# A Clinician's Tool for Analyzing Non-compliance

**David Maxwell Chickering*  and Judea Pearl**

Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
*dmax@cs.ucla.edu*
*judea@cs.ucla.edu*

### Abstract

We describe a computer program to assist a clinician with assessing the efficacy of treatments in experimental studies for which treatment assignment is random but subject compliance is imperfect. The major difficulty in such studies is that treatment efficacy is not "identifiable", that is, it cannot be estimated from the data, even when the number of subjects is infinite, unless additional knowledge is provided. Our system combines Bayesian learning with Gibbs sampling using two inputs: (1) the investigator's prior probabilities of the relative sizes of subpopulations and (2) the observed data from the experiment. The system outputs a histogram depicting the posterior distribution of the average treatment effect, that is, the probability that the average outcome (e.g., survival) would attain a given level, had the treatment been taken uniformly by the entire population. This paper describes the theoretical basis for the proposed approach and presents experimental results on both simulated and real data, showing agreement with the theoretical asymptotic bounds.

## Introduction

Standard clinical studies in the biological and medical sciences invariably invoke the instrument of randomized control, that is, subjects are assigned at random to various groups (or treatments or programs) and the mean differences between participants in different groups are regarded as measures of the efficacies of the associated programs. For example, to determine if a new drug is useful for treating some disease, subjects will be divided (at random) into a control group and a treatment group. The members of the control group are given a placebo and the members of the treatment group are given the drug in question. For each group, the clinician records the fraction of subjects that recover from the disease. By comparing these fractions the clinician can derive a quantitative measure of effectiveness of the drug for treating the disease. In particular, if $f_c$ and $f_t$ are the fractions of subjects that recovered from the control group and treatment group respectively, then the difference $E = f_c - f_t$ is an indication of the effectiveness of the drug.

The major source of difficulty in managing and analyzing such experiments has been subject noncompliance. For example, a subject in the treatment group may experience negative side effects and will stop taking the drug. Alternatively, if the experiment is testing a drug for a terminal disease, a subject suspecting that he is in the control group may obtain the drug from other sources. Imperfect compliance poses a problem because simply comparing the fractions as above may provide a misleading estimate for how effective the drug would be if applied uniformly to the population. For example, if those subjects who refused to take the drug are precisely those who would have responded adversely, the experiment might conclude that the drug is more effective than it actually is. It can be shown, in fact, that treatment effectiveness in such studies is *non-identifiable*. That is, in the absence of additional modeling assumptions, treatment effectiveness cannot be estimated from the data without bias, even as the number of subjects in the experiment approaches infinity, and even when a record is available of the action and response of each subject (Pearl 1995).

In a popular compromising approach to the problem of imperfect compliance, researchers perform an *intent-to-treat* analysis, in which the control and treatment group are compared without regard to whether the treatment was actually received[1]. The result of such an analysis is a measure of how well the treatment *assignment* effects the disease, as opposed to the desired measure of how well the treatment itself effects the disease. Estimates based on intent-to-treat analysis are valid only as long as the experimental conditions perfectly mimic the conditions prevailing in the eventual usage of the treatment. In particular, the experiment should mimic subjects' incentives for receiving each treatment. In situations where field incentives are more compelling than experimental incentives, as is usually the case when drugs receive the approval of a government agency, treatment effectiveness may

*Supported by NSF Grant No. IRI-9119825, and a grant from Rockwell International.

[1]This approach is currently used by the FDA to approve new drugs.

vary significantly from assignment effectiveness. For example, imagine a study in which (a) the drug has an adverse effect on a large segment of the population and (b) only those members of the segment who drop from the treatment arm recover. The intent-to-treat analysis will attribute these cases of recovery to the drug since they are part of the intent-to-treat arm, while in reality these cases have recovered by avoiding the treatment.

Another approach to the problem is to use a correction factor based on an "instrumental variables" formula [Angrist et al., 1993], according to which the intent-to-treat measure should be divided by the fraction of subjects who comply with the treatment assigned to them. Angrist et al. (1993) have shown that, under certain conditions, the corrected formula is valid for the subpopulation of "responsive" subjects, that is, subjects who would have changed treatment status if given a different assignment. Unfortunately, this subpopulation cannot be identified and, more seriously, it cannot serve as a basis for policies involving the entire population because it is instrument dependent—individuals who are responsive in the study may not remain responsive in the field, where the incentives for obtaining treatment differ from those used in the study.

Using a graphical model with latent variables, Balke and Pearl (1994) derive bounds, rather than point estimates, for the treatment effect, while making no assumptions about the relationship between subjects' compliance and subjects' physical response to treatment. However, the derived bounds are "asymptotic", i.e., they ignore sampling variations by assuming that the proportions measured in the experiment are representative of the population as a whole, a condition which is valid only when the number of subjects is large. This large-sample assumption may be problematic when the study includes a relatively small number of subjects.

In this paper we describe a system that provides an assessment of the actual treatment effect and is not limited to studies with large samples. The system uses the graphical model of Balke and Pearl (1994) to learn the the treatment effect using Bayesian updating combined with Gibbs sampling. The system takes as input (1) the investigator's prior knowledge about subject compliance and response behaviors and (2) the observed data from the experiment, and outputs the posterior distribution of the treatment effect. The use of graphical models and Gibbs' methods for deriving posterior distributions in such models are both well known. The main contribution of this paper is a description of how these techniques can be applied to the causal analysis of clinical trials, and a presentation of experimental results of a practical system applied to various simulated and real data. While the basic idea of estimating causal effects using Bayesian analysis goes back to Rubin (1978), and was further used by Imbens and Rubin (1994) to estimate the correctional formula advocated by Angrist et al. (1993), we believe that this is the first time an assumption-free assessment of the *average treatment effect* is made available to the clinical research community.

The paper is organized as follows. First, we introduce a graphical, causal model that represents a prototypical clinical trial with partial compliance, and define *treatment effect* in terms of the model. Next, we describe an equivalent graphical model, using potential-response variables (Balke and Pearl, 1994), that allows the compliance and response behavior to be represented more efficiently. Next, we describe the general Bayesian-learning and Gibbs-sampling methods that were used to derive the posterior parameter densities in the graphical model. Finally, we describe experimental results obtained when our system is applied to various simulated and real data sets. We include results obtained when the system is modified to answer counterfactual queries about specific individuals, e.g., "what if Joe (who died with no treatment) were to have taken the treatment?"

## The Graphical Model

Graphical models are convenient tools for representing causal and statistical assumptions about variables in a domain. In this section, we describe the graphical model of Figure 1, which is used to represent a prototypical clinical trial with partial compliance. We use $Z$, $D$ and $Y$ to denote observed binary variables from the experiment, where $Z$ represents the treatment assignment, $D$ represents the treatment received, and $Y$ represents the observed outcome. To facilitate the notation, we let $z$, $d$, and $y$ represent, respectively, the values taken by the variables $Z$, $D$, and $Y$, with the following interpretation: $z \in \{z_0, z_1\}$, $z_1$ asserts that the treatment has been assigned ($z_0$ its negation); $d \in \{d_0, d_1\}$, $d_1$ asserts that the treatment has been administered ($d_0$ its negation); and $y \in \{y_0, y_1\}$, $y_1$ asserts a positive observed response ($y_0$ its negation). We use $U$ to denote all characteristics, both observed and unobserved, that influence the value of $D$ and $Y$ for the subjects. The domain of $U$ is left unspecified, and in general will combine the spaces of several random variables, both discrete and continuous.

Let $\nu_e$ denote the *physical probability* of the event $E = e$, or equivalently, the fraction of subjects in the population for which $E = e$. The graphical model explicitly represents two independence assumptions about the joint physical probability distribution $\nu_{z,d,y,u}$. First, the model asserts that the treatment assignment $Z$ does not influence $Y$ directly, but only through the actual treatment $D$. That is, $Z$ and $Y$ are conditionally independent given $D$ and $U$. Second, the model asserts that $Z$ and $U$ are marginally independent. This second independence is partly ensured through the randomization of $Z$, which rules out both (1) the existence of a common cause for both $Z$ and $U$, and (2) the possibility that $U$ has causal influence

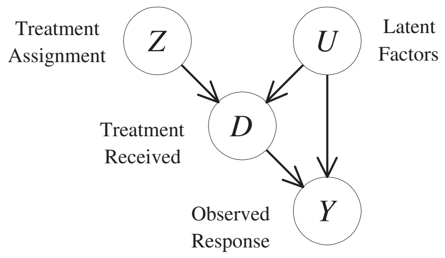



Figure 1: Graphical model for a prototypical clinical trial with partial compliance

on $Z$. The absence of a directed path from $Z$ to $U$ represents the assumption that a person's disposition to comply with or deviate from a given assignment is not affected by the assignment; any such effect can be viewed as part of the disposition itself. The two independence assumptions together induce the following decomposition of the joint distribution:

$$\nu_{z,d,y,u} = \nu_z \nu_u \nu_{d|z,u} \nu_{y|d,u}$$

In addition to the independence assumptions, the graphical model also encodes causal assumptions (e.g., that $Z$ does not effect $Y$ directly) which permit one to predict how the joint probability will change in light of *exogenous* local interventions [Pearl, 1995]. In particular, the absence of any direct link (or any spurious path) from $Z$ to $Y$ implies that $\nu_{y|d,u}$ is the same regardless if $d$ is measured in an observational study, or dictated by some (exogenous) public policy. Consequently, if we wish to predict the distribution $\nu^*_{y|d}$ of $Y$, under a new condition where the treatment $D = d$ is applied uniformly to the population, we should calculate

$$\nu^*_{y|d} = E_u[\nu_{y|d,u}]$$

where $E_u$ denotes the expectation with respect to $\nu_u$. Likewise, if we are interested in the average *change* in $Y$ due to treatment, we use the *average causal effect*, denoted $\text{ACE}(D \rightarrow Y)$, as defined by Holland (1988):

$$\text{ACE}(D \rightarrow Y) = E_u[\nu_{y_1|d_1,u} - \nu_{y_1|d_0,u}] \qquad (1)$$

Let $\mathcal{D}$ denote the observed collection of triples $\{z, d, y\}$, one for each subject, that we obtain from the experiment. Given $\mathcal{D}$, the objective of our system is to derive the posterior Bayesian probability distribution $p(\text{ACE}(D \rightarrow Y) \,|\mathcal{D})$.

## The Potential-Response Model

The graphical model presented in the previous section is attractive for representing the assumptions that underlie a given experimental design, but may not be convenient for computation. For example, the graph of Figure 1 represents explicitly the assumptions that $Z$ is randomized and that $Z$ does not affect Y directly, while making no assumption about the relationship between compliance and the way subjects would respond to the treatment. However, leaving the domain of the unobserved variable $U$ unspecified makes it difficult to derive the distribution of interest, namely, $p(\text{ACE}(D \rightarrow Y) \,|\mathcal{D})$.

As is done by Balke and Pearl (1994), we exploit the observation of Pearl (1994) that $U$ can always be replaced by a single discrete and finite variable such that the resulting model is equivalent with respect to all observations and manipulations of $Z$, $D$, and $Y$. In particular, because $Z$, $D$, and $Y$ are all binary variables, the state space of $U$ divides into 16 equivalence classes: each equivalence class dictates two functional mappings; one from $Z$ to $D$, and the other from $D$ to $Y$. To describe these equivalence classes, it is convenient to regard each of them as a point in the joint space of two four-valued variables $C$ and $R$. The variable $C$ determines the compliance behavior of a subject through the mapping:

$$d = F_D(z, c) = \begin{cases} d_0 & \text{if} \quad c = c_0 \\ \\ d_0 & \text{if} \quad c = c_1 \quad \text{and} \quad z = z_0 \\ d_1 & \text{if} \quad c = c_1 \quad \text{and} \quad z = z_1 \\ \\ d_1 & \text{if} \quad c = c_2 \quad \text{and} \quad z = z_0 \\ d_0 & \text{if} \quad c = c_2 \quad \text{and} \quad z = z_1 \\ \\ d_1 & \text{if} \quad c = c_3 \end{cases} \qquad (2)$$

Imbens and Rubin (1994) call a subject with compliance behavior $c_0$, $c_1$, $c_2$ and $c_3$, respectively, a *never-taker*, a *complier*, a *defier* and an *always-taker*. Similarly, the variable $R$ determines the response behavior of a subject through the mapping:

$$y = F_Y(d, r) = \begin{cases} y_0 & \text{if} \quad r = r_0 \\ \\ y_0 & \text{if} \quad r = r_1 \quad \text{and} \quad d = d_0 \\ y_1 & \text{if} \quad r = r_1 \quad \text{and} \quad d = d_1 \\ \\ y_1 & \text{if} \quad r = r_2 \quad \text{and} \quad d = d_0 \\ y_0 & \text{if} \quad r = r_2 \quad \text{and} \quad d = d_1 \\ \\ y_1 & \text{if} \quad r = r_3 \end{cases} \qquad (3)$$

Following Heckerman and Shachter (1995), we call the response behavior $r_0$, $r_1$, $r_2$ and $r_3$, respectively, *never-recover*, *helped*, *hurt* and *always-recover*.

Let $CR$ denote the variable whose state space is the cross-product of the states of $C$ and $R$. We use $cr_{ij}$, with $0 \leq i, j \leq 3$ to denote the state of $CR$ corresponding to compliance behavior $c_i$ and response behavior $r_j$. Figure 2 shows the graphical model that results from replacing $U$ from Figure 1 by the 16-state variable $CR$. A state-minimal variable like $CR$ is called a *response variable* by Balke and Pearl (1994) and a *mapping variable* by Heckerman and Shachter (1995), and its states correspond to the *potential response* vectors in Rubin's model (Rubin 1978).

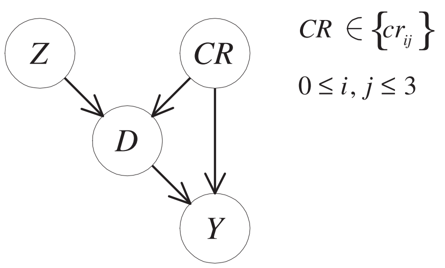



Figure 2: Potential-response model invoking a 16 state variable $CR$

Applying the definition of $\mathrm{ACE}(D \to Y)$ given in Equation 1, it follows that using the model of Figure 2 we have:

$$\mathrm{ACE}(D \to Y) = \left[\sum_i \nu_{cr_{i1}}\right] - \left[\sum_i \nu_{cr_{i2}}\right] \quad (4)$$

Equivalently, $\mathrm{ACE}(D \to Y)$ is the difference between the fraction of subjects who are helped by the treatment ($R = r_1$) and the fraction of subjects who are hurt by the treatment ($R = r_2$).

## Learning the Causal Effect

Given the observed data $\mathcal{D}$ from the experiment, as well as a prior distribution over the unknown fractions $\nu_{CR}$, our system uses the potential-response model defined in the previous section to derive the posterior distribution for $\mathrm{ACE}(D \to Y)$ . In this section, we describe how this computation can be done. To simplify discussion, we introduce the following notation. Assume there are $m$ subjects in the experiment. We use $z^i$, $d^i$ and $y^i$ to denote the observed value of $Z$, $D$ and $Y$, respectively, for subject $i$. Similarly, we use $cr^i$ to denote the (unobserved) compliance and response behavior for subject $i$. We use $\mathcal{D}^i$ to denote the triple $\{z^i, d^i, y^i\}$.

The posterior distribution of the causal effect can be derived using the graphical model shown in Figure 3, which explicitly represents the independences that hold in the joint (Bayesian) probability distribution defined over the variables $\{\mathcal{D}, \nu_{CR}, \mathrm{ACE}(D \to Y)\}$ . The model can be understood as $m$ realizations of the potential-response model, one for each triple in $\mathcal{D}$, connected together using the node representing the unknown fractions $\nu_{CR}$. The model explicitly represents the assumption that, given the fractions $\nu_{CR}$, the probability of a subject belonging to any of the compliance-response subpopulations does not depend on the compliance and response behavior of the other subjects in the experiment. From Equation 4, $\mathrm{ACE}(D \to Y)$ is a deterministic function of $\nu_{CR}$ and consequently $\mathrm{ACE}(D \to Y)$ is independent of all other variables in the domain once these fractions are known.

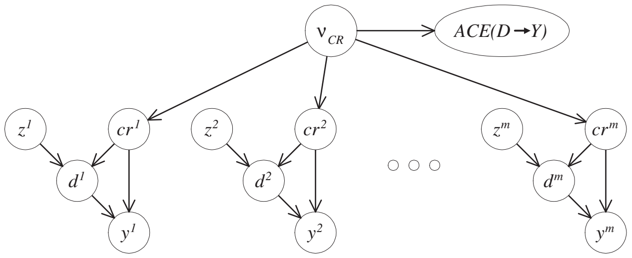



Figure 3: Model used to represent the independences in $p(\{\mathcal{D}\} \cup \{\nu_{CR}\} \cup \{\mathrm{ACE}(D \to Y)\})$

Determining the posterior probability for a node using a graphical model is known as performing *inference* in that model. In many cases, the independences of the model can be exploited to make the process of inference efficient. Unfortunately, because the $cr^i$ are never observed, deriving the posterior distribution for $\mathrm{ACE}(D \to Y)$ is not tractable even with the given independences. To obtain the posterior distribution, our system applies an approximation technique known as Gibbs sampling, which we describe in the following section.

## Gibbs Sampling

Gibbs sampling is a well-known Markov chain sampling method that can be used to approximate the expected value of a function. The method can easily be applied to approximate the posterior density of $\mathrm{ACE}(D \to Y)$ by exploiting the independences in the model from Figure 3.

Suppose we are interested in the expected value of some function $f(X)$ with respect to the distribution $p(X|Y)$:

$$E_{X|Y}[f] = \int_X f(X)p(X|Y)dX$$

In many cases, it may not be easy to solve the above integral analytically. However, we can approximate $E_{X|Y}[f]$ by repeatedly sampling values for $X$ from the distribution $p(X|Y)$, and then taking an average. Assuming that $N$ samples are taken and letting $X^i$ denote the value for $X$ on the $ith$ sample we have:

$$E_{X|Y}[f] \approx \frac{1}{N}\sum_{i=1}^{N} f(X^i) \quad (5)$$

In practice, sampling points directly from $p(X|Y)$ may be difficult. The Gibbs sampling method draws points from the distribution by repeatedly sampling from the *conditional* distributions $p(X_i|X \setminus X_i, Y)$, which are often very easy to derive in closed from. After initially instantiating all the values of $X$, the algorithm repeatedly uninstantiates a single component $X_i$, and re-samples that component according to the conditional distribution $p(X_i|X \setminus X_i, Y)$. It can be shown that as the number of iterations of the Gibbs sampler grows large, the sampled values for $X$ are distributed as $p(X|Y)$[2].

We can use a Gibbs sampler to approximate the posterior distribution of $\mathrm{ACE}(D \to Y)$ as follows. Let $f_{a,b}(\nu_{CR})$ denote the indicator function that is 1 if $a \leq \mathrm{ACE}(D \to Y) \leq b$ and 0 otherwise. Then we have:

$$\begin{aligned} & p(a \leq \mathrm{ACE}(D \to Y) \leq b|\mathcal{D}) \\ &= E_{\nu_{CR}|\mathcal{D}}[f_{a,b}(\nu_{CR})] \\ &= \int f_{a,b}(\nu_{CR}) \cdot p(\nu_{CR}|\mathcal{D}) \cdot d\nu_{CR} \end{aligned}$$

After expanding the integral to include the unobserved compliance and response behavior for each of the subjects we have:

$$\begin{aligned} & p(a \leq \mathrm{ACE}(D \to Y) \leq b|\mathcal{D}) \\ &= \int f_{a,b}(\nu_{CR}) \cdot p(\nu_{CR}, cr^1, \ldots, cr^m|\mathcal{D}) \\ & \quad \cdot d\nu_{CR} \cdot dcr^1 \cdot \ldots \cdot dcr^m \end{aligned}$$

Thus we can use the approximation of Equation 5 in conjunction with the Gibbs sampler to estimate the probability that $\mathrm{ACE}(D \to Y)$ falls within any interval $[a, b]$. The conditional distributions from which we sample are easily derived in light of the independences depicted in Figure 3. In particular, letting $X = \{\nu_{CR}, cr^1, \ldots, cr^m\}$, we have:

$$p(cr^i|X \setminus cr^i, \mathcal{D}) = \alpha \cdot p(d^i, y^i|z^i, cr^i) \cdot \nu_{cr^i}$$

where $\alpha$ is the normalization constant. $p(d^i, y^i|z^i, cr^i)$ is either one or zero, depending on whether the observed values of $z^i$, $d^i$ and $y^i$ agree with the given compliance and response behavior. Note that we have used the fact that if the fractions $\nu_{CR}$ are known, then the probability of $cr^i$ is simply $\nu_{cr^i}$.

To update $\nu_{CR}$ we sample from the posterior distribution:

$$p(\nu_{CR}|X \setminus \nu_{CR}, \mathcal{D}) = \beta \prod_{i=0}^{3} \prod_{j=0}^{3} {\nu_{cr_{ij}}}^{N_{cr_{ij}}} \cdot p(\nu_{CR})$$

where $\beta$ is the normalization constant and $N_{cr_{ij}}$ is the number of times $cr_{ij}$ occurs in $X$.

One choice of the functional form for $p(\nu_{CR})$ is particularly convenient for our application. In particular, if the prior $p(\nu_{CR})$ is a *Dirichlet* distribution, then both efficiently computing the posterior distribution in closed form and sampling from that distribution are easy. Assuming that the prior distribution for $\nu_{CR}$ is Dirichlet implies there exists exponents $N'_{cr_{00}}, \ldots, N'_{cr_{33}}$ such that

$$p(\nu_{CR}) = \gamma \prod_{i=0}^{3} \prod_{j=0}^{3} {\nu_{cr_{ij}}}^{N'_{cr_{ij}}-1}$$

[2]The resulting Markov chain must be ergodic for this result to hold, a property that can be easily established for our application.

where $\gamma$ is the normalization constant. Let $N'_{CR} = \sum_{i=0}^{3} \sum_{j=0}^{3} N'_{cr_{ij}}$. Having the given Dirichlet prior can be thought of as at some point being ignorant about the fractions $\nu_{CR}$, and then observing the compliance and response behavior of $N'_{CR}$ subjects, $N'_{cr_{ij}}$ of which have behavior $cr_{ij}$. Using this simplifying assumption, we update $\nu_{CR}$ by sampling from the following Dirichlet distribution:

$$p(\nu_{CR}|cr^1, \ldots, cr^n) = \gamma\beta \prod_{i=0}^{3} \prod_{j=0}^{3} {\nu_{cr_{ij}}}^{N_{cr_{ij}} + N'_{cr_{ij}} - 1}$$

For accurate results, the Gibbs sampler is typically run in two distinct phases. In the first phase, enough samples are drawn until it is reasonable to assume that the resulting Markov chain has converged to the correct distribution. These initial samples are commonly referred to as the *burn-in* samples, and the corresponding values of the function being estimated are ignored. In the second phase, the values of the function are recorded and are used in the approximation of Equation 5. There are countless techniques for determining when a series has converged, and no single method has become universally accepted among researchers. Another complication of the Gibbs sampler is that successive samples in the second phase are inherently dependent, yet we use these samples to approximate independent samples from the distribution. As a consequence of the many different methods to address these problems, tuning a Gibbs sampler for the best results tends to be more of an art than a science.

The approach we took for the results presented in the next section can be explained as follows. We ran the Gibbs sampler for enough iterations to ensure a relatively smooth estimate of the distribution, always discarding a large number of the initial points sampled. We then repeated the same schedule, starting with a different random seed, and compared the resulting outputs. If the distributions were reasonably distinct, we repeated the process using more samples. We emphasize that the any one of the many methods of data analysis can readily be applied to the output of our system.

## Experimental Results

We have applied the Gibbs sampling algorithm to the model of Figure 3 for various real and simulated data sets. Our system takes as input (1) the observed data $\mathcal{D}$, expressed as the number of cases observed for each of the 8 possible instantiations of $\{z, d, y\}$, and (2) a Dirichlet prior over the unknown fractions $\nu_{CR}$, expressed as the 16 exponents $N'_{CR}$. The system outputs the posterior distribution of $\mathrm{ACE}(D \to Y)$ , expressed in a histogram.

To investigate the effect of the prior distribution on the output, we have run all our experiments using two different priors as input. The first is a flat (uniform) distribution over the 16-vector $\nu_{CR}$, and is

Table 1: Population fractions resulting in an identifiable ACE($D \rightarrow Y$)

| $z$ | $d$ | $y$ | $\nu_{z,d,y}$ |
|---|---|---|---|
| 0 | 0 | 0 | 0.275 |
| 0 | 0 | 1 | 0.0 |
| 0 | 1 | 0 | 0.225 |
| 0 | 1 | 1 | 0.0 |
| 1 | 0 | 0 | 0.225 |
| 1 | 0 | 1 | 0.0 |
| 1 | 1 | 0 | 0.0 |
| 1 | 1 | 1 | 0.275 |

commonly used to express ignorance about the domain. The second prior is skewed to represent a dependency between the compliance and response behavior of the subjects. Figure 4 shows the distribution of ACE($D \rightarrow Y$) induced by these two prior distributions. Note that the skewed prior of Figure 4b assigns almost all the weight to negative values of ACE($D \rightarrow Y$) .

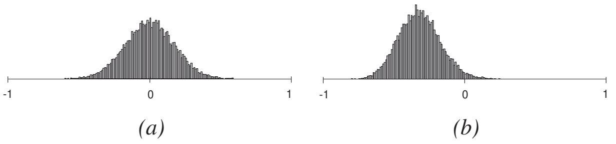



Figure 4: (a) The prior distribution of ACE($D \rightarrow Y$) induced by flat priors over the parameters $\nu_{CR}$, and (b) the distribution for ACE($D \rightarrow Y$) induced by skewed priors over the parameters.

In the following sections, we present the output of our system using (1) a simulated data set for which the causal effect is identifiable, (2) a real data set from an experiment designed to determine the effect of cholestyramine on reduced cholesterol level, and (3) a real data set from a study to determine the effect of vitamin A supplementation on childhood mortality.

## Simulated Data Example: Identifiable Causal Effect

As we noted in the introduction, Balke and Pearl (1994) have derived the tightest bounds for ACE($D \rightarrow Y$) under the large-sample assumption. They show that for some distributions of $Z$, $D$ and $Y$, the resulting upper and lower bounds collapse to a single point. We say that ACE($D \rightarrow Y$) is *identifiable* in this case. In this section, we show the output of our system when run on data sets derived from a distribution for which ACE($D \rightarrow Y$) is identifiable. One such distribution is shown in Table 1, yielding ACE($D \rightarrow Y$) $= 0.55$.

Figure 5 shows the the output of our system when applied to data sets of various sizes drawn from the distribution shown in Table 1, using both the flat and the skewed prior. As expected, as the number of cases increases, the posterior distributions become increasingly concentrated near the value 0.55. In general, because the skewed prior for ACE($D \rightarrow Y$) is concentrated further from 0.55 than the uniform prior, more cases are needed before the posterior distribution converges to the value 0.55.

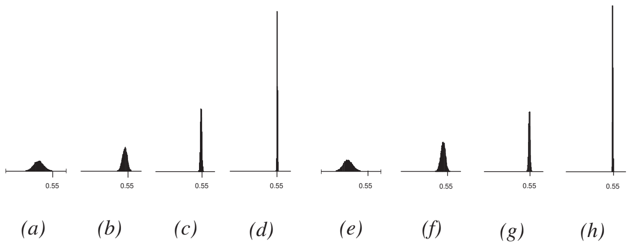



Figure 5: Output histograms for identifiable treatment effect using two priors. (a), (b), (c) and (d) show the posteriors for ACE($D \rightarrow Y$) using the flat prior and a data set consisting of 10, 100, 1000 and 10000 subjects, respectively. (e), (f), (g) and (h) show the posteriors for ACE($D \rightarrow Y$) using the skewed prior with the same respective data sets.

## Real Data Example: Effect of Cholestyramine on Reduced Cholesterol

Consider the Lipid Research Clinics Coronary Primary Prevention data described in [Lipid, 1984]. A portion of this data consisting of 337 subjects was analyzed by Efron and Feldman (1991) using a model that incorporates subject compliance as an explanatory variable; this same data set is the focus of this section.

A population of subjects was assembled and two preliminary cholesterol measurements were obtained: one prior to a suggested low-cholesterol diet and one following the diet period. The initial cholesterol level was taken as a weighted average of these two measures. The subjects were randomized into two groups: in the first group all subjects were prescribed cholestyramine ($z_1$), while the subjects in the other group were prescribed a placebo ($z_0$). During several years of treatment, each subject's cholesterol level was measured multiple times, and the average of these measurements was used as the post-treatment cholesterol level. The compliance of each subject was determined by tracking the quantity of prescribed dosage consumed.

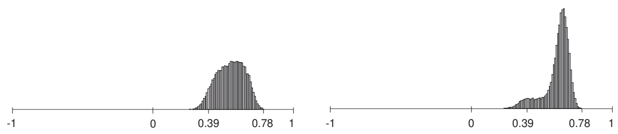
We transformed the (continuous) data from the Lipid study to the binary variables $D$ and $Y$ using the same approach as Balke and Pearl (1994). The resulting data set is shown in Table 2. Using the large-sample assumption, Balke and Pearl (1994)

Table 2: Observed data for the Lipid study and the Vitamin A study

| z | d | y | Lipid Study Observations | Vitamin A Study Observations |
|---|---|---|---|---|
| 0 | 0 | 0 | 158 | 11514 |
| 0 | 0 | 1 | 14 | 74 |
| 0 | 1 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 |
| 1 | 0 | 0 | 52 | 2385 |
| 1 | 0 | 1 | 12 | 34 |
| 1 | 1 | 0 | 23 | 9663 |
| 1 | 1 | 1 | 78 | 12 |

use the given data to derive the bounds $0.39 \leq \text{ACE}(D \rightarrow Y) \leq 0.78$.

Figure 6 shows posterior densities for $\text{ACE}(D \rightarrow Y)$ given the data. The density of Figure 6a corresponds to flat priors over the parameters and the density of Figure 6b corresponds to skewed priors over the parameters. Rather remarkably, even with only 337 cases in the data, both posterior distributions are highly concentrated within the large-sample bounds.



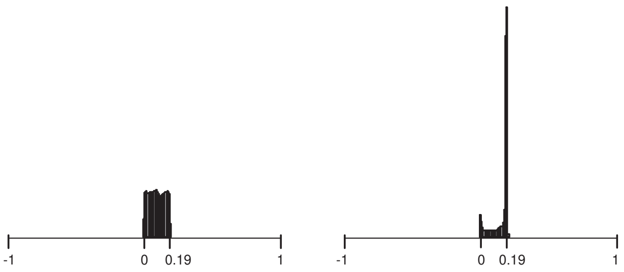

(a) (b)

Figure 6: Output histograms for the Lipid data. (a) Using flat priors and (b) using skewed priors.

## Real Data Example: Effect of Vitamin A Supplements on Child Mortality

In this section, we consider an experiment described by Sommer et al. (1986) designed to determine the impact of vitamin A supplementation on childhood mortality. In the study, 450 villages in northern Sumatra were randomly assigned to participate in a vitamin A supplementation scheme or serve as a control group for one year. Children in the treatment group received two large doses of vitamin A ($d_1$), while those in the control group received no treatment ($d_0$). After the year had expired, the number of deaths $y_0$ were counted for both groups. The results of the study are shown in Table 2.

Under the large-sample assumption, the method of Balke and Pearl (1994) yields the bounds: $-0.01 \leq \text{ACE}(D \rightarrow Y) \leq 0.19$. Figure 7 shows posterior densities for $\text{ACE}(D \rightarrow Y)$ given the data. The density of Figure 7a corresponds to flat priors over the parameters and the density of Figure 7b corresponds to skewed priors over the parameters.




*(a)* *(b)*

Figure 7: Output histograms for the Vitamin A Supplementation data. (a) Using flat priors and (b) using skewed priors.

It is interesting to note that for this study, the choice of the prior distribution has a significant effect on the posterior. This suggests that if the clinician is not very confident in the prior, a sensitivity analysis should be performed.

## A Counterfactual Query

In addition to assessing the average treatment effect, the system is also capable (with only minor modification) of answering a variety of counterfactual queries concerning individuals with specific characteristics. In this section, we show the result of our system when modified to answer the following query: What is the probability that Joe would have had an improved cholesterol reading had he taken cholestyramine, given that (1) Joe was in the control group of the Lipid study, (2) Joe took the placebo as prescribed, and (3) Joe's cholesterol level did not improve.

We can answer the above query by running the Gibbs' sampler on a model identical to that shown in Figure 3, except that the function $\text{ACE}(D \rightarrow Y)$ (Equation 4) is replaced by another function of $\nu_{CR}$, one that represents our query. If Joe was in the control group and took the placebo, that means that he is either a complier or a never-taker. Furthermore, because Joe's cholesterol level did not improve, Joe's response behavior is either never-recover or helped. Consequently, Joe must be a member of one of the following four compliance-response populations: $\{cr_{01}, cr_{02}, cr_{11}, cr_{12}\}$. Joe would have improved had he taken cholestyramine if his response behavior is either helped ($r_1$) or always-recover ($r_3$). It follows that the query of interest is captured by the function

$$f(\nu_{CR}) = \frac{\nu_{cr_{01}} + \nu_{cr_{11}}}{\nu_{cr_{01}} + \nu_{cr_{02}} + \nu_{cr_{11}} + \nu_{cr_{12}}}$$

Figure 8a and Figure 8b show the prior distribution over $f(\nu_{CR})$ that follows from the flat prior and the skewed prior, respectively. Figure 8c and Figure 8d show the posterior distribution $p(f(\nu_{CR}|\mathcal{D}))$ obtained by our system when run on the Lipid data, using the flat prior and the skewed prior, respectively. From the bounds of Balke and Pearl (1994), it follows that under the large-sample assumption, $0.51 \leq f(\nu_{CR}|\mathcal{D}) \leq 0.86$.

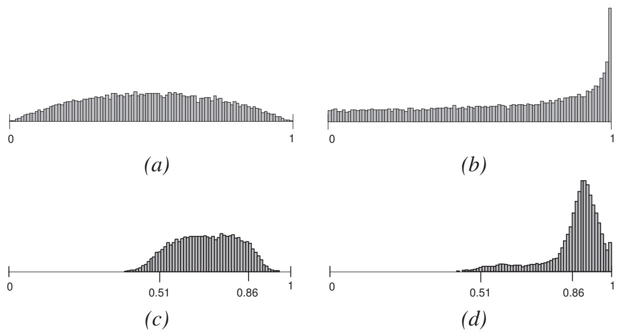



Figure 8: Prior (a, b) and posterior (c,d) distributions for a subpopulation $f(\nu_{CR}|\mathcal{D})$ specified by the counterfactual query "Would Joe have improved had he taken the drug, given that he did not improve without it". (a) corresponds to the flat prior, (b) to the skewed prior.

Thus, despite 39% non-compliance in the treatment group, and despite having just 337 subjects, the study strongly supports the conclusion that, given Joe's specific history, he would have been better off taking the drug. Moreover, the conclusion holds for both priors.

## Conclusion

This paper identifies and demonstrates a new application area for network-based inference techniques – the management of causal analysis in clinical experimentation. These techniques, which were originally developed for medical diagnosis, are shown capable of circumventing one of the major problems in clinical experiments – the assessment of treatment efficacy in the face of imperfect compliance. While standard diagnosis involves purely probabilistic inference in fully specified networks, causal analysis involves partially specified networks in which the links are given causal interpretation and where the domain of some variables are unknown.

The system presented in this paper provides the clinical research community, we believe for the first time, an assumption-free, unbiased assessment of the average treatment effect. We offer this system as a practical tool to be used whenever full compliance cannot be enforced and, more broadly, whenever the data available is insufficient for answering the queries of interest to the clinical investigator.

## References

Angrist, J.; Imbens, G.; and Rubin, D. 1993. Identification of causal effects using instrumental variables. Technical Report 136, Department of Economics, Harvard University, Cambridge, MA. Forthcoming, Journal of American Statistical Association, 1996.

Balke, A., and Pearl, J. 1994. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence,* Seattle, WA, 46–54. Morgan Kaufman.

Efron, B., and Feldman, D. 1991. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association* 86(413):9–26.

Holland, P. W. 1988. Causal inference, path analysis, and recursive structural equations models. In Clogg, C., ed., *Sociological Methodology.* Wachington, DC: American Socialogical Association. chapter 13, 449–484.

Imbens, G., and Rubin, D. 1994. Bayesian inference for causal effects in randomized experiments with noncompliance. Technical report, Harvard University.

Lipid Research Clinic Program. 1984. The lipid research clinics coronary primary prevention trial results, parts i and ii. *Journal of the American Medical Association* 251(3):351–374. January.

Pearl, J. 1994. From bayesian networks to causal network. In *Proceedings of the UNICOM Seminar on Adaptive Computing and Information Processing,* Brunel University, London, 165–194. Also in A. Gammerman (Ed.), Bayesian Networks and Probabilistic Reasoning, Alfred Walter Ltd., London, 1-31, 1995.

Pearl, J. 1995. Causal inference from indirect experiments. *Artificial Intelligence in Medicine Journal* 7(6):561–582.

Rubin, D. 1978. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* 7:34–58.

Sommer, A.; Tarwotjo, I.; Djunaedi, E.; West, K. P.; Loeden, A. A.; Tilden, R.; and Mele, L. 1986. Impact of vitamin A supplementation on childhood mortality: A randomized controlled community trial. *The Lancet* i:1169–1173.