Clustered Encouragement Designs with Individual Noncompliance: Bayesian Inference with Randomization, and Application to Advance Directive Forms.

CONSTANTINE E. FRANGAKIS*

Department of Biostatistics, The Johns Hopkins University

Baltimore, MD 21205, U.S.A.

DONALD B. RUBIN

Department of Statistics, Science Center 709, Harvard University

Cambridge, MA 02138, U.S.A.

and

XIAO-HUA ZHOU

Division of Biostatistics, Indiana University School of Medicine

Indianapolis, IN 46202, U.S.A.

September 17, 2001

* For correspondence: 615 N. Wolfe St., Baltimore, MD, 21205, tel: (410) 955-3505; fax: (410) 955-0958; email: cfrangak@jhsph.edu

SUMMARY

In many studies comparing a new "target treatment" with a control target treatment, the received treatment does not always agree with assigned treatment – that is, the compliance is imperfect. An obvious example arises when ethical or practical constraints prevent even the randomized assignment of receipt of the new target treatment but allow the randomized assignment of the encouragement to receive this treatment. In fact, many randomized experiments where compliance is not enforced by the experimenter (e.g., with non-blinded assignment) may be more accurately thought of as randomized encouragement designs. Moreover, often the assignment of encouragement is at the level of clusters (e.g., doctors) where the compliance with the assignment varies across the units (e.g., patients) within clusters. We refer to such studies as "clustered encouragement designs" (CED) and they arise relatively frequently (e.g., Sommer and Zeger, 1991; McDonald, Hiu, and Tierney, 1992; Dexter et al., 1998). Here, we propose Bayesian methodology for causal inference for the effect of the new target treatment vs. the control target treatment in the randomized CED with all-or-none compliance at the unit level, which generalizes the approach of Hirano, Imbens, Rubin, and Zhou (2000) in important and surprisingly subtle ways, to account for the clustering, which is necessary for statistical validity. We illustrate our methods using data from a recent study exploring the role of physician consulting in increasing patients' completion of Advance Directive forms.

Keywords: Advance Directive; Causal Inference; Clustering; Noncompliance; Phenomenological Bayesian Model; Rubin Causal Model.

1. INTRODUCTION AND PURPOSE.

1·1. *Motivating studies and data features.*

When evaluating treatment options, direct assignment and enforcement of treatment receipt may not be ethical or feasible. In such cases, it is more realistic to view the design as involving the randomization of encouragement, as opposed to receipt, of the two target treatments, new and standard, where in some designs the encouragement is explicit and no enforcement is even attempted. Commonly, moreover, this encouragement is applied to clusters (e.g., physicians or villages) of subjects (e.g., patients). An example of such a "clustered-encouragement-design" (CED) was reported by Sommer and Zeger (1991) where investigators randomized villages in Indonesia to offer or not vitamin A supplements to all their infants, but not all infants in the villages assigned to get vitamin A actually received it. Another example of the CED was a study to evaluate a vaccine for influenza, where any randomized withholding of the vaccine was considered unethical (McDonald, Hiu, and Tierney, 1992; Hirano, Imbens, Rubin, and Zhou, 2000); for this reason, investigators randomized physicians to receive or not receive encouragement to vaccinate their patients, but many patients of the encouraged doctors did not receive a flu shot, and some patients of the not-encouraged doctors did receive the shot.

A more recent example of a CED was conducted on Advance Directive (AD) forms (Dexter et al., 1998), which are intended to be completed by patients to allow them to make early decisions about medical treatments at the late stages of life (instructional directives), and designate a representative decision maker (proxy directives) (Wenger et al., 1993). Dexter et al. (1998) randomized physicians to receive or not receive encouragement to discuss AD with patients; the outcome was patient completion of AD, and the original study addressed the effect of encouragement on AD completion (Dexter et al., 1998). For our purpose, however, an equally important substantive research goal is to assess the effect of physicians' discussion of AD as the new target treatment for potentially increasing patients' completion rates of the forms relative to the control target treatment of no such discussion (e.g., Miles et al., 1996).

2

Generally, CED studies share two specific data-structure aspects. First, there is frequent noncompliance of individual subjects - not clusters – for the new target treatments within randomized encouragement arms. Second, the distribution of noncompliance and outcomes frequently varies within and between clusters, which are the units of randomization, rather than the individual subjects. We consider CED studies where the compliance for target treatments is by definition (or for practical purposes) all or none (for extensions, see Sec. 5). This type of noncompliance means that there exist, essentially, two subgroups of patients who are not fully identifiable from the data: those who would not change their actual behavior concerning the target treatment no matter what their physician's assignment – the noncompliers, and those who would comply under both assignment – the compliers (e.g., Imbens and Rubin, 1994; Baker and Lindeman, 1994; Angrist, Imbens, and Rubin, 1996; Baker, 1998; Frangakis and Rubin, 1999). These definitions are local to this particular study and do not suggest compliance or not in other studies.

An intention-to-treat (ITT) analysis is especially appropriate when the randomized intervention is the scientific intervention - the target treatment – of interest. However, the CED uses randomized encouragement only as a surrogate to induce the new target treatment, and ITT analysis is not as appropriate for two reasons. First, the noncompliers arguably do not carry information about the comparison between the target treatments (e.g., biological efficacy or side effects) because, by definition of this group, the randomization cannot change receipt of target treatment; for relevant discussion between explanatory and pragmatic comparisons, see Sheiner and Rubin (1996) and Armitage (1998). Second, the noncompliers may experience effects of encouragement. For example, in the study on flu shots (McDonald, Hiu, and Tierney, 1992), it is possible that, for noncomplying physician-patient pairs, the encouragement has triggered physicians to suggest to their patients a number of other precautions against flu in addition to vaccination, and which might not have been taken in the absence of the encouragement; These are pure "encouragement" effects that confound the effect of vaccination if the noncompliers are included in the ITT analysis.

The second aspect common in CED studies, the clustered structure of units, also has methodological implications. Because the assigned encouragement is at the cluster level, assignment is ignorable (Rubin, 1978) only conditionally on the clusters. And, because noncompliance and outcomes can vary both within and between clusters, the interactions between clustering, noncompliance and outcomes need to be addressed.

1·2. *Addressing clustering with noncompliance at the individual level.*

The problem of noncompliance has received increasing attention recently. In particular, it is now generally recognized that the approach of focusing on the compliers, who are not generally fully identifiable from observed data (e.g., Sommer and Zeger, 1991; Baker and Lindeman, 1994; Angrist et al., 1996), is critically different from approaches that use the treatment actually received as if it were randomized, such as "as-treated" or "per-protocol" approaches, whose bias has been well documented (e.g., Rubin, 1991; Mark and Robins, 1993; Robins and Greenland, 1994; Sheiner and Rubin, 1996). Moreover, implicit assumptions in the standard "instrumental variables" analyses, such as the a priori exclusion of effects of assignment, have now formal expressions (Angrist et al., 1996), thereby allowing researchers to avoid such exclusion assumptions when they are not plausible. In related work without the exclusion restriction, Robins (1989) derived estimated bounds for treatment effects, Imbens and Rubin (1997a) developed an appropriate Bayesian approach for distinct patient-physician pairs, and, for the latter case, recently, Hirano et al. (2000) have modeled covariate information.

Research on clustered data, on the other hand, has a long history in interconnected literatures including: survey methodology, dating back at least to Neyman (1934), and Hansen and Hurwitz (1943); random effects, dating to Hartley and Rao (1967), Harville (1976) and Laird and Ware (1982); estimating equations methods (e.g., Liang and Zeger, 1986); hierarchical Bayesian and Empirical Bayesian methods dating to James and Stein (1961), Efron and Morris (1973), Rubin (1981), and others. In more recent work on clustered randomization with noncompliance, Frangakis, Rubin, and Zhou (1998) relaxed the exclusion restrictions but offered

4

limited information on the role of covariates and on the influence of prior distributions, whereas Korhonen et al. (2000) focused on analyses under the exclusion restrictions.

Here, we investigate the broader combined problem of clustered encouragement followed by individual noncompliance, and thereby propose general methodology for causal inference in studies where these two data structures are present together, and where structural exclusion restrictions are relaxed. In the next section we introduce notation and formalize our goal. In Sec. 3 we discuss models and methodology: within an abstract phenomenological Bayesian model (Rubin, 1978), we introduce an appropriate framework for causal inference with clustered data suffering from noncompliance. We discuss the critical role of clustering and covariates, and propose a flexible submodel. In Sec. 4 we illustrate our methods by analyzing data from the study on AD forms. The final section offers concluding remarks. The appendix gives details on our models.

2. CLUSTERED ENCOURAGEMENT DESIGN

2·1. *Setting.*

Consider a hospital serving a group of patients, $i = 1, ..., N$, the ith patient with physician Q_i, where $Q_i = 1, ..., M \leq N$, so that each physician may serve more than one patient.

In order to compare two target treatments, a new one versus a control one, assume that the hospital considers two possible actions for each physician: (i) encouraging the physician to administer the new target treatment, and (ii) no encouragement. In either case, however, patients within a physician may not comply with their physician's assignment. To allow for this, we adopt the formulation of Angrist et al. (1996) for all-or-none compliance, although we discuss extensions in Sec.5. We assume that patient i will either receive the new target treatment, indicated by $D_i(z) = 1$, or the control one, $D_i(z) = 0$, when Q_i is assigned action z, where $z = 1$ for encouragement for the new target treatment and 0 otherwise. Similarly let $Y_i(z)$ be patient i's outcome of interest, e.g., occurrence/absence of the disease, when physician Q_i is assigned action z. Covariate information about patient i and physician Q_i is collectively

5

denoted by a p-dimensional vector $\underline{X}_i$ that includes the vector of ones to allow for an intercept. Note that if compliance behavior were always the same within physicians, it would be more relevant to formulate the problem with the physicians as units defining variables, in addition to their being units of design.

Assume, for simplicity, that physicians' assignments are decided by complete randomization, where we let $Z_i = 1$ if patient i's physician, Q_i, is randomized to encouragement, and $Z_i = 0$ otherwise, and note that, since randomization is in clusters, $Z_i = Z_j$ whenever $Q_i = Q_j$. The Bayesian analysis is unchanged if the randomization depended on fully observed covariates that describe physicians or their patients. Finally, note that only the values $D_i^{obs} := D_i(Z_i)$ and $Y_i^{obs} := Y_i(Z_i)$ are observed; the values under the alternative assignment, $D_i^{mis} := D_i(1 - Z_i)$ and $Y_i^{mis} := Y_i(1 - Z_i)$ are missing.

2·2. *Compliance principal strata.*

Each patient i belongs to one of the following four groups: $C_i = c$ for a complier, defined by $\underline{D}_i = (D_i(0), D_i(1)) = (0, 1)$, that is, a patient who would comply with respect to the target treatment under both assignments of physician Q_i; $C_i = n$ for a never-taker, that is, a patient who, in this study, would never take the new target treatment no matter the physician's assignment, so that $\underline{D}_i = (0, 0)$; $C_i = a$ for an always-taker, one who, in this study, would always take the new target treatment, so that $\underline{D}_i = (1, 1)$; and $C_i = d$ for a defier, one who would act opposite to the assignment, so that $\underline{D}_i = (1, 0)$ (e.g., Imbens and Rubin, 1994; Pearl, 1994; Baker and Lindeman, 1994; Angrist et al., 1996; Baker, 1998; Barnard et al., 2001). Because membership to those four strata, unlike membership to observed compliance strata, is unaffected by encouragement, we call them "compliance principal strata" (for manifestations of principal strata in more general settings, see Frangakis and Rubin, 2000).

Although the encouragement to use the new target treatment may or may not induce its actual receipt for some patients, we assume that, in this context, the encouragement would not reverse a decision to take the new target treatment. This assumption was termed monotonicity

by Imbens and Angrist (1994), and allows only for the first three principal strata, that is, defiers are not allowed.

2·3. *Goal.*

By definition, the principal strata of noncompliers – the never-takers and always-takers, are those for whom different assignment in this experiment would not change their behavior with respect to the target treatments, and therefore those groups are not relevant for comparing the target treatments (e.g., Sommer and Zeger, 1991; Sheiner and Rubin, 1996). Moreover, if the randomized treatment is encouragement for vaccination, then other effects may be present if the encouraged physicians suggest to patients alternative preventive measures, such as recommendations to reduce the patients' exposure, prescriptions of other medicines, or earlier taking of the vaccine. Such "pure encouragement" effects are arguably more dominant for noncompliers than for compliers, because compliers experience both the effect of encouragement and the effect of the target treatment. Also, discerning "pure encouragement" effects among compliers requires assumptions not verifiable even with knowledge of all memberships to the compliance principal strata in the study. Nevertheless, because any effects of assignment Z_i on outcome Y for noncompliers must be due to sources other than the target treatment D, such effects for these patients can be removed simply by focusing analyses on the principal stratum of compliers.

For these reasons, we distinguish between the effect of encouragement on outcomes among noncompliers and among compliers. Using notation consistent with Imbens and Rubin (1997a), we let $l(t) = \{i : C_i = t\}$, the subset of subjects with compliance status $t = c, n,$ or a, and let N_t be the number of patients in that group. The standard ITT estimand $\sum_i [Y_i(1) - Y_i(0)]/N$ equals the mixture $\sum_{t \in \{c,a,n\}} N_t \, \text{ITT}^{(t)}/N$, where

$$\text{ITT}^{(t)} = \frac{1}{N_t} \sum_{i \in l(t)} Y_i(1) - Y_i(0), \quad t = c, a, n, \tag{2·1}$$

are the average effects of assignment within compliance principal strata. In the remainder of our

7

discussion, we focus on estimating the principal strata-specific ITT effects (2·1), and, in particular, the ITT effect among the compliers, $ITT^{(c)}$, the complier average causal effect (CACE).

3.   Modeling in the CED

3·1.   *Role of clustering in the phenomenological Bayesian approach.*

All potentially observable data can be expressed by the matrix $\mathbf{H} = (\underline{\mathbf{D}}, \underline{\mathbf{Y}}, \underline{\mathbf{X}}, \mathbf{Q}, \mathbf{Z})$, whose ith row is the vector $(\underline{D}_i, \underline{Y}_i, \underline{X}_i, Q_i, Z_i)$. Here, $\underline{D}_i$ denotes the pair of potential receipts $(D_i(0), D_i(1))$ for patient i, and $\underline{Y}_i$ is the pair of potential outcomes $(Y_i(0), Y_i(1))$. Although the quantities $(\underline{\mathbf{D}}, \underline{\mathbf{Y}}, \underline{\mathbf{X}}, \mathbf{Q})$ in $\mathbf{H}$ may be assumed fixed over hypothetical replications of the experiment (as in a permutation-based analysis, Rubin, 1990), we will more generally allow them to be considered random, with a distribution $pr(\underline{\mathbf{D}}, \underline{\mathbf{Y}}, \underline{\mathbf{X}}, \mathbf{Q})$. The joint probability distribution of $\mathbf{H}$ that is induced by the distribution $pr(\underline{\mathbf{D}}, \underline{\mathbf{Y}}, \underline{\mathbf{X}}, \mathbf{Q})$ and the clustered randomization, $pr(\mathbf{Z} | \underline{\mathbf{D}}, \underline{\mathbf{Y}}, \underline{\mathbf{X}}, \mathbf{Q})$, will still be denoted by pr. We assume that the matrix $(\underline{\mathbf{D}}, \underline{\mathbf{Y}}, \underline{\mathbf{X}}, \mathbf{Q})$ contains all the information on observable data and on the design, so that the rows of $\mathbf{H}$ are exchangeable (Rubin, 1978). Following results on exchangeability (de Finetti, 1974), we may write essentially without loss of generality, write

$$pr(\underline{\mathbf{D}}, \underline{\mathbf{Y}}, \underline{\mathbf{X}}, \mathbf{Q}) = \int \prod_i pr\left\{(\underline{D}_i, \underline{Y}_i, \underline{X}_i, Q_i) | \theta\right\} pr(\theta) d\theta, \qquad (3·1)$$

for some distributions $pr(\theta)$ and $pr(\underline{D}_i, \underline{Y}_i, \underline{X}_i, Q_i | \theta)$, where θ can be thought of as representing the characteristics of a larger reference population from which the study units, physicians and patients, are drawn.

We stress that, although the clustered randomization on $\mathbf{Q}$ is likely to give less precise estimates of effects than a complete randomization at the patient level, the clustered randomization does not affect the joint exchangeability in (3·1). Rather, the clustered randomization is related to the assignment mechanism: because the assignment is at the level of physicians, $\mathbf{Q}$, we have that $pr(\mathbf{Z} | \underline{\mathbf{D}}, \underline{\mathbf{Y}}, \underline{\mathbf{X}}, \mathbf{Q}) = pr(\mathbf{Z} | \mathbf{Q})$, so that assignment is ignorable (Rubin, 1976, 1978) with

but not without conditioning on $\mathbf{Q}$. Consequently, inference on the potential outcomes also needs to be conditional on $\mathbf{Q}$, which, in this case, is expected to reduce precision because no physician has patients in both assignment arms.

If the compliance principal strata C_i and the potential outcomes $Y_i(z)$ were known for all i and z, then the principal strata-specific effects $\mathrm{ITT}^{(t)}$ could be computed from definition (2·1). Although the values $\mathbf{D^{obs}} := \{D_i^{obs}\}$ and $\mathbf{Y^{obs}} := \{Y_i^{obs}\}$ are known, the values $\mathbf{D^{mis}} := \{D_i^{mis}\}$ and $\mathbf{Y^{mis}} := \{Y_i^{mis}\}$ are unknown (defined at the end of Section 2.1). From (3·1), the missing values $(\mathbf{Y^{mis}}, \mathbf{D^{mis}})$ have a joint posterior predictive distribution

$$\mathrm{pr}(\mathbf{Y^{mis}}|\mathbf{D^{mis}}, \mathbf{H^{obs}}, \theta)\,\mathrm{pr}(\mathbf{D^{mis}}, \theta|\mathbf{H^{obs}}), \qquad (3\cdot2)$$

where $\mathbf{H^{obs}} := (\mathbf{D^{obs}}, \mathbf{Y^{obs}}, \underline{\mathbf{X}}, \mathbf{Q}, \mathbf{Z})$, the observed data. Then, Bayesian inference on the estimands $\mathrm{ITT}^{(t)}$ follows from their posterior predictive distributions induced by (3·2).

3·2. *Role of covariates for relaxing exclusion restrictions.*

Because membership to the compliance principal strata is not fully identifiable from observed data, it is important to understand how information on the principal strata-specific causal effects is recoverable.

For simplicity, assume $Y_i(z)$ is binary, e.g., 1 for occurrence of disease, 0 otherwise, and let $\mathrm{ITT}^{(t,\theta)} := \mathrm{pr}(Y_i(1) = 1|C_i = t, \theta) - \mathrm{pr}(Y_i(0) = 1|C_i = t, \theta)$, the average causal effect within compliance principal stratum $t = c, n, a$ in the reference population θ defined by (3·1) (unconditionally on physicians). The estimand $\mathrm{ITT}^{(c,\theta)}$ is estimable consistently, as N grows, under the so-called "exclusion restrictions" (e.g., Bloom, 1984; Sommer and Zeger, 1991; Angrist et al. 1996). However, for the reasons in Sec. 1.1, we do not wish to impose a priori these assumptions for the CED. In the absence of exclusion restrictions and of covariates, the effects $\mathrm{ITT}^{(t,\theta)}$ are not consistently estimable, but bounds are (e.g., Robins, 1989; Manski, 1990; Balke and Pearl, 1997) under the standard asymptotics with sample size N growing. Never-

9

theless, when covariates are available, Frangakis (1999, PhD thesis) shows that the predictive model $\mathrm{pr}(C_i = t | X_i, \theta)$ of compliance principal strata from the covariates is estimable with no exclusion restrictions, and that for the effects $\mathrm{ITT}^{(t,\theta)}$, asymptotic bounds are narrower if the covariates are used than if they are not used, so that $\mathrm{ITT}^{(t,\theta)}$ are consistently estimable under an asymptotic argument that has the number of covariates growing in an appropriate sequence in addition to sample size.

In practice, the above discussion does not change the way Bayesian inference is drawn, but helps critically in understanding the role of covariates in recovering information. Under (3·1), Bayesian inference is based directly on the posterior predictive distribution of the target estimands $\mathrm{ITT}^{(t)}$ induced by (3·2), where (i) in the absence of covariates, the spread of the posterior distribution will reflect the relatively large uncertainty in predicting compliance principal strata, whereas (ii) when a covariate that predicts compliance principal strata is modeled, the spread of the posterior distribution of the estimands will be narrower.

In the remaining part of this paper, we describe a specific Bayesian model and apply it to the study of AD introduced in Sec. 1.1. Asymptotic inference using estimated covariate-adjusted bounds and comparisons with small sample Bayesian inference will be discussed in detail in a subsequent article.

<center>3·3. *Specific model.*</center>

We model the joint distribution $\mathrm{pr}(\underline{D}_i, \underline{Y}_i, \underline{X}_i, Q_i, \theta)$ of (3·1) by a sequence of conditional distributions. First, because we focus on the finite study estimands $\mathrm{ITT}^{(c)}$ and, because $\underline{X}_i$ and Q_i are known for all i, we take the marginal distribution of $\underline{X}_i$ and Q_i to be the observed distribution. Conditionally on $\mathbf{X}, \mathbf{Q}$, we assume the following structure for the other model components.

Compliance Principal Strata.

Define the vector $\underline{W}_i$ to be an $r \times 1$ subset of $\underline{X}_i$ that includes the vector of ones but excludes characteristics that are constant across patients clustered within physician Q_i, because those are

<center>10</center>

aliased with the intercept for that physician. We model the compliance principal strata with two probit submodels,

$$\text{pr}(C_i = n | \underline{X}_i, Q_i, \theta) = 1 - \Phi(\underline{X}_i' \alpha^{(C,1)} + \underline{W}_i' b_{Q_i}^{(C,1)}),$$

$$\text{pr}(C_i = c | \underline{X}_i, Q_i, \theta) = \{1 - \text{pr}(C_i = n | \underline{X}_i, Q_i, \theta)\}\{1 - \Phi(\underline{X}_i' \alpha^{(C,2)} + \underline{W}_i' b_{Q_i}^{(C,2)})\}, \tag{3.3}$$

and $\text{pr}(C_i = a | \underline{X}_i, Q_i, \theta) = 1 - \text{pr}(C_i = c | \underline{X}_i, Q_i, \theta) - \text{pr}(C_i = n | \underline{X}_i, Q_i, \theta)$. In these expressions, $\alpha^{(C,1)}, \alpha^{(C,2)}$ are $p \times 1$ parameter vectors that model the association between compliance principal strata and covariates. The corresponding parameters $b_{Q_i}^{(C,1)}, b_{Q_i}^{(C,2)}$ are $r \times 1$ vectors specific to physician Q_i, and these model the association between compliance principal strata and physicians. The function $\Phi()$ is the standard normal cumulative distribution.

To facilitate computations later, we augment the above model with two latent variables for each person, V_i, and U_i, defined via the relations

$$C_i = n \text{ if } C_i^n \equiv \underline{X}_i' \alpha^{(C,1)} + \underline{W}_i' b_{Q_i}^{(C,1)} + V_i \le 0,$$

$$C_i = c \text{ if } C_i^n > 0 \text{ and } C_i^c \equiv \underline{X}_i' \alpha^{(C,2)} + \underline{W}_i' b_{Q_i}^{(C,2)} + U_i \le 0,$$

and where $V_i \sim N(0,1)$ and $U_i \sim N(0,1)$ independently.

Potential Outcomes. For our application in Sec. 4 we have binary potential outcomes so we posit the probit model,

$$\text{pr}(Y_i(z) = 1 | C_i = t, \underline{X}_i, Q_i, \theta) = \Phi\{f^{(1)}(\underline{X}_i, t, z)\alpha^{(Y)} + f^{(2)}(\underline{W}_i, t)b_{Q_i}^{(Y)}\}, \tag{3.4}$$

for $t = c, n, a$ and $z = 0, 1$. Here, $f^{(1)}, f^{(2)}$ are link vector functions of dimensions p_f, r_f, respectively; $\alpha^{(Y)}$ is a $p_f \times 1$ parameter vector; and $b_{Q_i}^{(Y)}$ are $r_f \times 1$ parameters specific to physician Q_i. This formulation assumes, for parsimony, that the second term $f^{(2)}$ in (3.4) is not a function of assignment arm. As noted by a referee, such issue of parsimony also arises when modelling many covariates.

Also, we assume that, given the covariates, physician indicators, compliance principal strata, and parameters, the two potential outcomes $Y_i(1)$ and $Y_i(0)$ are independent. This model, say M-ind, can be extended to a model, say M-dep, to allow conditional association between the two potential outcomes, but the parameters of such a model that control that association, e.g., the partial correlation between $Y_i(1)$ and $Y_i(0)$, do not appear in the observed data likelihood of model M-dep given the parameters, or in the estimands $\text{ITT}^{(t,\theta)}$ in the larger reference population, regardless of the assumed or the true population distribution. For these reasons, the posterior distribution for these estimands based on model M-ind is identical to the posterior distribution based on the models M-dep; for details, see Imbens and Rubin, 1997a, Sec. 3, paragraphs 6-7. For the finite population estimands, $\text{ITT}^{(t)}$, there will generally be some sensitivity of the posterior distribution to the choice of model M-dep, which, to focus on main points here, can be explored in applications.

As with the model for compliance, to facilitate computations, we augment the outcome model as follows:

$$Y_i(z) = 1 \text{ if } Y_i(z)^* \equiv f^{(1)}(\underline{X}_i, C_i, z)\alpha^{(Y)} + f^{(2)}(\underline{W}_i, C_i)b_{Q_i}^{(Y)} + S_i(z) \geq 0$$

where, for each z, $S_i(z) \sim N(0,1)$ independently of U_i, V_i for $i = 1, ..., N$. Furthermore, our assumption of conditional independence between $Y_i(0)$ and $Y_i(1)$ at this stage implies independence between $S_i(0)$ and $S_i(1)$ (computations need only consider the term $S_i = S_i(Z_i)$; see appendix, item 4). We emphasize that this independence is not related to independence (or dependence) between outcome and compliance principal strata.

Parameters.

The parameter θ includes $\{\alpha^{(C,1)}; \alpha^{(C,2)}; \alpha^{(Y)}; b_q^{(C,1)}, b_q^{(C,2)}, b_q^{(Y)}, q = 1, ..., M; D^{(b)}\}$, where $D^{(b)}$ is a variance-covariance matrix parameter defined below. We choose the prior distributions for θ to be proper but diffuse, in order to ensure proper posterior distributions and relatively fast convergence of the fitting algorithms, but, at the same time, to be relatively noninformative for

our application (Sec. 4). Given the physician-specific parameters, we posit prior distributions

$$\alpha^{(C,1)} \sim N(\alpha_0^{(C,1)}, \mathbf{I}\xi), \quad \alpha^{(C,2)} \sim N(0, \mathbf{I}\xi), \quad \alpha^{(Y)} \sim N(0, \mathbf{I}\xi), \tag{3·6}$$

independently. Here, $\mathbf{I}$ is the identity matrix, and ξ is an inflating factor, which is set by the analyst. The component of $\alpha_0^{(C,1)}$ that corresponds to the intercept is set to $-\Phi(\frac{1}{3})\sqrt{\xi}$ and the remaining components are set to 0 to represent a prior proportion of approximately 33% for each of the three compliance principal strata.

The physician-specific parameters are assumed random with

$$(b_q^{(C,1)}, b_q^{(C,2)}, b_q^{(Y)}) | D^{(b)} \sim N(0, D^{(b)}), \tag{3·7}$$

independently across physicians $q = 1, ..., M$. In our application of Sec. 4 we had a relatively low proportion of always-takers, so we constrained $D^{(b)}$ so that $(b_q^{(Y)}, b_q^{(C,1)})$ and $b_q^{(C,2)}$ were a priori independent. For the other components of $D^{(b)}$, we assume that the inverses of $D^{(1)} := \text{var}(b_q^{(Y)}, b_q^{(C,1)})$ and $D^{(2)} := \text{var}(b_q^{(C,2)})$ a priori have Wishart distributions with scale matrices $(\beta^{(1)} R^{(1)})^{-1}$ and $(\beta^{(2)} R^{(2)})^{-1}$, respectively, and degrees of freedom, $\beta^{(1)}$ and $\beta^{(2)}$ respectively, equal to the dimensions of $(b_q^{(C,1)}, b_q^{(Y)})$ and $b_q^{(C,2)}$ (in order to introduce relatively little prior information, e.g., Wakefield et al., 1994; for $R^{(1)}$ and $R^{(2)}$ see appendix), although it would also be interesting to study more informative priors distributions. Informative priors distributions can be fitted, for example, using the same model but after introducing pseudo-subjects in the dataset (e.g., Hirano, et al., 2000), where we can express the prior information by attaching weights to the pseudo-subjects' contribution in the likelihood based on the configuration in their covariates, cluster indicators, compliance principal strata, assignment, and outcomes.

Using model (3·3)-(3·7), inference for the estimands of interest, $\text{ITT}^{(t)}, t = c, a, n$, in the finite study population is based on (3·2). The appendix outlines an algorithm for simulating these distributions for our models.

13

4. APPLICATION TO ADVANCE DIRECTIVES

4·1. *Procedures.*

We now return to the study of AD forms introduced in Sec. 1.1. AD forms are designed to increase patient's autonomy, and although they enjoy some support by ethicists and physicians (Hughes and Singer, 1992), few patients complete them in practice, and few physicians discuss the role of these forms with their patients. It has been hypothesized that if physicians briefly discussed the role of AD forms with their patients, this would cause completion rates to substantially increase (Miles et al. 1996). The discussion effect is very important because, if shown large, could help convince physicians to spend the brief time needed to discuss AD forms with even more of their eligible patients. So, our goal is to address this hypothesis with a valid analysis of a CED. The problem of more flexible alternative designs in this application is studied by Frangakis and Baker (2001).

The data we use are a subset from the study on AD forms by Dexter et al. (1998), who analyzed the data by ITT analyses. In that study the researchers randomly divided eligible physicians of an urban hospital into four groups: one group routinely received computer reminders to discuss instructional directives with their patients; another group received reminders on proxy directives; a third group received both reminders, and a fourth group received no reminders. Here, the subset of data is from the extreme arms: the control group, denoted by ($Z = 0$), and the group receiving reminders for both AD forms, denoted by ($Z = 1$), where only patients eligible for AD discussion and completion were included. To use these data in our framework, let D_i^{obs} be the indicator for actual discussion of AD, equal to 1 if patient i's physician discusses either of the two AD forms with patient i (the new target treatment), 0 otherwise (the control target treatment), and let Y_i^{obs} be the indicator for completion of AD, equal to 1 if the patient completes either AD form, 0 otherwise. Patient age data are also of interest because age has been previously shown to be associated with discussions of AD forms (Duffield and Podzamsky, 1996; Boyd et al., 1996; Hakim et al., 1996). Moreover, because in a preliminary

analysis, we found none of the available covariates other than age to be as useful in predicting compliance status, here we use age as the only covariate. Table 1 gives some basic characteristics of our data that can be summarized easily using theory for finite population cluster sampling (Cochran, 1963).

Table 1 about here

Table 1 shows that 3% of patients completed AD forms under control $(Z = 0)$ versus 14% when their physicians were encouraged to discuss them $(Z = 1)$. However, approximately three quarters of the physician-patient pairs did not discuss the forms when encouraged, and so, following Sec. 2.2, they must be never-discussants in the sense that they would not discuss the forms whether encouraged or not in this study. Analogously, approximately 5% of the pairs are always-discussants. The estimated remaining approximately one fifth of physician-patient who are neither always-discussants nor never-discussants are discussion-compliers, in the sense that they would discuss AD if and only if encouraged. Therefore, although the modest 11 % ITT effect of encouragement on completion rates may suggest to physicians that discussing AD forms with their patients has no practical effect, the majority of physician-patient counted in that estimate are not relevant to the effect of AD discussion on AD completion. That is, evidence for the effect of AD <u>discussion</u> on AD <u>completion</u> needs to be sought among discussion compliers.

Generally, it is not known whether physicians or patients initiated the discussions. Nevertheless, focusing on the discussion-complier pairs almost ensures that the discussions under encouragement are initiated by physicians, because we know they would not have occurred without encouraging the physicians. Therefore, we focus on estimating the effect of encouragement on completion among the subset of discussion-compliers, $\text{ITT}^{(c)}$, which, here, we call the effect of discussion on form completion. To estimate $\text{ITT}^{(c)}$, we use two procedures based on models of the form $(3 \cdot 3)$-$(3 \cdot 7)$.

The first model-based procedure uses the specifications $(3 \cdot 3)$-$(3 \cdot 7)$ with no covariates. For this model, $\underline{X}_i$ and $\underline{W}_i$ are 1. The outcome link function $f^{(1)}$ saturates the patterns of compliance principal strata by treatment arm: $f^{(1)}(\underline{X}_i, C_i, z) = [I(C_i = n) * z, I(C_i = n) * (1 -$

15

$z), I(C_i = c) * z, I(C_i = c) * (1 - z), I(C_i = a) * z, I(C_i = a) * (1 - z)]$. We take the vector function $f^{(2)}(\underline{W}_i, C_i)$ to be $[I(C_i = n), I(C_i = c), I(C_i = a)]$. The hyperparameter ξ in (3·6) is set here to 5, although other values were also tried (see also Sec. 4.3).

The second procedure uses patients' age in the model components (3·3)-(3·7). Specifically, for this model, both $\underline{X}_i$ and $\underline{W}_i$ are set to $[1, \text{age}_i]$, where age_i is the patient's age in standardized log-log scale. The link function $f^{(1)}$ now fits separate intercepts and slopes on age_i for each of the six combinations of compliance principal strata C_i crossed by assignment arm z. The link function $f^{(2)}$ fits separate intercepts and slopes on age_i for each compliance principal strata. The values of the hyperparameters are as in the model without age. For further details on the fitting methods, see the appendix.

4·2. *Results.*

Table 2 about here

In Table 2, we report estimates of the estimands: compliance principal strata–specific percentages of AD completion under control, under encouragement, and the between-arm (encouragement - control) difference $\text{ITT}^{(t)}, t = c, n, a$ using the two model-based procedures. For each model, we obtained the posterior predictive distributions of these estimands as induced by (3·2). The table summarizes these distributions by means (within treatment arms, the posterior predictive means are the posterior predictive probabilities of AD completion averaged over X_i, Q_i and θ from its posterior distribution), standard deviations, and 95% intervals (2.5% and 97.5% quantiles).

The model that accounts for clustering but does not model age gives, mostly, unhelpfully broad answers, consistent with the role of clustering and covariates discussed, respectively, in Secs. 3.1 and 3.2. An exception in this dataset is inference for the never-discussants, which occurs because the observed completion rate for the non-discussants under control arm (Table 1) is zero. This is, generally, a mixture of completion rates of discussion-compliers and never-discussants assigned control, and therefore, here, both of these rates are zero. However, most

16

of the other 95% intervals, including for the effect on discussion-compliers, are too wide for practical use or interpretation with this model.

In contrast, the model that uses age in both the compliance and the outcome components provides 95% posterior intervals that are quite usefully narrower than those of the model without age. In particular, among compliers, the effect of assignment on completion rates has a posterior mean of 62% and is most likely at least 34% (2.5% posterior quantile). These estimates are also substantially higher than those reported by the ITT analyses in Table 1, and reflect in a principled way the uncertainty from the different sources of missing information.

The other results are generally not surprising, except perhaps the negative estimates for the effect of encouragement among always-discussants. Nevertheless, because the zero effect is well within the posterior interval in both models, this result is consistent with random fluctuation. Moreover, because in this application the proportion of always-takers is low, imposing a priori the exclusion restriction would not substantially change the results. In other applications where this assumption would be plausible and with larger proportions of always-takers, it would be beneficial to formulate this assumption explicitly in the model.

Table 3 about here

It is relevant to check the extent to which the increased precision of the second model-based procedure can be attributed to information in the data, including the ability of age to predict compliance principal strata, or information supplied by the prior distribution. In Table 3 we report the probabilities of completion rates under control, encouragement and their difference as induced solely by the prior distributions for each model. For example, to get a draw $Y(0)$ from the prior distribution for compliers, we (i) draw a subject's $\underline{X}_i, Q_i$ from $\mathrm{pr}(\underline{X}_i, Q_i)$ (by assumption here, the observed distribution), (ii) draw θ from the defining model $\mathrm{pr}(\theta|\underline{X}_i, Q_i)(= \mathrm{pr}(\theta))$, (iii) draw compliance principal stratum C_i from the model $\mathrm{pr}(C_i = t|\underline{X}_i, Q_i, \theta), t = c, n, a$ of (3·3), and (iv) if the stratum is "complier", we calculate $\mathrm{pr}(Y_i(0) = 1|C_i = c, \underline{X}_i, Q_i, \theta)$ from (3·4) and, with this probability, draw a Bernoulli outcome $Y(0)$. The distributions for the other entries are derived analogously. None of these distributions appears particularly informative for

17

the estimands summarized in Table 2, suggesting that the increased precision in the model with age in Table 2 is not particularly influenced by the chosen prior distributions.

Moreover, the posterior distributions of the model parameters showed evidence that the probability of being a never-discussant decreased with age (mean, $[2.5\%, 97.5\%]$ quantiles for $\alpha_{age}^{(C,1)}$: 0.052 [0.025, 0.094]), and most of the shift was to being a complier ($\alpha_{age}^{(C,2)}$: -1.293 [-2.275,-0.122]). We also calculated the posterior predictive distribution of the probability that the next new patient seen has each compliance principal stratum as a function of that patient's age. Means and standard deviations from this distribution conditionally on age are displayed in Fig. 1. Each draw from this distribution is obtained as the defining probability of compliance principal stratum in (3·3), where $\alpha^{(C,1)}$ and $\alpha^{(C,2)}$ are drawn from their joint posterior distribution, and $b_q^{(C,1)}$ and $b_q^{(C,2)}$ are drawn from their defining model (3·7) after having drawn $D^{(b)}$ from its posterior distribution.

Figure 1 about here

Based on these results, the estimates obtained from the second model-based procedure are expected to be more appropriate for the compliance principal stratum-specific effects than either the estimated bounds or the first model estimates. Taking the effect of encouragement on the discussion-compliers to be the relevant estimand for the effect of AD discussions on AD completion, these results give support to, and considerably strengthen, the hypothesis that physician-discussion can substantially increase patient-completion of AD forms.

4·3.   *Other models.*

More than twenty other models of the form (3·3)-(3·7) were fit, were we varied the link functions $f^{(1)}$ and $f^{(2)}$, the functional forms for age, and the inflating factor ξ in (3·6). In addition, in a preliminary effort in this work (Frangakis et al., 1998), we had also tried logistic mixed effects analogues to the probit models reported here. Those models gave results mostly similar to the ones presented in Table 2 here (see, for example, Table 2 of Frangakis et al., 1998),

18

although with varying performance in the criteria of (i) degree in which the prior distributions influenced the results in the sense of the measures in Table 3; and (ii) convergence diagnostics. For example, in contrast to the models in Sec. 3.3, the simulation stage for the logistic mixed effects model, after incorporation of Metropolis-Hastings adjustments to address lack of conjugacy, did not pass the convergence diagnostics of Gelman and Rubin (1992) (see also appendix) within a satisfactory time. Among the models we tried, the two reported in the previous section were the most acceptable with respect to these two criteria. We did not consider models with residual dependence between the potential outcomes conditionally on stage (3·4), or with assignment-arm differences in the term $f^{(2)}$ for physician-specific random parameters.

5.   REMARKS

We described methodology for causal inference in studies with randomization in clusters but noncompliance at the individual level. The proposed method improves upon current procedures, which face limitations with respect to either validity or precision in estimation.

Our procedure is designed to be Bayesian for the input prior. In general, of course, a posterior interval does not automatically share the property of a confidence interval just as the latter does not automatically share the property of the former. It would be interesting to study frequency calibration of Bayesian procedures in this problem, for example, by asymptotics that allow information from covariates to grow with samples size, as mentioned in Sec. 3.2, or by mixing permutation distributions with the Bayesian model (e.g., Rubin, 1998).

Our methods assume all-or-none observed compliance. For situations where observed compliance is continuous or multilevel, an approach that would discretize compliance to two (or few) levels can still be practically useful, as is often common practice with continuous variables (e.g., age simplified to young/old) where appropriate. Alternatively, direct modeling of the multiple compliance principal strata can be done by using appropriate assumptions on the parameters.

For an example, suppose that from a study's continuous compliance measure we create

D_i^{obs} to have three ordered levels, labeled 2, 1, and 0, roughly representing, respectively, full dose, half dose, and, no dose of the new treatment. Then, there are generally 9 principal strata $\underline{D}_i = (D_i(0), D_i(1))$. Suppose further that the behavioral or pharmacological context of that study makes the following three assumptions plausible. First, the exclusion restriction holds, i.e., $Y_i(1) = Y_i(0)$ if $D_i(1) = D_i(0)$. Second, encouragement to the new treatment increases receipt of dose of the new treatment both within and across subjects in the following sense:

Multilevel Monotonicity.

(a) (within subjects): $D_i(0) \leq D_i(1)$, and

(b) (across subjects): If $D_i(0) < D_j(0)$ then $D_i(1) \leq D_j(1)$.

Multilevel monotonicity allows for a total of 5 principal strata, three with treatment receipt unchangeable by encouragement ($\underline{D}_i$=(0,0), (1,1), or (2,2)), and two with treatment receipt increasing by 1 dose with encouragement ($\underline{D}_i$=(0,1) or (1,2)). A third assumption, then, can be that for the last two groups, the effect of encouragement on outcomes is the same, e.g., in the scale of relative risk:

$$\frac{\text{pr}(Y_i(1) = 1 | \underline{D}_i = (0,1))}{\text{pr}(Y_i(0) = 1 | \underline{D}_i = (0,1))} = \frac{\text{pr}(Y_i(1) = 1 | \underline{D}_i = (1,2))}{\text{pr}(Y_i(0) = 1 | \underline{D}_i = (1,2))} = R.$$

Then it can be proven that, under the above conditions, the causal effect R, as well as all other unknown components of the distributions, are consistently estimable. As in Sec. 3.2, some of the assumptions can be relaxed with the help of covariates or by substitution with alternative assumptions. It is then relevant to explore the plausibility of such assumptions in the study context, and compare results to alternative methods. Details and applications of this approach to multilevel compliance will be discussed in the future.

More generally, it is precisely the emphasis of this approach on the existence of principal strata of compliance that allows the researcher to input into analyses explicit scientific assumptions.

A MODEL FITTING

Computation of the posterior distribution (3·2) of the missing compliance principal strata, say, $\mathbf{C}^{\mathbf{mis}}$, missing potential outcomes $\mathbf{Y}^{\mathbf{mis}}$, and parameters θ were based on simulations from a Gibbs sampler (Geman and Geman, 1984), which draws, in this order: the missing compliance principal strata $\mathbf{C}^{\mathbf{mis}}$; the missing potential outcomes $\mathbf{Y}^{\mathbf{mis}}$; the latent variables C_i^n and C_i^c for the current set of never-takers, compliers, and always-takers; the latent variables $Y_i^* \equiv Y_i(Z_i)^*$ for the outcome model; the parameters $\alpha^{(C,1)}$, $\alpha^{(C,2)}$ for the compliance model; the outcome model parameters $\alpha^{(Y)}$; the cluster-specific parameters $b_q^{(Y)}, b_q^{(C,1)}, b_q^{(C,2)}$, for $q = 1, ..., M$; and the variance matrices $D^{(1)}, D^{(2)}$. For all steps, drawing is done cyclically and each step conditions on all other unknowns, with the following exceptions: the first step must exclude C_i^n and C_i^c from the conditioning to allow the draws of $\mathbf{C}^{\mathbf{mis}}$ to vary over their sample space; also, at this step, the conditional distribution on Y_i^{obs} is relatively easy to simulate from, and, so, replaces the conditional distribution on Y_i^* for algorithmic efficiency; and, the potential outcomes $\mathbf{Y}^{\mathbf{mis}}$, drawn at step 2 to calculate the estimands (2·1), are not included in any other conditioning, for algorithmic efficiency. The distributions involved in the Gibbs sampler are as follows.

1. Any missing compliance principal stratum is drawn at this step from $\mathrm{pr}(C_i | Y_i^{obs}, D_i^{obs}, \underline{X}_i, Q_i, Z_i, \theta)$. This distribution is obtained from the joint distribution $\mathrm{pr}(C_i, Y_i^{obs}, D_i^{obs} | \underline{X}_i, Q_i, Z_i, \theta)$. For example, a subject with $Z_i = D_i^{obs} = 0$ can be either a complier or a never-taker, and the conditional Bernoulli distribution of C_i is proportional to

$$\{l(c, Z_i, \underline{X}_i, Q_i, Y_i^{obs}, \theta)\}^{I(C_i=c)} \{l(n, Z_i, \underline{X}_i, Q_i, Y_i^{obs}, \theta)\}^{I(C_i=n)}$$

21

where we define $l(t_0, z_0, x_0, q_0, y_0, \theta)$ to be

$$\mathrm{pr}(C_i = t_0 | \underline{X}_i = x_0, Q_i = q_0, \theta) \, \mathrm{pr}(Y_i(z_0) = y_0 | C_i = t_0, \underline{X}_i = x_0, Q_i = q_0, \theta).$$

Therefore, the conditional probability of the subject being a complier at this step is

$$l(c, Z_i, \underline{X}_i, Q_i, Y_i^{obs}, \theta) \{ l(c, Z_i, \underline{X}_i, Q_i, Y_i^{obs}, \theta) + l(n, Z_i, \underline{X}_i, Q_i, Y_i^{obs}, \theta) \}^{-1}.$$

The drawing of C_i for subjects with $Z_i = D_i^{obs} = 1$ is done in a similar way.

2. The missing potential outcome, $Y_i^{mis} = Y_i(1 - Z_i)$, of each person is drawn from the Bernoulli distribution with probability

$$\mathrm{pr}(Y_i(z') = 1 | C_i = t, \underline{X}_i, Q_i, Y_i(1 - z'), \theta) = \mathrm{pr}(Y_i(z') = 1 | C_i = t, \underline{X}_i, Q_i, \theta),$$

that is, the defining model (3·4), where z' is set to $1 - Z_i$.

3. The drawing of C_i^n is from $\mathrm{pr}(C_i^n | \underline{X}_i, Q_i, C_i, \theta)$. This distribution is the same as the defining model $\mathrm{pr}(C_i^n | \underline{X}_i, Q_i, \theta)$ but truncated either to the left or to the right of zero depending on C_i. The drawing of the truncated normal is done using its inverse distribution function, which is readily calculable. For subjects that, in the previous cycle of the algorithm, had been imputed as always-takers or compliers, the drawing of C_i^c is done in a similar way.

4. The drawing of Y_i^* is from $\mathrm{pr}(Y_i(z)^* | \underline{X}_i, Q_i, Y_i^{obs}, \theta)$, where z is set to Z_i. This distribution is the same as the defining model $\mathrm{pr}(Y_i(z)^* | \underline{X}_i, Q_i, \theta)$ except that it is truncated to the right or left of zero depending on Y_i^{obs}. The drawing is as with the compliance latent normals C_i^n and C_i^c.

5. The drawing of the coefficients $\alpha^{(C,1)}$ is from $\mathrm{pr}(\alpha^{(C,1)} | \{ \text{ all } C_i^n, \underline{X}_i, Q_i, b_{Q_i}^{(C,1)} \})$. This distribution is a Bayesian linear regression based on the defining likelihood and prior with offsets $\underline{W}_i' b_{Q_i}^{(C,1)}$ known at this step. The drawing of the coefficients $\alpha^{(C,2)}$ is from $\mathrm{pr}(\alpha^{(C,2)} | \{ \text{ all } C_i^c, \underline{X}_i, Q_i, b_{Q_i}^{(C,2)} : C_i = a \text{ or } c \})$, and the drawing of the coefficient $\alpha^{(Y)}$ is from $\mathrm{pr}(\alpha^{(Y)} | \{ \text{ all } Y_i^*, \underline{X}_i, Q_i, Z_i, C_i, b_{Q_i}^{(Y)} \})$, both of which are Bayesian linear regressions with offsets, respectively, $\underline{W}_i' b_{Q_i}^{(C,2)}$

22

and $f^{(2)}(\underline{W}_i, C_i) b_{Q_i}^{(Y)}$.

6. Cluster-specific parameters $b_q^{(Y)}$ are independently drawn for each cluster $q = 1, ..., M$ from $\mathrm{pr}(b_q^{(Y)}|\{$ all $Y_i^*, \underline{X}_i, Z_i, C_i : Q_i = q\}, \alpha^{(Y)})$ which is a Bayesian regression with off-sets $f^{(1)}(\underline{X}_i, C_i, Z_i)\alpha^{(Y)}$. Then, the prior mean and variance matrix for $\{b_q^{(C,1)}\}$ are adjusted to the conditional prior mean and variance matrix given the drawn values of $b_q^{(Y)}, q = 1, ..., M$ and used as prior distributions in the Bayesian regression $\mathrm{pr}(b_q^{(C,1)}|\{$ all $C_i^n, \underline{X}_i : Q_i = q\}, \alpha^{(C,1)})$, with offsets $\underline{X}_i' \alpha^{(C,1)}$, to draw $b_q^{(C,1)}$. The parameters $b_q^{(C,2)}$ are drawn analogously, except that the prior mean and variance matrix are not adjusted for the drawing of $b_q^{(Y)}, b_q^{(C,1)}$ because of the assumed prior independence.

7. The drawing of $\{D^{(1)}\}^{-1}$ is from the Wishart distribution with $M + \beta^{(1)}$ degrees of freedom and scale matrix $[\sum_q (b_q^{(Y)}, b_q^{(C,1)})(b_q^{(Y)}, b_q^{(C,1)})' + \beta^{(1)} R^{(1)}]^{-1}$ (Gelman et al., 1995). The drawing of $\{D^{(2)}\}^{-1}$ is from the Wishart distribution with $M + \beta^{(2)}$ degrees of freedom and scale matrix $[\sum_q b_q^{(C,2)} b_q^{(C,2)'} + \beta^{(2)} R^{(2)}]^{-1}$.

For each model, three chains were run. Independently among chains and across subjects, any unknown compliance principal stratum was initialized to a Bernoulli draw, conditionally on assignment arm, and with probabilities of compliance principal strata obtained from the sampling point estimates derived from Table 1. Using the initialized principal strata for each chain, parameter estimates were subsequently initialized based on generalized linear models estimates: for $\alpha^{(C,1)}$ and $\alpha^{(Y)}$, based on all subjects; for $\alpha^{(C,2)}$, based on the initialized set of compliers and always-takers; and for the physician-specific parameters, based on physicians with corresponding full rank design matrices. The initialized physician-specific parameters were used to set the values of $R^{(1)}$ and $R^{(2)}$, to represent, respectively, preliminary estimates of $D^{(1)}$ and $D^{(2)}$ (e.g., Wakefield et al., 1994). The matrices $D^{(1)}$ and $D^{(2)}$ were also initialized to the values of $R^{(1)}$ and $R^{(2)}$, respectively. Each chain was run for 25000 iterations. At 12500 iterations, and based on the three chains for each model, the potential scale reduction statistic (Gelman and Rubin, 1992) was computed for $\mathrm{ITT}^{(t)}, t = c, n, a$ giving $1.00, 1.02, 1.02$ respectively, suggesting no evidence against convergence. Inference for each model of Table 2

23

is based on the remaining 37500 iterations, combining the three chains.

REFERENCES

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Am. Statist. Assoc.* **91**, 444–72.

Armitage, P. (1998). Attitudes in clinical trials. *Statist. Med.*, **17**, 2675–83.

Baker, S. G. (1998). Analysis of survival data from a randomized trial with all-or-none compliance: estimating the cost-effectiveness of a cancer screening program. *J. Am. Statist. Assoc.* **93**, 929–34.

Baker, S. G. and Lindeman, K. S. (1994). The paired availability design: a proposal for evaluating epidural analgesia during labor. *Statist. Med.* **13**, 2269–78.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Am. Statist. Assoc.* **92**, 1171–6.

Barnard, J., Frangakis, C. E., Hill, J., and Rubin, D. B. (2001). School Choice in NY City: A Bayesian Analysis of an Imperfect Randomized Experiment. Forthcoming in *Case Studies in Bayesian Statistics* (with discussion), C. Gatsonis et al. (eds). New York: Springer-Verlag.

Bloom, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* **8**, 225-46.

Boyd, K., Teres, D., Rapoport, J., and Lemeshow, S. (1996). The relationship between age and the use of DNR orders in critical care patients. *Arch. Intern. Med.* **156**, 1821–6.

Cochran, W. (1963). Sampling techniques. New York: Wiley.

de Finetti, B. (1974). Theory of probability. New York: Wiley.

Dexter, P., Wolinsky, F., Gramelspacher, G., Zhou, X.-H., Eckert, G., Waisburd, M., and Tierney, W. (1998). Effectiveness of computer-generated reminders for increasing discussions about Advance Directives and completion of Advance Directives. *Ann. Intern. Med.* **128**, 102–10.

Duffield, P., and Podzamsky, J. E. (1996). The completion of Advance Directives in primary care. *J. Am. Statist. Assoc.* **42**, 378–84.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *J. Am. Statist. Assoc.* **68**, 117–30.

Frangakis, C. E. (1999). Coexistent complications with noncompliance with study-protocols, and implications for statistical analysis. PhD thesis (Part 2), Department of Statistics, Harvard University.

Frangakis, C. E. and Baker, S. G. (2001). Compliance sub-sampling designs for comparative research estimation and optimal planning. Forthcoming in *Biometrics*.

Frangakis, C. E., and Rubin, D. B. (2000). The defining role of "principal causal effects" in addressing post-treatment variables and surrogate endpoints in causal inference. In *Proc. Epidem. Sect., Am. Statist. Assoc.* (with discussion), pp. 23–33.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86**, 365–79.

Frangakis, C. E., Rubin, D. B., and Zhou, X.-H. (1998). The clustered-encouragement-design. *Proc. Biom. Sect., Am. Statist. Assoc.*, pp. 71–9.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. B. (1995). Bayesian data analysis. London: Chapman and Hall.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7**, 457–11.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **6**, 721–41.

Hakim, R. et al. (1996). Factors associated with do-not-resuscitate orders: patients' preferences, prognoses, and physicians' judgments. SUPPORT investigators. *Ann. Intern. Med.* **125**, 284–93.

Hansen, M. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.* **14**, 333-62.

Hartley, H. O., and Rao, J. N. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93-108.

Harville, D. A. (1976). Extensions of Gauss-Markov theorem to include the estimation of random effects. *Ann. Statist.* **4**, 384–95.

Hirano, K., Imbens, G., Rubin, D. B., and Zhou, X.-H. (2000). Estimating the effect of an influenza vaccine in an encouragement design. *Biostatistics*, **1**, 69–88.

Hughes, D. L. and Singer, P. A. (1992). Family physicians' attitudes toward advance directives. *Canadian Medical Association Journal* **146**, 1937–44.

Imbens, G. and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–76.

Imbens, G. W. and Rubin, D. B. (1994). Causal inference with instrumental variables. Discussion paper

1676. Cambridge, MA: Harvard Institute of Economic Research.

Imbens, G. W. and Rubin, D. B. (1997a). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25**, 305–27.

Imbens, G. W. and Rubin, D. B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies* **64**, 555–574.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), University of California Press, Berkeley, 361–79.

Korhonen, P, Loeys, T., Goetghebeur, E., and Palmgren, J. (2000). Vitamin A and infant mortality: beyond intention-to-treat in a randomized trial. *Lifetime Data Analysis* **6**, 107–21.

Laird, N. M. and Ware J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–74.

Liang, K–Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Manski, C. F. (1990). Non-parametric bounds on treatment effects. *Am. Econ. Rev., Papers & Proc.* **80**, 319–23.

Mark, S. D. and Robins, J. M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Statist. Med.* **12**, 1605–28.

McDonald, M., Hiu, S., and Tierney, W. (1992). Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *Clinical Computing* **21**, 1087–91.

Miles, S., Koepp, P. R, and Weber E. P. (1996). Advance end of life treatment planning. A research review. *Arch. Intern. Med.* **156**, 1062–8.

Neyman, J. (1934). On the different aspects of the representative method; a method of stratified sampling and a method of purposive selection. *J. R. Statist. Soc.*, A, **97**, 558–606.

Pearl, J. (1994). Causal inference from indirect experiments. In *Symposium Notes of the* 1994 *AAAI Spring Symposium on Artificial Intelligence in Medicine*, Stanford, CA.

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). pp. 113–59. Washington, DC: National Center for Health Services Research, U.S. Public Health Service.

Robins, J. M. and Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial. *J. Am. Statist. Assoc.* **89**,

737–49.

Robins, J. M. and Greenland, S. (1996). Comment on: "Identification of causal effects using instrumental variables" by Angrist, J. D., Imbens, G. W., and Rubin, D. B., *J. Am. Statist. Assoc.* **91**, 456–8.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–92.

Rubin, D. B. (1978). Bayesian inference for causal effects. *Ann. Statist.* **6**, 34–58.

Rubin, D. B. (1981). Using empirical Bayes techniques in the law school validity studies. *J. Am. Statist. Assoc.* **75**, 801–27.

Rubin, D. B. (1990). Comment on: "Neyman (1923) and causal inference in experiments and observational studies." *Statist. Sci.* **5**, 472–80.

Rubin, D. B. (1990). Comment on: "Compliance as an explanatory variable in clinical trials." by B. Efron, and D. Feldman. *J. Am. Statist. Assoc.* **86**, 9-26.

Rubin, D. B. (1998). More powerful randomization-based p-values in double-blind trials with noncompliance. *Statist. Med.* **17**, 371–87.

Schafer, J. L. (1996). Analysis of incomplete multivariate data. London: Chapman and Hall.

Sheiner, L. B. and Rubin, D. B. (1995). Intention-to-treat analysis and the goal of clinical trials. *Clin. Pharmacol. Therapeutics* **56**, 6–10.

Sommer, A. and Zeger, S. (1991). On estimating efficacy from clinical trials. *Statist. Med.* **10**, 45–52.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Statist. Assoc.* (with discussion). **82**, 528–50.

Wakefield, J., Smith, A. F. M., Racine-Poon, A., and Gelfand, A. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics* **43**, 201–21.

Wenger, N. S. et al. (1994). Prior capacity of patients lacking decision making ability early in hospitalization: implications for advance directive administration. The SUPPORT principal investigators. *Journal of General Internal Medicine* **9**, 539–43.

Wright, S. (1928). Appendix. In *The tariff on animal and vegetable oils* by P. G. Wright. Mcmillan, New York.

Zelen, M. (1979). A new design for randomized clinical trials. *N. Engl. J. Med.* **300**, 1242–5.

Table 1: *Patient characteristics in our sample from the study of Advance Directives. Estimates are based on sampling-theory ratio estimation for cluster sampling of finite population (Cochran, 1963, p. 30); (data modified from Frangakis and Baker, 2001, to report results of clustering for finite study population).*

	Control assignment		Encouragement assignment		Difference between assignments		
Doctors (no.)	26		25				
Patients (no.)	158		175				
Age	64.6	(0.6)	64.1	(0.4)	-0.5	(1.0)	$[-0.5]$
Discussing AD, %	5.1	(1.2)	25.7	(3.9)	21.6	(5.8)	$[3.7]$
Completing AD, %							
among discussants	62.5	(8.6)	51.1	(5.6)			
among non-discussants	0.0	–	1.5	(0.8)			
all	3.2	(0.9)	14.3	(3.2)	11.1	(4.6)	$[2.4]$

Estimates are mean (se). Numbers in brackets are ratios of estimated mean over standard error. Difference is encouragement minus control assignment.

Table 2: *Completion rates* (%) *of Advance Directives for compliance principal strata and assignment arms: (i) the model of Sec. 3.3 without covariates, and (ii) the model including age in both the compliance principal stratum and potential outcomes component. Reported results from the models are means (standard deviations) and* 95% *intervals* [2.5% and 97.5% quantiles] *of the corresponding posterior predictive distributions.*

AD Completion, %	Discussion-Compliers	Never-Discussants	Always-Discussants
Control assignment			
Model (w/o age)	3.1 (6.4)	0.2 (0.4)	62.0 (14.4)
	[0.0, 21.6]	[0.0, 1.6]	[31.7, 86.7]
Model (w/ age)	3.3 (4.0)	0.2 (0.4)	46.7 (12.2)
	[0.0, 13.7]	[0.0, 1.2]	[26.9, 71.6]
Encouragement assignment			
Model (w/o age)	47.9 (21.3)	1.9 (1.2)	47.9 (31.1)
	[4.5, 89.7]	[0.8, 5.1]	[0.0, 100.0]
Model (w/ age)	65.3 (9.9)	2.1 (1.2)	34.8 (9.7)
	[41.9, 80.9]	[0.8, 5.2]	[17.2, 54.2]
Difference between assignments			
Model (w/o age)	44.7 (22.4)	1.7 (1.4)	-14.1 (33.9)
	[0.0, 86.2]	[-0.4, 5.1]	[-68.9, 46.7]
Model (w/ age)	62.0 (11.2)	1.9 (1.3)	-11.8 (14.4)
	[34.7, 79.5]	[0.0, 4.9]	[-42.1, 14.7]

Table 3: *Prior distributions. Induced probabilities of completion* (%) *of Advance Directives for compliance principal strata and assignment arms using: (i) the prior distribution for the model of Sec. 3.3 without covariates, and (ii) the prior distribution of the model that included age. Reported results are means (standard deviations) and* 95% *intervals* [2.5% *and* 97.5% *quantiles*] *of the corresponding prior distributions.*

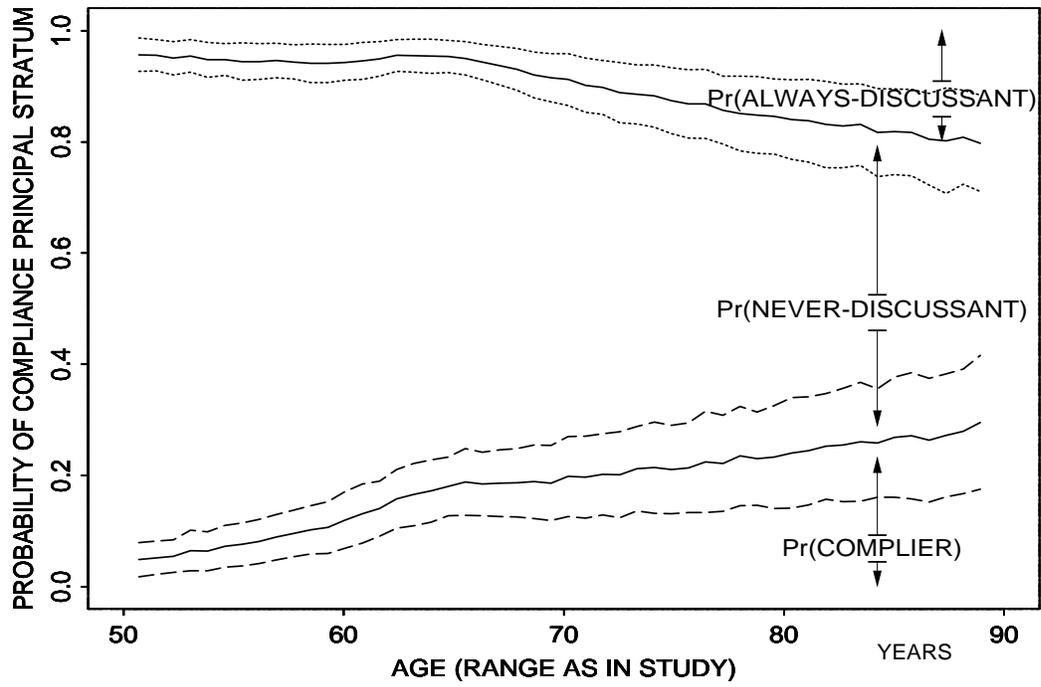| AD Completion, %	Discussion-Compliers	Never-Discussants	Always-Discussants
Control assignment			
Model (w/o age)	49.0 (50.0)	48.4 (50.0)	49.8 (50.0)
	[0.0, 100.0]	[0.0, 100.0]	[0.0, 100.0]
Model (w/ age)	50.3 (48.4)	48.7 (48.1)	53.5 (48.7)
	[0.0, 100.0]	[0.0, 100.0]	[0.0, 100.0]
Encouragement assignment			
Model (w/o age)	49.8 (50.0)	49.2 (50.0)	49.2 (50.0)
	[0.0, 100.0]	[0.0, 100.0]	[0.0, 100.0]
Model (w/ age)	50.2 (48.0)	49.0 (48.1)	52.5 (48.6)
	[0.0, 100.0]	[0.0, 100.0]	[0.0, 100.0]
Difference between assignments			
Model (w/o age)	0.9 (66.6)	0.8 (68.7)	-0.6 (63.9)
	[-100.0, 100.0]	[-100.0, 100.0]	[-100.0, 100.0]
Model (w/ age)	-0.1 (59.8)	0.8 (65.0)	-1.1 (52.2)
	[-100.0, 100.0]	[-100.0, 100.0]	[-100.0, 100.0]

Figure 1. Age and compliance principal stratum in the study on Advance Directive forms. Posterior predictive distribution of probabilities of being a discussion-complier (mean is height of lower solid curve), a never-discussant (mean is height between solid curves), and an always-discussant (mean is height between upper solid curve and 1.0). The two dotted lines around the lower (upper) solid curve are ± 1 posterior standard deviation of the probability of complier (always-discussant) (see also Sec. 4.2).