at least in principle. Imbens and Rubin (1997) present a detailed description of a particular approach to estimation. Here we frame the task in the context of Bayesian graphical models (Spiegelhalter and Lauritzen, 1990, Madigan and York, 1995) which simplifies the procedures and makes extensions to models involving covariates, multiple compliance indicators, and missing data direct and transparent, at least in principle. We emphasize that we are not departing from the conceptual framework of Imbens and Rubin and indeed our analysis of their examples produces similar results to theirs.

In the first instance, consider a situation in which the response variables $Y(0)$ and $Y(1)$ are binary. Our goal is to model the joint posterior distribution of $D(0)$, $D(1)$, $Y(0)$, and $Y(1)$, and thence the posterior distribution of the CACE. In this instance, the data, if complete, would comprise a 2X2X2X2 contingency table. If we confine ourselves to either decomposable log-linear models or acyclic directed graphical models (often called "Bayesian networks") and adopt conjugate prior distributions on the model parameters, then prior-to-posterior analysis with complete data is available in closed form. Thus we can select from a variety of available Monte Carlo algorithms to compute the requisite posterior distribution in a straightforward fashion. The essence of these algorithms is to alternately sample from the conditional distribution of the missing data given values for the parameters and the conditional distribution of the parameters given values for the missing data. Madigan and York (1995) and York et al. (1995) provide a detailed description of the application of such Monte Carlo methods to graphical models with missing data and/or latent variables, and describe a series of applications.

Several different graphical models might be plausible for a given analysis and Figure 1 presents three possibilities. Figure 1(a) presents an unrestricted model and is equivalent to the saturated log-linear model. This model imposes no restrictions on the joint distribution of $D(0)$, $D(1)$, $Y(0)$, and $Y(1)$ and has as many parameters as there are configurations of the four variables (i.e., 16 in this binary case). Figure 1(b) has just two edges and embodies the assumption that $D(0)$ and $Y(0)$ are independent of $D(1)$ and $Y(1)$. In other words, knowing the value of $D(0)$ and/or $Y(0)$ for a particular subject provides no extra information about likely values of $D(1)$ and $Y(1)$ for that subject, and vice versa. Figure 1(c) relaxes the model of Figure 1(b) by allowing for a dependence between $D(0)$ and $D(1)$. This model says that, in general, knowing the value of $D(0)$ for a particular subject is informative about the value of $D(1)$ for that subject, and vice versa.

However the model also implies that $Y(0)$ and $Y(1)$ are conditionally independent given either $D(0)$ or $D(1)$.

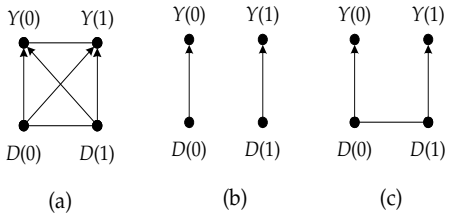

Figure 1: Graphical Models for the IRC with No Covariates

We wish to highlight four particular points. First, without further restrictions on the model parameters, the CACE is "unidentifiable" in these models. This presents problems for a frequentist analysis, but a Bayesian analysis with proper priors does result in proper posteriors. Second, the posterior inference about the CACE in these models may be sensitive to the choice of prior distributions on the model parameters. Third, we rely extensively on Markov Chain Monte Carlo methods to carry out the computations. Because of the potentially large amounts of missing data, numerical and convergence problems may arise. Fourth, somewhat ironically, the edges in our graphical model formulation of the IRC model do not necessarily have a causal interpretation. We are using graphical models merely to encode conditional independencies and provide a convenient and transparent framework for the requisite multivariate analysis.

3.3 BAYESIAN INFERENCE WITH COVARIATES

The benefits of the graphical model approach become apparent when we extend the models to include covariates. Imbens and Rubin (1997) suggest this development in their concluding remarks and the graphical model framework both facilitates the extension and highlights a potential pitfall. Imbens and Rubin (1997) argue that including covariates makes inference conditional and therefore more precise, and covariates allow a more precise partitioning of the sample into compliers, always-takers, never-takers, and defiers when covariates are good predictors of compliance status.

Including covariates (possibly with missing values) in the graphical models of Figure 1 is, in principle, a simple extension and Madigan and York (1995) provide a detailed description. The pitfall that presents itself is that inference about the CACE can be very sensitive to the modeling assumptions concerning the covariates. For example, the two models of Figure 2 can lead to quite different posterior distributions for the CACE. In model (a) both potential health outcomes, Y(0) and Y(1), are conditionally independent of Sex given the potential treatment statuses, D(0) and D(1). Model (b) does not imply this independence. Essentially model (a) says that Sex is directly related to compliance behavior but only indirectly related to health outcome. Model (b) says that Sex is directly related to both compliance behavior and health outcome.

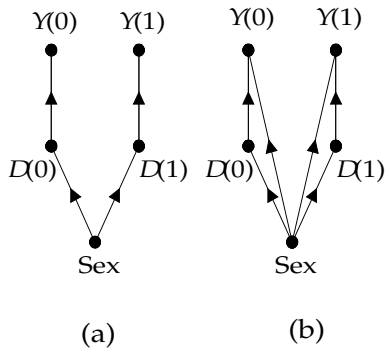

Figure 2: Graphical Models for the IRC Model with Sex Covariate.

3.4 MODEL SELECTION AND MODEL AVERAGING

Calculation of Bayes factors is central to both model selection and model averaging for IRC models. Kass and Raftery (1995, Section 4.3) review various approaches to calculation, including several methods which directly utilize posterior simulations.

We applied the models of Figures 1 and 2 to the example in Section 4.1 below and the resultant causal inferences differ substantially. The standard approach to statistical modeling is to select a single model that maximizes some criterion (e.g. the model with the highest posterior probability). The resulting inferences will, however, be over-precise since they fail to account for model uncertainty (Draper, 1995, Madigan and Raftery, 1994). Bayesian model averaging provides a particular solution to this problem and York et al. (1995) describe a Markov Chain Monte Carlo algorithm that allows for incomplete data and is directly applicable to IRC models.

3.5 RELATIONSHIP TO INSTRUMENTAL VARIABLE MODELS

Glickman and Normand (1995) summarize the four assumptions routinely adopted in the instrumental variables literature. The first is the SUTVA assumption mentioned above. This is the only one of the assumptions we adopt by default. The second assumption is the "exclusion restriction." Different versions of the

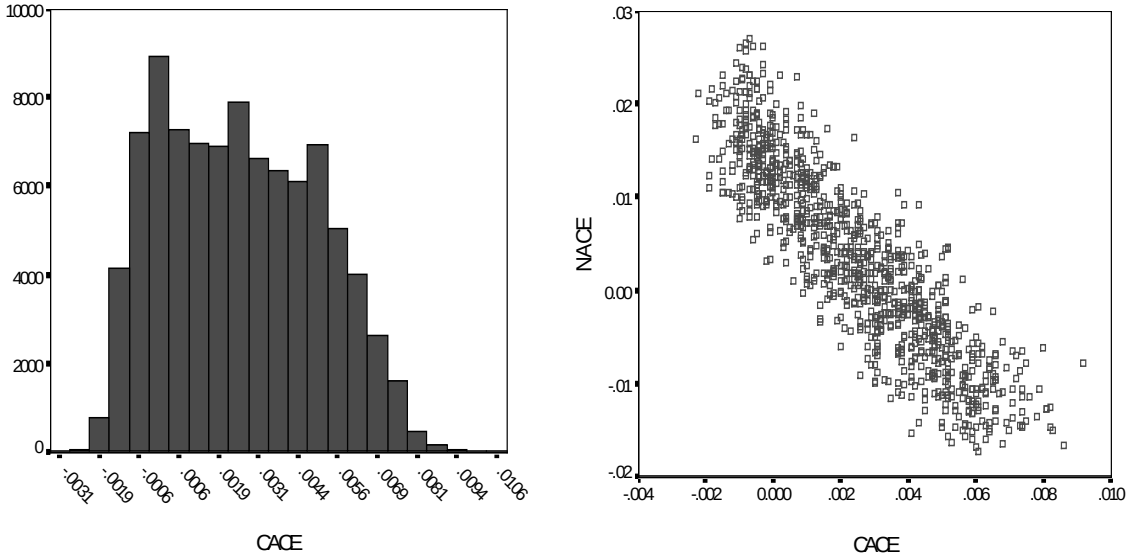

Figure 3: Histogram of the CACE and Scatterplot of NACE Versus CACE in the Vitamin A Example. No Covariate

exclusion restriction exist in the literature. The "weak exclusion principle" of Imbens and Rubin (1997), for instance, states that $Y_i(1) = Y_i(0)$ for all i such that $D_i(1) = D_i(0)$. That is, if for subject i, treatment assignment Z_i has no effect on treatment status, D_i, it also has no effect on health outcome Y_i, so that NACE = AACE = 0. Stronger variants of this assumption appear in the econometrics literature. The third assumption, "monotonicity", states that $D_i(1) \geq D_i(0)$ for all i, with inequality for at least one subject. These three assumptions are sufficient to ensure the identifiability of the CACE.

Finally, in order to make causal inferences, it is necessary to assume that the mechanism that generates Z_i is "ignorable" (Rubin, 1978). If no covariate data are recorded, then the mechanism that generates the Z_i is ignorable if the Z_i can be viewed as being randomized to subjects. Given observed covariate data, the mechanism that generates Z_i is ignorable if the distribution of the Z_i does not depend on unobserved data, but possibly on observed covariate data. Since we only consider randomized studies here, this assumption is trivially satisfied.

4 TWO EXAMPLES

This section presents two examples. The first addresses the Indonesian Vitamin A trial data analyzed by Imbens and Rubin (1997) and involves a binary outcome variable. We introduce an artificial covariate and investigate the modeling consequences. The second example concerns the educational experiment of Schaffner et al. (1997.) This involves a continuous outcome, multiple compliance measures, and covariates.

4.1 THE INDONESIAN VITAMIN A TRIAL

Sommer and Zeger (1991) report results from a trial that randomly assigned villages in Northern Sumatra to receive or not to receive vitamin supplements for a one-year period. No subjects in villages not assigned to receive the supplements in fact received them, but some subjects in villages assigned to the supplements did not receive them. As in the Imbens and Rubin description, we have $D_i(0)=0$, but $D_i(1)=1$ or 0 for all i (this is an example where the monotonicity assumption holds). We note that a correct analysis of these data would account for the within-village dependence, but we were unable to secure the original data from the authors. Table 1 shows the available data.

A Markov Chain Monte Carlo analysis of these data using uniform priors on all the parameters and using model (c) of Figure 1 (albeit with $D(0)\equiv 0$), produces inferences similar to those of Imbens and Rubin (1997). The posterior mean and standard deviation of the CACE are 0.0025 and 0.0024 respectively. (A C program to compute these results is available from the author). This corresponds to an increase in survival rate of 2.5 per 1,000 subjects. The overall survival rate in the sample is 994.9 per 1,000. Figure 3 shows a histogram of the CACE draws which is essentially flat in the region -0.001 to 0.007. Note that there is a non-negligible posterior probability that the CACE is in fact negative. Imbens and Rubin (1997) also show the joint posterior distribution of the CACE and the NACE - Figure 3 shows essentially the same plot. Despite the inherent under-identification of the

Table 1: Sommer-Zeger Vitamin Supplement Data

Type	Assig-nment Z	Vitamin? D	Surv-ival? Y	Number (Total= 23,682)
Complier or Never-Taker	0	0	0	74
Complier or Never-Taker	0	0	1	11,514
Never-Taker	1	0	0	34
Never-Taker	1	0	1	2,385
Complier	1	1	0	12
Complier	1	1	1	9,663

models, the analysis suggests that if the CACE is negative, then the NACE would have to be positive. So, if you believe that the CACE is negative, this necessitates that you also believe that the effect of treatment assignment is positive. Imbens and Rubin (1997) go on to demonstrate the sharper inferences that result from imposing the exclusion restriction.

Sommer et al. (1986) in the original report of this trial noted that "The impact of vitamin A supplementation seemed to be greater in boys than in girls." We simulated several versions of a sex covariate to investigate the potential impact of such a covariate on the causal inferences. In the first instance, we simulated a sex covariate that was marginally independent of the other four variables. Not surprisingly, this has little impact on the causal inferences irrespective of the model chosen. Next we simulated a sex covariate that was highly correlated with treatment status (i.e., *D*) but which was almost conditionally independent of health outcome (i.e., *Y*) given treatment status. Figure 4 shows the results using this covariate and model (a) of Figure 2 and Figure 5 shows the results with the same covariate and model (b) of Figure 2.

Since the covariate here is fictitious, we cannot make substantive points about the causal effects. However, we wish to highlight the sensitivity of the analysis both to the covariate and to the particular selected model. In the analysis without the covariate (Figure 3) there is uncertainty about whether or not the CACE is positive, but the negative correlation with NACE provides useful insights. In the analysis with covariate and model (a), we are now essentially certain that the CACE is positive, the NACE is negative, and the correlation between NACE and CACE has almost disappeared. Using model (b) however, we draw inferences that are more similar to the model without the covariate, although the posterior variability of the CACE has increased substantially. The point here is that causal inferences can be highly sensitive to the treatment of covariates.

All the results are based on runs of length 100,000 with burn-in of 10,000. This exceeds the run lengths suggested

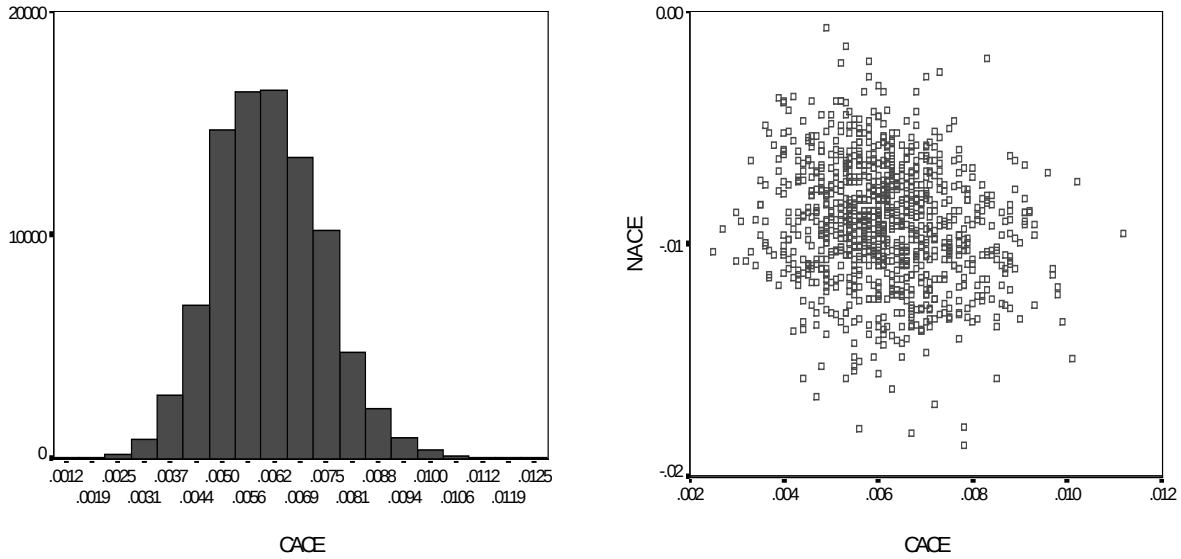

Figure 4: Histogram of the CACE and Scatterplot of NACE Versus CACE in the Vitamin A Example. Sex Covariate Related to Treatment Status. This Analysis uses Model (a) of Figure 2 with $D(0) \equiv 0$.

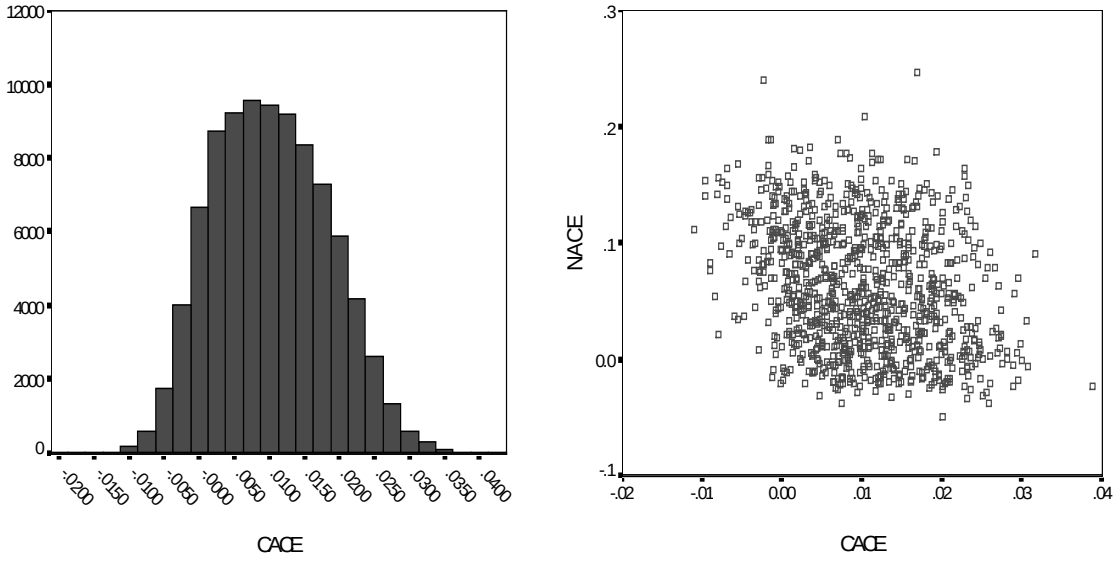

Figure 5: Histogram of the CACE and Scatterplot of NACE Versus CACE in the Vitamin A Example. Sex Covariate Related to Treatment Status. This Analysis Uses Model (b) of Figure 2 with $D(0) \equiv 0$.

by Raftery-Lewis diagnistics by a factor of four. The scatterplots present a random sample from the MCMC output for display purposes.

4.2 EDUCATIONAL EXPERIMENT

This section presents an example concerning the educational experiment of Schaffner et al. (1997). This involves a continuous outcome, multiple compliance measures, and covariates. We carried out the calculations using the program BUGS (Spiegelhalter et al., 1995) and the corresponding BUGS code is available from the author.

Schaffner et al. (1997) describe a randomized experiment to evaluate a set of educational interventions in the context of undergraduate introductory statistics. The experiment took place during a three-week period of a ten-week course. 70 students (34 female, 36 male) participated in the experiment. During the first week of the quarter, the students completed an in-class multiple-choice pre-test. During the third week of the quarter each student was randomly assigned to either a treatment group (n=38) or a control group (n=32). The randomization blocked on section time (8:30 or 12:30) and gender. The two groups (treatment and control) met in separate classrooms with instructors alternating between the classrooms. Both groups were assigned the same homework problems and both were assigned to carry out exercises online (discussed in more detail below).

The lecture portions of the two classes systematically differed. The control group followed a traditional didactic lecture style whereas the cooperative/constructive (treatment) group used the same overhead notes, but the instructor encouraged the class to generate many of the ideas on the notes before they were displayed. In addition to the different styles of lectures, the students in the two groups participated in different online activities. Each Monday, the experiment assigned the groups a new "DIANA" assignment and a new "Web" assignment. DIANA is a simple intelligent tutoring system. Control group students received a reduced version of DIANA with simple correct-incorrect feedback, whereas treatment group students received elaborate student-specific feedback. For the Web assignment, control group students simply filled out a form describing their action plans for a particular statistical problem. The treatment group students worked in subgroups of 6-8 students to solve the same problem, but with discussion via the Web extending over a week, and with instructor intervention.

All students (treatment and control) were graded on a participation-only basis for both the DIANA and Web assignments, receiving a separate score of zero, one, two, or three for the DIANA component and for the Web component. At the conclusion of the three-week experiment, the groups reconvened in one classroom to take a post-test.

Since Schaffner et al. (1997) found the pre-test was essentially independent of the post-test, we ignore it in our analysis. Table 2 describes the random variables for student i.

Table 2: Random variables for the educational experiment

Variable Name	Possible Values	
Z_i	0,1	Random assignment
$D_i(j)$	0,1,2,3	Number of completed DIANA assignments, all students assigned to treatment j
$W_i(j)$	0,1,2,3	Number of completed Web assignments, all students assigned to treatment j
$Y_i(j)$	0-11	Score on the post-test, all students assigned to treatment j
G_i	Male, Female	Gender
S_i	8:30 or 12:30	Section

Both *W* and *D* are measures of compliance, but since the intervention also included some special classroom activities, the exclusion assumption would not be reasonable *a priori*. Several causal effects are of interest here. Denote by CACE(*i*,*j*) the average causal effect for students who complete *i* DIANA assignments and *j* Web assignments. So, CACE(0,0) measures the causal effect due to classroom component of the intervention. CACE(3,3) measures the causal effect for students who fully comply with all aspects of the intervention. CACE(3,0) measures the causal effect without the Web component. CACE(0,3) measures the causal effect without the DIANA component. Figure 6 shows a particular model for these data.

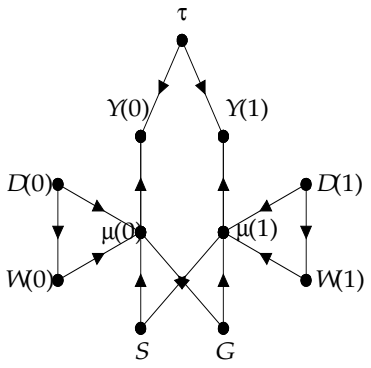

Figure 6: Graphical Model for the Educational Experiment. This Model Implies that Section and Gender are Independent of DIANA and Web Compliance.

Specifically, this model assumes that the covariates and compliance variables enter linearly as follows:

$$Y_i(j) \sim N(\mu_i(j),\tau),\ i=1,\ldots,n,\ j=1,2.$$

$$\mu_i(j) \sim \alpha_i(j) + \beta_i(j)D_i(j) + \gamma_i(j)W_i(j) + a(j)G_i + b(j)S_i$$

At the next level in the model's hierarchy we have:

$$\alpha_i(j) \sim N(\mu_\alpha(j),\tau_\alpha(j))$$

$$\beta_i(j) \sim N(\mu_\beta(j),\tau_\beta(j))$$

$$\gamma_i(j) \sim N(\mu_\gamma(j),\tau_\gamma(j))$$

$$D_i(j) \sim \text{Bin}(p_D(j), 3)$$

$$W_i(j) \sim \text{Bin}(p_W(D(j)), 3)$$

Finally, $\mu_\alpha(j)$, $\mu_\beta(j)$, $\mu_\gamma(j)$, $a(j)$, $b(j)$ are normally distributed *a priori* with mean zero and precision 0.0001, $\tau_\alpha(j)$, $\tau_\beta(j)$, $\tau_\gamma(j)$, and τ are gamma(0.001,0.001) *a priori*, and the binomial probabilities are uniformly distributed *a priori*. These prior distributions are intended to be reasonably flat in the regions where the likelihood is non-negligible. Figure 7 shows the corresponding causal effect histograms and Table 3 shows the posterior means and standard deviations.

Table 3: Posterior Means and Standard Deviations for the Educational Example

Causal Effect	No. of DIANA Assignments	No. of Web Assignments	Mean	SD
CACE(0,0)	0	0	-0.06	1.5
CACE(0,3)	0	3	+1.41	1.4
CACE(3,0)	3	0	-0.44	1.9
CACE(3,3)	3	3	+1.03	0.7

There is considerable uncertainty associated with each of the causal effects and posterior 95% intervals include zero for all four effects. Focusing on the posterior means, this analysis suggests that the causal effect of the Web assignments is positive, but that the causal effect of the DIANA assignments is actually negative. The causal effect associated with the classroom

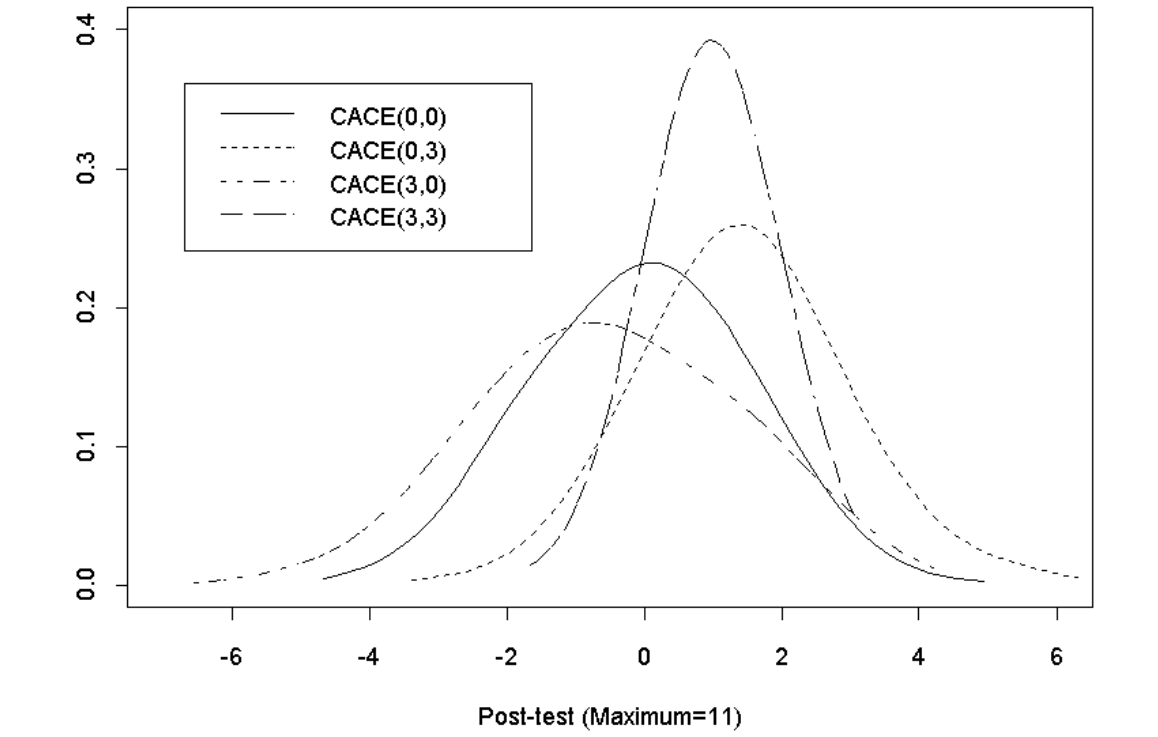

Figure 7: Samples from the Posterior Densities of Various Causal Effects in the Educational Example.

activities alone seems to be negligible. Overall, the point estimate of the causal effect for the complete intervention (i.e., CACE(3,3)) is about 9%, a result consistent with that of Schaffner et al. (1997).

Clearly, there is an arbitrariness concerning the model we have used for this analysis, and variants on the model do lead to somewhat different inferences (although less dramatically different than the previous example).

Again we used run lengths of 100,000 with a burn-in of 10,000. This exceeded the run lengths suggested by Raftery-Lewis diagnostics by a factor of between two and four.

5 CONCLUSIONS

We have described a Bayesian graphical modeling approach to the Rubin causal model. The analysis of the Educational Experiment in particular shows how the graphical model framework greatly facilitates generalizations of the original IRC model and highlights the important role of model uncertainty in causal inferences.

Noncompliance in the "real world" is a complex phenomenon and there are many issues we have not addressed. These include unobserved or partially observed compliance, mixed randomized/non-randomized studies, and longitudinal studies. Robbins (1998) surveys an extensive and important body of work that deals with many of these issues, albeit from a classical perspective.

Acknowledgements

A grant from the National Science Foundation supported this work (DMS 9704573). Thanks to David Draper, Ed George, David Hand, Martha Nason, Phil Neal, Adrian Raftery, Thomas Richardson and Lawrence Schall for helpful discussions.

References

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B)*, **57**, 45-97.

Drew, W.L. (1992). Cytomegalovirus infection in patients with AIDS. *Journal of Infectious Diseases*, **143**, 188-92.

Drew, W.L., Ives, D., Lalezari, J.P., et al. (1995). Oral ganciclovir as maintenance treatment for cytomegalovirus retinitis in patients with AIDS. *New England Journal of Medicine*, **333**, 615-620.

FDA (1988). Food and Drug Administration. *Guideline for the Format and Content of the Clinical and Statistical Sections of new Drug Applications*, FDA, US Department of Health and Human Services, Rockville, MA, USA.

Fisher, L.D., Dixon, D.O., Herson, J., Frankowski, R.K., Hearron, M.S., and Peace, K.E. (1990). Intention to treat in clinical trials. In: *Statistical Issues in Drug Research and Development*, (Ed. K.E. Peace), Marcel Dekker, pp.331-350.

Fisher, M. and Barton, S. (1996). Oral ganciclovir: A new option for patients with CMV retinitis. *International Journal of STD and AIDS*, **7**, 1-3.

Glickman, M.E. and Normand, S-L.T. (1995). The derivation of a latent threshold instrumental variable model. Technical Report #HCP-1995-5, Dept of Health Care Policy, Harvard Medical School.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81**, 945-970.

Imbens, G. and Rubin, D.B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, **25**, 305-327.

Jones, B., Jarvis, P., Lewis, J.A., and Ebbutt, A.F. (1996). Trials to assess equivalence: the importance of rigorous methods. *BMJ*, **313**, 36-9.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Lee, Y.J., Ellenberg, J.H., Hirtz, D.G., and Nelson, K.B. (1991). Analysis of clinical trials by treatment actually received: is it really an option? *Statistics in Medicine*, **10**, 1595-605.

Lewis, J.A. (1995). Statistical issues in the regulation of medicines. *Statistics in Medicine*, **14**, 127-36.

Lewis, J.A. and Machin, D. (1993). Intention to treat - who should use ITT? *British Journal of Cancer*, **68**, 647-650.

Madigan, D. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's Window. *Journal of the American Statistical Association*, **89**, 1535-1546.

Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215-232.

McCarthy, M. (1995). Oral ganciclovir fails to prevent CMV in HIV trial. *The Lancet*, **346**, 895.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9, [Translated in *Statistical Science*, **5**, 465-480, 1990.

Peduzzi, P., Wittes, J., Detre, K., and Holford, T. (1993). Analysis as-randomized and the problem of non-adherence: an example from the veterans affairs randomized trial of coronary artery bypass surgery. *Statistics in Medicine*, **12**, 1185-1195.

Pocock, S.J. and Abdalla, M. (1998). The hope and the hazards of using compliance data in randomized controlled trials. *Statistics in Medicine*, **17**, 303-317.

Robins, J.M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine*, **17**, 269-302.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educaitonal Psychology*, **66**, 688-701.

Rubin, D.B. (1978). Bayesian inference for causal effects. *Annals of Statistics*, **6**, 34-58.

Rubin, D.B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, **25**, 279-292.

Salsburg, D. (1994). Intent to treat: The reductio ad absurdum that became gospel. *Pharmacoepidemiology and Drug Safety*, **3**, 329-335.

Schaffner, A., Madigan, D., Hunt, E.B., and Minstrell, J. (1997). Virtual benchmark instruction: Cooperative learning for undergraduate statistics education. *Journal of Statistics Education*, under revision.

Sheiner, L.B. and Rubin, D.B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology and Therapeutics*, **57**, 6-15.

Sommer, A., Tarwotjo, I., Djunaedi, E., West, K.P.,Jr., Loeden, A.A., Tilden, R., and Mele, L. (1986). Impact of vitamin A supplementation on childhood mortality. A randomised controlled community trial. *The Lancet*, 1169-73.

Sommer, A. and Zeger, S. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine*, **10**, 45-52.

Spector, S.A., McKinley, G.F., Lalezari, J.P., et al. (1996). Oral ganciclovir for the prevention of cytomegalovirus disease in persons with AIDS. *New England Journal of Medicine*, **334**, 1491-7.

Spiegelhalter, D.J. and Lauritzen, S.L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**, 579-605.

Spiegelhalter, D.J., Fredman, L.S., and Parmar, M.K.B. (1994). Bayesian approaches to randomized trials (with discussion). *Journal of the Royal Statistical Society (Series A)*, **157**, 357-416.

Spiegelhalter, D.J., Thomas, A., Best, N.G., and Gilks, W.R. (1995). *BUGS: Bayesian inference using Gibbs sampling, Version 0.50*. MRC Biostatistics Unit, Cambridge.

York, J., Madigan, D., Heuch, I., and Lie, R.T. (1995). Birth defects registered by double sampling: a Bayesian approach incorporating covariates and model uncertainty. *Applied Statistics*, **44**, 227-242.