IMHOTEP

AFRICAN JOURNAL OF PURE AND APPLIED MATHEMATICS

Imhotep Mathematical Proceedings
Volume 1, Numéro 1, (2014),
pp. 1 – 24.

Optimal compliance prediction models for estimating causal effects

Lang'o Odondi
MRC ConDuCT Hub, School of Social and Community Medicine, University of Bristol, UK
Lango.Odondi@bristol.ac.uk

Abstract

Understanding the causal relationship between intervention and outcome is at the heart of most research in the health sciences, and a variety of statistical methods have been developed to address causality. However, noncompliance with treatment assignment is a key source of complication for causal inference. Estimation of causal effects is likely to be compounded by the presence of noncompliance in both treatment arms of clinical trials where the intention-to-treat (ITT) analysis produces a biased estimator for the true causal estimate even under homogeneous treatment effects assumption. Principal stratification method has been developed to address such posttreatment complications by stratifying the population into partially latent classes (principal strata) based on potential values observed after randomization (e.g. noncompliance) under each of the levels of randomized intervention. The present work combines the two strategies of model selection and principal stratification with a novel application to a real data from a trial conducted to ascertain whether or not unopposed oestrogen (hormone replacement therapy - HRT) reduced the risk of further cardiac events in postmenopausal women who survive a first myocardial infarction. The causal model links the resulting two marginal prediction models with a user-defined sensitivity parameter which is a function of the correlation between the two compliance behaviours. The method's key assumption of conditional prediction is verified for our data via sensitivity analysis comparing results of causal estimates using different sets of predictors of compliance. We adjust for noncompliance in both treatment arms under a Bayesian framework to produce causal risk ratio estimates for each principal stratum. The results suggested better efficacy for HRT among those who would comply with it compared to those who would comply with either HRT or placebo: compliance with HRT treatment only and with either treatment allocation would reduce the risk for death (reinfarction) by 47%(25%) and 13%(60%) respectively.

Proceedings of the 2nd Strathmore International Mathematics Conference (SIMC 2013), 12 - 16 August 2013, Strathmore University, Nairobi, Kenya.

IMHOTEP - Math. Proc. **1** (2014), 1–24
1608-9324/010001-0, DOI 13.1007/s00009-003-0000

IMHOTEP Mathematical Proceedings

Optimal compliance prediction models for estimating causal effects

Lang'o Odondi

Abstract. Understanding the causal relationship between intervention and outcome is at the heart of most research in the health sciences, and a variety of statistical methods have been developed to address causality. However, noncompliance with treatment assignment is a key source of complication for causal inference. Estimation of causal effects is likely to be compounded by the presence of noncompliance in both treatment arms of clinical trials where the intention-to-treat (ITT) analysis produces a biased estimator for the true causal estimate even under homogeneous treatment effects assumption. Principal stratification method has been developed to address such posttreatment complications by stratifying the population into partially latent classes (principal strata) based on potential values observed after randomization (e.g. noncompliance) under each of the levels of randomized intervention. The present work combines the two strategies of model selection and principal stratification with a novel application to a real data from a trial conducted to ascertain whether or not unopposed oestrogen (hormone replacement therapy - HRT) reduced the risk of further cardiac events in postmenopausal women who survive a first myocardial infarction. The causal model links the resulting two marginal prediction models with a user-defined sensitivity parameter which is a function of the correlation between the two compliance behaviours. The method's key assumption of conditional prediction is verified for our data via sensitivity analysis comparing results of causal estimates using different sets of predictors of compliance. We adjust for noncompliance in both treatment arms under a Bayesian framework to produce causal risk ratio estimates for each principal stratum. The results suggested better efficacy for HRT among those who would comply with it compared to those who would comply with either HRT or placebo: compliance with HRT treatment only and with either treatment allocation would reduce the risk for death (reinfarction) by 47%(25%) and 13%(60%) respectively.

Paper presented at the 2nd Strathmore International Mathematics Conference (SIMC 2013), 12 - 16 August 2013, Strathmore University, Nairobi, Kenya.

I. Introduction

Valid causal inference is a central motivation in the analysis of data from randomised controlled trials when comparing two or more interventions. Effective randomization of subjects between the treatment groups plays a key role in permitting such statistical comparison [22, 19]. However, the presence of intermediate variables in the intervention-outcome causal pathway is likely to complicate estimation of causal effects by introducing selection bias since they often manifest themselves as non-random phenomena [41, 24]. Noncompliance with treatment assignment one such phenomenon which may often manifests itself as treatment discontinuation, switching or subject dropout from the study. Intention-to-treat (ITT) is the standard analysis for estimating causal effects under perfect compliance with treatment assignment. By comparing treatment groups as assigned, the ITT analysis preserves the baseline comparability between treatment groups. However, using ITT results in presence of treatment noncompliance are likely to underestimate the treatment efficacy by mixing the effects of treatment compliers and non-compliers [43]. Since noncompliance mostly manifests itself as a non-random phenomenon, it is a challenge adjusting for its corresponding informative characteristics.

Applying standard regression methods to adjust for intermediate variables produce estimates which lack causal interpretation [29]. One reason for failure in causal analysis is because such methods make the very strong but often implausible assumption of unmeasured confounders between the intermediate variable and outcome. In the presence of noncompliance, per-protocol and as-treated analyses are commonly used to supplement ITT in evaluating efficacy [42, 20]. But these post-hoc analyses lack the benefits of randomization, for example, selection bias in these methods may be evident in different compliance behaviour in different treatment arms. Efron and Feldman [9] in their seminal work used compliance as a covariate in a regression adjustment for a placebo-controlled clinical trial evaluating the efficacy of Cholestyramine in lowering serum cholesterol levels towards reducing the risk of coronary heart disease. However, their method has been criticized for the implicit strong assumption of comparability in compliance between the active treatment and placebo arms [1, 3]. In the presence of selectivity effects, many methods have been developed to account for noncompliance in more than one treatment arm.

Frangakis and Rubin [11] developed the principal stratification as a general framework to adjust for intermediate variables observed post-randomization. The method basically stratifies the population into partially latent classes (principal strata) based on potential values of posttreatment variable, like a noncompliance status. Principal strata comprise units having the same values of the intermediate potential outcomes and are not affected by treatment assignment hence retains the tenets of randomization and so provide valid and well-defined causal effect estimates for selected subgroup/strata.

Principal stratification is a flexible method for causal modelling which may be extended to adjusting for noncompliance in more than one treatment arm. But owing to the latent nature of principal strata, causal inference using the method often requires making structural assumptions to allow parameter identification. According to Cole and Frangakis [6], causal estimates are generally identifiable under three sufficient assumptions: exchangeability (no unmeasured confounders), positivity (existence of a non-zero probability to receive treatment) and consistency (relating observed data to counterfactual data). In the presence of more than one active treatment, a joint analysis may provide additional analytical insights than pairwise efficacy comparisons [4]. In general, comparing more than one active treatment compounds the challenge of identification of causal estimates due to possible multiple forms/degrees of noncompliance with treatment assignment [30, 21].

Crucial to parameter identification under principal stratification method is selection of good baseline covariates which are predictive of the intermediate status (e.g. noncompliance). The implicit challenge of model selection is not only a statistical problem [17] but may be compounded when applied to intermediate variables occurring on the causal pathway and observed post-randomisation. An efficient selection of plausible predictors of intermediate variables can be used to effectively address identification problem of causal estimands by reducing bias in addition to relaxing implicit causal assumptions [20]. From a clinical perspective, adequate knowledge about predictors of treatment compliance may be a valuable tool to inform treatment decisions. Shrinkage principle is a common strategy of reducing regression coefficients to improve the quality of predictions through bias-variance tradeoff so as to produce stable models in the presence of many predictors [34]. For example, Tibshirani [38] introduced the least absolute shrinkage and selection operator (Lasso) as an efficient model selection method which performs the twin tasks of variable selection and regression coefficient estimation simultaneously.

To adjust for noncompliance with treatment assignment, Roy et al. [30] proposed a principal stratification framework for trials to compare two active treatments using baseline covariates to address identification problem. Long et al. [21] also proposed a likelihood estimation method to provide point causal estimands for a three-armed trial by using Bayesian methods to model the arm-specific compliances directly while treating the principal compliance status as missing. And by using Bayesian methods in a potential outcome framework, Zigler and Belin [44] recently used a key covariate predictive of compliance for causal effect estimation in an active-control trial. On the other hand, Fischer et al. [10] proposed a structural mean modelling approach using baseline covariates predictive of compliance in each arm to obtain compliance-adjusted efficacy in a randomized controlled trial comparing two-active treatments.

Central to the application of Roy et al. [30] model is the conditional prediction assumption that potential outcomes are statistically independent of the set of covariates predictive of

compliance given a stratum and treatment assignment. In the presence of many recorded baseline covariates and given this defining assumption, this underscores the crucial role of selecting efficient and meaningful predictors of compliance with treatment assignment for each trial arm. Using a Bayesian approach, the present work modifies and extends the principal stratification method of Roy et al. [30] to integrate optimal model selection procedures using plausible separate predictors of compliance in each arm and apply it to analyze survival data in terms of causal risk ratio estimates for each principal stratum of the Esprit study.

The rest of the paper is organized as follows: Section 2 describes the motivating data from the Esprit study. In Section 3, we describe the relevant causal modeling assumptions. Section 4 provide a brief outline of the methods of analysis by first presenting model selection predicting arm-specific compliance followed by the causal model framework linking the marginal compliance models and the resulting Bayesian inference. Section 5 present an application and results from analysis of the Esprit data. Finally, Section 6 presents a broad discussion.

II. Motivating data: The Esprit study

The onset of menopause is often characterized by diminishing production of oestrogen hormones due to a decline in ovarian function whose unpleasant symptoms (e.g. vasomotor, insomnia, fatigue and depression) can impact negatively on the body leading to low quality of life among such women for the better part of the last third of their lives [27]. Hormone replacement therapy (HRT) is a treatment for oestrogen-deficiency symptoms which is mainly administered in two broad forms depending on whether a woman has her uterus intact or not: unopposed oestrogen (oestrogen taken by itself) for those who have had hysterectomy (removal of the uterus) or oestrogen with progestin for the non-hysterectomized. The addition of progestin is meant to counteract the effects of estrogen on the uterus like endometrial cancer. Although observational studies conducted in the last quarter of last Century showed benefits of HRT in lowering rates of coronary heart disease [12], follow-up clinical trials failed to confirm such beneficial effects among postmenopausal women.

The oEStrogen in the Prevention of ReiInfarction Trial (Esprit) study was one of the trials whose ITT results revealed no HRT benefit among postmenopausal women [5]. Esprit was a two-armed placebo-controlled double blind clinical trial conducted to ascertain whether or not unopposed oestrogen reduces the risk of further cardiac events in postmenopausal women aged between $50-69$ years who survived a first myocardial infarction in England and Wales. The study comprised a total of 1017 subjects: 513 and 504 women were randomized to HRT treatment and placebo arms respectively and monitored over 24 months period. The primary outcomes were cardiac deaths and all-cause mortality or reinfarction. Although ITT analysis

of the data has been previously published, however the analysis took no account of compliance data which we use in this paper to estimate true causal effects.

III. Notation and assumptions

We consider a parallel two-armed clinical trial set up. Let $Z \in \{0,1\}$ denote a randomization indicator: $Z=1$ indicate randomization to the new treatment and $Z=0$ indicates randomization to control. In our application, 1 ($Z=1$) and 0 ($Z=0$) will represent randomization to HRT tablets and placebo treatment respectively. We define $Y \in \{0,1\}$ to be the outcome of interest (e.g. death). Also let $A \in \{0,1\}$ denote compliance with assigned treatment. For the Esprit study we define compliance as actual taking of assigned treatment up to a day before experiencing event of interest (death/reinfarction) or end of study, whichever occurred first. Although this all-or-nothing compliance definition may appear restrictive, it was considered adequate and plausible since any potential treatment switches are assumed to occur soon after randomization and HRT tablets are presumed to have no carryover effects, i.e. assuming no residual effect of treatment once a subject is classified a non-complier. We note that each subject has two potential compliance levels A_0 and A_1 (compliance with placebo and HRT treatment respectively) and two potential outcomes Y_0 and Y_1 (outcome under placebo and HRT treatment respectively). But the observed compliance and outcomes are respectively given by $A=ZA_1+(1-Z)A_0$ and $Y=ZY_1+(1-Z)Y_0$.

Analysis under the principal stratification formulation utilizes baseline covariates X to modify the standard assumptions (a)-(e) for causal modelling [18, 2] together with a new assumption (f):

(a) Randomization: $Z \perp \{Y_0, Y_1, A_0, A_1, X\}$, i.e. ignorable treatment assignment assumption.

(b) Stable unit-treatment value assumption (SUTVA), i.e. no interference between treatment units.

(c) Treatment access restriction: which posits no treatment switches among subjects.

(d) Exclusion restriction: $\Pr(Y_1|A_Z, X) = \Pr(Y_0|A_Z, X)$, i.e. treatment assignment has no effect on outcome except only through treatment received (knowledge of treatment assignment alone has no effect on outcome).

(e) Monotonicity: $\Pr(A_1 = 1|A_0 = 1, X) \geq \Pr(A_1 = 1|A_0 = 0, X)$, i.e. the probability of compliance with treatment assigned by $Z=1$ is higher among those who would comply with treatment assigned by $Z=0$, compared to those who would not.

(f) Conditional prediction: $Y \perp X|S, Z$, i.e potential outcome is statistically independent (ignorable) of the set of covariates predictive of compliance for a given principal stratum and treatment assignment.

Any switching of treatment is assumed to have occurred soon after randomization such that a subject is assumed to have completely taken HRT or placebo treatment. Assuming no other form of treatment interruptions among patients and no carryover effects of treatment, the all-or-nothing compliance definition above may be considered plausible with respect to the exclusion restriction assumption. The monotonicity assumption as applied here helps tighten the bounds of causal effects, i.e. ensures compliance type is observable when $Z \neq A$ [4, 30, 37]. In addition to the basic monotonicity assumption (no treatment defiers), the 'extended' monotonicity assumption posits similar compliance behaviour for both treatment arms. Plausibility of the 'extended' version of monotonicity assumption for the Esprit study may be discerned from the fact that there was no preference for one treatment over the other, i.e compliance with HRT treatment would be more prevalent among those who would comply with placebo. In our application, this assumption is reflected through a user-defined positive correlation (sensitivity parameter ϕ) between A_0 and A_1.

The conditional prediction assumption (f) is crucial for parameter identification in the Roy et al. [30] model. The assumption underscores an integral component of the method which involves selecting suitable predictors of compliance. The first part of this paper will address this challenge through a comprehensive model selection of the Esprit study to obtain optimal arm-specific predictors compliance. We will use penalized maximum likelihood (shrinkage) procedures to select plausible separate predictors of compliance for HRT treatment and placebo arms. Although this is an untestable assumption, we will compare results from different sets of predictors as a form of sensitivity analysis.

Each subject is assumed to belong to one of four basic principal strata defined by unique combinations of (A_0, A_1) where the principal strata comprise the set $\mathcal{S} = \{(0,0),(1,0),(0,1),(1,1)\}$. The causal inference of interest (Section IV.2) will be to seek the joint distributions $[(Y_0, Y_1)|S = s] \; \forall \; s \in \mathcal{S}$ which provides principal effects of interest for each stratum.

IV. Methods for analysis

IV.1. Compliance prediction models and validation

We use the logistic models to predict compliance to treatment allocation for each arm separately given a selected set of predictors of compliance $\mathrm{x}_0 = 1$ and $\mathrm{x}_1, \ldots, \mathrm{x}_n$:

$$\text{logit}\,[\mu_j(\mathrm{x})] = \left(\sum_{i=0}^{n} \gamma_{ji} \mathrm{x}_i \right), \; j = 0, 1, \tag{1}$$

where $\mu_j(\mathrm{x})$ is the probability of compliance with allocated treatment j given a set of covariates X: the estimated probabilities of complying with arm-specific treatment allocation may then

be obtained using

$$\hat{\mu}_j(\mathrm{x}) = \left[1 + \exp\left(-\sum_{i=0}^{n} \hat{\gamma}_{ji}\mathrm{x}_i\right)\right]^{-1}, \; j = 0, 1, \tag{2}$$

where γ represent the log odds ratio estimates of compliance.

How the two compliance behaviours are correlated is a crucial issue. Following Roy et al. [30], we define a non-negative sensitivity parameter ϕ as a function of the correlation ρ between compliances to treatment allocation (0/1): if $\hat{\mu}_0(\mathrm{x}) > \hat{\mu}_1(\mathrm{x})$ then

$$\phi = \rho\sqrt{\frac{\bar{\hat{\mu}}_0(\mathrm{x})[1 - \bar{\hat{\mu}}_1(\mathrm{x})]}{\bar{\hat{\mu}}_1(\mathrm{x})[1 - \bar{\hat{\mu}}_0(\mathrm{x})]}}. \tag{3}$$

The compliance models for each treatment arm provided by Equations (1) and (2) may be obtained through comprehensive model selection for predictive covariates, i.e. selecting both clinically sensible and statistically concise predictors of compliance with treatment assignment for each trial arm. With many covariates, the classical stepwise model selection procedures are likely to produce suboptimal prediction models [39, 13]. In comparison, penalized maximum likelihood methods have been shown to perform relatively better in selecting optimal predictors [15]. But a selected statistical prediction model also needs validation to evaluate its predictive ability, for example, external validation enables assessment of the performance of prediction in new data [34, 32]. But in the absence of further data, bootstrap validation provides reliable results by allowing calculation of predicted probabilities from a model which can be compared with the actually observed outcomes [8].

Percentage of optimism and both calibration and discrimination indices are among the most effective and commonly used measures of validation performance [14]. While calibration is a reliability measure of how well the model predictions compare with the observed outcomes, discrimination refers to the ability of the model to distinguish between subjects with positive or negative outcomes (e.g. the ability of a model to distinguish compliers with treatment allocation from non-compliers). Calibration is often quantified in practice by the calibration slope [7] obtainable from the validation plot which is a plot of observed probabilities against the predicted probabilities. On the other hand, discrimination is commonly measured using the concordance c-statistic as widely expressed in terms of the Somers rank correlation [33]: $D_{\mathrm{xy}} = 2(c - 0.5)$. This is a measure of the difference between concordance and discordance probabilities [13], such that $c = 0.5$ (1) implies random predictions (perfect discriminations). We can readily discern that larger values of calibration and discrimination (concordance) indicate better prediction and such models should indicate lower percentage of optimism.

IV.2. Causal model joining the marginal compliance models

The main causal inference of interest is obtainable from the joint distributions $[(Y_0, Y_1)|S = s]$. Following Roy et al. [30], the joint probability distribution of compliance to the standard

treatment 0 and compliance to new treatment 1 is a function of the arm-specific marginal compliance probabilities and ϕ and can be estimated for a given value of ϕ. We however note that ϕ is unknown, in general. Specifically if $U(\mathrm{x}) = \min\{1, \frac{\hat{\mu}_1(\mathrm{x})}{\hat{\mu}_0(\mathrm{x})}\}$ then Roy et al. [30] showed that the joint probabilities are given by

$$\begin{aligned}
\hat{\mu}_{11}(\mathrm{x}) &= \Pr(A_0 = 1|X)P(A_1 = 1|A_0 = 1, X) \\
&= \hat{\mu}_0(\mathrm{x})\hat{\mu}_1(\mathrm{x}) + \phi\hat{\mu}_0(x)[U(\mathrm{x}) - \hat{\mu}_1(\mathrm{x})], \\
\hat{\mu}_{01}(\mathrm{x}) &= \Pr(A_0 = 0|X)P(A_1 = 1|A_0 = 0, X) \\
&= \hat{\mu}_1(\mathrm{x}) - \hat{\mu}_0(\mathrm{x})\hat{\mu}_1(\mathrm{x}) - \phi\hat{\mu}_0(\mathrm{x})[U(\mathrm{x}) - \hat{\mu}_1(\mathrm{x})], \\
\hat{\mu}_{10}(\mathrm{x}) &= \Pr(A_0 = 1|X)P(A_1 = 0|A_0 = 1, X) \\
&= \hat{\mu}_0(\mathrm{x}) - \hat{\mu}_0(\mathrm{x})\hat{\mu}_1(\mathrm{x}) - \phi\hat{\mu}_0(\mathrm{x})[U(\mathrm{x}) - \hat{\mu}_1(\mathrm{x})], \\
\hat{\mu}_{00}(\mathrm{x}) &= \Pr(A_0 = 0|X)P(A_1 = 0|A_0 = 0, X) \\
&= 1 - \hat{\mu}_0(\mathrm{x}) - \hat{\mu}_1(\mathrm{x}) + \hat{\mu}_0(\mathrm{x})\hat{\mu}_1(\mathrm{x}) + \phi\hat{\mu}_0(\mathrm{x})[U(\mathrm{x}) - \hat{\mu}_1(\mathrm{x})],
\end{aligned} \tag{4}$$

where X is the set of covariates predictive of compliance in both arms and $A_1(A_0)$ is an indicator of compliance to HRT treatment (placebo) and $\hat{\mu}_{ij}(\mathrm{x})$ denote the probability of being in the compliance subgroup ij (i.e. estimated proportion of compliance per stratum).

Following the principal stratification framework developed by Frangakis and Rubin [11], the possible values of A_0 and A_1 define a stratification factor S for the population of patients. For a defined outcome variable Y (mortality/reinfarction for the Esprit study), let Y_0 and Y_1 refer to potential outcomes under placebo and HRT treatments respectively. There are four possible realizations of (Y_0, Y_1) at each level of S (for example $\varpi_{11} = \Pr[Y_0 = 1, Y_1 = 1]$). The joint distribution of potential outcomes (Y_0, Y_1) for each stratum $S, f(Y_0, Y_1|S, \varpi)$ may be assumed to be multinomial with probabilities $\varpi(S) = \Pr(Y_0 = y_0, Y_1 = y_1|S = s)$. And by the exclusion restriction, the expression for stratum $S = (0, 0)$ differs from the others: $\varpi_{10}(0, 0) = \varpi_{01}(0, 0) = 0$.

After reparameterizing in terms of π (probability of experiencing event, e.g. death or myocardial reinfarction) and $\beta = f(\gamma, \phi))$ (log odds ratio of compliance for specified sensitivity), Roy et al. showed that the observed-data likelihood is

$$L(\pi, \beta|Y, A, Z, X) = \sum_{s=0}^{3} [\pi_Z^{S=s}]^Y [1 - \pi_Z^{S=s}]^{1-Y} \Pr(S = s|X, \beta) G(s, A, Z), \tag{5}$$

where $\pi_Z^{S=s}$ is the probability that observed $Y = 1$, given $S = s$ and allocation to arm Z, and

$$\begin{aligned}
G(s, A, Z) &= \mathrm{I}(s = 0)\{1 - A\} + \mathrm{I}(s = 1)\{A(1 - Z) + (1 - A)Z\} \\
&\quad + \mathrm{I}(s = 2)\{AZ + (1 - A)(1 - Z)\} + \mathrm{I}(s = 3)A.
\end{aligned}$$

Now Y and A are observed values; e.g. $A = A_1$ (A_0) if allocated to active treatment (placebo). We can decompose the expression (5) for each stratum:

$$
\begin{aligned}
L(\pi,\beta|Y=1,A=1,Z=1,X) &= \pi_1^{s=(0,1)}\Pr\left(S=(0,1)|X,\beta\right)\\
&+ \pi_1^{s=(1,1)}\Pr\left(S=(1,1)|X,\beta\right),\\
L(\pi,\beta|Y=0,A=1,Z=1,X) &= [1-\pi_1^{s=(0,1)}]\Pr\left(S=(0,1)|X,\beta\right)\\
&+ [1-\pi_1^{s=(1,1)}]\Pr\left(S=(1,1)|X,\beta\right),\\
L(\pi,\beta|Y=1,A=0,Z=1,X) &= \pi_1^{s=(0,0)}\Pr\left(S=(0,0)|X,\beta\right)\\
&+ \pi_1^{s=(1,0)}\Pr\left(S=(1,0)|X,\beta\right),\\
L(\pi,\beta|Y=0,A=0,Z=1,X) &= [1-\pi_1^{s=(0,0)}]\Pr\left(S=(0,0)|X,\beta\right)\\
&+[1-\pi_1^{s=(1,0)}]\Pr\left(S=(1,0)|X,\beta\right),\\
L(\pi,\beta|Y=1,A=1,Z=0,X) &= \pi_0^{s=(1,0)}\Pr\left(S=(1,0)|X,\beta\right)\\
&+ \pi_0^{s=(1,1)}\Pr\left(S=(1,1)|X,\beta\right),\\
L(\pi,\beta|Y=0,A=1,Z=0,X) &= [1-\pi_0^{s=(1,0)}]\Pr\left(S=(1,0)|X,\beta\right)\\
&+ [1-\pi_0^{s=(1,1)}]\Pr\left(S=(1,1)|X,\beta\right),\\
L(\pi,\beta|Y=1,A=0,Z=0,X) &= \pi_0^{s=(0,0)}\Pr\left(S=(0,0)|X,\beta\right)\\
&+ \pi_0^{s=(0,1)}\Pr\left(S=(0,1)|X,\beta\right),\\
L(\pi,\beta|Y=0,A=0,Z=0,X) &= [1-\pi_0^{s=(0,0)}]\Pr\left(S=(0,0)|X,\beta\right)\\
&+ [1-\pi_0^{s=(0,1)}]\Pr\left(S=(0,1)|X,\beta\right).
\end{aligned}
\tag{6}
$$

By the exclusion restriction $\pi_1^{s=(0,0)} = \pi_0^{s=(0,0)}$, i.e.the risk of experiencing event of interest (e.g. death) is independent of the arm of allocation among the people who would comply with neither allocation. Writing

$$
\begin{aligned}
&\pi_1 = \pi_1^{s=(0,1)}, \pi_2 = \pi_1^{s=(1,1)}, \pi_3 = \pi_1^{s=(0,0)}, \pi_4 = \pi_1^{s=(1,0)},\\
&\pi_5 = \pi_0^{s=(1,0)}, \pi_6 = \pi_0^{s=(1,1)}, \pi_7 = \pi_0^{s=(0,1)},
\end{aligned}
\tag{7}
$$

we obtain 7 parameters captured by π from the likelihoods above using logistic models:

$$
\begin{aligned}
&\Pr[Y=1|A=1,Z=1] = \pi_1\mu_{01} + \pi_2\mu_{11},\\
&\Pr[Y=1|A=0,Z=1] = \pi_3\mu_{00} + \pi_4\mu_{10},\\
&\Pr[Y=1|A=1,Z=0] = \pi_5\mu_{10} + \pi_6\mu_{11},\\
&\Pr[Y=1|A=0,Z=0] = \pi_3\mu_{00} + \pi_7\mu_{01}.
\end{aligned}
\tag{8}
$$

We then obtain the stratum-specific relative risks for experiencing an event (death/reinfarction) as

$$
\tau_{11} = \frac{\pi_1^{s=(1,1)}}{\pi_0^{s=(1,1)}} = \frac{\hat{\pi}_2}{\hat{\pi}_6}, \quad \tau_{01} = \frac{\pi_1^{s=(0,1)}}{\pi_0^{s=(0,1)}} = \frac{\hat{\pi}_1}{\hat{\pi}_7}, \quad \tau_{10} = \frac{\pi_1^{s=(1,0)}}{\pi_0^{s=(1,0)}} = \frac{\hat{\pi}_4}{\hat{\pi}_5}. \tag{9}
$$

The τ_{ij} above provides the desired principal (causal) effects in terms of causal risk ratios obtained as medians of posterior relative risks of event (death or reinfarction) for each stratum defined by compliance type:

(i) τ_{11}: causal risk ratio of event due to compliance with HRT treatment relative to placebo among the subgroup of patients who would comply with either treatment allocation, i.e. $S=(1,1)$,

(ii) τ_{01}: causal risk ratio of event due to compliance with HRT treatment only among women who would comply if allocated to it, i.e. $S=(0,1)$ and

(iii) τ_{10}: causal risk ratio of event due to compliance with placebo treatment only among the subgroup who would comply if allocated to it, i.e. $S=(1,0)$.

The parameters above (Equation 9) can be estimated using Bayesian methods with suitable priors. Using uninformative (flat) priors $\pi \sim U(0,1)$, for example, may be satisfactory for our analyses given the likelihood of a typical trial data to dominate such priors and the fact that randomized trials are principally designed to be conclusive [16].

To extend the methods which adjust for noncompliance in one treatment arm to adjusting for noncompliance in two treatment arms, we will use a Bayesian approach to apply principal stratification using the Roy et al. [30] model reviewed above for survival data but which was originally proposed for binary data. Specifically we perform a comprehensive model selection to obtain arm-specific optimal predictors of compliance and develop a causal model linking the two marginal models from which we obtain causal effects for each stratum.

V. Application to the Esprit study

ITT analysis of the Esprit data showed no statistical difference between HRT and placebo treatment with hazard ratio results (HR$=\exp(\hat{\psi})=0.795$, p-value$=0.335$, $95\%\text{CI}: 0.498, 1.268$) suggesting a beneficial effect of HRT over placebo with respect to death. However, the ITT analysis took no account of compliance data. The rate of observed compliance was higher in the placebo (63%) arm compared to the HRT (42%) treatment arm which may be attributed to noncompliance as a result of possible unpleasant symptoms like bleeding. Utilizing compliance data, we consider two outcomes (all-cause mortality and myocardial reinfarction) for causal analysis using the methods described in the previous section.

When applying the Roy et al. [30] method for survival data, we use relative risks to approximate hazard ratios. This is justifiable for our analysis given that under short follow-up times (monthly) and small event rates conditions (death and myocardial reinfarctions), relative risk has been shown to be an algebraic approximation of hazard ratio, i.e. $\exp(\bar{\hat{\psi}}) \cong \tilde{\hat{\tau}}$ [36].

V.1. Predicting compliance in each arm

In predicting compliance for each arm, we first choose a full (saturated) model consisting of all potential predictors of compliance to treatment allocation on the basis of both clinical and statistical plausibility. In addition we consider penalized maximum likelihood estimation regression versions of this (saturated) model and also Lasso (Least Absolute Shrinkage and Selection Operator) model obtained by the method which performs the twin tasks of parameter shrinkage and model selection [38]. To evaluate the predictive performance of the selected models, we used calibration slope, percentage of optimism and calculated discrimination's concordance c-statistics from the reported Somers D_{xy} statistics. These validation measures will be recorded for each individual arm in five models:

(i) Original saturated model with all the 9 predictors without any selection:

$$\begin{aligned}\text{logit}(\mu_j) = \quad & \gamma_0 + \gamma_1 \text{Hysterectomy} + \gamma_2 \text{Smoking status} + \gamma_3 \text{Social-class} + \\ & \gamma_4 \text{Age} + \gamma_5 \text{CVD Risks} + \gamma_6 \text{Diabetes} + \gamma_7 \text{Fracture} + \gamma_8 \text{Alcohol} + \gamma_9 \text{HRT},\end{aligned}$$

where μ is the probability of compliance with treatment (placebo/HRT) allocation and histories of hysterectomy, cerebrovascular disease (CVD) risks, diabetes, fracture, alcohol, HRT use together with smoking status are taken as binary 0/1 predictors.

(ii) Reduced model obtained from (i) above by stepwise backward elimination procedures using Akaike information criterion (AIC) stopping rule and 0.10 significance level for a variable to be retained in a model.

(iii) Model fitted with the retained predictors in reduced model (ii) above but the predictors assumed pre-specified (following suggestion by Harrell et al. [14]).

(iv) Intermediate model composed of 6 variables constructed using penalized maximum likelihood estimation with modified AIC ($\chi^2 > 2\text{df}$).

(v) Least Absolute Shrinkage and Selection Operator (Lasso) model selection from the original (i) above.

V.2. Validation: evaluating performance of selected models

We used enhanced bootstrap on all aspects of models development (selection and estimation procedures) to revalidate on samples taken with replacement from the whole sample and apply on the five models specified above. The reduced model was obtained from the original model using stepwise backward elimination procedures using AIC stopping rule and 0.10 significance level for retaining a predictor in a model. The variables selected for the reduced model were consistently (90%) selected across bootstrap resamples. These were the same predictors deemed important by the backward elimination algorithm.

Table 1 provide results showing the performance of the 5 prediction models in terms of validation indices outlined earlier (see Section IV.1). The saturated (original) model consisting

TABLE 1. Validation performance of 5 models in terms of calibration, concordance and optimism (Section IV.1)

Model	Selected predictors	Calibration slope	Optimism (%)	Concordance *c*-statistics
(i) Original				
HRT		0.818	6.1	0.639
Placebo		0.671	8.6	0.573
(ii) Reduced				
HRT	(hyst+smk+CVD)	0.827	5.8	0.620
Placebo	(hyst+smk+alc)	0.667	8.1	0.566
(iii) Reduced†				
HRT	(hyst+smk+CVD)	0.961	1.4	0.642
Placebo	(hyst+smk+alc)	0.950	2.0	0.597
(iv) Intermediate (6 predictors)				
HRT		0.879	4.1	0.636
Placebo		0.766	6.0	0.580
(v) Lasso				
HRT	(hyst+smk+age+CVD)	0.935	2.3	0.647
Placebo	(hyst+smk+alc)	0.925	2.3	0.595

hyst≡hysterectomy; smk≡ smoking status; alc≡alcohol; CVD≡cerebrovascular disease; †model assumed pre-specified

of 9 predictors produced better predictions of compliance to HRT than placebo. Here predicting compliance to HRT and placebo, the original models would be overfitted by 18% and 33% respectively. Also these models would be optimistic by 6% and 9% respectively in predicting compliance to HRT and placebo. We note that although the observed rates of compliance for placebo were higher than for HRT, the relatively 'poor' performance of the former compared to the later predictive models may be attributed to poor quality of compliance data owing to the common practice to monitor compliance with active treatment more accurately than placebo treatment.

Compared to saturated model, the reduced model would perform relatively better in predicting compliance to HRT compared to predicting placebo: specifically the reduced model

predicting compliance to HRT would perform better at distinguishing compliers from non-compliers (concordance $c = 0.620$) than the reduced model predicting compliance to placebo ($c = 0.566$). Reduced models predicting compliance to placebo would be more optimistic (8%) than those predicting compliance to HRT (6%). Predictions for compliance to HRT using the reduced model would be equally well calibrated (slope $= 0.83$) compared to predictions from the original full model (slope$= 0.82$). As expected, the model with 3 predictors if assumed pre-specified performed 'best' in terms of both calibration and discrimination among the 5 models considered in predicting compliance to both HRT and placebo. These models also produced least optimistic fits for predicting compliance to both arms. Specifically predictions of compliance to both HRT and placebo using the 3 predictors assumed pre-specified were almost perfectly calibrated (0.96 and 0.95) and least optimistic (1% and 2%).

Validation of the model with 6 predictors showed adequate performance with intermediate measures between the saturated models composed of 9 predictors and the Lasso models. We observe that predictions of compliance to both HRT and placebo using the intermediate model performed relatively 'better' than both predictions of compliance using the reduced model. For example, predictions of compliance to placebo using the intermediate model is now equally optimistic (6%) as predictions of compliance to HRT using the reduced model, a result which may make the assumption of 'no preference to one treatment over the other' (extended monotonicity) plausible for the Esprit study.

Besides the reduced model fitted with predictors if assumed pre-specified, Lasso models produced the best calibrated and discriminative models predictive of compliance to both HRT and placebo (Table 1, lower panel). Predictions of compliance to both HRT and placebo using the Lasso models were the least optimistic (2%) and almost perfectly calibrated (slope$= 0.93$). Although the Lasso prediction models performed 'best' compared to predictions from the other four models, we note that the method uses the same tuning parameter for all coefficients. The drawback of shrinking all coefficients by a constant, even for those non-zero coefficients, may result in suboptimal choice of covariates with the potential to exclude potential predictors, i.e. data wastage. Moreover, Lasso is known to fail efficient model selection in the presence of correlated variable where it tends to select one variable from a group and ignore the others [45]. Overall, the intermediate models provided substantially improved predictions of compliance to both HRT and placebo in terms of calibration and optimism without affecting the capability to discriminate between compliers and non-compliers.

Table 2 provides estimated median compliance proportion for each of the four strata for both all-cause mortality and myocardial reinfarction outcomes. On average, the estimated median probabilities of compliance was higher among those patients allocated to placebo for both all-cause mortality (myocardial reinfarction) outcomes ($\bar{\hat{\mu}}_0(\mathrm{x}) = 0.567$ (0.565)) compared to

TABLE 2. Median compliance proportion per principal stratum for different values of ϕ

Stratum	Outcome: All-cause mortality					Outcome: Reinfarction				
	ϕ					ϕ				
	0	0.2	0.5	0.8	1	0	0.2	0.5	0.8	1
$\hat{\mu}_{11}(\mathrm{x})$	0.264	0.296	0.353	0.406	0.460	0.266	0.303	0.359	0.417	0.456
$\hat{\mu}_{01}(\mathrm{x})$	0.197	0.165	0.105	0.047	0	0.200	0.164	0.111	0.059	0.023
$\hat{\mu}_{10}(\mathrm{x})$	0.303	0.262	0.211	0.155	0.105	0.300	0.263	0.206	0.142	0.102
$\hat{\mu}_{00}(\mathrm{x})$	0.236	0.277	0.330	0.391	0.435	0.235	0.270	0.325	0.382	0.419

those on HRT tablets ($\bar{\hat{\mu}}_1(\mathrm{x})=0.461$ (0.470)), i.e. the ratio $U(\mathrm{x})=\min\{1, \frac{\bar{\hat{\mu}}_1(\mathrm{x})}{\bar{\hat{\mu}}_0(\mathrm{x})}\}=0.795$ (0.810). We note a likelihood that a higher prevalence (proportion) of placebo compliance compared to HRT may be a limitation of the model to effectively evaluate active HRT efficacy.

For all the four strata at mild value of sensitivity parameter ($\phi=0.2$), the group with the highest prevalence was patients who would comply with either treatment ($\bar{\hat{\mu}}_{11}=0.296$ (0.303)) and the group with the lowest prevalence is those who would only comply with HRT tablets ($\bar{\hat{\mu}}_{01}=0.165$ (0.164)). The median proportion of compliance among those patients who would comply with placebo only and those who would not comply with either treatment allocation were $\bar{\hat{\mu}}_{10}=0.262$ (0.263) and $\bar{\hat{\mu}}_{00}=0.277$ (0.270) respectively. Overall the estimated rates of potential compliance were generally similar in each stratum for a given value of ϕ (except for perfect correlation $\phi=1$). The apparent independence between the potential compliance rates and outcome may be an indication of the plausibility of conditional compliance assumption with respect to the Esprit data.

V.3. Causal risk ratio inference

We estimated the causal risk ratio parameters τ_{ij} (Eq. 9) in a Bayesian setting using non-informative priors for all log odds ratio parameters γ_j for potential predictors of compliance. We specified uniform $(0,1)$ priors for the $\pi_Z^{\mathcal{S}=s}$ $(\pi_i, i=1,\ldots,7)$ parameters, $z=0,1$, $\mathcal{S}$=(0,0),(1,0),(0,1),(1,1) and set the sensitivity parameter $\phi=0, 0.2, 0.5$ and 0.8. The choice of ϕ were motivated by the need to explore all possible compliance behaviours including conditional independence ($\phi=0$) and almost-perfect correlation ($\phi=0.8$). We ran three chains: null starting values for chain one, mean and median values from a trial run for chains two and three respectively. For convergence assessment, we ran simulation for $101,000$ iterations for each of the three chains and excluded the first $1,000$ as burn-in. Posterior median relative risks provided Bayesian point estimates for each stratum.

TABLE 3. Causal risk ratio estimates (means of median posterior relative risks) for mortality and reinfarction (mean median 95% CI) for each stratum for different values of ϕ: (a) All-cause mortality (b) Myocardial reinfarction.

	Compliance with both HRT and placebo		Compliance with HRT only		Compliance with placebo only	
ϕ	$\hat{\pi}_2(\pi_1^{s=(1,1)})$	$\hat{\pi}_6(\pi_0^{s=(1,1)})$	$\hat{\pi}_1(\pi_1^{s=(0,1)})$	$\hat{\pi}_7(\pi_0^{s=(0,1)})$	$\hat{\pi}_4(\pi_1^{s=(1,0)})$	$\hat{\pi}_5(\pi_0^{s=(1,0)})$
(a)						
0	0.011	0.012	0.053	0.099	0.051	0.055
	$\tau_{11}=0.941$	(0.026,34.349)	$\tau_{01}=0.534$	(0.101, 1.347)	$\tau_{10}=0.933$	(0.340,3.0129)
0.2	0.011	0.012	0.063	0.118	0.056	0.060
	$\tau_{11}=0.867$	(0.023,31.489)	$\tau_{01}=0.533$	(0.083,1.522)	$\tau_{10}=0.938$	(0.267,4.038)
0.5	0.011	0.012	0.092	0.168	0.064	0.073
	$\tau_{11}=0.878$	(0.025,30.539)	$\tau_{01}=0.533$	(0.068,2.310)	$\tau_{10}=0.874$	(0.126,5.702)
0.8	0.012	0.012	0.179	0.243	0.053	0.097
	$\tau_{11}=0.974$	(0.032,30.698)	$\tau_{01}=0.723$	(0.065,10.399)	$\tau_{10}=0.560$	(0.024,5.423)
(b)						
0	0.012	0.021	0.106	0.143	0.094	0.107
	$\tau_{11}=0.561$	(0.061,19.989)	$\tau_{01}=0.733$	(0.293,1.406)	$\tau_{10}=0.881$	(0.474,2.182)
0.2	0.010	0.026	0.130	0.172	0.105	0.113
	$\tau_{11}=0.397$	(0.011,13.839)	$\tau_{01}=0.752$	(0.301,1.561)	$\tau_{10}=0.932$	(0.448,3.804)
0.5	0.008	0.044	0.200	0.244	0.125	0.109
	$\tau_{11}=0.204$	(0.006,5.479)	$\tau_{01}=0.835$	(0.432,2.881)	$\tau_{10}=1.138$	(0.351,17.590)
0.8	0.008	0.069	0.419	0.256	0.119	0.059
	$\tau_{11}=0.121$	(0.004,1.186)	$\tau_{01}=1.594$	(0.465,45.238)	$\tau_{10}=1.905$	(0.100,69.149)

[a]All-cause mortality; [b]Myocardial reinfarction

Table 3 provide causal risk ratio estimates (Bayesian principal effects) obtained from mean posterior median relative risks for each stratum and corresponding mean 95% credible intervals for different values of sensitivity parameter ϕ. Here a posterior relative risk τ was obtained as the ratio of two probabilities of experiencing an event due to compliance with one treatment allocation relative to another in a stratum. Most of results (not all) show posterior median relative risks of less than one for all values of ϕ which indicates lower risks for mortality

and myocardial reinfarction for those women randomized to HRT who would be highly compliant with their treatment allocation. A primary interest is the quantity $\tau_{01} = [\pi_1^{S=(0,1)}][\pi_0^{S=(0,1)}]^{-1}$, i.e. the posterior (causal) relative risk for mortality/reinfarction among the subgroup who would comply with HRT treatment only. The results shows that the mean median 95% credible intervals widened with increase in ϕ values, an indication of less correlation between HRT treatment and placebo compliances. Overall, the results indicated that HRT tablets reduced risks for myocardial reinfarction more than the reduction in risks for all-cause mortality.

For a mild correlation value ($\phi = 0.2$), the results suggest that compliance with HRT tablets only would substantially reduce the risk of death by about 47%, i.e. causal risk ratio $\tau_{01} = 0.533$, 95% CI : $0.083, 1.522$. Also for this value of sensitivity parameter, the results suggest that compliance with HRT treatment compared to taking placebo among those who would comply with either treatment reduced the risk for all-cause mortality by 13%, i.e. causal risk ratio $\tau_{11} = 0.867$, 95% CI : $0.023, 31.489$. However, compliance with placebo treatment only would marginally reduce the risk of death by about 6%, i.e. causal risk ratio $\tau_{01} = 0.938, 95\%$ CI : $0.267, 4.038$. In general we note that compliance with HRT treatment only consistently suggested reduction of risk for death at all sensitivity parameter values ϕ. For example, when $\phi = 0.8$, while compliance with HRT treatment compared to taking placebo among those who would comply with either treatment would essentially have no effect ($\tau_{11} = 0.974, 95\%$ CI : $0.032, 30.698$), compliance with placebo treatment only would reduce the risk of death by about 44%, i.e. causal risk ratio $\tau_{10} = 0.560, 95\%$ CI : $0.024, 5.423$.

The size of causal (principal) effects varied according to the value of sensitivity parameter. Risks for myocardial reinfarction among those who would comply with placebo treatment only increased with increase in the value of sensitivity parameter ϕ. On the other hand results show reduction in risks for myocardial reinfarction among those who would comply with either placebo or HRT treatment as the value of ϕ increased. As expected the risks for both death and myocardial reinfarction outcomes were higher among those who would comply with placebo only compared to those who would comply with HRT only for any chosen value of the sensitivity parameter ϕ.

V.4. Sensitivity analysis

As outlined in assumptions (Section III), application of the Roy et al. [30] method is premised on plausibility of the crucial, but untestable, conditional compliance assumption which posits that the potential outcome is independent of the set of covariates predictive of treatment compliance given a compliance type and treatment assignment. Hence the task of selecting suitable predictors of treatment compliance constitutes an integral part of ensuring plausibility of this assumption. A sensitivity analysis using different models (i.e sets of predictors) may be used to assess how the causal estimates depend on departures from this crucial assumption [35].

TABLE 4. Sensitivity analysis using 3, 6 and 9 predictors of compliance: Causal risk ratio (median 95% CI)

	All-cause mortality			Myocardial reinfarction		
ϕ	Comply HRT and placebo	Comply HRT only	Comply placebo only	Comply HRT and placebo	Comply HRT only	Comply placebo only
A						
0	1.126	0.546	0.872	0.539	0.739	0.874
	(0.030,44.588)	(0.106,1.417)	(0.281,2.260)	(0.014,19.650)	(0.296,1.643)	(0.462,2.522)
0.2	1.169	0.543	0.833	0.419	0.759	0.904
	(0.031,42.058)	(0.088,1.633)	(0.210,2.428)	(0.012,14.459)	(0.298,1.839)	(0.419,3.684)
0.5	1.247	0.590	0.716	0.250	0.839	1.004
	(0.036,41.790)	(0.070,3.136)	(0.067,2.316)	(0.007,6.184)	(0.326,4.006)	(0.248,9.907)
0.8	1.460	0.854	0.315	0.175	1.772	0.857
	(0.050,44.267)	(0.068,14.310)	(0.012,1.769)	(0.006,1.862)	(0.447,52.647)	(0.030,40.108)
B	(τ_{11})	(τ_{01})	(τ_{10})	(τ_{11})	(τ_{01})	(τ_{10})
0	0.941	0.534	0.933	0.561	0.733	0.881
	(0.026,34.349)	(0.101,1.347)	(0.340,3.012)	(0.061,19.989)	(0.293,1.406)	(0.474,2.182)
0.2	0.867	0.533	0.938	0.397	0.752	0.932
	(0.023,31.489)	(0.083,1.522)	(0.267,4.038)	(0.011,13.839)	(0.301,1.561)	(0.448,3.804)
0.5	0.878	0.533	0.874	0.204	0.835	1.138
	(0.025,30.539)	(0.068,2.310)	(0.126,5.702)	(0.006,5.479)	(0.342,2.881)	(0.351,17.590)
0.8	0.974	0.723	0.560	0.121	1.594	1.905
	(0.032,30.698)	(0.065,10.399)	(0.024,5.423)	(0.004,1.186)	(0.465,45.238)	(0.100,69.149)
C						
0	1.264	0.533	0.889	0.707	0.737	0.841
	(0.032,47.799)	(0.102,1.323)	(0.331,2.161)	(0.018,26.389)	(0.306,1.396)	(0.455,1.702)
0.2	1.341	0.530	0.859	0.562	0.748	0.859
	(0.035,47.004)	(0.081,1.523)	(0.254,2.151)	(0.014,21.870)	(0.288,1.575)	(0.427,2.243)
0.5	1.413	0.553	0.782	0.356	0.810	0.900
	(0.038,49.229)	(0.065,2.290)	(0.112,2.268)	(0.010,10.390)	(0.330,2.621)	(0.296,4.480)
0.8	1.707	0.706	0.463	0.214	1.374	0.812
	(0.055,49.708)	(0.056,9.777)	(0.022,1.856)	(0.007,3.491)	(0.423,30.800)	(0.043,17.929)

Model comprising [A]3 (Lasso); [B]6 (Intermediate) and [C]9 (Saturated) predictors of compliance

Table 4 show results in terms of causal relative risks for models predicting compliance using three sets of predictors considered earlier: using $3, 6$ and 9 predictors from Lasso, intermediate and all plausible predictors respectively. For mild values of sensitivity parameters, model selection using penalized maximum procedures (6 predictors) produced 'best' causal risk ratio estimates showing reduction of risks for both all-cause mortality and myocardial reinfarction. Specifically HRT treatment was consistently effective (reduced risks) among those who would comply with HRT only (τ_{01}). The efficacy corresponding to this stratum was not dependent on the chosen value of sensitivity parameter. On the other hand the causal risk ratio estimates using 3 and 9 sets of predictors were comparable for all strata considered. In general, given a set of predictors, the results show same trend in principal effects with respect to change in magnitude of the sensitivity parameter ϕ for both outcomes (mortality and myocardial reinfarction). Surprisingly these results using 3 and 9 predictors of compliance now suggested harmful effects (increased risks) of HRT treatment relative to placebo among those who would comply with either treatment.

Results from the sensitivity analysis above (Table 4) may be a useful demonstration of the phenomenon that causal (principal) effects are dependent on the choice of covariates predicting compliance. This may be an indication that the advantages of classical model selection are transferable to the Roy et al. [30] method via use of optimal marginal compliance models, i.e. comprehensive model selection may be useful in providing optimal predictors of compliance to enrich principal stratification. However, we note that while selecting plausible predictors of compliance is an integral component of the method, model selection only acts as an intermediate step that provides covariates for marginal compliance prediction models which are then joined into a causal model using the crucial but unknown sensitivity parameter.

In general we observe from results in both Tables 3 and 4 that for a given stratum and set of selected covariates predicting compliance, the change in resulting causal risk ratio estimates were more pronounced for the reinfarction compared to all-cause mortality outcome. This apparent interaction of outcome with sensitivity parameter ϕ may be attributed to features of the two different outcomes. A possible explanation may be the fact that the choice of optimal predictors of compliance to make the conditional compliance assumption (f) valid might depend on outcome (death/reinfarction). Such an association may make conditional prediction assumption questionable for the Esprit data especially with regard to history of hysterectomy and cerebrovascular risks which are likely to be associated with better treatment compliance and subsequently favourable outcome.

VI. Discussion

By using optimal predictors of treatment compliance at mild values of the sensitivity parameter ϕ, compliance with HRT tablets showed reduction in risks for both all-cause mortality and myocardial reinfarction. Compliance with HRT treatment compared to taking placebo among those who would comply with either treatment also indicated beneficial effects in reducing risks for both outcomes. Compliance with HRT treatment suggested beneficial effects compared to placebo for all other values of ϕ among the subgroup who would comply with HRT treatment only and those who would comply with either treatment allocation. However, causal risk ratios estimating efficacy of compliance to either treatment (τ_{11}) had relatively wider mean 95% credible intervals compared to estimates for efficacy of compliance with HRT only (τ_{01}) or placebo only (τ_{10}). In general, the risk for myocardial reinfarction reduced with increase in the value of sensitivity parameter ϕ among those women who would comply with either treatment. On the other hand the risk of myocardial reinfarction increased with increase in ϕ among those who would comply with placebo treatment only. Overall as expected, for any chosen value of the sensitivity parameter ϕ, the risks for both death and myocardial reinfarction were higher among those who would comply with placebo only compared to those who would comply with HRT only.

The variation in the HRT efficacy estimates from the Roy et al. [30] model may be an indication of difference in compliance behaviour between those allocated to placebo and HRT treatment. By adjusting for noncompliance in both arms, the Roy et al. method perhaps accounts for potential correlation between compliance behaviours in respective arms through the chosen value of the sensitivity parameter which implicitly makes the results depend on ϕ. The fact that the results vary a lot with ϕ and yet we do not know its value suggests benefits of HRT treatment among those who comply when allocated it, i.e. strong monotonicity assumption (strong correlation in compliance behaviour between the two arms). Finally our analyses with flat priors may be considered adequate given the likelihood of a typical trial data to dominate such priors in addition to the fact that randomized trials are principally designed to be conclusive [16].

Noncompliance with treatment assignment in both arms of a clinical trial is likely to complicate efficacy estimation. Here the ITT provide a biased estimator for the true causal estimate even under homogeneous treatment effects assumption. Extending and applying the Roy et al. model to survival data (Esprit study) may be suitable by utilizing key covariates predictive of arm-specific compliance models to develop causal models linking the two marginal models. The resulting principal effects provides efficacy estimates for the different subgroups defined by compliance types. The method performed relatively 'better' than specialist randomization-based causal methods adjusting for noncompliance in one treatment arm only [25]. Simulation studies

indicated satisfactory performance of the method however, the results were heavily dependent on the choice of sensitivity parameters and hence may not be recommended in the presence of known heterogeneous treatment effects which produced large bias and wider corresponding 95% credible intervals. As a result, the method may only be recommended in the presence of sufficient information about compliance behaviours in respective arms.

Model selection in regression may be correctly considered as one of the most significant challenges in modern statistics [17]. Hitherto this challenge has not been extended to include prediction of compliance with treatment assignment in causal modelling. There are presently limited studies integrating model selection for compliance prediction in causal inference. While principal stratification has independently been demonstrated to provide better alternative identification strategies compared to selection model [23], integrating the two strategies may produce even more flexible models under relaxed assumptions. A record of plausible predictors of compliance can be used to effectively address identification problem of causal estimands by reducing bias and relaxing the implicit assumptions [20]. From a clinical perspective, knowledge about predictors of compliance may be a valuable tool to inform treatment decisions. As a result, there is need to adopt comprehensive model selection methods for accurate prediction (of compliance). After model selection, there is further need to use suitable validation indices (e.g. optimism, calibration and discrimination indices) to evaluate performance of selected models. With potentially many recorded baseline covariates, using penalized regression techniques may be recommended for building compliance prediction models.

The merits of standard model selection procedures are transferable to the principal stratification method adjusting for noncompliance in two treatment arms by linking the respective optimal marginal compliance models into an association model [26]. However, application of the method is premised on the plausibility of a defining assumption that potential outcome is independent of the set of covariates predicting treatment compliance for a given stratum. This assumption may be questionable for the Esprit data especially with regard to history of hysterectomy and cerebrovascular risks which potentially have a higher likelihood to be associated with treatment compliance leading to possible efficacy. For example, while the unpleasant experience of bleeding may affect treatment compliance negatively, those with history of cerebrovascular risks may comply with their treatment allocation with a hope to derive potential protective benefits. Further sensitivity analysis on the departure of conditional prediction assumption implicit in the Roy et al. model may be conducted using alternative methods which incorporate less stringent assumptions. For example, by adopting the Bayesian framework introduced by Long et al. [21] to model the principal compliances directly in multitreatment arms and for more general outcomes by treating the principal compliance status as missing data instead of joining them with a user-defined sensitivity parameter ϕ.

Although principal stratification provides a powerful framework which is extendible to analyse complex surrogate outcomes like 'truncation by death' where death occurs before a primary outcome of interest is recorded hence resulting in censored records [31], the method's application and usefulness may be limited to intermediates with fewer categories (e.g. binary) [40]. Although the all-or-nothing compliance suitably applied to the Esprit data, principal stratification in general produce inconsistent causal estimate for a truly continuous stratification variable but which has been coarsened for analysis [28]. As a result, policy informed by analysis based on principal stratification should be implemented with caution owing to the fact that the principal strata themselves remain unidentified.

References

[1] J. M. Albert and D. L. DeMets, *On a model-based approach to estimating efficacy in clinical trials*, Statistics in Medicine, **13**, (1994), 2323 – 2335.

[2] J. D. Angrist, G. W. Imbens D. B. and Rubin, *Identification of causal effects using instrumental variables,* Journal of the American Statistical Association, **91**, (1996) 44 – 455.

[3] P. Armitage, *Attitudes in clinical trials*, Statistics in Medicine, **17**,(1998) 2675–2683.

[4] J. Cheng and D. S. Small, *Bounds on causal effects in three-arm trials with non-compliance*, Journal of the Royal Statistical Society, B, **68(5)**, (2006) 815 – 836.

[5] N. Cherry, K. Gilmour, P. Hannaford, A. Heagerty, M. A. Khan, H. Kitchener, R. McNamee, M. Elstein, C. Kay, M. Seif and H. Buckley, *Oestrogen therapy for prevention of reinfarction in postmenopausal*, women: A randomised placebo controlled trial. Lancet, **360**, (2002) 2001 – 2008.

[6] S. R. Cole and C. E. Frangakis, *The consistency statement in causal inference: A definition or an assumption?*, Epidemiology, **20(1)**, (2009) 3 – 5.

[7] D. R. Cox, *Two further applications of a model for binary regression,* Biometrika, **45**, (1958) 562 – 565.

[8] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications*, Cambridge University Press, Cambridge, (1997).

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

[10] K. Fischer, E. Goetghebeur, B. Vrijens and I. R. White, *A structural mean model to allow for noncompliance in a randomized trial comparing* 2 *active treatments*, Biostatistics, **12(2)**, (2011) 247 – 257.

[11] C. E. Frangakis and D. B. Rubin, *Principal stratification in causal inference*, Biometrics, **58**, (2002) 21 – 29.

[12] D. Grady, S. M. Rubin, D. B. Petitti, C. S. Fox, D. Black, B. Ettinger, V. L. Ernster and S. R. Cummings, *Hormone therapy to prevent disease and prolong life in postmenopausal women*, Annals of Internal Medicine, **117(12)**, (1992) 1016 – 1037.

[13] F. E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis*, Springer Verlag, New York, (2001).

[14] F. E. Harrell, K. L. Lee and D. B. Mark, *Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors*, Statistics in Medicine, **15**, (1996) 361 – 387.

[15] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Verlag, New York, 2nd. edition, (2009).

[16] D. F. Heitjtan, A. J. Moskowitz and W. Whang, . *Bayesian estimation of cost-effectiveness ratios from clinical trials*, Health Economics, **8**, (1991) 191 – 201.

[17] T. Hesterberg, N. H. Choi, L. Meier and C. Fraley, *Least angle and ℓ_1 penalized regression: A review*, Statistics Surveys, **2**, (2008) 61 – 93.

[18] H. Jin and D. B. Rubin, *Principal stratification for causal inference with extended partial compliance*, Journal of the American Statistical Association, **103(481)**, (2008) 101 – 111.

[19] J. M. Lachin, *Statistical considerations in the intent-to-treat principle*, Controlled Clinical Trials, **21**,(2000) 167 – 189.

[20] R. J. A. Little, Q. Long and X. Lin, *A comparison of methods for estimating the causal effect of a treatment inrandomized clinical trials subject to noncompliance*, Biometrics, **65**, (2009) 640 – 649.

[21] Q. Long, R. J. A. Little and X. Lin, . *Estimating causal effects in trials involving multitreatment arms subject to non-compliance: a Bayesian framework*, Journal of the Royal Statistical Society, C, **59(3)**, (2010) 513 – 531.

[22] D. Machin and P. M. Fayers, . *Randomized Clinical Trials: Design, Practice and Reporting*, Wiley-Blackwell, Chichester (2010).

[23] F. Mealli and B. Pacini, *Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness*, Computational Statistics & Data Analysis, **53(2)**, (2008) 507 – 516.

[24] N. Nagelkerke, V. Fidler, R. Bernsen and M. Borgdorff, *Estimating treatment effects in randomized clinical trials in the presence of non-compliance*, Statistics in Medicine, **19**, (2000) 1849 – 1864.

[25] L. Odondi and R. McNamee, *Performance of statistical methods for analysing survival data in the presence of non-random compliance*, Statistics in Medicine, **29**, (2010) 2994 – 3003.

[26] L. Odondi and R. McNamee, *Applying optimal model selection in principal stratification for causal inference*, Statistics in Medicine, **32**,(2013) 1815 – 1828.

[27] S. Pecorelli and L. Fallo, *Hormone replacement therapy in gynecological cancer survivors*, Critical Reviews in Oncology/Hematology, **27**, (1998) 1 – 10.

[28] J. M. Robins, A. Rotnitzky and S. Vansteelandt, *Discussions of Principal stratification designs to estimate input data missing due to death*, Biometrics, **63**, (2007) 650 – 654.

[29] P. R. Rosenbaum, . *The consequences of adjustment for a concomitant variable that has been affected by the treatment*, Journal of the Royal Statistical Society A., **147(5)**, (1984) 656 – 666.

[30] J. Roy, J. W. Hogan and B. H. Marcus, *Principal stratification with predictors of compliance for randomized trials with* 2 *active treatments*, Biostatistics, **9(2)**, (2008) 277 – 289.

[31] D. B. Rubin, . *Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death*, Statistical Science, **21(3)**, (2006) 299 – 309.

[32] S. J. Sheather, *A Modern Approach to Regression with R*, Springer, New York (2009).

[33] R. H. Somers, A new asymmetric measure of association for ordinal variables, American Sociological Review, **27**, (1962) 799 – 811.

[34] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, Springer, New York (2009).

[35] R. Stone, *The assumptions on which causal inferences rest*, Journal of the Royal Statistical Society, B., **55(2)**, (1993) 455 – 466.

[36] M. J. Symons and D. T. Moore, *Hazard rate ratio and prospective epidemiological studies*, Journal of Clinical Epidemiology, **55**, (2002) 893 – 899.

[37] L. Taylor and X. H. Zhou, *Multiple imputation methods for treatment noncompliance and nonresponse in randomized clinical trials*, Biometrics, **65**, (2009) 88 – 95.

[38] R. J. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society B., **58(1)**, (1996) 267 – 288.

[39] J. C. Houwelingen, *Shrinkage and penalized likelihood methods to improve diagnostic accuracy*, Statistic Nederlica, **55**, (2001) 17 – 34.

[40] T. J. VanderWeele, *Causal mediation analysis with survival data*, Epidemiology, **22(4)**, (2011) 582 – 585.

[41] H. T. Weisberg, *Bias and Causation: Models and Judgment for Valid Comparisons*, Wiley, New York (2010).

[42] I. R. White, *Uses and limitations of randomization-based efficacy estimators*, Statistical Methods in Medical Research, **14**, (2005) 327 – 347.

[43] I. R. White and S. J. Pocock, *Statistical reporting of clinical trials with individual changes from allocated treatment*, Statistics in Medicine, **15**, (1996) 249–262.

[44] C. M. Zigler and T. R. Belin, . *The potential for bias in principal causal effect estimation when treatment received depends on a key covariate*, The Annals of Applied Statistics, **5(3)**, (2011) 1876 – 1892.

[45] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society B., **67(2)**, (2005) 301 – 320.

Lang'o Odondi

e-mail: Lango.Odondi@bristol.ac.uk

MRC ConDuCT Hub, School of Social and Community Medicine, University of Bristol, UK