The Annals of Statistics
1997, Vol. 25, No. 1, 305–327

BAYESIAN INFERENCE FOR CAUSAL EFFECTS IN RANDOMIZED EXPERIMENTS WITH NONCOMPLIANCE[1]

BY GUIDO W. IMBENS AND DONALD B. RUBIN

Harvard University

For most of this century, randomization has been a cornerstone of scientific experimentation, especially when dealing with humans as experimental units. In practice, however, noncompliance is relatively common with human subjects, complicating traditional theories of inference that require adherence to the random treatment assignment. In this paper we present Bayesian inferential methods for causal estimands in the presence of noncompliance, when the binary treatment assignment is random and hence ignorable, but the binary treatment received is not ignorable. We assume that both the treatment assigned and the treatment received are observed. We describe posterior estimation using EM and data augmentation algorithms. Also, we investigate the role of two assumptions often made in econometric instrumental variables analyses, the exclusion restriction and the monotonicity assumption, without which the likelihood functions generally have substantial regions of maxima. We apply our procedures to real and artificial data, thereby demonstrating the technology and showing that our new methods can yield valid inferences that differ in practically important ways from those based on previous methods for analysis in the presence of noncompliance, including intention-to-treat analyses and analyses based on econometric instrumental variables techniques. Finally, we perform a simulation to investigate the operating characteristics of the competing procedures in a simple setting, which indicates relatively dramatic improvements in frequency operating characteristics attainable using our Bayesian procedures.

1. Introduction. For most of this century, randomization has been a cornerstone of scientific experiments, especially those dealing with humans as experimental units. The theories of inference based on randomization, due to Fisher (1925) and Neyman (1923), reviewed and compared in Rubin (1990a) and extended to general observational studies in Rubin (1977), Rosenbaum and Rubin (1983) and Rosenbaum (1995), formally require that all experimental units adhere to their treatment assignments. In practice, however, noncompliance is relatively common in randomized experiments with human subjects. The standard approach to noncompliance, although sometimes sharply criticized [e.g., Salsburg (1994) and Sheiner and Rubin (1994)] is to rely on the same randomization distributions as if compliance had been perfect, and thus to compare average outcomes by assignment, ignoring information on com-

Received January 1995; revised April 1996.

[1]Supported by NSF Grants SBR-92-07456, DMS-94-04479 and SBR-95-11718.

AMS 1991 *subject classifications*. 62A10, 62A15, 62B15, 62C10, 62F15, 62K99, 62P99.

Key words and phrases. Intention-to-treat analysis, instrumental variables, EM algorithm, data augmentation, Gibbs sampler, likelihood-based inference, maximum likelihood estimation, Rubin causal model, compliers, exclusion restriction.

pliance behavior [Breslow (1982), Fisher, Dixon, Herson et al. (1990), Lee, Ellenberg, Hirtz and Nelson (1991) and Meier (1991)]; this is often referred to as an intention-to-treat analysis.

Here we use the "phenomenological Bayesian" approach of Rubin (1978a, b) to address the problem of noncompliance in randomized experiments. In the general approach, labeled the Rubin causal model (RCM) by Holland (1986) for work starting with Rubin (1974, 1975), causal inference problems are framed in terms of potentially observable outcomes, which are the responses of all units to all treatments. For the Bayesian, parametric models then serve as technical tools to generate posterior inferences for unobserved potential outcomes. Thus, Bayesian inference for causal effects involves the calculation of the posterior predictive distribution for responses to treatments not received, conditionally given responses to treatments received (and other observed quantities), which generates the posterior distribution of causal estimands—comparisons of these responses. Inferences across models with different parametric structures can be compared directly because these inferences are all driven by the posterior predictive distribution of the same causal estimands defined by the potentially observable outcomes.

Using this framework, the Bayesian is formally clear about the role played by the randomization of treatment assignment and the complications that arise from the nonrandom receipt of treatment due to noncompliance. Although the Bayesian never computes randomization distributions or design-based standard errors, the random assignment of treatment plays a critical role (Rubin, 1978a, 1990b), as does the compliance behavior of the units. The Bayesian approach also clarifies what can be learned in the noncompliance problem when causal estimands are intrinsically not fully "identified." In particular, issues of identification are quite different from those in the frequentist perspective because with proper prior distributions, posterior distributions are always proper. The effect of adding or dropping assumptions, such as those that are used in the instrumental variables literature [Bowden and Turkington (1984), Heckman and Robb (1984), Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996)], is directly addressed in the phenomenological Bayesian approach by examining how the posterior predictive distributions for causal estimands change.

This article is organized as follows. Section 2 defines the estimands of interest, and Section 3 describes the structure of Bayesian inference for these estimands in the presence of noncompliance. In Section 4 we present methods for posterior inference, both maximum likelihood estimation through EM [Dempster, Laird and Rubin (1977)] and simulation inference through data augmentation [Tanner and Wong (1987)]. In Section 5 we discuss the possible incorporation of two assumptions commonly invoked in econometric instrumental variables analyses, the exclusion restriction and the monotonicity assumption. Section 6 presents an analysis of a data set from Sommer and Zeger (1991) with a binary outcome, and in Section 7 we illustrate with artificial data how the analysis proceeds with continuous outcomes. Section 8 concludes with some discussion of extensions.

2. Causal estimands when confronted with noncompliance. Consider a hypothetical evaluation of the effect of a new drug (D) on some health outcome (Y) in a population of N units. Our objective is to estimate the effect of D (drug versus no drug) on Y, where we assume that the drug is either taken or not, thereby disallowing partial doses. The actual taking of the drug D is assumed to be beyond the control of the researcher. Instead, the researcher controls the assignment (i.e., the intention to treat), indicated by the variable Z; $Z_i = 1$ indicates that patient i is assigned to the treatment group, which is to receive the drug, whereas $Z_i = 0$ indicates that patient i is assigned to the control group, which is not to receive the drug. Let $\mathbf{Z}$ be the N component column vector of assignments with ith element Z_i. We make the stable unit treatment value assumption [SUTVA; Rubin (1980, 1990a)], which allows us to write the potential outcomes for unit i as a function of Z_i rather than the entire vector $\mathbf{Z}$. Thus, we let $D_i(z)$ be the binary indicator for the treatment that unit i actually would receive given the assignment z for $z = 0, 1$; $D_i(z) = 1$ indicates that unit i would take the drug if assigned z, and $D_i(z) = 0$ indicates that unit i would not take the drug if assigned z. In an ideal research environment, $D_i(z)$ would equal z for all i and z; that is, the treatment assigned would equal the treatment received for all units. In practice, $D_i(z)$ can differ from z for various reasons: individuals might accidentally receive the incorrect drug, or they might obtain the new drug despite being assigned to the control group [e.g., in AIDS trials; Robins (1989)], or individuals in the treatment group might fail to take the assigned drug because of disinterest or fear of potential side effects. Define $\underline{D_i} = (D_i(0), D_i(1))$ to be the row vector of potential treatment outcomes for unit i and $\underline{\mathbf{D}}$ to be the $N \times 2$ matrix with ith row equal to $\underline{D_i}$. We assume throughout that $D_i(Z_i)$, the treatment actually received, is observed.

Similar to the definition of $D_i(z)$, we define $Y_i(z, D_i(z))$ to be the outcome for unit i if exposed to treatment $D_i(z)$ after being assigned treatment z. The double-argument notation is, in principle, redundant because Y_i is actually a function of z alone, but is useful as will become apparent later. Define $\underline{Y_i} = (Y_i(0, D_i(0)), Y_i(1, D_i(1)))$ to be the row vector with the potential health outcomes for unit i under assignment to control and drug, and define $\underline{\mathbf{Y}}$ to be the $N \times 2$ matrix with ith row equal to $\underline{Y_i}$. We refer to $\underline{D_i}$ and $\underline{Y_i}$ as "potential outcomes," similar to Neyman's (1923) notion of "potential yields" in randomized agricultural experiments as discussed in Rubin (1990a).

The ITT (intention-to-treat) causal effect of Z on D for unit i is defined to be the difference $D_i(1) - D_i(0)$, and the ITT causal effect of Z on Y for unit i is $Y_i(1, D_i(1)) - Y_i(0, D_i(0))$. The average ITT causal effects are the averages of these unit-level causal effects over the population. The average effect of Z on D is $\mathrm{ITT}_D = \sum_{i=1}^{N}[D_i(1) - D_i(0)]/N$, and the average effect of Z on Y is $\mathrm{ITT}_Y = \sum_{i=1}^{N}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]/N$.

For unit i, $\underline{D_i}$ describes the compliance behavior. Because this is critical in our analysis, we use an indicator to partition the population of units into four types—compliers, never-takers, always-takers and defiers—based on their

compliance behavior. For unit i,

$$C_i = \begin{cases} c \text{ (i.e., unit } i \text{ is a complier)}, & \text{if } D_i(z) = z, \text{ for } z = 0, 1, \\ n \text{ (i.e., unit } i \text{ is a never-taker)}, & \text{if } D_i(z) = 0, \text{ for } z = 0, 1, \\ a \text{ (i.e., unit } i \text{ is an always-taker)}, & \text{if } D_i(z) = 1, \text{ for } z = 0, 1, \\ d \text{ (i.e., unit } i \text{ is a defier)}, & \text{if } D_i(z) = 1 - z, \text{ for } z = 0, 1. \end{cases}$$

We let $\mathscr{C}(t) = \{i | C_i = t\}$ for $t \in \{c, n, a, d\}$; $\mathbf{C}$ is the N component vector with ith element C_i, and N_t is the number of units of type t.

The population average ITT effects can therefore be decomposed as

$$\text{ITT}_D = \big(N_c 1 + N_n 0 + N_a 0 + N_d(-1)\big)/N = \big(N_c - N_d\big)/N$$

and

$$\text{ITT}_Y = \sum_{t \in \{c, n, a, d\}} N_t \, \text{ITT}_Y^{(t)}/N,$$

where, for $t \in \{c, n, a, d\}$,

$$\text{ITT}_Y^{(t)} = \sum_{i \in \mathscr{C}(t)} [Y_i(1, D_i(1)) - Y_i(0, D_i(0))]/N_t$$

is the average ITT effect of Z on Y for each of the four subpopulations defined by compliance behavior.

Of the four subpopulation ITT effects, two, $\text{ITT}_Y^{(n)}$ and $\text{ITT}_Y^{(a)}$, clearly do not address causal effects of the receipt of treatment because the former compares outcomes both with no drug, and the latter compares outcomes both with drug; neither never-takers nor always-takers can, at least in the context of this experiment, be induced to switch treatments. For compliers, assignment of treatment agrees with receipt of treatment, and $\text{ITT}_Y^{(c)}$ compares outcomes with drug to outcomes without drug. For such units it can, at least in some situations, be reasonable to attribute the effect on Y of assignment of treatment to the effect of receipt of treatment. This attribution is, in fact, what is typically done in randomized trials with full compliance. Even in that context, however, this attribution is not innocuous, and attempts to make it more plausible include the use of placebos and practices such as blinding and double blinding. For defiers, assignment to control leads to receipt of treatment and vice versa, and so $\text{ITT}_Y^{(d)}$ also compares outcomes with no drug to outcomes with drug. In some cases it may therefore be reasonable to attribute the effect of assignment to control versus treatment to the receipt of treatment versus control, although this may, in general, be less compelling than the attribution to receipt of treatment for compliers. To capture these attributions, we define the "attributed" causal effect of D on Y for a complier to be $Y_i(1, D_i(1)) - Y_i(0, D_i(0)) = Y_i(1, 1) - Y_i(0, 0)$, and for a defier to be $-(Y_i(1, D_i(1)) - Y_i(0, D_i(0))) = Y_i(0, 1) - Y_i(1, 0)$, with the "complier average causal effect" of D on Y denoted by CACE$= \text{ITT}_Y^{(c)}$, and the "defier average causal effect" denoted by DACE$= -\text{ITT}_Y^{(d)}$.

TABLE 1
Unit-level causal effects of assignment and treatment

Type of unit C_i	Potential treatment outcomes $D_i(0)$	$D_i(1)$	ITT causal effect of Z on D	ITT causal effect of Z on Y	"Attributed" causal effect of D on Y
c	0	1	1	$Y_i(1,1) - Y_i(0,0)$	$Y_i(1,1) - Y_i(0,0)$
n	0	0	0	$Y_i(1,0) - Y_i(0,0)$	—
a	1	1	0	$Y_i(1,1) - Y_i(0,1)$	—
d	1	0	−1	$Y_i(1,0) - Y_i(0,1)$	$Y_i(0,1) - Y_i(1,0)$

For an alternative motivation for the focus on complier and defier average causal effects, consider the analysis if full data on compliance behavior were actually available; that is, suppose that for all units C_i were observed. In that case, treating C_i as a covariate or pretreatment variable, the standard analysis of causal effects with strongly ignorable assignment given C_i [Rubin (1977) and Rosenbaum and Rubin (1983)] suggests discarding all units with either zero probability of receiving treatment (never-takers with $C_i = n$) or zero probability of receiving control (always-takers with $C_i = a$) and focusing solely on average effects for compliers and defiers.

We stress that although these arguments help to motivate interest in the ITT effects for compliers and defiers, they are not necessary for the statistical analyses that we present for the causal estimands $\mathrm{ITT}_Y^{(t)}$ for $t = c, n, a, d$.

Table 1 summarizes the definitions of the unit-level ITT effects and the attributed causal effects of treatment on the outcome for each type of unit defined by compliance behavior.

3. The structure of Bayesian inference for causal estimands. Five quantities are associated with each individual: Z_i, $D_i(0)$, $D_i(1)$, $Y_i(0, D_i(0))$ and $Y_i(1, D_i(1))$; a sixth quantity, the type C_i, is a function of $D_i(0)$ and $D_i(1)$. Three of these five quantities are observed: the treatment assigned, $Z_{\mathrm{obs},i} = Z_i$; the treatment received given the assigned treatment, $D_{\mathrm{obs},i} = D_i(Z_{\mathrm{obs},i})$; and the outcome under the assigned and received treatments, $Y_{\mathrm{obs},i} = Y_i(Z_{\mathrm{obs},i}, D_{\mathrm{obs},i})$. The two missing quantities are the treatment received under the other treatment assignment, $D_i(1 - Z_{\mathrm{obs},i})$, and the outcome under the other treatment assignment and the associated treatment received, $Y_i(1 - Z_{\mathrm{obs},i}, D_i(1 - Z_{\mathrm{obs},i}))$.

Bayesian inference considers the observed values of these quantities to be realizations of random variables and the unobserved values to be unobserved random variables. For the N units, the random variables in Bayesian inference are thus the N-vector $\mathbf{Z}$ and the two $N \times 2$ matrices $\underline{\mathbf{D}}$ and $\underline{\mathbf{Y}}$. Letting $f(\mathbf{Z}, \underline{\mathbf{D}}, \underline{\mathbf{Y}})$ be the joint probability (density) function of these random variables, we can write

$$f(\mathbf{Z}, \underline{\mathbf{D}}, \underline{\mathbf{Y}}) = f(\underline{\mathbf{D}}, \underline{\mathbf{Y}}|\mathbf{Z}) f(\mathbf{Z}) = f(\underline{\mathbf{D}}, \underline{\mathbf{Y}}) f(\mathbf{Z}), \tag{1}$$

where the second equality follows from the assumption of random assignment of $\mathbf{Z}$. Bayesian inference for the causal estimands, functions of $\underline{\mathbf{Y}}$ and $\underline{\mathbf{D}}$, follows from their joint posterior distribution, that is, their conditional distribution given observed values as derived from (1), as outlined in Rubin (1978a).

Assuming that all the potentially observable information about the units that will be modeled in this investigation is contained in the variables $(\mathbf{Z}, \underline{\mathbf{D}}, \underline{\mathbf{Y}})$, the distribution of $(\underline{\mathbf{D}}, \underline{\mathbf{Y}})$ is unit exchangeable, that is, invariant under a permutation of the unit indices. Therefore, appealing to deFinetti's theorem, we can assume, with essentially no loss of generality, that the rows of $(\underline{\mathbf{D}}, \underline{\mathbf{Y}})$ are independent and identically distributed random variables given a parameter vector π with prior distribution $p(\pi)$. Thus, we write

$$f(\underline{\mathbf{D}}, \underline{\mathbf{Y}}) = \int \prod_{i=1}^{N} f(\underline{D}_i, \underline{Y}_i|\pi) p(\pi)\, d\pi, \tag{2}$$

and the posterior distribution of π can be written as

$$\begin{aligned} p(\pi|\mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}}) &\propto p(\pi) \int\int \left[\prod_{i=1}^{N} f(\underline{D}_i, \underline{Y}_i|\pi) \right] d\mathbf{Y}_{\text{mis}}\, d\mathbf{D}_{\text{mis}} \\ &= p(\pi) \left[\prod_{i=1}^{N} \int\int f(\underline{D}_i, \underline{Y}_i|\pi)\, dY_{\text{mis},i}\, dD_{\text{mis},i} \right], \end{aligned} \tag{3}$$

where $\mathbf{D}_{\text{mis}}$ and $\mathbf{Y}_{\text{mis}}$ are the missing, or unobserved, components of $\underline{\mathbf{D}}$ and $\underline{\mathbf{Y}}$, respectively, and (3) is evaluated at the observed values $\mathbf{Z}_{\text{obs}}$, $\mathbf{D}_{\text{obs}}$ and $\mathbf{Y}_{\text{obs}}$; the constant of proportionality in (3) is the integral of the right-hand side over π.

For notational convenience, we factor the distribution $f(\underline{D}_i, \underline{Y}_i|\pi)$ into (i) the distribution of $\underline{D}_i$ given π, where the population probability of type t units is $\omega_t(\pi)$, and (ii) the conditional distribution of $\underline{Y}_i$ given $(\underline{D}_i, \pi)$, where we let $g_{tz}(y|\eta_{tz}(\pi))$ be the distribution of $Y_i(z, D_i(z))$ for units of type t for $z = 0, 1$ and $t = c, n, a, d$; this distribution depends on the parameter vector π through $\eta_{tz}(\pi)$. The complete parameter vector is $\pi = (\omega_c, \omega_n, \omega_a, \omega_d, \eta_{c0}, \eta_{c1}, \eta_{n0}, \eta_{n1}, \eta_{a0}, \eta_{a1}, \eta_{d0}, \eta_{d1}, \eta_{c01}, \eta_{n01}, \eta_{a01}, \eta_{d01})$, where the final four parameters, $\pi_{\text{assoc}} = (\eta_{c01}, \eta_{n01}, \eta_{a01}, \eta_{d01})$, refer to the association parameters in the joint distribution of outcomes for never-takers, always-takers, compliers and defiers, respectively. Letting $\delta(t, \underline{D}_i)$ equal 1 if $\underline{D}_i$ implies type t and 0 otherwise, we can write

$$\begin{aligned} f(\underline{D}_i, \underline{Y}_i|\pi) = &\sum_{t\in\{c,n,a,d\}} \delta(t, \underline{D}_i)\omega_t \\ &\times \sum_{t\in\{c,n,a,d\}} \delta(t, \underline{D}_i) g_{t0}(Y_i(0, D_i(0))|\eta_{t0}) \\ &\times g_{t1}(Y_i(1, D_i(1))|\eta_{t1}) h_t(\underline{Y}_i|\eta_{t01}), \end{aligned} \tag{4}$$

where $h_t(\underline{Y}_i|\eta_{t01})$ is defined such that the product of the last three factors is the joint distribution of $\underline{Y}_i$ given $\underline{D}_i$ and π.

To perform the integration in (3), we need to consider the structure of the missing data. There are four possible patterns of missing and observed data in $(\underline{D}_i, \underline{Y}_i)$, corresponding to the four possible values for $(Z_{\text{obs},i}, D_{\text{obs},i})$: $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$. Indicate the subsets of units exhibiting each pattern by $\mathcal{S}(0,0)$, $\mathcal{S}(0,1)$, $\mathcal{S}(1,0)$ and $\mathcal{S}(1,1)$ with cardinality N_{00}, N_{01}, N_{10} and N_{11}, respectively. In addition, let $\mathcal{S}(z,\cdot) = \mathcal{S}(z,0) \cup \mathcal{S}(z,1)$ be the set of units i with $Z_{\text{obs},i} = z$ for $z = 0, 1$. For $i \in \mathcal{S}(0,0)$, both $D_i(1)$ and $Y_i(1, D_i(1))$ are missing; the integration over $D_i(1)$ eliminates terms in (4) with $t = a$ or $t = d$, and the integration over $Y_i(1, D_i(1))$ eliminates the factors in the remaining terms in (4) involving g_{t1} and h_t with $t = c$ or n. That is, for $i \in \mathcal{S}(0,0)$, the units are generally a mixture of compliers and never-takers, and the observations $Y_{\text{obs},i}$ are either from g_{n0} or from g_{c0}. Hence, for $i \in \mathcal{S}(0,0)$ we obtain

$$\int\int f(\underline{D}_i, \underline{Y}_i|\pi)\, dY_{\text{mis},i}\, dD_{\text{mis},i} = \omega_c g_{c0}(Y_{\text{obs},i}|\eta_{c0}) + \omega_n g_{n0}(Y_{\text{obs},i}|\eta_{n0}).$$

Analogous expressions hold for units in $\mathcal{S}(0,1)$, $\mathcal{S}(0,1)$ and $\mathcal{S}(1,1)$.

Letting $g^i_{tz} = g_{tz}(Y_{\text{obs},i}|\eta_{tz}(\pi))$ for $t = c, n, a, d$ and $z = 0, 1$, from (3) and (4) the posterior distribution of π is thus

$$\begin{aligned} p(\pi|\mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}}) &\propto p(\pi) \qquad (5)\\ &\times \prod_{i\in\mathcal{S}(0,0)} (\omega_c g^i_{c0} + \omega_n g^i_{n0}) \prod_{i\in\mathcal{S}(0,1)} (\omega_a g^i_{a0} + \omega_d g^i_{d0}) \qquad (6)\\ &\times \prod_{i\in\mathcal{S}(1,0)} (\omega_n g^i_{n1} + \omega_d g^i_{d1}) \prod_{i\in\mathcal{S}(1,1)} (\omega_c g^i_{c1} + \omega_a g^i_{a1}). \qquad (7)\end{aligned}$$

Because the association parameter π_{assoc} does not enter the likelihood function, specified by the four products (6)–(7), the posterior distribution of π_{assoc} equals its prior distribution if π_{assoc} is a priori independent of the other components of π, as we assume in the remainder of this article.

One function of π that is free of π_{assoc} and is particularly relevant is the superpopulation CACE

$$\int y g_{c1}(y|\eta_{c1})\, dy - \int y g_{c0}(y|\eta_{c0})\, dy, \qquad (8)$$

which is the average causal effect of D on Y in the hypothetical superpopulation of compliers from which the N_c complying units in the current experiment can be conceptualized as having been drawn. Analogously, we can define the superpopulation DACE and other superpopulation causal estimands such as the ITT effects of Z on Y for never-takers and always-takers, $\text{ITT}_Y^{(n)}$ and $\text{ITT}_Y^{(a)}$; none of these involves π_{assoc}. Inference for the finite population causal estimands for the units in the study (e.g., CACE) follows from the posterior distribution of π by predictive Bayesian inference; these generally do involve π_{assoc} [see Rubin (1990a), Section 7, for a specific example]. Henceforth, we

focus on the superpopulation estimands and so ignore π_{assoc} and let π denote the remaining parameters.

4. Computation of the posterior distribution of causal estimands. The superpopulation causal estimands CACE, $\text{ITT}_Y^{(n)}$, $\text{ITT}_Y^{(a)}$ and DACE are functions of pairs of parameters (η_{c0}, η_{c1}), (η_{n0}, η_{n1}), (η_{a0}, η_{a1}) and (η_{d0}, η_{d1}), respectively, whose estimation is complicated by the fact that they can only be indirectly estimated through the observation of mixtures of the distributions [cf. expressions (6) and (7)], which involve four proportions, ω_c, ω_n, ω_a and ω_d, and eight distributions of Y_i: $g_{tz}(y|\eta_{tz})$ for $t = c, n, a, d$ and $z = 0, 1$. Modern methods of computational statistics, however, including methods for iterative maximization such as the EM [Dempster, Laird and Rubin (1977)], ECM [Meng and Rubin (1991, 1993)] and ECME [Liu and Rubin (1994)] algorithms and methods for iterative simulation including Markov chain Monte Carlo methods such as data augmentation [DA; Tanner and Wong (1987)] and the Gibbs sampler [e.g., Geman and Geman (1984), Gelfand, Hills, Racine-Poon and Smith (1990) and Gelman and Rubin (1992)], can make inference relatively straightforward. These methods are advantageous because they exploit the fact that with C_i known for all units, inference for each of the four causal estimands would involve only the data from its associated subpopulation with no mixture components.

We now outline the general structure of the EM and DA algorithms for our problem, by first considering the imputation of $\mathbf{C}$ and second discussing the "complete-data" analysis, where "complete-data" means complete *compliance* data, that is, $(\mathbf{C}, \mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}})$, or, equivalently, $(\mathbf{Z}_{\text{obs}}, \mathbf{D}, \mathbf{Y}_{\text{obs}})$, but *not* the full set of potential outcomes $(\mathbf{Z}_{\text{obs}}, \mathbf{D}, \mathbf{Y})$.

The first step of both an EM and a DA iteration involves imputation of $\mathbf{C}$, either its conditional expectation with EM or a draw from its conditional distribution with DA, both given observed values and a current value of π. Given $(\mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}}, \pi)$, the C_i are independent indicators of type t dependent on the data only through $(Z_{\text{obs},i}, D_{\text{obs},i}, Y_{\text{obs},i})$. Table 2 presents the conditional probabilities for each type $t = c, n, a, d$ given π and the observed data. Sam-

TABLE 2

$\Pr(C_i = t|Z_{\text{obs},i}, D_{\text{obs},i}, Y_{\text{obs},i}, \pi)$, *conditional probability of subject* i *being type* t *given observed data* $Z_{\text{obs},i}$, $D_{\text{obs},i}$, $Y_{\text{obs},i}$, *and parameters* π*: numerator is table entry and denominator is row total*

		Subject type t				
$Z_{\text{obs},i}$	$D_{\text{obs},i}$	$t = c$	$t = n$	$t = a$	$t = d$	**Row total**
0	0	$\omega_c g_{c0}^i$	$\omega_n g_{n0}^i$	0	0	$\omega_c g_{c0}^i + \omega_n g_{n0}^i$
0	1	0	0	$\omega_a g_{a0}^i$	$\omega_d g_{d0}^i$	$\omega_a g_{a0}^i + \omega_d g_{d0}^i$
1	0	0	$\omega_n g_{n1}^i$	0	$\omega_d g_{d1}^i$	$\omega_n g_{n1}^i + \omega_d g_{d1}^i$
1	1	$\omega_c g_{c1}^i$	0	$\omega_a g_{a1}^i$	0	$\omega_c g_{c1}^i + \omega_a g_{a1}^i$

pling from the distribution of $\mathbf{C}$ given $(\mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}}, \pi)$ for DA therefore only involves independent drawing from binomial distributions. Similarly, the E-step in EM simply replaces each C_i by its expectation represented as a four-component indicator of probabilities.

The second step of both an EM and a DA iteration involves an analysis of the complete-data posterior distribution of π. The conditional posterior distribution of π given $(\mathbf{C}, \mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}})$ has a much simpler structure than the actual posterior distribution of π given $(\mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}})$, because the parameters of the eight outcome distributions, $(\eta_{c0}, \eta_{c1}, \eta_{n0}, \eta_{n1}, \eta_{a0}, \eta_{a1}, \eta_{d0}, \eta_{d1})$, all appear in separate factors of the likelihood, each with an i.i.d. structure:

$$p(\pi|\mathbf{C}, \mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}}) \propto p(\pi) \prod_{z=0,1} \prod_{t\in\{c,n,a,d\}} \left[\prod_{i\in(\mathscr{C}(t)\cap\mathscr{S}(z,\cdot))} \omega_t g^i_{tz} \right], \tag{9}$$

where, as before, $g^i_{tz} = g_{tz}(Y_{\text{obs},i}|\eta_{tz})$. To capitalize on this structure, assume prior joint independence of $(\omega_c, \omega_n, \omega_a, \omega_d)$, η_{c0}, η_{c1}, η_{n0}, η_{n1}, η_{a0}, η_{a1}, η_{d0}, η_{d1}, so that the prior distribution of π and the posterior distribution of π given $\mathbf{C}$ both factor into the nine components, one for the quadrinomial probabilities of type:

$$p(\omega_c, \omega_n, \omega_a, \omega_d|\mathbf{C}, \mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}}) \propto p(\omega_c, \omega_n, \omega_a, \omega_d)\omega_c^{N_c}\omega_n^{N_n}\omega_a^{N_a}\omega_d^{N_d},$$

and one for each of the eight outcome distributions, defined by assignment and compliance status;

$$p(\eta_{tz}|\mathbf{C}, \mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}}) \propto p(\eta_{tz}) \prod_{i\in(\mathscr{C}(t)\cap\mathscr{S}(z,\cdot))} g^i_{tz} \tag{10}$$

for $z = 0, 1$ and $t = c, n, a, d$. Evaluating the complete-data posterior distribution for η_{tz} at most involves integrals with dimension equal to the dimension of η_{tz}. For common choices of $g_{tz}(\cdot)$, such as the binomial distribution for binary outcomes used in the example in Section 6 or the normal distribution for continuous outcomes used in the example in Section 7, analysis of these complete-data posterior distributions is direct for conventional choices of prior distributions. The analysis of the posterior distribution of $(\omega_n, \omega_a, \omega_c, \omega_d)$ is straightforward when either finding its mode or drawing from it with the conventional conjugate Dirichlet prior distribution.

5. Two commonly invoked assumptions. Here we formulate two assumptions that, although not necessary to apply our analysis, are often plausible and facilitate inference for CACE. These assumptions, discussed in detail in Angrist, Imbens and Rubin (1996), connect our work to that on *instrumental variables* (with the random assignment indicator interpreted as the *instrumental variable*), which has a long tradition in econometrics, dating back to Wright (1928, 1934) and Reiersol (1941), reviewed in Bowden and Turkington (1984) and applied to treatment evaluations in Heckman and Robb (1985). A recent influential application where the posited instrument is explicitly ran-

domized is given in Angrist (1990). In the last two decades, empirical studies using similar assumptions have also appeared in other literatures, often in the context of randomized experiments with noncompliance [Zelen (1979, 1990), Bloom (1984), Holland (1988), Hearst, Newman and Hulley (1986), Permutt and Hebel (1989), Robins (1989), Sommer and Zeger (1991), Balke and Pearl (1994), Baker and Lindeman (1994) and McClellan and Newhouse (1994)].

The first assumption, the *weak exclusion restriction*, requires that the treatment assignment is unrelated to potential outcomes for never-takers and always-takers: for all i such that $D_i(0) = D_i(1)$, $Y_i(0, D_i(0)) = Y_i(1, D_i(1))$; that is, if for unit i, treatment assignment Z_i has no effect on treatment status D_i, it has no effect on outcome Y_i either, so that $\text{ITT}_Y^{(n)} = \text{ITT}_Y^{(a)} = 0$. In our distributional notation, $g_{n0}(y|\eta_{n0}) = g_{n1}(y|\eta_{n1})$ and $\eta_{n0} = \eta_{n1}$, and similarly $g_{a0}(y|\eta_{a0}) = g_{a1}(y|\eta_{a1})$ and $\eta_{a0} = \eta_{a1}$, implying $g_{n0}^i = g_{n1}^i$ and $g_{a0}^i = g_{a1}^i$. The only part of EM or the Gibbs sampler that is affected is the complete-data analysis, where the parameters of the two always-taker and the two never-taker distributions are now the same, respectively; the product in (10) then becomes $\prod_{i\in\mathscr{C}(t)} g_t^i$. A stronger version of the exclusion restriction appearing in Imbens and Rubin (1994) and Angrist, Imbens and Rubin (1996) also asserts that the unit-level effect of $Z = 1$ versus $Z = 0$ for compliers is solely due to the exposure to $D = 1$ versus $D = 0$, and asserts that for defiers the unit-level effect of $Z = 1$ versus $Z = 0$ is solely due to the exposure to $D = 0$ versus $D = 1$. This restriction therefore combines the causal attribution in the last column of Table 1 with the weak exclusion restriction for never-takers and always-takers. Versions of the exclusion restriction underlie all instrumental variables inferences in econometrics, although typically stated using formulations involving regression function disturbances that link the basic exclusion restriction with functional form and independence assumptions [e.g., Heckman and Robb (1985)].

The second assumption, *strict monotonicity*, restricts the patterns of compliance behavior in the population. Strict monotonicity of treatment assignment on treatment received requires $D_i(1) \geq D_i(0)$ for all $i = 1, \ldots, N$, with inequality for at least one unit i. This assumption rules out the presence of defiers and requires the presence of compliers; that is, ignoring measure theoretic details, it requires $\omega_d = 0$ and $\omega_c > 0$. It is called *strict monotonicity* as it combines Assumptions 4 and 5 of Angrist, Imbens and Rubin (1996), which separate the basic monotonicity assumption $D_i(1) \geq D_i(0)$ from the condition that $D_i(1) \neq D_i(0)$ for at least one unit. Balke and Pearl (1994) call this the *no-defiance* assumption, as it rules out the existence of defiers. Although more widely applicable [see, e.g., the discussion in Angrist, Imbens and Rubin (1996)], monotonicity is especially plausible in the context of a randomized trial of a drug whose access is restricted to units assigned to take it, implying $D_i(0) = 0$ for all i, as in the application in Section 6 and in Zelen (1979) and Bloom (1984). Given monotonicity, there are no defiers, $\mathscr{C}(d)$ is empty, and consequently the two distributions g_{d0} and g_{d1} are irrelevant, and DACE is not defined. The correct Bayesian analysis follows from (5)–(10) with $\omega_d = 0$ and $\mathscr{C}(d)$ empty.

Given the four assumptions (SUTVA, random assignment, the weak exclusion restriction and strict monotonicity), the assignment indicator is called an instrumental variable in the econometric literature, and in this case there is a simple relation between CACE and the two simple ITT effects: CACE=$\text{ITT}_Y/\text{ITT}_D$. This relation between the population ITT estimands and CACE under the four assumptions, and its role in econometric instrumental variables procedures, is the focus of Angrist, Imbens and Rubin (1996), which also discusses alterations in this relation due to violations of these assumptions. An important advantage of our Bayesian analysis is that neither the exclusion restriction nor the monotonicity assumption is essential, and consequently violations of these assumptions are easily addressed, as illustrated in the next section.

6. An application with real data and binary outcomes. In this section we apply our analysis to data from a randomized community trial of the impact of vitamin A supplements on children's survival. The data set is the same as in Sommer and Zeger (1991) and is displayed in Table 3. In this trial, villages in Indonesia were randomly assigned to receive or not to receive vitamin supplements. Although no subjects from the villages assigned not to receive the supplements in fact received them, a number of subjects from villages assigned to receive the supplements did not receive them. In our notation, $D_i(0) = 0$ but $D_i(1) = 0$ or 1. Monotonicity is therefore satisfied and there are only compliers and never-takers. Although taking account of the clustering resulting from randomization at the village rather than the individual level is straightforward using hierarchical extensions of our basic model, because indicators for the village are not available to us we do not model the dependence between individuals from the same village. The outcome is binary—death is 0 and survival is 1.

With binary outcomes and no defiers or always-takers, there are five relevant scalar parameters in total (ignoring, as discussed in Section 3, the association parameters): four probabilities for the outcome distributions $(\eta_{c0}, \eta_{c1}, \eta_{n0}, \eta_{n1})$ and one probability for the distribution of the type indicator, $\omega = \omega_c = \Pr(C_i = c|\pi)$; the probability of never-takers is $\omega_n = 1 - \omega$.

TABLE 3
Sommer–Zeger vitamin supplement data

Type	Assignment $Z_{\text{obs},i}$	Vitamin supplements $D_{\text{obs},i}$	Survival $Y_{\text{obs},i}$	Number of units (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

The estimand of primary interest is the superpopulation complier average causal effect $\eta_{c1} - \eta_{c0}$, but we shall also be interested in the superpopulation analogue of $\mathrm{ITT}_Y^{(n)}$, $\eta_{n1} - \eta_{n0}$.

Because $D_i(0) = 0$ for all i, implying $\omega_d = \omega_a = 0$, the posterior distribution of π in (5)–(7) becomes

$$
\begin{aligned}
&p(\pi|\mathbf{Z}_{\mathrm{obs}}, \mathbf{D}_{\mathrm{obs}}, \mathbf{Y}_{\mathrm{obs}}) \\
&\quad \propto p(\pi) \prod_{i \in \mathscr{S}(1,0)} \omega_n g_{n1}^i \prod_{i \in \mathscr{S}(1,1)} \omega_c g_{c1}^i \prod_{i \in \mathscr{S}(0,0)} \left(\omega_c g_{c0}^i + \omega_n g_{n0}^i\right).
\end{aligned} \tag{11}
$$

Assuming prior independence of the parameters, the complete-data posterior distribution of π, given in (9) for the general case, can be written as the product of five distributions, one for ω,

$$
p(\omega|\mathbf{C}, \mathbf{Z}_{\mathrm{obs}}, \mathbf{D}_{\mathrm{obs}}, \mathbf{Y}_{\mathrm{obs}}) \propto p(\omega)\omega^{N_c}(1-\omega)^{N_n}, \tag{12}
$$

and four for the distributions indexed by $t = n, c$ and $z = 0, 1$,

$$
p(\eta_{tz}|\mathbf{C}, \mathbf{Z}_{\mathrm{obs}}, \mathbf{D}_{\mathrm{obs}}, \mathbf{Y}_{\mathrm{obs}}) \propto p(\eta_{tz}) \prod_{(\mathscr{C}(t) \cap \mathscr{S}(z,\cdot))} \eta_{tz}^{Y_{\mathrm{obs},i}} (1-\eta_{tz})^{(1-Y_{\mathrm{obs},i})}, \tag{13}
$$

where we substituted $\eta_{tz}^{Y_{\mathrm{obs},i}}(1-\eta_{tz})^{(1-Y_{\mathrm{obs},i})}$ for g_{tz}^i. All five conditional posterior distributions are easy to draw from for conjugate Beta prior distributions. The conditional type probabilities given observed variables are, simplifying the results in Table 2 to reflect the absence of both defiers and always-takers,

$$
\begin{aligned}
&\Pr(C_i = c|Z_{\mathrm{obs},i}, D_{\mathrm{obs},i}, Y_{\mathrm{obs},i}, \pi) \\
&\qquad = \begin{cases} 0, & \text{if } i \in \mathscr{S}(1,0), \\ 1, & \text{if } i \in \mathscr{S}(1,1), \\ \omega_c g_{c0}^i/(\omega_c g_{c0}^i + \omega_n g_{n0}^i), & \text{if } i \in \mathscr{S}(0,0), \end{cases}
\end{aligned}
$$

and $\Pr(C_i = n|Z_{\mathrm{obs},i}, D_{\mathrm{obs},i}, Y_{\mathrm{obs},i}, \pi) = 1 - \Pr(C_i = c|Z_{\mathrm{obs},i}, D_{\mathrm{obs},i}, Y_{\mathrm{obs},i}, \pi)$.

First, we calculate the maximum likelihood estimate (MLE) of π and $\mathrm{CACE} = \eta_{c1} - \eta_{c0}$ for the data in Table 1 without imposing the exclusion restriction. The likelihood function is maximized at $\omega = (9663+12)/(9663+12+2385+34) = 0.800$, $\eta_{n1} = 2385/(34+2385) = 0.986$, $\eta_{c1} = 9663/(12+9663) = 0.999$, $\eta_{c0} \in [0.800 \times 11514 - 0.200 \times 74)/(0.800 \times (74+11514)), 1] = [0.992, 1]$ and $\eta_{n0} \in [(0.200 \times 11514 - 0.800 \times 74)/(0.200 \times (74+11514)), 1] = [0.968, 1]$. There is no unique solution for η_{n0} and η_{c0}, but rather a region of values at which the likelihood function is maximized. The MLE for $\eta_{c1} - \eta_{c0} = [-0.001, 0.007]$.

Second, in Figure 1 we approximate the posterior distribution for the CACE, with uniform prior distributions over the legitimate parameter spaces. The vertical lines in this figure indicate the 90% interval $(-0.0009, 0.0070)$ based on the histogram estimate of the exact posterior distribution obtained using 1000 iterations from each of 20 independent runs of the DA algorithm, with the first 500 iterations discarded. The starting values for all parameters for

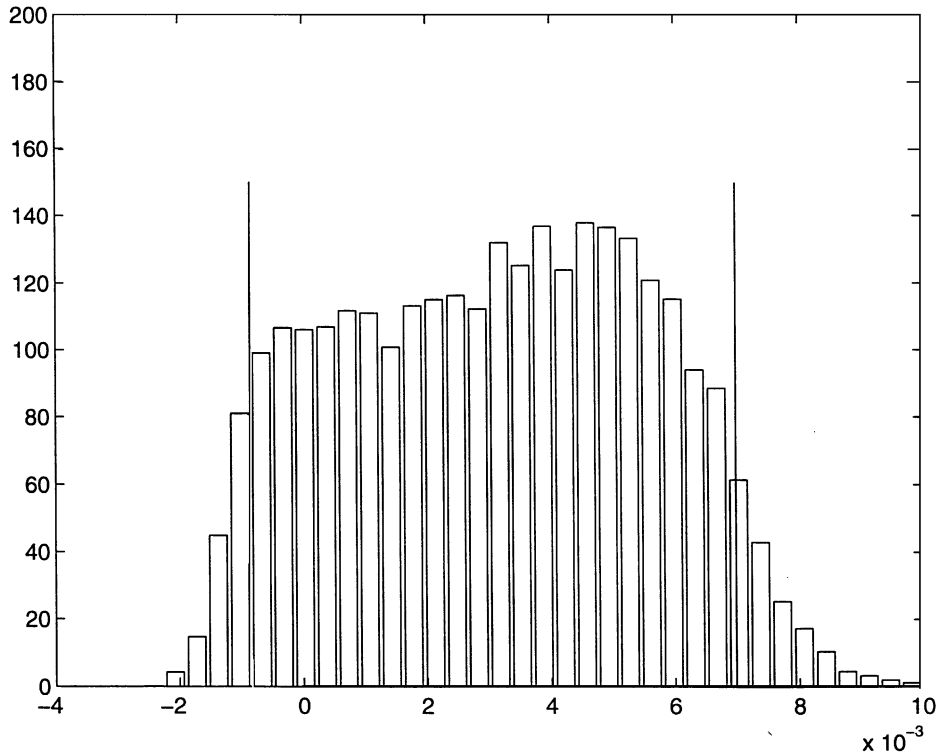

FIG. 1. *Histogram of CACE without exclusion restriction* (*data from Table* 3).

each run were drawn from uniform distributions over the appropriate parameter space. The analysis and inspection of the iterations were based on the Gelman–Rubin (1992) criteria for convergence. The shape of the posterior distribution suggests that the CACE is likely to be in the range -0.0012 to 0.0071, but that the data are not informative about the relative likelihood of values within that range, which is expected because the MLE of CACE is the interval $[-0.001, 0.007]$. In Figure 2 we also approximate the posterior distribution for $\text{ITT}_Y^{(n)}$.

Third, we impose the exclusion restriction, which requires that $\eta_n = \eta_{n0} = \eta_{n1}$. The only part of the Gibbs sampler that is affected is the replacement of the two conditional distributions for never-takers by the single distribution

$$p(\eta_n|\mathbf{C}, \mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}}) \propto p(\eta_n) \prod_{i \in \mathscr{C}(n)} \eta_n^{Y_{\text{obs},i}} (1-\eta_n)^{(1-Y_{\text{obs},i})}. \tag{14}$$

The posterior distribution given the exclusion restriction is approximated by the histogram in Figure 3. Now the posterior distribution is not only much tighter but also well approximated by a normal distribution. In this case the unique MLE for CACE is 0.0032 (3.2 per 1000), and the 90% interval is $(0.0012, 0.0051)$. For comparison purposes, mortality in the entire sample

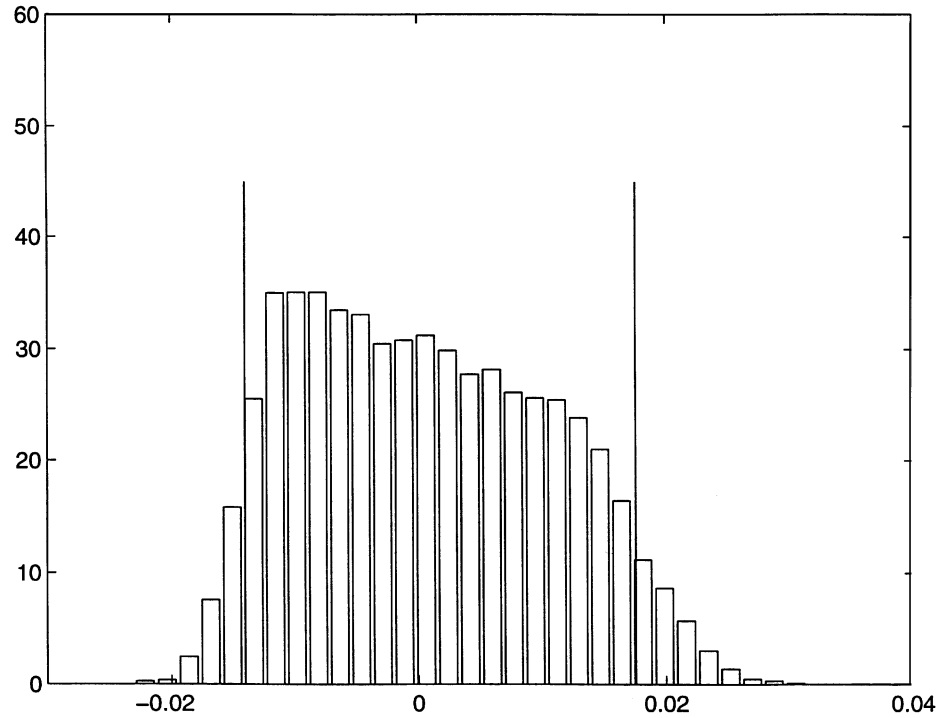

FIG. 2. *Histogram of* $\mathrm{ITT}_Y^{(n)}$ *without exclusion restriction* (*data from Table* 3).

is 5.1 per 1000, and the posterior mean of mortality for compliers with and without vitamin A is 1.2 and 4.5 per 1000, respectively, implying that the point estimates under the exclusion restriction suggest a reduction in mortality of 65%. The solid line in this figure represents the normal approximation to the posterior distribution around the MLE using the information matrix as the basis of the variance. Sommer and Zeger call CACE in this example the "attributable risk."

Table 4 summarizes the results for CACE with and without the exclusion restriction. We also present the posterior distribution for the superpopulation $\mathrm{ITT}_Y^{(n)}$, the average ITT effect of Z on Y for never-takers. Under the exclusion restriction, $\mathrm{ITT}_Y^{(n)}$ is forced to be equal to 0, but without the exclusion restriction, $\mathrm{ITT}_Y^{(n)}$ has a nondegenerate posterior distribution, which turns out to be centered around 0 for these data, lending credibility to the exclusion restriction. Moreover, $\mathrm{ITT}_Y^{(n)}$ has a joint posterior distribution with CACE, displayed in Figure 4, which suggests that, even without the exclusion restriction, in order to believe that *receipt* of vitamin A has a *negative* effect on survival for compliers, we must believe that *assignment* to receive vitamin A must have a strong *positive* effect for never-takers. This combination of hypotheses appears implausible, suggesting that even without making the exclusion re-

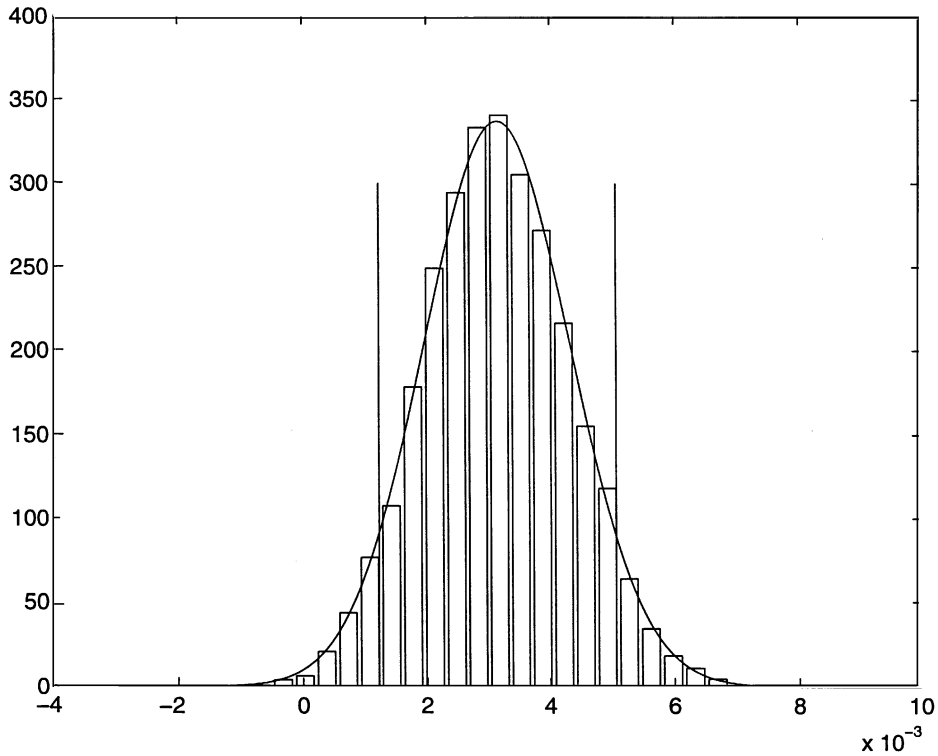

FIG. 3. *Histogram of CACE with exclusion restriction* (*data from Table* 3).

striction, one can be confident that receipt of vitamin A has a positive effect on survival.

Following a suggestion by a referee, we calculated for comparison purposes the bounds on the population average causal effect of *receipt* of treatment proposed by Robins (1989), Manski (1990) and Balke and Pearl (1994), which are obtained by letting the outcomes for never-takers given receipt of vitamin A range from "all survived" to "all died." For the Sommer–Zeger data, given

TABLE 4

Posterior distribution for superpopulation CACE and $\mathrm{ITT}_Y^{(n)}$ *for Sommer–Zeger data set: increase in survival rates per* 1000 *units* (*overall survival rate in sample is* 994.9 *per* 1000, *i.e.*, *mortality rate is* 5.1 *per* 1000)

Estimand	Exclusion restriction	Mean	Standard deviation	Median	5th percentile	95th percentile
CACE	No	3.1	2.5	3.2	−0.9	7.0
$\mathrm{ITT}_Y^{(n)}$	No	0.5	10.1	0.2	−14.1	17.5
CACE	Yes	3.1	1.2	3.1	1.2	5.1

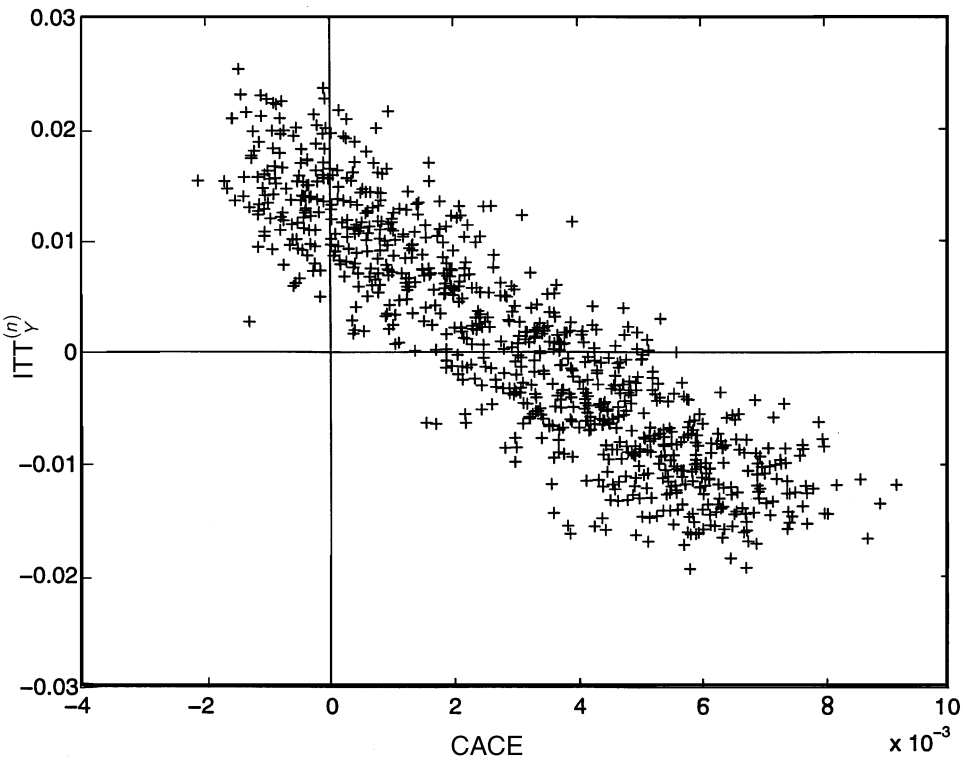

FIG. 4. *Joint posterior distribution of CACE and* $\text{ITT}_Y^{(n)}$ (*data from Table* 3).

SUTVA, random assignment, the exclusion restriction and monotonicity, point estimates of the Robins–Manski–Balke–Pearl lower and upper bounds on the effect of vitamin A on survival rates are -0.1946 and 0.0054, respectively, implying that administering vitamin A to the entire population could reduce mortality by 5.4 per 1000 or increase it by as much as 194.6 per 1000. Under exactly the same assumptions, our point estimate for the average effect for compliers, given in Table 4, is a 3.1 per 1000 reduction in mortality, with a 90% posterior interval of (1.2, 5.1). The reason for the enormous width of the range encompassed by these bounds, and for its centering at an extremely high level of mortality, is the focus on the population rather than the complier average combined with the lack of restrictions on mortality for the approximate 20% of the population who are estimated to be never-takers in this experiment.

7. An artificial example with continuous outcomes. We now consider the case with continuous outcomes, assuming normal distributions with density indicated by $\phi(\cdot)$. The approach, however, is as described in Section 4: we use EM and DA to capitalize on the straightforward complete-data analysis. In this section we focus on the properties of the posterior distribution under both monotonicity and the exclusion restriction, thereby assuming $\omega_d = 0$ and

$g_n^i = g_{n0}^i = g_{n1}^i$ and $g_a^i = g_{a0}^i = g_{a1}^i$. These assumptions, as well as normality, are assumed to hold in the population as well as in the model used for analyzing the sample.

The parameter π is, given monotonicity and the exclusion restriction, $(\omega_c, \omega_n, \omega_a, \eta_{c0\mu}, \eta_{c0\sigma}, \eta_{c1\mu}, \eta_{c1\sigma}, \eta_{n\mu}, \eta_{n\sigma}, \eta_{a\mu}, \eta_{a\sigma})$, where the subscripts μ and σ indicate the mean and standard deviation, respectively. The posterior distribution of π is as in (5)–(7) with $g_{cz}^i = \phi((Y_{\text{obs},\,i} - \eta_{cz\mu})/\eta_{cz\sigma})/\eta_{cz\sigma}$ for $z = 0, 1$, and $g_t^i = \phi((Y_{\text{obs},\,i} - \eta_{t\mu})/\eta_{t\sigma})/\eta_{t\sigma}$ for $t = n, a$; and $\omega_d = 0$. For computation, we turn to the complete-data posterior distribution of π, which, assuming appropriate prior independence of the parameters, can be written as the product of five factors:

$$p(\omega_c, \omega_n, \omega_a|\mathbf{C}, \mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}}) \propto p(\omega_c, \omega_n, \omega_a)\omega_c^{N_c}\omega_n^{N_n}\omega_a^{N_a},$$

$$p(\eta_{cz}|\mathbf{C}, \mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}}) \propto p(\eta_{cz}) \prod_{i \in \mathscr{C}(c) \cap \mathscr{S}(z,\cdot)} g_{cz}^i \quad \text{for } z = 0, 1,$$

$$p(\eta_t|\mathbf{C}, \mathbf{Z}_{\text{obs}}, \mathbf{D}_{\text{obs}}, \mathbf{Y}_{\text{obs}}) \propto p(\eta_t) \prod_{i \in \mathscr{C}(t)} g_t^i \quad \text{for } t = n, a.$$

All five posterior distributions are easy to draw from for conventional conjugate prior distributions because they involve only Dirichlet, normal and inverse chi-squared posterior distributions. Also in this normally distributed outcome case, the components of the conditional distribution of $\mathbf{C}$ given $\mathbf{Z}_{\text{obs}}$, $\mathbf{D}_{\text{obs}}$ and $\mathbf{Y}_{\text{obs}}$ are easy to draw from.

We illustrate this analysis in two ways. First, we analyze a specific data set from a known population. In Table 5 we give the underlying parameters of a hypothetical infinite population where CACE is the difference in means of the two complier distributions, $0.9 - 0.1 = 0.8$. We drew a single data set containing 100 observations, generated according to the joint distribution in Table 5, subject to a completely randomized design, 50 observations with $Z_i = 1$ and 50 with $Z_i = 0$. Table 6 presents the global joint MLE's of the parameters, which imply $\widehat{\text{CACE}} = \hat{\eta}_{c1\mu} - \hat{\eta}_{c0\mu} = 0.955 - (-0.054) = 1.009$.

In Figure 5 we present a histogram estimate of the posterior distribution of CACE for this artificial data set; also presented are two alternative estimates of this posterior distribution. The solid line represents the normal approximation to the posterior distribution based on the information matrix calculated using second derivatives of the logarithm of the likelihood function at the

TABLE 5
Hypothetical population distribution

t	$P(C_i = t\|\pi)$	$D_i(0)$	$D_i(1)$	$Y_i\|C_i = t,\ Z_i = 0, \pi$	$Y_i\|C_i = t,\ Z_i = 1, \pi$
c	$\omega_c = 0.25$	0	1	$N(0.1, 0.16)$	$N(0.9, 0.49)$
n	$\omega_n = 0.45$	0	0	$N(1.0, 0.25)$	$N(1.0, 0.25)$
a	$\omega_a = 0.30$	1	1	$N(0.0, 0.36)$	$N(0.0, 0.36)$
d	$\omega_d = 0.00$	1	0	—	—

TABLE 6

Maximum likelihood estimates for a data set from the population distribution in Table 5 *with* 50 *units assigned* $Z_i = 0$ *and* 50 *assigned* $Z_i = 1$ *under the monotonicity condition and the exclusion restriction*

C_i	$P(C_i \mid \pi)$	$D_i(0)$	$D_i(1)$	$Y_i(0) \mid C_i, \pi$	$Y_i(1) \mid C_i, \pi$
c	$\omega_c = 0.264$	0	1	$N(-0.054, 0.007)$	$N(0.955, 0.124)$
n	$\omega_n = 0.482$	0	0	$N(0.752, 0.395)$	—
a	$\omega_a = 0.254$	1	1	—	$N(-0.054, 0.357)$

global MLE: $N(1.009, 0.144)$. The dashed line represents the normal approximation of the posterior distribution of the ratio $\text{ITT}_Y/\text{ITT}_D$ around the ratio of the ITT estimates based on treatment–control average differences. More formally, define $\overline{Y_1} = \sum Y_{\text{obs},i} Z_{\text{obs},i}/N_1$ and $\overline{Y_0} = \sum Y_{\text{obs},i}(1 - Z_{\text{obs},i})/N_0$ with estimated variances $\text{Var}(\overline{Y_1}) = \sum Z_{\text{obs},i}(Y_{\text{obs},i} - \overline{Y_1})^2/N_1^2$ and $\text{Var}(\overline{Y_0}) = \sum (1 - Z_{\text{obs},i})(Y_{\text{obs},i} - \overline{Y_0})^2/N_0^2$, and analogously for $\overline{D_1}$ and $\overline{D_0}$, where $N_1 = \sum Z_{\text{obs},i}$ and $N_0 = \sum (1 - Z_{\text{obs},i})$ are the number of observations assigned to treatment and control, respectively. Then $\widehat{\text{IVE}} = \widehat{\text{ITT}}_Y/\widehat{\text{ITT}}_D$, where the two

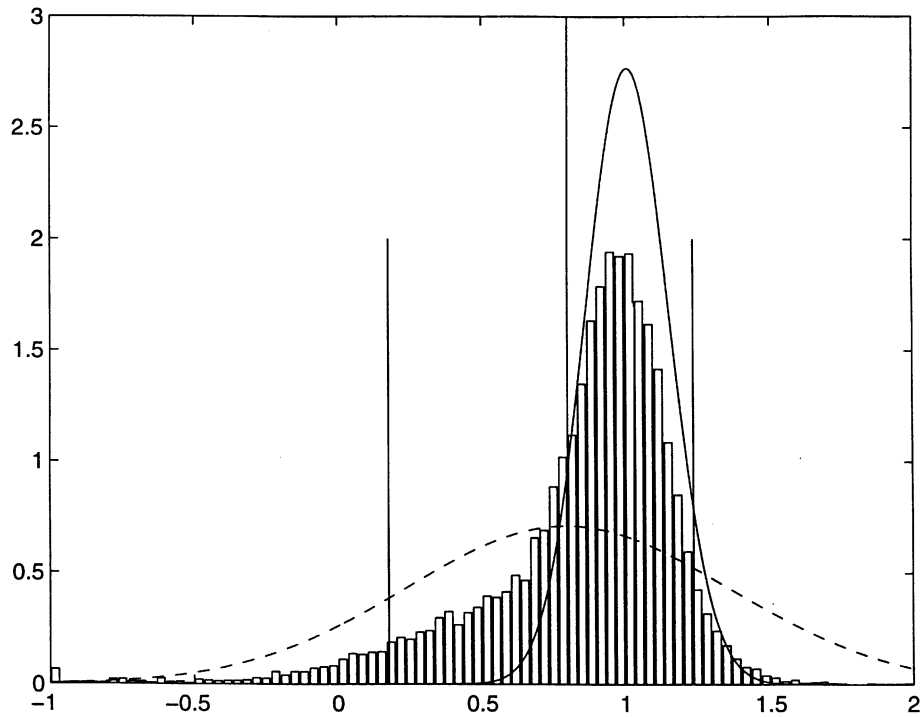

FIG. 5. *Estimates of the posterior distribution of CACE under exclusion restriction and monotonicity condition* (*data analyzed in Table* 6): *histogram is based on simulation, solid line is normal approximation based on mle, dashed line is normal approximation based on* $\widehat{\text{IVE}}$.

intention-to-treat estimates are $\widehat{\text{ITT}}_Y = \overline{Y_1} - \overline{Y_0}$ and $\widehat{\text{ITT}}_D = \overline{D_1} - \overline{D_0}$, with estimated variance for the large-sample approximation to the distribution of $\widehat{\text{IVE}}$ equal to

$$\text{Var}(\widehat{\text{IVE}}) = \Big(\text{Var}(\widehat{\text{ITT}}_Y)\widehat{\text{ITT}}_D^2 + \text{Var}(\widehat{\text{ITT}}_D)\widehat{\text{ITT}}_Y^2 - 2\,\text{Cov}(\widehat{\text{ITT}}_Y, \widehat{\text{ITT}}_D)\widehat{\text{ITT}}_Y\widehat{\text{ITT}}_D\Big)\Big/\widehat{\text{ITT}}_D^4,$$

where

$$\text{Var}(\widehat{\text{ITT}}_Y) = \text{Var}(\overline{Y_1}) + \text{Var}(\overline{Y_0}),$$
$$\text{Var}(\widehat{\text{ITT}}_D) = \text{Var}(\overline{D_1}) + \text{Var}(\overline{D_0})$$

and

$$\text{Cov}(\widehat{\text{ITT}}_Y, \widehat{\text{ITT}}_D) = \sum Z_{\text{obs},i}(Y_{\text{obs},i} - \overline{Y_1})(D_{\text{obs},i} - \overline{D_1})/N_1^2 + \sum(1 - Z_{\text{obs},i})(Y_{\text{obs},i} - \overline{Y_0})(D_{\text{obs},i} - \overline{D_0})/N_0^2.$$

The $\widehat{\text{IVE}}$ with its associated standard error is widely used in the econometric literature where it is known there as the instrumental variables estimate [e.g., Bowden and Turkington (1984)].

The two normal approximations presented in Figure 5, and in fact any such normal approximations to the posterior distribution of CACE, are poor. The normal approximation around the MLE fits the local region around the mode of the posterior distribution well, but cannot cope with the thick left tail of the actual posterior distribution. The 90% interval based on the normal approximation is (0.77, 1.25), including considerably less than 90% of the actual posterior distribution. The 90% interval based on the normal approximation to the posterior distribution of $\text{ITT}_Y/\text{ITT}_D$ is much wider $(-0.12, 1.72)$, but also fails to correspond accurately to the central 90% posterior interval.

We illustrate our analysis in a second way by presenting an evaluation of the repeated sampling operating characteristics of our proposed Bayesian procedures, including a comparison with the two alternatives from Figure 5. Repeatedly, we drew a sample of size 100 from the population distribution of Table 5, with 50 units randomly assigned $Z_i = 1$ and 50 assigned $Z_i = 0$. For each sample we also calculated the MLE and constructed a large-sample 90% confidence interval based on the normal approximation to the sampling distribution already used in constructing the solid line in Figure 3, as well as the econometric instrumental variables estimator and a 90% confidence interval based on the normal approximation to its sampling distribution using the same procedure as used in constructing the dashed line in Figure 3. For each sample we also calculated the posterior mean, median and the central 90% probability interval, based on a single Gibbs run of length 5000, started at the MLE's (earlier work based on multiple runs supported the propriety of the single-run approach in this case). For each estimator we then calculated over the repeated samples, its average, its median, its root mean squared error, its median absolute error and the coverage rate of its associated central 90% probability interval. Table 7 presents the results for 1000 replications.

TABLE 7

Operating characteristics of various procedures under the monotonicity condition and the exclusion restriction for replications from the population of Table 5 *with* 50 *units assigned* $Z_i = 0$ *and* 50 *assigned* $Z_i = 1$

					90% interval	
Estimator	**Mean bias**	**Median bias**	**Root mean squared error**	**Median absolute error**	**Coverage rate**	**Median width**
Posterior mean	−0.10	−0.07	0.48	0.30	0.91	1.61
Posterior median	−0.08	−0.06	0.51	0.32		
MLE	−0.14	−0.12	0.51	0.31	0.74	1.11
IVE	0.55	0.13	2.31	0.54	0.91	2.78

The posterior mean and median are clearly superior to the standard IV estimator in terms of accuracy and width of their associated 90% interval estimates. Compared to the MLE, the posterior mean and median are both slightly more accurate, but a more substantial advantage is the dramatically superior frequency coverage rate of their associated nominal 90% interval.

8. Conclusions. In this paper we apply the phenomenological Bayesian approach of Rubin (1978a) to develop a framework for obtaining Bayesian causal inferences in a randomized experiment with noncompliance. We demonstrate that our proposed method for inference for causal effects can proceed with and without additional assumptions on the compliance behavior (the monotonicity assumption) or on the effect of assignment on outcome for those whose treatment status is not affected by assignment (the exclusion restriction). Without these assumptions, inference, although straightforward in our approach, can be imprecise even in large samples. With these assumptions, one can estimate the complier average causal effect more accurately using our approach than using the standard econometric instrumental variables approach or other methods previously presented, as illustrated by our simulation experiment.

Although the illustrations provided are necessarily limited, the general developments are more widely applicable, as we view randomized trials with noncompliance as a bridge between randomized trials and observational studies. This view appears to be shared by Breslow, who writes "... the most important use of causal analysis may lie in the interpretation of results from randomized intervention trials that have substantial noncompliance" [Breslow (1996), page 26]. A number of extensions to the basic model are likely to be particularly relevant in practice.

First, typically researchers will have observations on covariates in addition to the outcome of interest, the treatment and the assignment. Covariates are incorporated into our model by making the outcome distributions $g_{tz}(y|\eta_{tz})$ and the probabilities ω_t depend on these covariates, thereby serving three purposes. First, covariates make inference conditional and therefore more precise as in any setting. Second, they make inferences more specific by estimating dif-

ferent average treatment effects for subpopulations indexed by the covariates. Third, they allow a more precise partitioning of the sample into compliers, always-takers, never-takers and defiers; when covariates are good predictors of compliance status, assignment is highly correlated with treatment status, conditional on this covariate, and sharper statements are possible concerning treatment effects in the subpopulation of compliers.

A second group of extensions involves modeling the clustering of units, which is often present because commonly the randomization takes place at a level different from the unit of observation. For example, McDonald, Hiu and Tierney (1992) consider a study where randomly selected doctors were encouraged to vaccinate at-risk patients against influenza. Clustering is modeled using common parameters for units in the same clusters, where the across-cluster parameters are linked together in a hierarchical model. Such an analysis would have been employed for the Sommer–Zeger data in Section 6 if clustering indicators had been available.

Third, the treatment received need not be binary: it is often the case that units take different dosages of the treatment, and so even if Z_i is binary, D_i is not. Efron and Feldman (1991) discuss such a case with both partial compliance and a binary assignment measured, but they assume that compliance under assignment to placebo reveals what compliance under assignment to the active treatment would have been. Avoiding this assumption, one can extend the basic model in Angrist, Imbens and Rubin (1996) along the lines of Angrist and Imbens (1995) and Angrist, Graddy and Imbens (1995) to allow for variable levels of the treatment while still maintaining or relaxing the assumption that, at the unit level, the only way the assignment affects the outcome of interest is through the level of the treatment. Such models will lead to more complex mixture structures than the one discussed in this paper, but ones where the payoff to using the phenomenological Bayesian approach could be even more substantial.

Acknowledgments. The authors are grateful to Gary Chamberlain, an Associate Editor and four reviewers for comments and suggestions.

REFERENCES

ANGRIST, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *American Economic Review* **80** 313–335.

ANGRIST, J. D., GRADDY, K. and IMBENS, G. W. (1995). Non-parametric demand analysis with an application to the demand for fish. Technical Report 178, National Bureau of Economic Research, Cambridge, MA.

ANGRIST, J. D. and IMBENS, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Amer. Statist. Assoc.* **90** 431–442.

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Amer. Statist. Assoc.* **91** 444–472.

BAKER, S. G. and LINDEMAN, K. (1994). The paired availability design: a proposal for evaluating epidural analgesia during labor. *Statistics in Medicine* **13** 2269–2278.

BALKE, A. and PEARL, J. (1994). Nonparametric bounds of causal effects from partial compliance data. Technical Report R-199-J, Dept. Computer Science, Univ. California, Los Angeles.

BLOOM, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* **8** 225–246.

BOWDEN, R. J. and TURKINGTON, D. A. (1984). *Instrumental Variables*. Cambridge Univ. Press.

BRESLOW, N. E. (1982). Clinical trials. In *Encyclopedia of Statistical Sciences* **2** 13–21. Wiley, New York.

BRESLOW, N. E. (1996). Statistics in epidemiology: the case-control study. *J. Amer. Statist. Assoc.* **91** 14–28.

DEMPSTER, A. P., LAIRD, N. and RUBIN, D. B. (1977). Maximum likelihood estimation from incomplete data using the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

EFRON, B. and FELDMAN, D. (1991). Compliance as an explanatory variable in clinical trials (with discussion). *J. Amer. Statist. Assoc.* **86** 9–26.

FISHER, L., DIXON, D., HERSON, J., FRANKOWSKI, R., HEARRON, M. and PEACE, K. (1990). Intention to treat in clinical trials. In *Statistical Issues in Drug Research and Development* (K. Peace, ed.) 331–350. Dekker, New York.

FISHER, R. A. (1925). *Statistical Methods for Research Workers*, 1st ed. Oliver and Boyd, Edinburgh.

GELFAND, A., HILLS, S., RACINE-POON, A. and SMITH, A. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Amer. Statist. Assoc.* **85** 398–405.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulations using multiple sequences. *Statist. Sci.* **7** 457–511.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence* **6** 721–741.

HEARST, N., NEWMAN, T. and HULLEY, S. (1986). Delayed effects of the military draft on mortality: a randomized natural experiment. *New England Journal of Medicine* **314** 620–624.

HECKMAN, J. and ROBB, R. (1985). Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data* (J. Heckman and B. Singer, eds.). Cambridge Univ. Press.

HOLLAND, P. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970.

HOLLAND, P. (1988). Causal inference, path analysis, and recursive structural equations models. In *Sociological Methodology*, Chap. 13. American Sociological Association, Washington, DC.

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–476.

IMBENS, G. W. and RUBIN, D. B. (1994). Estimating outcome distributions for compliers in instrumental variables models. Working Paper 1545, Harvard Institute of Economic Research.

LEE, Y., ELLENBERG, J., HIRTZ, D. and NELSON, K. (1991). Analysis of clinical trials by treatment actually received: is it really an option? *Statistics in Medicine* **10** 1595–1605.

LIU, C. and RUBIN, D. B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrica* **81** 633–648.

MANSKI, C. F. (1990). Non-parametric bounds on treatment effects. *American Economic Review, Papers and Proceedings* **80** 319–323.

MCCLELLAN, M. and NEWHOUSE, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? *Journal of the American Medical Association* **272** 859–866.

MCDONALD, C., HIU, S. and TIERNEY, W. (1992). Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *Clinical Computing* **9** 304–312.

MEIER, P. (1991). Comment on "Compliance as an explanatory variable in clinical trials" by B. Efron and D. Feldman. *J. Amer. Statist. Assoc.* **86** 19–22.

MENG, X. and RUBIN, D. B. (1991). Using EM to obtain asymptotic variance–covariance matrices: the SEM algorithm. *J. Amer. Statist. Assoc.* **86** 899–909.

MENG, X. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80** 267–278.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. [Translated in *Statist. Sci.* **5** (1990) 465–480.]

PERMUTT, T. and HEBEL, J. (1989). Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics* **45** 619–622.

REIERSOL, O. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* **9** 1–24.

ROBINS, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS* (L. Sechrest, H. Freeman and A. Bailey, eds.). NCHSR, U.S. Public Health Service, Washington, DC.

ROSENBAUM, P. (1995). *Observational Studies*. Springer, New York.

ROSENBAUM, P. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

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

RUBIN, D. B. (1975). Bayesian inference for causality: the importance of randomization. In *Proceedings of the Social Statistics Section of the American Statistical Association* 233–239. Amer. Statist. Assoc.

RUBIN, D. B. (1977). Assignment to treatment on the basis of a covariate. *Journal of Education Statistics* **2** 1–26.

RUBIN, D. B. (1978a). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **6** 34–58.

RUBIN, D. B. (1978b). The phenomenological Bayesian perspective in sample surveys from finite populations: foundations. In *Imputation and the Editing of Faulty or Missing Survey Data* 10–18. U.S. Department of Commerce, Washington, DC.

RUBIN, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by D. Basu. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (1990a). Comment on "Neyman (1923) and causal inference in experiments and observational studies." *Statist. Sci.* **5** 472–480.

RUBIN, D. B. (1990b). Formal modes of statistical inference for causal effects. *J. Statist. Plann. Inference* **25** 279–292.

SALSBURG, D. (1994). Intent to treat: the reduction and absurdum that became gospel. *Pharmacoepidemiology and Drug Safety* **3** 329–335.

SHEINER, L. B. and RUBIN, D. B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology and Therapy* **57** 6–10.

SOMMER, A. and ZEGER, S. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine* **10** 45–52.

TANNER, M. and WONG, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

WRIGHT, S. (1928). Appendix. In *The Tariff on Animal and Vegetable Oils* by P. G. Wright. Macmillan, New York.

WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Statist.* **5** 161–215.

ZELEN, M. (1979). A new design for randomized clinical trials. *New England Journal of Medicine* **300** 1242–1245.

ZELEN, M. (1990). Randomized consent designs for clinical trials: an update. *Statistics in Medicine* **9** 645–656.

DEPARTMENT OF ECONOMICS
LITTAUER CENTER 117
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138
AND NBER
E-MAIL: gimbens@harvard.edu

DEPARTMENT OF STATISTICS
SCIENCE CENTER 709
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138
E-MAIL: rubin@stat.harvard.edu