Published in final edited form as:

Clin Trials. 2013 ; 10(3): 363–377. doi:10.1177/1740774513479522.

Accommodating Missingness When Assessing Surrogacy Via Principal Stratification

Michael R. Elliott, PhD^{a,b}, Yun Li, PhD^a, and Jeremy M.G. Taylor, PhD^a

^aDepartment of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

^bSurvey Methodology Program, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, USA

Abstract

Background—When an outcome of interest in a clinical trial is late-occurring or difficult to obtain, surrogate markers can extract information about the effect of the treatment on the outcome of interest. Understanding associations between the causal effect of treatment on the outcome and the causal effect of treatment on the surrogate is critical to understanding the value of a surrogate from a clinical perspective.

Purpose—Traditional regression approaches to determine the proportion of the treatment effect explained by surrogate markers suffer from several shortcomings: they can be unstable, and can lie outside of the 0–1 range. Further, they do not account for the fact that surrogate measures are obtained post-randomization, and thus the surrogate-outcome relationship may be subject to unmeasured confounding. Methods to avoid these problem are of key importance.

Methods—Frangakis C, Rubin DM. Principal stratification in causal inference. *Biometrics* 2002; 58:21–9 suggested assessing the causal effect of treatment within pre-randomization “principal strata” defined by the counterfactual joint distribution of the surrogate marker under the different treatment arms, with the proportion of the overall outcome causal effect attributable to subjects for whom the treatment affects the proposed surrogate as the key measure of interest. Li Y, Taylor JMG, Elliott MR. Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* 2010; 66:523–31 developed this “principal surrogacy” approach for dichotomous markers and outcomes, utilizing Bayesian methods that accommodated non-identifiability in the model parameters. Because the surrogate marker is typically observed early, outcome data is often missing. Here we extend Li, Taylor, and Elliott to accommodate missing data in the observable final outcome under ignorable and non-ignorable settings. We also allow for the possibility that missingness has a counterfactual component, a feature that previous literature has not addressed.

Results—We apply the proposed methods to a trial of glaucoma control comparing surgery versus medication, where intraocular pressure (IOP) control at 12 months is a surrogate for IOP control at 96 months. We also conduct a series of simulations to consider the impacts of non-

ignorability, as well as sensitivity to priors and the ability of the Decision Information Criterion to choose the correct model when parameters are not fully identified.

Limitations—Because model parameters cannot be fully identified from data, informative priors can introduce non-trivial bias in moderate sample size settings, while more non-informative priors can yield wide credible intervals.

Conclusions—Assessing the linkage between causal effects of treatment on a surrogate marker and causal effects of a treatment on an outcome is important to understanding the value of a marker. These causal effects are not fully identifiable: hence we explore the sensitivity and identifiability aspects of these models and show that relatively weak assumptions can still yield meaningful results.

Keywords

Causal Inference; Surrogate Marker; Bayesian Analysis; identifiability; Non-response; Counterfactual

Introduction

Given the time required to obtain clinical endpoints such as survival, there is interest in using surrogate endpoints such as disease-free survival at early follow-up periods [1] or biomarkers such as CD4 counts for AIDS [2] to assess the effectiveness of a treatment regime in clinical trial settings. The demand for “surrogate markers” in clinical research has led to the development of a large number of statistical methods to evaluate the effectiveness of such measures [3]. Prentice’s [4] foundational paper defined “perfect” surrogacy in the hazard regression setting as occurring when an outcome T is independent of treatment Z conditional on the surrogate measure S . Freedman, Graubard, and Schatzkin [5] pointed out that this is an extremely strict criterion rarely met in practice, and suggested as an alternative measure in the absence of a surrogate-treatment interaction the proportionate reduction of the treatment effect on a binary outcome when the surrogate measure is included in a logistic regression model. This “proportion-explained” estimator itself had shortcomings, including the fact that it can be unstable, and can lie outside of the 0–1 range [6] [7]. Thus a meta-analytic approach to assess surrogacy has been suggested by Buyse et al. [8], who distinguish between trial-level and individual-level surrogacy, and use random effects models to assess the variance of a predicted association between treatment and outcome in a new trial at both the individual and trial level when only the effect of treatment on a surrogate has been measured. Small variances in these associations (i.e., large coefficients of determinations) at both the individual and trial level are indicative of a good surrogate measure.

An alternative approach to assessing surrogacy uses causal inference, with the goal of understanding causal links between treatment, surrogate, and outcome. Traditional regression models that condition on surrogacy measures to assess the fraction of the treatment effect explained cannot be viewed as causal if there are confounders between the surrogate marker and the final outcomes, since the surrogate marker is observed post-randomization [9]. Thus Robins and Greenland [10] define direct and indirect effects in

mediation analysis in the potential outcomes framework. In this framework, we assume that we can in principle observe the values of the outcome under all possible treatment assignments for a given individual. The targets of inference become the differences in the values of the potential outcomes under different treatment assignments within the individual, averaged over the population. Assuming that the surrogate marker can be manipulated independently from the outcome, Robins and Greenland [10] define (natural) direct effects as the expected value of the difference in the potential outcomes under different treatment assignments when the value of the marker is held constant, and indirect effects as the expected difference in the potential outcomes under treatment when the marker is changed to the value it would have taken under treatment and under control. Assumptions required to obtain identified estimates of direct and indirect effects such as randomization, monotonicity (the treatment is never harmful), and no treatment-mediator interactions can be relaxed in part by sensitivity analyses [11] [12] [13]. An alternative “principal stratification” approach to assess surrogacy was proposed by Frankgakis and Rubin [14]. Principal strata are defined by the joint potential outcomes of the surrogate marker, thus forming a “pre-randomization” variable that can be conditioned on while retaining causal interpretations of randomized treatment effects. The causal effects of interest become the differences in the potential outcomes under treatment and control within the strata in which the surrogate changes as a result of the treatment assignment. This approach has been explored for binary outcomes in more detail by Gilbert and Hudgens [15] and Li, Taylor, and Elliott [16], with the former considering categorical and continuous (possibly subject to censoring) markers and focusing on the setting where the marker under the control is fixed and known (allowing identification of model parameters), and the latter on the setting where the marker is also binary (generally allowing identification of parameter boundaries only). Joffe and Green [17] provide a summary of both the standard regression-based and the two major causal approaches (which they term respectively that causal effects [CE] and causal association [CA] paradigms), showing that some of the models developed under the CE paradigm can be recast as CA models, and vice versa.

At the risk of oversimplification, much of the argument between these two approaches revolves around a tradeoff between assuming manipulable variables in the causal pathway between treatment and outcome to allow inference about true causal mechanisms (the CE approach) and sacrificing that mechanistic interpretation for conceptualizations that may be more congruent with the observable data (the CA approach). Recognizing the lively debate about the relative merits of the CE versus CA approaches (see, for example, [18], [19], or [20]) in various settings, we believe the key aspect in surrogacy is the causal association. We need to know that the causal effect of the treatment on the outcome and the causal effect of the treatment on the marker are “in sync”, so that the marker can predict or replace the outcome in treatment effect estimation. Thus we proceed with the principal stratification (CA) approach in this manuscript, extending the work of Li, Taylor, and Elliott [16] to accommodate missing data in the (observable) outcome measure.

Such missing data is a common occurrence in clinical trials since the value of the surrogate variable is typically to provide information in advance of the outcome measure of interest. We utilize the machinery of the missing data literature [21], focusing on developing a nonignorable missing mechanism for the final outcome of interest on the assigned treatment

arm that is based on the assumption of latent ignorability [22] [23] [24] [25]. Latent ignorability assumes an ignorable missingness mechanism *conditional* on membership in a not fully-observable stratum – here the principal strata based on the joint values of the surrogate marker under the differing treatment assignments. In contrast to much of the previous work that assumed latent ignorability in the context of non-compliance in randomized clinical trials, we cannot make “exclusion restriction” assumptions to force identifiability, since such an assumption would “assume” the very quantity we are trying to estimate, the degree to which subjects who had no causal impact of treatment on the surrogate outcome have a causal impact on the true endpoint of interest. Instead, we have an extra complication in that the parameters of interest are not fully identifiable even in the absence of missing outcomes, since only the marker and the outcome under the actual treatment assignment are observed. We also allow for the possibility that missingness has a counterfactual component, one that might differ between the treatment and control due to differential dropout, a feature that previous literature has not addressed to our knowledge.

The manuscript proceeds as follows. Section 2 develops surrogacy assessment using principal stratification, including the definitions of surrogacy measures of interest. Section 3 extends the principal stratification model for surrogacy to account for missing data using both latent ignorable and missing completely at random (MCAR) models, and discusses the use of the decision information criterion (DIC) measure to choose between the models. Section 4 considers the reduction in bias and mean square error when missing data methods are employed via a simulation study. Section 5 applies the proposed methods to a study of intraocular pressure (IOP) in glaucoma patients, using early measures of IOP as a surrogate marker for later measures of IOP to assess the effect of surgical vs. drug treatment. We conclude with a summary discussion in Section 6.

Assessing Surrogacy Via Principal Stratification

Notation

We denote treatment assignment by Z_l , the potential outcome for the surrogate under each of the treatment assignments for the l th subject by $S_l(Z_l)$, and the potential outcome for the true endpoint under each of the treatment assignments by $T_l(Z_l)$. We assume the surrogate is fully observed but the true endpoint under the assigned treatment arm is missing (for example, due to insufficient follow-up time or dropout), and denote $R_l(Z_l) = \{0, 1\}$ corresponding to missing and observed true endpoints under each of the treatment assignments respectively. Assuming dichotomous treatment assignments, surrogate markers, and true endpoints, the support for the joint distribution of the potential outcomes of the true endpoints is given by $\{(0, 0), (0, 1), (1, 1), (1, 0)\}$, corresponding respectively to failure under both arms, failure under control and success under treatment, success under both arms, and success under control and failure under treatment. Potential surrogate markers and responses have similar support corresponding to success/failure or observed/missing associated with each treatment arm. We denote the probability of a subject belonging to a cell in the resulting $4 \times 4 \times 4$ contingency table by $P((S(0), S(1)) = i, (T(0), T(1)) = j, (R(0), R(1)) = k) = \pi_{ijk}$, where $i, j, k = \{1, 2, 3, 4\}$, corresponding to the 4 support points (see Table 1). Thus the “complete data” are $\{(S(0), S(1))_l, (T(0), T(1))_l, (R(0), R(1))_l, l = 1, \dots, n\}$. The observed data for the l th

subject is given by (z_l, r_l, s_l, t_l) , where $r_l = R(Z_l = z_l)$, $s_l = S(Z_l = z_l)$, and

$$t_l = \begin{cases} T(Z_l = z_l) & \text{if } r_l = 1 \\ . & \text{if } r_l = 0 \end{cases} \text{ where '.' indicates a missing value.}$$

Surrogacy Measures of Interest

The principal strata correspond to the categories associated with the distribution of the potential surrogate markers, with (0,0) termed never responsive, (0,1) responsive, (1,1) always responsive, and (1,0) harmed where without loss of generality 0 corresponds to a “poor health” surrogate marker and 1 to a “good health” surrogate marker. Under the assumption of “monotonicity,” corresponding to no one harmed with respect to either the surrogate or outcome, the overall causal effect of treatment (CE) is given by $E(T_l(1) - T_l(0)) = \pi_{+2} = \pi_{12} + \pi_{22} + \pi_{32}$. (Here and below, we denote $\pi_{ij} \equiv \pi_{ij+} \equiv \sum_k \pi_{ijk}$, the joint distribution of the potential surrogate marker and potential outcome marginalized across the missingness patterns.) Frangakis and Rubin [14] proposed associative and dissociative effects corresponding respectively to the fraction of patients on which the treatment changed both the surrogate marker and the final outcome $AE = \pi_{22}$ and the fraction of patients on which the treatment changed the surrogate marker but not the final outcome $DE = \pi_{12} + \pi_{32}$. A good surrogate will have a large AE, indicating that the causal effect on the surrogate is highly associated with a causal effect on the final outcome of interest. Similarly, a good surrogate will have a small DE, indicating that the causal effect on the outcome is small when the causal effect on the surrogate is zero. (Note that, since the surrogate and outcome are not manipulated separately in the principal stratification framework, associative effects cannot be interpreted as indirect effects from the treatment through the surrogate as defined by [10].) Because AE and DE are constrained to sum to CE, interpreting “large” and “small” is perhaps most useful relative to the CE, hence Taylor, Wang, and Thiebaut [11] defined associative proportions $AP = AE/CE$ and dissociative proportions $DP = DE/CE$ as the fraction of the overall treatment effect partitioned between the associative and dissociative effects. Li, Taylor, and Elliott [16] proposed another surrogacy measure, common associative

proportion (CAP) = $\frac{\pi_{22}}{\pi_{12} + \pi_{21} + \pi_{22} + \pi_{23} + \pi_{32}}$, suggested by the concept of “perfect surrogacy” [14] which would occur when there is no causal effect on T unless there is also a causal effect on S (i.e., $\pi_{12} = \pi_{32} = 0$). Without the monotonicity assumption, $CE = \pi_{+2} - \pi_{+4}$ (the net treatment effect corresponding to the fraction responsive to the treatment minus the fraction harmed), $AE = \pi_{22} + \pi_{42} - (\pi_{24} + \pi_{44})$ (net treatment effect on patients whose surrogate was responsive to treatment), and $DE = \pi_{12} + \pi_{32} - (\pi_{14} + \pi_{34})$ (net treatment effect on patients whose surrogate was not responsive to treatment) [26], and AP and DP are unchanged. The CAP does not have a clear analog when monotonicity is relaxed, although larger values of π_{22}/π_{2+} compared with π_{12}/π_{1+} , π_{32}/π_{3+} , and π_{42}/π_{4+} would generally be required for a good surrogate [26].

Principal Stratification Model for Surrogacy Accounting for Nonresponse in the True Endpoint

Model assumptions

Factoring the joint distribution of the complete data (i.e., assuming we could observe treatment assignments, surrogate markers, true endpoints, and response behavior on both treatment arms), we obtain

$$p(T(\mathbf{Z}), S(\mathbf{Z}), R(\mathbf{Z})|\mathbf{Z}) = p(T(\mathbf{Z})|S(\mathbf{Z}), R(\mathbf{Z}), \mathbf{Z})p(S(\mathbf{Z}), R(\mathbf{Z}), |\mathbf{Z})p(R(\mathbf{Z})|\mathbf{Z}))$$

for $T(\mathbf{Z}) = (T_1(\mathbf{Z}), \dots, T_n(\mathbf{Z}))$ where $T_l(\mathbf{Z})$ refers to the set of potential outcomes for the l th subject associated with all possible treatment assignments $\mathbf{Z}$ in the sample, and similarly for $S(\mathbf{Z})$ and $R(\mathbf{Z})$.

We make the following three assumptions throughout the remainder of this manuscript:

1. Randomization: treatment assignment is made independently of the potential outcomes for the surrogate markers, so that

$$p(T(\mathbf{Z}), S(\mathbf{Z}), R(\mathbf{Z})|\mathbf{Z}) = p(T(\mathbf{Z}), S(\mathbf{Z}), R(\mathbf{Z}))$$

2. Stable Unit Treatment Assignment [27]: treatment assignment for subject i is independent of $(S_j(Z_j), T_j(Z_j), R_j(Z_j))$ for $j \neq i$, so that

$$p(T(\mathbf{Z}), S(\mathbf{Z}), R(\mathbf{Z})) = \prod_l p(T_l(Z_l), S_l(Z_l), R_l(Z_l));$$

also the observed surrogate marker is equal to the potential outcome under the observed treatment arm ($s_l = z_l S_l(z_l) + (1 - z_l)S_l(0)$), and similarly for T_l .

3. Latent ignorability of missing data [22] [23] [24] [25]: conditional on the joint distribution of the surrogate markers under both treatment assignments $S_l(Z_l)$, the joint distribution of the true endpoint under both treatment assignments $T_l(Z_l)$ is independent of the counterfactual missingness pattern $R_l(Z_l)$. Thus we have

$$p(T_l(Z_l), S_l(Z_l), R_l(Z_l)) = p(T_l(Z_l)|S_l(Z_l), R_l(Z_l))p(S_l(Z_l), R_l(Z_l)) = p(T_l(Z_l)|S_l(Z_l))p(S_l(Z_l), R_l(Z_l)).$$

We also consider a missing completely at random (MCAR) model in which missingness is independent of both the surrogate and the true outcome:

$$p(T_l(Z_l), S_l(Z_l), R_l(Z_l)) = p(T_l(Z_l)|S_l(Z_l))p(S_l(Z_l))p(R_l(Z_l))$$

Note also that we do not make the compound exclusion restriction (CER). Under CER, $S_l(0) = S_l(1)$ implies that $R_l(0) = R_l(1)$ and $T_l(0) = T_l(1)$ [22] [23]. Doing so would imply that subjects who had no causal impact of treatment on the surrogate outcome would have no causal impact of treatment on either the true endpoint of interest or on their response

behavior, thereby assuming away key issue that we would like the data to speak to in our analysis.

We also consider the models with and without the monotonicity assumption for surrogate marker and true endpoint. Under monotonicity, subjects will never have a negative causal effect of treatment, so that $P(S_l(0) = 1, S_l(1) = 0) = P(T_l(0) = 1, T_l(1) = 0) = 0$. Here we also consider a more limited form of monotonicity, which we term “stochastic monotonicity,” that only assumes the treatment is more likely to be helpful than harmful ($\pi_{2++} > \pi_{4++}$) for the surrogate measures, and that, within the non-harmful principal strata, the treatment is more likely to be helpful than harmful ($\pi_{j2+} > \pi_{j4+}, j = 1, 2, 3$) for the final outcome [28].

Finally, the counterfactual missingness patterns might be restricted by knowledge about the missingness mechanism. For example, if dropout is entirely due to administrative censoring, it might be reasonable to assume that the missingness value observed on the assigned treatment arm might be equivalent to the missingness value on the unassigned arm, so $R(Z) \in \{(0, 0), (1, 1)\}$ (i.e., $\pi_{++2} = \pi_{++4} = 0$) and thus the missingness pattern is fully observed. We do not pursue these restrictions further.

Model Estimation

Latent Ignorability Model—Factoring

$$P((S(0), S(1))=i, (T(0), T(1))=j, (R(0), R(1))=k) = \pi_{ijk} \text{ as}$$

$$P((T(0), T(1))=j | (S(0), S(1))=i, (R(0), R(1))=k) P((S(0), S(1))=i, (R(0), R(1))=k) = \pi_{j|i} \pi_{i+k}$$

we have under the latent ignorability assumptions that $\pi_{jik} \equiv \pi_{j|i} \pi_{i+k}$ for all k , reducing the number of free parameters in the complete data from 63 to 27: 12 free parameters for $\pi_{j|i}$, and 15 free parameters for π_{i+k} . (Under monotonicity, $\pi_{j=4|i} = 0$ for all i and $\pi_{4+k} = 0$ for all k , further reducing the number of free parameters in the complete data to 17.)

However, there are only 10 sufficient statistics in the observed data: 6 for the observed $S \times T$ tables stratified by Z when $R = 1$, 2 for the observed S stratified by Z when $R = 0$, and 2 for the observed $R \times Z$ table. Hence we use a fully Bayesian approach to cope with the non-identifiability in the observed data likelihood [29] [30] [31].

The complete data likelihood is given by

$$\prod_i \prod_j \prod_k \pi_{ijk}^{n_{ijk}} = \prod_i \prod_j \pi_{j|i}^{n_{ij+}} \prod_i \prod_k \pi_{i+k}^{n_{i+k}}$$

We assume a Dirichlet prior for the cell probabilities:

$$p(\pi_{j|i}) \sim \text{DIR}(a_{j|i})$$

$$p(\pi_{i+k}) \sim \text{DIR}(b_{i+k})$$

Details of the Gibbs sampler are given in the Appendix.

Missing Completely at Random (MCAR) Model—Under the missing completely at random assumption, $p((S(0), S(1)) = i, (R(0), R(1)) = k) = p((S(0), S(1)) = i)p((R(0), R(1)) = k)$ and thus $\pi_{ijk} = \pi_{j/i}\pi_{i++}\pi_{++k} = \pi_{jj+}\pi_{++k}$. The distribution of R is thus independent of S and T and thus can be ignored in both the data augmentation step and the draw of the parameters conditional on the complete data for estimation of the surrogacy effects of interest (although required to obtain model fit estimates discussed in Section 3.3). We again assume a Dirichlet prior for the cell probabilities:

$$p(\pi_{ij+}) \sim \text{DIR}(a_{ij+})$$

$$p(\pi_{++k}) \sim \text{DIR}(b_{++k})$$

Details of the Gibbs sampler are given in the Appendix.

Choosing between the Missingness Mechanisms

To choose between the latent ignorable and MCAR missingness mechanisms, we can compute the Deviance Information Criterion (DIC) of Spiegelhalter et al. [32]. The DIC measure accounts for the fact that, in a hierarchical framework, the number of effective parameters may be unclear: the random effects associated with each subject may “count” as approximately one parameter if the between-variance estimates are large (small degree of shrinkage), and as nearly zero parameters if the between-variance estimates are small (large degree of shrinkage). DIC estimates the number of effective parameters by

$pD = \overline{D(\pi)} - D(\bar{\pi})$ where $\overline{D(\pi)} = E_{\pi|y} D(\pi)$ and $D(\bar{\pi}) = D(E(\pi | y))$ for the observed likelihood deviance $D(\pi)$ given in the Appendix. The DIC measure is then given by $D(\bar{\pi}) + 2pD$. Although we do not entertain a Bayesian hierarchical model here, we have a similar issue in that the number of parameters is unclear given that not all are fully identified. We will assess the effectiveness of DIC in choosing between the latent ignorable and MCAR missingness mechanisms via simulation.

Simulation Study

Study Design

Although our estimation approach is Bayesian, assessing its frequentist properties is important for model assessment [33]. Hence we conduct a simulation study using a $2 \times 2 \times 2$ design, where the surrogate is either poor or good as measured by the associative and common associative proportions, where the principal strata are either independent of or associated with missingness (corresponding to ignorable and non-ignorable missingness mechanisms), and where missingness is either independent of treatment or more likely if assigned to treatment. For each simulation, we fit both the latent ignorable model and the MCAR model as well as a model that uses the complete cases only, and compute the bias and root mean square error of the posterior mode, nominal 90% coverage, and mean 90%

credible interval length, for the causal effect, associative proportion, and common associative proportion. We use 10,000 draws from an MCMC chain after 1,000 draws for burn-in. Each simulation contains 500 observations, and 200 simulations are run for each scenario in the design. The missingness mechanism is designed to provide approximately 50% missing for each scenario. For each scenario, we also compute the DIC under both the latent ignorable and MCAR model. We restrict our simulations to the monotonicity setting to minimize the complexity of the simulation design and focus on the properties of interest.

Table 2 provides details of the simulation design. Under the ignorable model, the principal strata are independent of missingness. Under the non-ignorable model, the odds of a subject belonging to an “always responsive” surrogate marker principal strata versus a “never responsive” surrogate marker principal strata are approximately 4 times greater when the outcome is observed under control than when it is missing under control. The marginal distribution of the missingness patterns is defined by $\pi_{++1} = \dots = \pi_{++4} = .25$ when missingness is independent of treatment, and $\pi_{++1} = .15$, $\pi_{++2} = .15$, $\pi_{++3} = .15$, $\pi_{++4} = .55$, when missingness is more likely under treatment.

We assume two forms of the prior: a uniform Dirichlet prior, with $a_{1|1} = \dots = a_{3|3} = 1$ and $b_{1+1} = \dots = b_{4+4} = 1$ for the latent ignorable model and $a_{11+} = \dots = a_{33+} = 1$ and $b_{++1} = \dots = b_{++4} = 1$ for the MCAR model, and a “Jeffreys-type” prior with $a_{1|1} = \dots = a_{3|3} = 1/2$ and $b_{1+1} = \dots = b_{4+4} = 1/2$ for the latent ignorable model and $a_{11+} = \dots = a_{33+} = 1/2$ and $b_{++1} = \dots = b_{++4} = 1/2$ for the MCAR model.

Results

Table 3 shows the results from data generated under the non-ignorable simulation, while Table 4 shows the results from data generated under the ignorable simulation. We summarize the results as follows:

- When the design is nonignorable, using the MCAR model leads to increased bias and root mean square error (RMSE) relative to the latent ignorable model for all three estimators considered. This is true even for the CE estimator, which is fully identified. This impact of model misspecification is strongest on CE for the “good” surrogate (large AP/CAP) and on AP for the “poor” surrogate (small AP/CAP).
- Using the latent ignorable model has little impact on bias or RMSE when design is ignorable.
- When the design is non-ignorable, the CE estimator has reduced bias and improved coverage under either of the missing data models relative to the complete-case analysis, with this difference being stronger when the surrogate is poor than when it is strong. Under the ignorable design, use of the latent ignorable model yields a slightly greater MSE than the complete-case analysis, although coverage is also improved.
- Coverage for AP and CAP tends to be conservative because of partial identification, use of the latent ignorable model helps ensure coverage does not become anticonservative. Coverage for AP and CAP is poor for the complete-case

analysis, especially under non-ignorable missingness with a good surrogate.¹ Although the CE is fully identified in the absence of missing data, the need to rely on incompletely identified parameters means that the 90% confidence interval under the missing data models is actually wider than an analysis using the fully-observed data only, with the coverage becoming highly conservative when the surrogate is poor.

- Lack of identification induces bias in the AP and CAP estimators when the surrogate is poor because posterior distributions are relatively at. Use of less-informative prior reduces bias, at the cost of increased RMSE and CI length.
- DIC always picks the latent ignorable model when correct. DIC picks the MCAR model the majority of the time (60–80%) when correct, with high values of AP/CAP allowing the MCAR model to be selected with greater probability when it was correct.

Application to the Collaborative Initial Glaucoma Treatment Study

We apply the missing data models to an analysis of the Collaborative Initial Glaucoma Treatment Study (CIGTS) [35]. Glaucoma is an eye disease caused by increased intraocular pressure (IOP) that can result in reduced vision or blindness. The CIGTS was a clinical trial that compared the effects of eye surgery (treatment) against the standard practice of medication (control) to reduce or stop visual field loss. Because visual field loss is caused by increased IOP, one of the major secondary outcomes of interest is reduction in IOP. Here we consider reduction in IOP below 18mmHg after 96 months of follow-up, based on previous work that has shown IOP of less than 18mmHg at every time point during at least six years of follow-up was associated with a reduced likelihood of visual field loss [36]. Because of the extensive follow-up time, we wish to determine if early reductions in IOP could serve as a marker for much late reductions, hence the surrogate marker was reduction in IOP below 18mmHg after 12 months of follow-up. However, such an analysis suffers from a substantial amount of missing outcome data due to the long follow-up period. Because the cause of the missingness is due to dropout for unknown reasons, we do not restrict the missingness patterns in the analysis.

The observed data is given in Table 5. Of 574 subjects with intraocular pressure (IOP) measured at 12 months, only 228 had fully observed data (IOP also measured at 96 months). For fully-observed subjects in the control arm (drug only), 66.7% had reduced IOP to below 18 mmHg at 96 months, 79.4% of fully-observed subjects on the treatment arm had reduced IOP at 96 months, giving an estimated causal effect of treatment (CET) of .127 (90% CI .030,.224). Reduced IOP at 12 months was observed for 58.4% of subjects who were fully observed on the control arm, versus 54.8% of subjects who did not have 96 month IOP

¹As noted in Gustafson and Greenland [34], Bayesian credible intervals have by definition correct coverage regardless of the sample size and model identification when the parameters are generated under the assumed prior. However, because non-identified models converge to regions of 0 and non-zero posterior probability rather than degenerate atoms [31], if we take a frequentist perspective where the true values of the data are fixed, the resulting credible intervals will have frequentist coverage properties that will approach 1 if the data-generating model is correct unless the true value of the interval is on the boundary of the non-zero posterior probability region [28].

measures. For subjects on the treatment arm, 80.4% of fully observed subjects had reduced IOP at 12 months, versus 81.4% of subjects without 96 month IOP measures.

We fit a MCAR model and a latent ignorable model, as well as a model for the fully-observed data, under the monotonicity and non-monotonicity assumption, as well as the “stochastic monotonicity” assumption that only assumes the treatment more likely to be helpful than harmful. Each model is fit using a single chain of 100,000 draws after a burn-in of 1,000. We consider uniform priors using the Dirichlet distribution with $a_{j|i} = 1$ for all i, j and $b_{i+k} = 1$ for all i, k for the latent ignorable model and $a_{ij+} = 1$ for all i, j and $b_{++k} = 1$ for all k for the MCAR model. We assess sensitivity to the prior by considering a Jeffreys-type prior using the Dirichlet distribution with $a_{1|1} = \dots = a_{4|4} = 1/2$ and $b_{1+1} = \dots = b_{4+4} = 1/2$ and $a_{11+} = \dots = a_{44+} = 1/2$ and $b_{++1} = \dots = b_{++4} = 1/2$. Results are given in Table 6. Table 7 provides the DIC measures for the latent ignorable and MCAR models.

Based on DIC, the best fit is provided by the latent ignorable model under stochastic monotonicity, similar fit is provided by the MCAR model under monotonicity and the uniform Dirichlet prior. Particularly poor fit is evidenced by the MCAR model under non-monotonic assumptions, as evidenced by the discrepancy between the CE estimator from the model and the identifiable estimate (.127) obtained from the fully-observed data, use of the Jeffreys-type prior improved the fit to some degree. The best-fitting latent ignorable model under stochastic monotonicity and MCAR model under monotonicity had little sensitivity to the prior assumptions, and gave broadly similar results. In particular, early reduction of IOP appears to be at best a modestly useful surrogate marker from a causal perspective, with only a weak association between the 8-year causal effect on IOP and the 12-month causal effect on IOP. The point estimate of the associative proportion is somewhat greater under the stochastic monotonicity assumption than under the full monotonicity assumption, although the 90% credible interval can and does include 0, indicating some evidence of interactions in the causal effect outcomes within the principal strata [28]. The best model fit was obtained under the latent ignorable model with the stochastic monotonicity assumption and the Jeffreys-type prior, Figure 1 shows the associated posterior distributions of the CE, AE, and AP under this model.

The overall degree of model fit is assessed via the posterior predictive distribution of functions T of the observed values of y_l given by

$$p(T(\mathbf{y}_l^{rep})|\mathbf{y}) = \int p(T(\mathbf{y}^{rep})|\pi, \mathbf{y})p(\pi|\mathbf{y})d\pi \quad (1)$$

We obtain draws from (1) and compare the predictive distribution $T(\mathbf{y}^{rep})$ with the observed values of $T(\mathbf{y}^{obs})$ [37]. In particular, we consider the eight cell counts of surrogate marker and outcome by treatment assignment given by $m_{ijz} = \sum_l I(s_l = i, t_l = j | z_l = z)$ for $i, j, z \in \{0, 1\}$; histograms of m_{ijz}^{rep} compared against m_{ijz}^{obs} in Figure 2 show that the predictive distributions are centered near the observed values, showing that the model is reasonable given the data.

In general the lack of large differences in the surrogate marker treatment effects between the outcome response categories suggests that failing to account for missingness in the outcome

should have modest impact, and indeed this is the case where model fit is reasonable (under monotonicity, and in the non-monotonic models with the Jeffreys-type priors). Estimates for the CE, AE, and AP appear to be smaller in the fully observed data than in the models that incorporate missing outcome data, consistent with the mechanisms that lead to negative bias for these estimates in the simulation setting. Intervals for models that incorporate missing outcome data are somewhat wider than intervals that discard this data, consistent with the fact that accounting for missingness requires estimation of not fully-identified parameters.

Discussion

This manuscript considers a principal stratification approach to assess surrogacy for dichotomous markers and outcomes when missing data is present for the outcome, extending the work of Li, Taylor, and Elliott [16] to accommodate missingness under latent ignorability assumptions. The principal strata are defined by the joint distribution of the surrogate marker under both treatment and control, with the quality of the surrogate being the causal effect of treatment that is associated with strata in which the surrogate marker is impacted by treatment. Latent ignorability assumes that the conditional distribution of the potential outcomes within the principal strata are independent of the outcome missingness but allows for the possibility that the marginal distribution of the surrogate marker or treatment outcome is associated with outcome missingness. This is a weaker assumption than missing completely at random in which missingness is independent of both the surrogate and the true outcome, and is identifiable at the “complete data” (counterfactual) level, in contrast to a fully non-ignorable model which would require postulating non-identified parameters for the unobserved outcomes at this complete-data level. A unique aspect of our approach (to our knowledge) is that we allow for the possibility that missingness has a counterfactual component, one that might differ between the treatment and control due to differential dropout.

Alternatives to the latent ignorability missingness mechanisms can be considered as well. A special case of latent ignorability would be a “fully ignorable” missingness mechanism in which missingness depends only on observed data, so that the complete data could be factored as $P((T(0), T(1)) | (S(0), S(1)))P((S(0), S(1)))P((R(0), R(1)) | S(z))$, where z is the observed treatment arm. Chen, Geng, and Zhou [24] propose a “complete-nonignorability” model which is identified under the complete data using the selection model decomposition, replacing $P(R_I(Z_I) | T_I(Z_I), S_I(Z_I)) = P(R_I(Z_I) | S_I(Z_I))$ under latent ignorability with $P(R_I(Z_I) | T_I(Z_I), S_I(Z_I)) = P(R_I(Z_I) | T_I(Z_I))$ to yield $p(T_I(Z_I) | S_I(Z_I))p(S_I(Z_I), R_I(Z_I)) = p(R_I(Z_I) | T_I(Z_I))p(T_I(Z_I), S_I(Z_I))$.

There are two major limitations in this approach that result from lack of full identifiability. First and foremost, we suffer from sensitivity to prior assumptions even in large datasets, since the posteriors do not converge in distribution to normal distributions centered at their likelihood modes. Second, posterior credible intervals are typically fairly wide and cannot shrink beyond asymptotic boundary conditions. Thus even mildly informative priors can introduce non-trivial bias in moderate sample size settings, while more “non-informative” priors can yield credible intervals of very large widths. Hence this manuscript has focused on exploring the sensitivity and identifiability aspects of these inherently non-identified

models. We also note that avoiding these limitations by using fully identified models requires making stronger assumptions that are not warranted by the inferential question we are addressing.

A variety of extensions to this work can be considered. Our focus is on assessing and ameliorating the effect of missingness on inference about binary surrogate measures and outcomes in a counterfactual setting, and in the process we have focused on a relatively simple Dirichlet prior formulation for the cell parameters. The work of Li, Taylor, and Elliott [16] and Li et al. [26] considered a log-linear parameterization that was capable of incorporating a priori assumptions about positive correlations between the surrogate marker and final outcome in a more refined fashion than the model considered here, particularly when the monotonicity assumption is relaxed. Extensions to continuous surrogate measures and outcomes are also possible and are the focus of current work.

Acknowledgments

This research was supported by National Institute of Mental Health Grant R01MH-078016 and National Cancer Institute Grant R01CA-129102.

References

1. Chen TT, Simon RM, Korn EL, et al. Investigation of disease-free survival as a surrogate endpoint for survival in cancer clinical trials. *Commun Stat: Theor Meth*. 1998; 27:1363–1378.
2. Lin DY, Fischl MA, Schoenfeld DA. Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in Human Immunodeficiency Virus clinical trials. *Stat Med*. 1993; 12:835–842. [PubMed: 8101011]
3. Burzykowski, T.; Molenberghs, G.; Buyse, M. *The Evaluation of Surrogate Endpoints*. New York: Springer-Verlag; 2005.
4. Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat Med*. 1989; 8:431–440. [PubMed: 2727467]
5. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med*. 1992; 11:167–178. [PubMed: 1579756]
6. Buyse M, Molenberghs G, Burzykowski T, et al. Statistical validation of surrogate endpoints: problems and proposals. *Drug Information Journal*. 2000; 34:447–454.
7. Molenberghs G, Buyse M, Geys H, et al. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials*. 2002; 23:607–625. [PubMed: 12505240]
8. Buyse M, Molenberghs G, Burzykowski T, et al. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*. 2000; 1:49–67. [PubMed: 12933525]
9. Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J R Stat Soc Ser A*. 1984; 147:656–666.
10. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992; 3:143–155. [PubMed: 1576220]
11. Taylor JMG, Wang Y, Thibaut R. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*. 2005; 61:1102–1111. [PubMed: 16401284]
12. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods*. 2010; 15:309–334. [PubMed: 20954780]
13. Vanderweele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*. 2010; 21:540–551. [PubMed: 20479643]
14. Frangakis C, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002; 58:21–29. [PubMed: 11890317]

15. Gilbert PB, Hudgens MG. Evaluating candidate principal surrogate endpoints. *Biometrics*. 2008; 64:1146–1154. [PubMed: 18363776]
16. Li Y, Taylor JMG, Elliott MR. A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics*. 2010; 66:523–531. [PubMed: 19673864]
17. Joffe MM, Greene T. Related causal frameworks for surrogate outcomes. *Biometrics*. 2009; 65:530–538. [PubMed: 18759836]
18. Rubin DB. Direct and Indirect Causal Effects via Potential Outcomes. *Scandinavian Journal of Statistics*. 2004; 31:161–170.
19. Pearl J. Principal Stratification a Goal or a Tool? *The International Journal of Biostatistics*. 2011; 7(1):20. [PubMed: 21556288]
20. Vanderweele TJ. Principal Stratification Uses and Limitations. *The International Journal of Biostatistics*. 2011; 7(1):28.
21. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd Ed. Wiley; 2002.
22. Frangakis C, Rubin DB. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*. 1999; 86:365–379.
23. Peng Y, Little RJA, Raghunathan TE. An extended general location model for causal inferences from data subject to noncompliance and missing values. *Biometrics*. 2004; 60:598–607. [PubMed: 15339281]
24. Chen H, Geng Z, Zhou X-H. Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data. *Biometrics*. 2009; 65:675–691. [PubMed: 18759847]
25. Taylor L, Zhou XH. Multiple imputation methods for treatment noncompliance and nonresponse in randomized clinical trials. *Biometrics*. 2009; 65:88–95. [PubMed: 18397338]
26. Li Y, Taylor JMG, Elliott MR, Sargent DJ. Causal assessment of surrogacy in a meta-analysis of colorectal cancer trials. *Biostatistics*. 2011; 12:478–492. [PubMed: 21252079]
27. Rubin DB. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*. 1990; 25:279–292.
28. Elliott MR, Raghunathan TE, Li Y. Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics*. 2010; 11:353–372. [PubMed: 20101045]
29. Gelfand AE, Sahu SK. Identifiability, improper priors, and gibbs sampling for generalized linear models. *J Am Stat Assoc*. 1999; 94:247–253.
30. Long Q, Little RJA, Lin X. Estimating casual effects in trials involving multitreatment arms subject to non-compliance: a Bayesian framework. *J R Stat Soc Ser C*. 2010; 59:513–531.
31. Gustafson P. Bayesian inference for partially identified models. *The International Journal of Biostatistics*. 2010; 6(2):17.
32. Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B*. 2002; 64:583–639.
33. Little RJA. Calibrated Bayes, for statistics in general, and missing data in particular. *Statist Sci*. 2011; 26:162–174.
34. Gustafson P, Greenland S. Interval estimation for messy observational data. *Statist Sci*. 2009; 24:328–342.
35. Musch DC, Lichter PR, Guire KE, et al. Bayesian measures of model complexity and fit. *Ophthalmology*. 1999; 106:653–662. [PubMed: 10201583]
36. AGIS Investigators. The Advanced Glaucoma Intervention Study (AGIS) 7: The relationship between control of intraocular pressure and visual field deterioration. *Am J Ophthalmol*. 2000; 130:429–440. [PubMed: 11024415]
37. Gelman A, Meng X-L, Stern H. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*. 1986; 6:733–807.

APPENDIX

Gibbs Sampler Under Latent Ignorability

The posterior distribution under latent ignorability and Dirichlet prior for the cell probabilities $p(\pi_{j|i}) \sim \text{DIR}(a_{j|i})$ and $p(\pi_{i+k}) \sim \text{DIR}(b_{i+k})$ is given by

$$\prod_i \prod_j \prod_k \pi_{ijk}^{n_{ijk}} = \prod_i \prod_j \pi_{j|i}^{n_{ij+} + a_{j|i}} \prod_i \prod_k \pi_{i+k}^{n_{i+k} + b_{i+k}}$$

where the complete data n_{ijk} corresponds to the cell counts where $((S(0), S(1)) = i, (T(0), T(1)) = j, (R(0), R(1)) = k)$. The Gibbs sampling algorithm proceeds as follows.

We first obtain draws from cell count parameters consistent with the latent ignorable assumption:

$$\pi_{1|i}, \pi_{2|i}, \pi_{3|i}, \pi_{4|i} | n_{111}, \dots, n_{444} \sim \text{DIR}(n_{i1+} + a_{1|i}, n_{i2+} + a_{2|i}, n_{i3+} + a_{3|i}, n_{i4+} + a_{4|i})$$

$$\pi_{1+1}, \dots, \pi_{4+4} | n_{111}, \dots, n_{444} \sim \text{DIR}(n_{1+1} + b_{1+1}, \dots, n_{4+4} + b_{4+4})$$

We then obtain draws of the complete data conditional on the cell count parameters and the observed data (s_l, t_l, r_l, z_l) :

$$P((R(0)_l, R(1)_l) = k | (S(0)_l, S(1)_l) = i, r_l, z_l, \pi_{i+1}, \dots, \pi_{i+4}) \propto$$

$$((1 - r_l)\pi_{i+1}, (1 - |r_l - z_l|)\pi_{i+2}, r_l\pi_{i+3}, |r_l - z_l|\pi_{i+4})$$

$$P((S(0)_l, S(1)_l) = i | (T(0)_l, T(1)_l) = j, (R(0)_l, R(1)_l) = k, s_l, z_l, \pi_{j|1}, \dots, \pi_{j|4}, \pi_{1+k}, \dots, \pi_{4+k}) \propto$$

$$((1 - s_l)\pi_{j|1}\pi_{1+k}, (1 - |s_l - z_l|)\pi_{j|2}\pi_{2+k}, s_l\pi_{j|3}\pi_{3+k}, |s_l - z_l|\pi_{j|4}\pi_{4+k})$$

$$P((T(0)_l, T(1)_l) = j | (s(0)_l, S(1)_l) = i, r_l = 1, t_l, z_l, \pi_{1|i}, \dots, \pi_{4|i}) \propto$$

$$((1 - t_l)\pi_{1|i}, (1 - |t_l - z_l|)\pi_{2|i}, t_l\pi_{3|i}, |t_l - z_l|\pi_{4|i})$$

$$P((T(0)_l, T(1)_l)=j | (S(0)_l, S(1)_l)=i, r_l=0, \pi_{1|i}, \dots, \pi_{4|i}) \propto$$

$$(\pi_{1|i}, \pi_{2|i}, \pi_{3|i}, \pi_{4|i})$$

Under the monotonicity assumption, we set $\pi_{i4k} = \pi_{4jk} = 0$ and update the complete data draws accordingly. Under stochastic monotonicity, we reject draws that fail to meet the stochastic monotonicity assumptions: $\pi_{2++} > \pi_{4++}$, and $\pi_{j2+} > \pi_{j4+}$ for $j = 1, 2, 3$.

Gibbs Sampler Under Missing Completely at Random

The Gibbs sampling algorithm under MCAR proceeds as follows.

We draw from the cell probabilities for joint distribution of the principal strata and potential outcome and from the the cell probabilities for missingness pattern independently:

$$\pi_{11+}, \dots, \pi_{44+} | n_{111}, \dots, n_{444}, \sim \text{DIR}(n_{11+} + a_{11}, \dots, n_{44+} + a_{44})$$

$$\pi_{++1}, \dots, \pi_{++4} | n_{111}, \dots, n_{444}, \sim \text{DIR}(n_{++1} + b_{++1}, \dots, n_{++4} + b_{++4})$$

The data augmentation step for $R(Z_l)$ and $S(Z_l)$ simplifies to

$$P((R(0)_l, R(1)_l)=k | r_l, z_l, \pi_{++1}, \dots, \pi_{++4}) \propto$$

$$((1 - r_l)\pi_{++1}, (1 - |r_l - z_l|)\pi_{++2}, r_l\pi_{++3}, |r_l - z_l|\pi_{++4})$$

$$P((S(0)_l, S(1)_l)=i | (T(0)_l, T(1)_l)=j, s_l, z_l, \pi_{1j+}, \dots, \pi_{4j+}) \propto$$

$$((1 - s_l)\pi_{1j+}, (1 - |s_l - z_l|)\pi_{2j+}, s_l\pi_{3j+}, |s_l - z_l|\pi_{4j+})$$

The data augmentation step for $T(Z_l)$ remains as in the latent ignorable selection model.

Observed Likelihood Deviance

The observed likelihood deviance $D(\pi)$ (up to a constant) is given by $-2(\sum_i \sum_j \sum_k m_{ijk} \log \theta_{ijk} + \sum_l \sum_k m_{lk} \log \gamma_{lk} + \sum_m \sum_k m_{mk} \log \psi_{mk})$, where m_{ijz} correspond to the observed cell counts with s and t fully observed by treatment assignment, $s = i$, $t = j$, $z = k$, m_{lk} to the observed cell counts with only s observed by treatment assignment, $s = i$, $z = k$, and m_{mk} to the

missing data indicator for t by treatment assignment, $r = m$, $z = k$. Under the latent ignorable model (note $\pi_{ij|k} = \pi_{j|i}\pi_{i|k}$)

$$\theta_{000} = \left[\sum_{k=3,4} (\pi_{1|1}\pi_{1+k} + \pi_{2|1}\pi_{1|k} + \pi_{1|2}\pi_{2+k} + \pi_{2|2}\pi_{2+k}) \right] / \psi_{10}$$

$$\theta_{100} = \left[\sum_{k=3,4} (\pi_{1|3}\pi_{3+k} + \pi_{2|3}\pi_{3|k} + \pi_{1|4}\pi_{4+k} + \pi_{2|4}\pi_{4+k}) \right] / \psi_{10}$$

$$\theta_{010} = \left[\sum_{k=3,4} (\pi_{1|3}\pi_{1+k} + \pi_{3|2}\pi_{2+k} + \pi_{4|1}\pi_{1+k} + \pi_{4|2}\pi_{2+k}) \right] / \psi_{10}$$

$$\theta_{110} = \left[\sum_{k=3,4} (\pi_{3|3}\pi_{3+k} + \pi_{4|3}\pi_{3+k} + \pi_{3|4}\pi_{4+k} + \pi_{4|4}\pi_{4+k}) \right] / \psi_{10}$$

$$\theta_{001} = \left[\sum_{k=2,3} (\pi_{1|1}\pi_{1+k} + \pi_{4|1}\pi_{1+k} + \pi_{1|4}\pi_{4+k} + \pi_{4|4}\pi_{4+k}) \right] / \psi_{11}$$

$$\theta_{101} = \left[\sum_{k=2,3} (\pi_{1|2}\pi_{2+k} + \pi_{1|3}\pi_{3+k} + \pi_{4|2}\pi_{2+k} + \pi_{4|3}\pi_{3+k}) \right] / \psi_{11}$$

$$\theta_{011} = \left[\sum_{k=2,3} (\pi_{2|1}\pi_{1+k} + \pi_{3|1}\pi_{1+k} + \pi_{2|4}\pi_{4+k} + \pi_{3|4}\pi_{4+k}) \right] / \psi_{11}$$

$$\theta_{111} = \left[\sum_{k=2,3} (\pi_{2|2}\pi_{2+k} + \pi_{3|2}\pi_{2+k} + \pi_{2|3}\pi_{3+k} + \pi_{3|3}\pi_{3+k}) \right] / \psi_{11}$$

$$\gamma_{00} = \left[\sum_{k=1,2} (\pi_{1+k} + \pi_{2+k}) \right] / \psi_{00}$$

$$\gamma_{10} = \left[\sum_{k=1,2} (\pi_{3+k} + \pi_{4+k}) \right] / \psi_{00}$$

$$\gamma_{01} = \left[\sum_{k=1,4} (\pi_{1+k} + \pi_{4+k}) \right] / \psi_{01}$$

$$\gamma_{11} = \left[\sum_{k=1,4} (\pi_{2+k} + \pi_{3+k}) \right] / \psi_{01}$$

$$\psi_{00} = \pi_{++1} + \pi_{++2}$$

$$\psi_{10} = \pi_{++3} + \pi_{++4}$$

$$\psi_{01} = \pi_{++1} + \pi_{++4}$$

$$\psi_{11} = \pi_{++2} + \pi_{++3}.$$

Under the MCAR model, the relationship between the missingness pattern and the surrogate marker is assumed away, and

$$\theta_{000} = \pi_{11+} + \pi_{12+} + \pi_{21+} + \pi_{22+}$$

$$\theta_{100} = \pi_{31+} + \pi_{32+} + \pi_{41+} + \pi_{42+}$$

$$\theta_{010} = \pi_{13+} + \pi_{23+} + \pi_{3214+} + \pi_{24+}$$

$$\theta_{110} = \pi_{33+} + \pi_{34+} + \pi_{43+} + \pi_{44+}$$

$$\theta_{001} = \pi_{11+} + \pi_{14+} + \pi_{41+} + \pi_{44+}$$

$$\theta_{101} = \pi_{21+} + \pi_{31+} + \pi_{24+} + \pi_{34+}$$

$$\theta_{011} = \pi_{12+} + \pi_{13+} + \pi_{42+} + \pi_{43+}$$

$$\theta_{111} = \pi_{22+} + \pi_{23+} + \pi_{32+} + \pi_{33+}$$

$$\gamma_{00} = \pi_{1++} + \pi_{2++}$$

$$\gamma_{10} = \pi_{3++} + \pi_{4++}$$

$$\gamma_{01} = \pi_{1++} + \pi_{4++}$$

$$\gamma_{11} = \pi_{2++} + \pi_{3++}$$

$$\psi_{00} = \pi_{++1} + \pi_{++2}$$

$$\psi_{10} = \pi_{++3} + \pi_{++4}$$

$$\psi_{01} = \pi_{++1} + \pi_{++4}$$

$$\psi_{11} = \pi_{++2} + \pi_{++3}$$

Figure 1.

Posterior distributions of causal effect, associative effect, and associative proportion for IOP analysis, under the latent ignorable model with the stochastic monotonicity assumption and Jeffreys-type prior.

IOP: intraocular pressure.

Figure 2.

Posterior predictive distributions of observable cell counts for surrogate marker and outcome by treatment assignment given by $m_{ijz} = \sum_l I(s_l = i, t_l = j \mid z_l = z)$ for $i, j, z \in \{0, 1\}$; densities represent the predictive distributions of the cell counts of the observed 12-month IOP and 96-month IOP measures by treatment arm assignment, solid lines indicate observed values (under latent ignorable model with the stochastic monotonicity assumption and Jeffreys-type prior).

Table 1

Joint distribution of potential surrogate marker, outcome, and missingness patterns.

$R(Z = 0), R(Z = 1)$									
$T(Z = 0), T(Z = 1)$		$(0,1)$		$(1,1)$		$T(Z = 0), T(Z = 1)$		$(1,0)$	
$T(Z = 0), T(Z = 1)$		$(0,0)$		$(0,1)$		$(1,1)$		$(1,0)$	
$S(Z = 0), S(Z = 1)$	$(0,0)$	$(0,1)$	$(1,1)$	$(1,0)$	$(0,0)$	$(0,1)$	$(1,1)$	$(0,0)$	$(1,0)$
$(0,0)$	π_{111}	π_{121}	π_{131}	π_{141}	π_{112}	π_{122}	π_{132}	π_{142}	π_{113}
$(0,1)$	π_{211}	π_{221}	π_{231}	π_{241}	π_{212}	π_{222}	π_{232}	π_{242}	π_{213}
$(1,1)$	π_{311}	π_{321}	π_{331}	π_{341}	π_{312}	π_{322}	π_{332}	π_{342}	π_{313}
$(1,0)$	π_{411}	π_{421}	π_{431}	π_{441}	π_{412}	π_{422}	π_{432}	π_{442}	π_{413}
									π_{423}
									π_{433}
									π_{443}
									π_{114}
									π_{124}
									π_{134}
									π_{144}
									π_{214}
									π_{224}
									π_{234}
									π_{244}
									π_{314}
									π_{324}
									π_{334}
									π_{344}
									π_{414}
									π_{424}
									π_{434}
									π_{444}

Table 2

Simulation design: marginal distribution of response pattern, conditional distribution of principal strata given response pattern, and conditional distribution of potential outcome given principal strata. All designs assume monotonicity for surrogate marker and final outcome. Under poor surrogate design, causal effect(CE)=.3400, associative proportion (AP)=.1765, common associative proportion (CAP)=.1035. Under good surrogate design, causal effect(CE)=.3380, associative proportion (AP)=.7101, common associative proportion (CAP)=.6030.

Response Pattern	R(0, 0)	R(0, 1)	R(1, 1)	R(1, 0)
Independent of Treatment	.25	.25	.25	.25
Dependent on Treatment	.15	.15	.15	.55
Principal Stratum	S(0, 0) R	S(0, 1) R	S(1, 1) R	S(1, 0) R
Ignorable missingness				
R(0, 0)	.3	.3	.4	0
R(0, 0)	.3	.3	.4	0
R(1, 1)	.3	.3	.4	0
R(1, 0)	.3	.3	.4	0
Non-ignorable missingness:				
R(0, 0)	.5	.3	.2	0
R(0, 0)	.5	.3	.2	0
R(1, 1)	.2	.3	.5	0
R(1, 0)	.2	.3	.5	0
Outcome	T(0, 0) S	T(0, 1) S	T(1, 1) S	T(1, 0) S
Poor Surrogate				
S(0, 0)	.4	.4	.2	0
S(0, 1)	.5	.2	.3	0
S(1, 1)	.2	.4	.4	0
Good Surrogate				
S(0, 0)	.8	.14	.06	0
S(0, 1)	.1	.8	.1	0
S(1, 1)	.06	.14	.8	0

Table 3

Simulation study results: non-ignorable missingness. Independent non-response=missingness equally likely under treatment and control, dependent non-response=missingness more likely under treatment; poor surrogate=small associative proportion, good surrogate=large associative proportion (see Table 2). RMSE=root mean square error. Top line gives result under uniform prior; lower line gives result under Jeffreys' type prior.

	Poor Surrogate						Good Surrogate					
	Independent Non-Response			Dependent Non-Response			Independent Non-Response			Dependent Non-Response		
	CE	AP	CAP	CE	AP	CAP	CE	AP	CAP	CE	AP	CAP
True value $\times 10^{-2}$	34.0	17.6	10.4	34.0	17.6	10.4	33.8	71.0	60.3	33.8	71.0	60.3
Latent Ignorable												
Bias $\times 10^{-2}$	-2.1	10.0	10.0	-1.8	10.3	3.7	-2.4	-0.6	-7.0	-2.4	-1.6	-6.9
	-1.5	10.0	1.9	-2.0	11.4	4.3	-4.0	-0.6	-8.6	-2.9	-1.4	-6.6
RMSE $\times 10^{-2}$	5.4	14.2	5.3	14.3	7.8	9.6	5.4	7.0	10.1	5.2	7.1	9.8
	4.8	14.3	6.9	5.1	14.4	8.2	6.3	7.3	12.3	5.2	7.3	9.5
90% Coverage	94	100	100	99	100	100	88	98	99	90	99	98
	99	100	100	100	100	100	86	99	98	92	99	99
Mean CI length	.21	.62	.48	.23	.63	.48	.22	.50	.58	.21	.48	.55
	.23	.61	.47	.24	.62	.47	.25	.62	.65	.24	.57	.62
DIC Selection %	100	100	100	100	100	100	100	100	100	99	99	100
	100	100	100	100	100	100	100	100	100	100	100	100
MCAR												
Bias $\times 10^{-2}$	-5.2	18.0	5.6	-4.0	20.0	7.7	-6.0	0.0	-7.6	-5.7	-1.1	-8.0
	-2.7	2.7	-2.3	-3.1	6.3	-1.1	-4.1	6.2	0.1	-3.4	4.0	-0.7
RMSE $\times 10^{-2}$	7.4	20.1	9.8	6.6	22.2	10.9	7.8	7.4	10.9	7.3	7.6	11.1
	5.9	23.5	9.9	6.1	27.0	12.0	6.4	10.1	8.6	5.8	8.9	8.5
90% Coverage	76	99	99	86	99	99	65	98	98	67	97	96
	90	100	100	90	99	99	84	97	99	67	97	96
Mean CI length	.18	.67	.53	.19	.66	.52	.16	.40	.44	.17	.40	.43
	.19	.84	.72	.20	.84	.72	.16	.42	.49	.17	.43	.49
DIC Selection %	0	0	0	0	0	0	0	0	0	0	0	0

	Poor Surrogate						Good Surrogate					
	Independent Non-Response			Dependent Non-Response			Independent Non-Response			Dependent Non-Response		
	CE	AP	CAP	CE	AP	CAP	CE	AP	CAP	CE	AP	CAP
	0			0			0			1		
Complete-Case												
Bias $\times 10^{-2}$	-6.5	-5.8	-4.2	-6.2	-3.4	-3.1	-15.9	-33.0	-36.6	-16.4	-31.4	-36.5
	-4.8	-14.9	-8.1	-6.0	-14.1	-7.9	-15.9	-37.0	-39.2	-15.8	-42.3	-43.4
RMSE $\times 10^{-2}$	8.6	10.2	5.6	8.4	11.0	6.1	16.6	37.2	39.3	17.1	35.6	39.2
	7.4	15.3	8.3	8.5	15.6	8.6	16.7	46.9	45.4	16.8	51.2	49.1
90% Coverage	69	100	100	76	100	100	8	50	32	12	54	36
	82	100	100	76	100	100	14	75	74	20	73	72
Mean CI length	.19	.53	.43	.20	.55	.44	.17	.61	.53	.17	.62	.54
	.19	.64	.56	.20	.68	.59	.18	.73	.66	.19	.73	.66

Table 4

Simulation study results: ignorable missingness. Independent non-response=missingness equally likely under treatment and control, dependent non-response=missingness more likely under treatment; poor surrogate=small associative proportion, good surrogate=large associative proportion (see Table 2). RMSE=root mean square error. Top line gives result under uniform prior; lower line gives result under Jeffreys' type prior.

Poor Surrogate						Good Surrogate							
	Independent Non-Response			Dependent Non-Response			Independent Non-Response			Dependent Non-Response			
	CE	AP	CAP	CE	AP	CAP	CE	AP	CAP	CE	AP	CAP	
True value $\times 10^{-2}$	34.0	17.6	10.4	10.4	34.0	17.6	10.4	33.8	71.0	60.3	33.8	71.0	60.3
Latent Ignorable													
Bias $\times 10^{-2}$	-1.8	11.8	2.6	-0.0	12.6	5.1	-2.4	-1.8	-7.8	-2.2	-1.4	-6.4	
	-1.1	10.9	2.1	-1.0	10.9	3.9	-2.8	-3.6	-9.0	-2.2	-2.5	-7.6	
RMSE $\times 10^{-2}$	5.5	16.1	7.4	5.4	16.8	9.6	5.4	7.4	10.8	5.2	7.3	9.2	
	4.8	14.9	7.2	5.3	15.4	8.5	6.0	8.4	12.3	6.0	9.0	11.7	
90% Coverage	96	100	100	97	99	99	88	98	98	92	99	98	
	99	100	100	100	99	99	90	97	98	88	98	98	
Mean CI length	.21	.64	.50	.22	.62	.49	.21	.49	.55	.22	.48	.55	
	.22	.62	.49	.24	.62	.49	.24	.59	.63	.25	.56	.62	
DIC Selection %	42				48			30			32		
	25				22			21			16		
MCAR													
Bias $\times 10^{-2}$	-3.9	20.8	8.0	-3.1	21.0	10.2	-5.4	-1.9	-8.4	-5.9	-0.0	-6.5	
	-2.0	4.5	-1.0	-2.5	6.8	0.3	-2.9	3.0	-1.0	-3.2	4.1	-0.7	
RMSE $\times 10^{-2}$	6.7	23.4	11.3	6.4	23.0	13.2	7.3	8.1	11.3	7.5	7.4	9.8	
	5.7	25.8	11.7	6.3	26.3	12.6	6.0	9.4	8.3	6.3	10.2	9.9	
90% Coverage	84	97	98	86	98	96	70	97	97	70	98	98	
	91	99	99	87	99	99	82	98	99	80	98	100	
Mean CI length	.18	.66	.53	.19	.64	.52	.16	.39	.42	.17	.39	.42	
	.18	.82	.72	.20	.83	.72	.17	.41	.48	.17	.42	.48	
DIC Selection %	58				52			70			68		

		Poor Surrogate						Good Surrogate					
		Independent Non-Response			Dependent Non-Response			Independent Non-Response			Dependent Non-Response		
		CE	AP	CAP	CE	AP	CAP	CE	AP	CAP	CE	AP	CAP
		75			78			79			84		
Complete-Case													
Bias $\times 10^{-2}$		-4.1	17.8	6.4	-2.4	21.8	10.4	-7.2	-4.8	-11.4	-7.8	-2.3	-9.4
		-2.2	1.2	-2.8	-2.3	6.3	-1.4	-4.3	0.3	-5.0	-4.3	2.0	-3.5
RMSE $\times 10^{-2}$		6.9	21.4	9.9	6.4	24.8	13.8	9.2	12.0	15.4	10.3	10.7	14.4
		5.8	24.8	10.4	6.4	28.5	12.8	7.6	14.1	15.7	8.7	16.7	16.3
90% Coverage		82	100	100	90	94	94	63	95	92	62	97	95
		90	100	100	90	94	94	80	98	98	75	98	100
Mean CI length		.18	.66	.53	.20	.66	.53	.19	.49	.48	.21	.49	.49
		.19	.82	.71	.20	.66	.53	.20	.52	.55	.21	.54	.56

Table 5

Collaborative Initial Glaucoma Treatment Study: Observed Data.

		Control			Treatment				
		Reduced IOP at 96 Months			Reduced IOP at 96 Months				
		No	Yes	Missing	No	Yes	Missing		
Reduced IOP	No	28	29	69	145	11	8	35	65
at 12 Months	Yes	14	55	97	147	10	73	144	216
		42	84	166	292	21	81	179	282

Table 6

Analysis of Collaborative Initial Glaucoma Treatment Study: posterior mean, posterior mode, and 90% credible intervals for causal treatment effect (CE), associative effect (AE), associative proportion (AP), and common associative proportion (CAP). First row for Dirichlet prior set uniformly to 1; second row for Dirichlet prior set uniformly to 1/2.

	<i>T</i> Fully Observed			MCAR			Latent Ignorability		
	Mean	Mode	CI	Mean	Mode	CI	Mean	Mode	CI
<i>Monotonicity</i>									
CE	.130	.125	.060–.209	.126	.118	.058–.200	.148	.141	.069–.233
	.120	.112	.044–.203	.115	.111	.042–.195	.133	.123	.039–.234
AE	.045	.009	.004–.109	.040	.008	.003–.097	.052	.030	.004–.118
	.042	.003	.000–.121	.035	.002	.000–.102	.050	.003	.000–.142
AP	.339	.192	.034–.731	.316	.119	.030–.698	.348	.322	.039–.726
	.339	.020	.004–.863	.308	.020	.003–.830	.362	.023	.005–.884
CAP	.134	.025	.011–.323	.127	.023	.010–.315	.159	.031	.013–.376
	.131	.009	.001–.383	.117	.008	.001–.354	.164	.013	.001–.492
<i>No Monotonicity</i>									
CE	–.024	–.022	–.119–.070	.015	.018	–.076–.106	.073	.080	–.054–.185
	.032	.030	–.064–.127	.060	.062	–.032–.151	.061	.085	–.121–.205
AE	–.027	–.018	–.130–.074	.025	.019	–.071–.121	.068	.084	–.066–.187
	.031	.036	–.080–.142	.066	.071	–.044–.171	.062	.094	–.109–.211
AP	.827	.915	–3.482–5.125	.517	.972	–4.379–5.793	.853	.849	–2.189–3.850
	.845	.979	–3.734–5.427	.987	.978	–2.446–4.430	.795	.901	–2.341–3.866
<i>Stochastic Monotonicity</i>									
CE	.023	.028	–.060–.107	.044	.048	–.037–.125	.116	.120	.031–.205
	.056	.061	–.034–.144	.079	.074	–.003–.163	.125	.124	.027–.232
AE	–.015	–.026	–.100–.072	.006	.002	–.075–.086	.055	.047	–.025–.141
	.010	.014	–.084–.103	.034	.029	–.056–.124	.057	.034	–.031–.156
AP	.650	.578	–4.060–5.902	.433	.690	–2.863–3.900	.426	.551	–.425–.884
	.400	.773	–2.744–3.486	.367	.791	–1.538–1.331	.409	.552	–.483–.912

Table 7

DIC measures for various models accounting for missing data. First row for Dirichlet prior set uniformly to 1; second row for Dirichlet prior set uniformly to 1/2.

	MCAR	Latent Ignorability
Monotonicity	63.68	67.64
	64.13	67.97
No Monotonicity	79.06	66.37
	69.19	66.62
Stochastic Monotonicity	77.53	63.66
	68.48	63.25