Published in final edited form as:
J R Stat Soc Ser C Appl Stat. 2013 November ; 62(5): 687–704. doi:10.1111/rssc.12012.

Assessing the heterogeneity of treatment effects via potential outcomes of individual patients

Zhiwei Zhang,
Food and Drug Administration, Silver Spring, USA

Chenguang Wang,
Johns Hopkins University School of Medicine, Baltimore, USA

Lei Nie, and
Food and Drug Administration, Silver Spring, USA

Guoxing Soon
Food and Drug Administration, Silver Spring, USA

Summary

There is growing interest in understanding the heterogeneity of treatment effects (HTE), which has important implications in treatment evaluation and selection. The standard approach to assessing HTE (i.e. subgroup analyses based on known effect modifiers) is informative about the heterogeneity between subpopulations but not within. It is arguably more informative to assess HTE in terms of individual treatment effects, which can be defined by using potential outcomes. However, estimation of HTE based on potential outcomes is challenged by the lack of complete identifiability. The paper proposes methods to deal with the identifiability problem by using relevant information in baseline covariates and repeated measurements. If a set of covariates is sufficient for explaining the dependence between potential outcomes, the joint distribution of potential outcomes and hence all measures of HTE will then be identified under a conditional independence assumption. Possible violations of this assumption can be addressed by including a random effect to account for residual dependence or by specifying the conditional dependence structure directly. The methods proposed are shown to reduce effectively the uncertainty about HTE in a trial of human immunodeficiency virus.

Keywords

Causal inference; Conditional independence; Copula; Counterfactual; Random effect; Sensitivity analysis

Address for correspondence: Zhiwei Zhang, Center for Devices and Radiological Health, Division of Biostatistics, Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA. zhiwei.zhang@yahoo.com.

1. Introduction

It is well recognized that treatment effects can be heterogeneous, i.e. that the same treatment can have different effects on different patients. Understanding the heterogeneity of treatment effects (HTE) is of increasing importance in treatment evaluation and selection. The standard approach to assessing HTE is subgroup analyses (or, more generally, regression analyses) based on known effect modifiers, which may be demographic variables, disease aetiology, certain baseline measurements and genetic markers (e.g. Peto (1982), Gail and Simon (1985), Russek-Cohen and Simon (1997) and Pocock *et al.* (2002)). A subgroup analysis comparing treatment effects on different subpopulations is informative about the HTE between subpopulations but not within. In fact, one could think of an individual patient's treatment outcomes as determined by a large set of prognostic factors and effect modifiers. Ideally, with all relevant information available and used correctly, one would be able to predict precisely the outcome of an individual patient under a given treatment. In reality, however, some effect modifiers may be unknown to the scientific community, resulting in residual HTE that cannot be explained by known effect modifiers.

It is perhaps more natural to think of HTE in terms of individual potential outcomes (Gadbury and Iyer, 2000; Gadbury *et al.*, 2001, 2004; Poulson *et al.*, 2012). Under the causal model of Rubin (1974), each patient has a potential outcome under each possible treatment, and the effect of an experimental treatment relative to a control can be assessed on each individual patient by comparing the corresponding potential outcomes. Consider, for example, a randomized, double-blinded, placebo-controlled, confirmatory clinical trial known as ‘MOTIVATE’ (maraviroc *versus* optimized therapy in viraemic antiretroviral treatment-experienced patients; Gulick *et al.* (2008)). Maraviroc is a CC chemokine receptor 5 antagonist and a new antiretroviral drug for treating human immunodeficiency virus type 1 (HIV-1). The MOTIVATE trial compares maraviroc with placebo, each combined with optimized background therapy (OBT), with respect to a success rate (virologic response at week 48 of treatment; see Section 4 for details). Because the outcome is binary, patients can be classified into four categories according to their potential outcomes under the two treatments, as shown in Table 1. The observed success rates are 57.5% and 22.5% for maraviroc and placebo respectively. Because the difference is highly significant, statistically and clinically, it is clear that the use of maraviroc can lead to improved outcomes on the population level. Moreover, the positive effect of maraviroc appears quite consistent across subpopulations (Fatkenheuer *et al.*, 2008). However, these findings do not imply that every patient would fare better with maraviroc and OBT than with OBT alone. Assuming that the observed success rates are the true rates, Fig. 1 shows possible values of the four cell probabilities in Table 1 as functions of an odds ratio. In particular, π_{10}, the proportion of patients who would fare better with OBT alone than with maraviroc plus OBT, varies over a wide range (0–22.5%). These are the patients who would be harmed if maraviroc were to be applied to the entire population in addition to OBT. In this situation, it seems natural to characterize HTE in terms of the cell probabilities in Table 1.

The objective of this paper is to develop methods for estimating HTE on the basis of potential outcomes, either in the entire population or in a subpopulation defined by known effect modifiers. This objective is directly relevant to treatment evaluation in regulatory settings and is potentially helpful in discovering new effect modifiers. An analysis of HTE based on potential outcomes may be complementary to a standard subgroup analysis based on known effect modifiers. The latter approach can be useful in medical practice by informing treatment selection. However, with limited knowledge of effect modifiers, it is usually impossible to identify subpopulations that are completely homogeneous with respect to the potential outcomes. Thus, for treatment evaluation, the need remains to understand the residual HTE in a subpopulation defined with the best available knowledge of effect modifiers. Moreover, a large amount of unexplained HTE, suggested by an analysis based on potential outcomes, may motivate scientists to search for new effect modifiers. Although scientists would naturally like to discover all kinds of biomarkers, a data-based motivation could be an important consideration in allocating limited resources. Further implications of individual level HTE are discussed by Poulson *et al.* (2012).

Previous work in this area includes derivation of bounds (Gadbury and Iyer, 2000; Gadbury *et al.*, 2004) and a sensitivity analysis approach (Gadbury *et al.*, 2001). Gadbury and co-workers recognized that the observed data are insufficient to identify all aspects of HTE, which depend on the joint distribution of potential outcomes under different treatments. Randomization in a clinical trial allows us to identify empirically the marginal distribution of each potential outcome but not their joint distribution, which is also known as the fundamental problem of causal inference (Holland, 1986). Our first step in dealing with this identifiability problem is to adjust for covariates. If the set of covariates is sufficient for explaining the (usually positive) dependence between potential outcomes (for different treatments applied to the same patient), the joint distribution of potential outcomes and hence all measures of HTE will then be identified by assuming conditional independence between potential outcomes given observed covariates. Possible violations of this assumption can be addressed by including a random effect to account for residual dependence or by specifying the conditional dependence structure directly. The latter approach is a considerable generalization of the sensitivity analysis approach that was proposed by Gadbury *et al.* (2001).

In the next section, we set up the notation and give a general rationale for the methods proposed. We then describe some specific methods for estimating HTE in Section 3 and apply them in Section 4 to real data from the HIV trial mentioned earlier. The paper ends with a discussion in Section 5.

The programs that were used to analyse the data can be obtained from http://www.blackwellpublishing.com/rss

2. Notation and rationale

Suppose that a randomized clinical trial is conducted to compare an experimental treatment (e.g. maraviroc) with a control treatment, which may be placebo or a standard treatment, with respect to a clinical outcome of interest. To fix ideas, we focus on a binary outcome (1 for success; 0 for failure) in most of this paper; extension to a continuous outcome is considered in Appendix C. The success criterion for an individual patient often has important implications on the study design. For example, the primary end point in the

MOTIVATE trial implies a longitudinal study that follows patients for at least 48 weeks. For ease of presentation, we shall be concerned with a general binary outcome, which may or may not be time dependent, until it becomes necessary to consider specific features of the study design. For a generic patient in the target population, let $Y(t)$ denote the potential outcome that will realize if the patient receives treatment t (0 for control; 1 for experimental). Note that the $Y(t)$, $t = 0, 1$, cannot both be observed on the same subject except in crossover trials under certain conditions, which we do not consider until Section 5. Let T denote the treatment assigned randomly to a study subject; thus T is a Bernoulli variable independent of all baseline variables. Without considering non-compliance, we assume that T is also the actual treatment given to the subject, and we write $Y = Y(T)$ for the actual outcome.

As indicated earlier, it is important in treatment evaluation to assess HTE in terms of the joint probabilities

$$\pi_{\mathrm{jk}}=P\left\{Y(0)=j, Y(1)=k\right\}, \qquad j, k=0, 1.$$

These probabilities are not empirically identifiable. Owing to randomization, it is straightforward to identify and estimate the marginal probabilities

$$\begin{array}{l}\pi_{1+}=\pi_{10}+\pi_{11}=P\left\{Y(0)=1\right\}=P(Y=1|T=0),\\ \pi_{+1}=\pi_{01}+\pi_{11}=P\left\{Y(1)=1\right\}=P(Y=1|T=0).\end{array}$$

Gadbury *et al.* (2004) showed that π_{10}, the proportion of patients who would be harmed by the new treatment relative to the control, is constrained by the marginal probabilities:

$$\max(0, \pi_{1+} - \pi_{+1}) \le \pi_{10} \le \min(1 - \pi_{+1}, \pi_{1+}). \quad (1)$$

Together with the marginal probabilities (π_{1+} and π_{+1}), the value of π_{10} (or any other cell probability), either assumed or identified under suitable assumptions, is sufficient to determine the other cell probabilities, which are subject to similar bounds. The lower bound in inequality (1) corresponds to maximal positive dependence between $Y(0)$ and $Y(1)$, and the upper bound corresponds to maximal negative dependence. If $\pi_{1+} < \pi_{+1}$, then maximal positive dependence means that $Y(0) \le Y(1)$ with probability 1. Although this may be plausible in some settings (e.g. Huang *et al.* (2012)), the use of OBT for all patients in the MOTIVATE trial makes it quite implausible to assume *a priori* that $Y(0) \le Y(1)$. Another special case, which is interesting though admittedly unrealistic, arises when $Y(0)$ is independent of $Y(1)$, in which case we have $\pi_{10} = \pi_{1+}(1 - \pi_{+1})$.

Of course, the potential outcomes $Y(0)$ and $Y(1)$ generally depend on each other because they arise from the same patient; in that sense they resemble repeated measurements, except that they cannot both be observed. Our first step in accounting for the dependence between $Y(0)$ and $Y(1)$ is to condition on relevant covariates that are associated with both outcomes.

Let **X** denote a vector of such covariates measured at baseline, which may include both prognostic factors (with only ‘main effects’) and effect modifiers (which interact with treatment). The distinction between prognostic factors and effect modifiers depends on the link function and therefore is not always clear cut for a binary outcome. In the MOTIVATE trial, **X** may include baseline measurements of HIV ribonucleic acid (RNA) and CD4 cells as well as genotypic and phenotypic sensitivity. Without considering specific methods yet, we note that **X** can be used to sharpen the bounds on the π_{jk}. Specifically, by applying inequality (1) to each stratum defined by **X**, we obtain

$$\max\left\{0, \pi_{1+|X}(\mathbf{X}) - \pi_{+1|X}(\mathbf{X})\right\} \le \pi_{10|X}(\mathbf{X}) \le \min\left\{1 - \pi_{+1|X}(\mathbf{X}), \pi_{1+|X}(\mathbf{X})\right\}, \quad (2)$$

where $\pi_{jk|\mathrm{X}}(\mathbf{X}) = P\{Y(0) = j, Y(1) = k|\mathbf{X}\}$ $(j, k = 0, 1)$, and $\pi_{1+|\mathrm{X}}(\mathbf{X})$ and $\pi_{+1|\mathrm{X}}(\mathbf{X})$ are defined similarly. The lower bound in inequality (2) corresponds to maximal positive dependence of $Y(0)$ and Y(1) given **X**, whereas the upper bound corresponds to maximal negative dependence. Taking expectations over **X** in inequality (2) leads to

$$E[\max\{0, \pi_{1+|X}(\mathbf{X}) - \pi_{+1|X}(\mathbf{X})\}] \le \pi_{10} \le E[\min\{1 - \pi_{+1|X}(\mathbf{X}), \pi_{1+|X}(\mathbf{X})\}]. \quad (3)$$

It is elementary to show that the lower bound in inequality (3) is generally higher than that in inequality (1), and they coincide when $\pi_{1+|\mathrm{X}}.(\mathbf{X}) - \pi_{+1|\mathrm{X}}(\mathbf{X})$ is always positive or always negative (i.e. when there is no qualitative interaction between T and **X**). Similarly, the upper bound in inequality (3) is generally lower than that in inequality (1). Analogous results hold for the other cell probabilities. Thus, conditioning on **X** may be a good starting point in understanding the π_{jk}.

In general, we propose to estimate $\pi_{jk} = E\{\pi_{jk|X}(\mathbf{X})\}$ by averaging over **X** an estimate of $\pi_{jk|X}(\mathbf{X})$. Note that $\pi_{jk|X}(\mathbf{X})$ is generally dependent on **X** because **X** is chosen to be associated with the potential outcomes. This is not a problem with estimating π_{jk} and can actually be helpful in sharpening the bounds, as shown in the preceding paragraph. However, when **X** contains strong effect modifiers, it may be necessary to consider restricted use of the new treatment to a subpopulation of patients with a favourable benefit–risk profile. If such a subpopulation is predefined, that subpopulation can be taken to be the entire population without loss of generality, and the methodology proposed remains applicable. Without a predefined subpopulation, we would have to use the data to identify a favourable subpopulation. Although not designed for that purpose, the methodology proposed can help the search for a favourable subpopulation, as we discuss in Section 4.

Estimates of $\pi_{jk|X}(\mathbf{X})$ may be obtained by using three different approaches, which are outlined below and further developed in Section 3.

2.1. Covariate adjustment based on conditional independence

If **X** is sufficient for explaining the dependence between $Y(0)$ and $Y(1)$, then we can expect that

$$Y(0) \perp Y(1) \big| \mathbf{X}, \quad (4)$$

i.e. that *Y*(0) and *Y*(1) are conditionally independent given **X**. This assumption is similar in spirit to the missingness at random assumption for missing data (Rubin, 1976) and the assumption of strongly ignorable treatment assignment in causal inference (Rosenbaum and Rubin, 1983); they all attempt to reduce a stochastic dependence of concern by conditioning on relevant covariates. Like the latter two assumptions, assumption (4) cannot be verified with the observed data and must be based on external information such as expert opinions. Methods based on assumption (4) are described in Section 3.1.

2.2. Sensitivity analysis based on a random-effect model

We might question the validity of assumption (4) because **X** may not explain all the dependence between *Y*(0) and *Y*(1). We therefore relax assumption (4) in Section 3.2 by including a latent variable to account for any residual dependence between *Y*(0) and *Y*(1). The relaxed assumption can be written as

$$Y(0) \perp Y(1) \big| (\mathbf{X}, U), \quad (5)$$

where *U* is a subject-specific random effect that is independent of **X**. In the MOTIVATE trial, *U* can be a suitable combination of all relevant characteristics of a patient that are predictive of the outcome and that are unmeasured in the study or yet unknown to the scientific community. In other words, *U* represents what is missing from **X** that makes assumption (4) break down. The assumption that *U* is independent of **X** is not as stringent as it may seem, because a candidate *U* could be orthogonalized with respect to **X**. Under assumption (5), assumption (4) corresponds to the special case that *U* is a constant. In general, the distribution of *U* is not identifiable if we observe only a random sample of (*T*, **X**, *Y*). We therefore propose in Section 3.2 a sensitivity analysis approach based on a range of assumptions about the variability of *U*. We also show how to gauge the variability of *U* from longitudinal data and thus narrow the range of the sensitivity analysis.

2.3. Sensitivity analysis based on an odds ratio

The inclusion of a random effect imposes a particular positive dependence structure for *Y*(0) and *Y*(1) given **X**. Although this structure may be plausible in some situations, it is not guaranteed to hold. To accommodate negative dependence as well as different forms of positive dependence between *Y*(0) and *Y*(1) given **X**, we also propose another sensitivity analysis approach based on the odds ratio

$$\rho(\mathbf{X}) = \frac{\pi_{11|X}(\mathbf{X})\pi_{00|X}(\mathbf{X})}{\pi_{01|X}(\mathbf{X})\pi_{10|X}(\mathbf{X})}. \quad (6)$$

Note that $\rho(\mathbf{X})$ captures the entire dependence structure for $Y(0)$ and $Y(1)$ given $\mathbf{X}$ and is not constrained by the conditional probabilities $\pi_{1+|X}(\mathbf{X})$ and $\pi_{+1|X}(\mathbf{X})$. This approach is described in Section 3.3.

3. Methodology

This section presents methods for estimating the joint probabilities π_{jk} from a random sample of subjects, with individual subjects denoted by the subscript i=1,…,n (attached to random variables). Our methodological discussion will be focused on point estimation and sensitivity analysis. For the methods proposed, asymptotic normality is usually immediate from standard M-estimation theory (e.g. van der Vaart (1998) and Stefanski and Boos (2002)), and asymptotic variance formulae are straightforward to derive though cumbersome to present. For ease of implementation, we recommend non-parametric bootstrap standard errors and confidence intervals for inference.

3.1. Covariate adjustment based on conditional independence

Under the conditional independence assumption (4), the cell probabilities can be identified as

$$\begin{aligned}\pi_{\mathrm{jk}} &= E[P\{Y(0)=j, Y(1)=k|\mathbf{X}\}]\\ &= E[P\{Y(0)=j|\mathbf{X}\}\ P\{Y(1)=k|\mathbf{X}\}]\\ &= E\{P(Y=j|T=0,\mathbf{X})\ P(Y=k|T=1,\mathbf{X})\},\end{aligned} \quad (7)$$

where the last step follows from randomization.

Assume first that $\mathbf{X}$ is discrete, taking values in $\{\mathbf{x}_1,\dots, \mathbf{x}_L\}$, say. Write $S_{tl} = \{i: T_i = t, \mathbf{X}_i = \mathbf{x}_l\}$ and $n_{tl} = |S_{tl}|$, where $|\cdot|$ denotes the size of a set. Then the conditional probability $p(y|t, \mathbf{x}) = P(Y = y|T = t, \mathbf{X} = \mathbf{x})$ can be estimated empirically by

$$\hat{p}(y|t,\mathbf{x}_l)=\frac{1}{n_{\mathrm{tl}}}\sum_{i\in S_{\mathrm{tl}}} I(Y_i=y),$$

and the corresponding estimate of π_{jk} is given by

$$\frac{1}{n}\sum_{l=1}^{L} n_{+l}\ \hat{p}(j|0,\mathbf{x}_l)\ \hat{p}(k|1,\mathbf{x}_l),$$

where $n_{+l} = n_{0l} + n_{1l}$.

For a general covariate vector $\mathbf{X}$, one could specify a regression model for $p(y|t, \mathbf{x})$, such as the generalized linear model (GLM)

$$p(1|t,\mathbf{x};\theta)=1-p(0|t,\mathbf{x};\theta)=\psi\{\theta_1+\theta_T t+\theta_X^{'}\mathbf{x}+\boldsymbol{\theta}_{TX}^{'}(t\mathbf{x})\}, \quad (8)$$

where ψ is an inverse link function. The exact form of model (8) is not important; for example, the interaction term can be omitted if **X** does not include effect modifiers. The regression parameter θ can be estimated by maximizing the likelihood $\Pi_{i=1}^{n} p(Y_i|T_i,\mathbf{X}_i;\theta)$ and the resulting estimate will be denoted by $\hat{\theta}$. Now equation (7) can be used to estimate π_{jk} by

$$\frac{1}{n}\sum_{i=1}^{n} p(j|0,\mathbf{X}_i;\hat{\theta})\, p(k|1,\mathbf{X}_i;\hat{\theta}). \quad (9)$$

3.2. Sensitivity analysis based on a random-effect model

Under assumption (5), it can be shown as in equation (7) that

$$\begin{aligned}\pi_{jk} &= E\{P(Y=j|T=0,\mathbf{X},U)\,P(Y=k|T=1,\mathbf{X},U)\}\\ &= \int\int p^*(j|0,\mathbf{x},u)p^*(k|1,\mathbf{x},u)\,\mathrm{d}F_U(u)\,\mathrm{d}F_X(\mathbf{x}),\end{aligned} \quad (10)$$

where $p^*(y|t, \mathbf{x}, u) = P(Y = y|T = t, \mathbf{X} = \mathbf{x}, U = u)$. Here and in what follows, F denotes a (conditional) distribution function with the subscript indicating the random variable(s) concerned. To fix ideas, consider the generalized linear mixed model (GLMM)

$$p^*(1|t,\mathbf{x},u;\boldsymbol{\theta}^*)=1-p^*(0|t,\mathbf{x},u;\boldsymbol{\theta}^*)=\psi\{\theta_1^*+\theta_T^* t+\boldsymbol{\theta}_{\mathbf{x}}^{*'}\mathbf{x}+\boldsymbol{\theta}_{TX}^{*'}(t\mathbf{x})+u\}, \quad (11)$$

with $U\sim N(0,\sigma_U^2)$. Then equation (10) suggests that π_{jk} may be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} p^*(j|0,\mathbf{X}_i,u;\boldsymbol{\theta}^*)\, p^*(k|1,\mathbf{X}_i,u;\boldsymbol{\theta}^*)\,\phi(u;0,\sigma_U^2)\,\mathrm{d}u, \quad (12)$$

where $\varphi(\cdot;\mu,\sigma^2)$ is the density function of $N(\mu,\sigma^2)$, with $(\boldsymbol{\theta}^*,\sigma_U^2)$ replaced by estimates or plausible values (as in a sensitivity analysis). The question is how to obtain such values. In general, the parameters $(\boldsymbol{\theta}^*,\sigma_U^2)$ are not completely identifiable from the $(T_i, \mathbf{X}_i, Y_i)$, i= 1,…, n; this is essentially fitting a GLMM to cross-sectional data. However, some relevant information may be available from repeated measurements, which are typically available at baseline and follow-up visits. In the MOTIVATE trial, for example, repeated measurements of virologic response are available at 11 time points from baseline to week 48 of treatment. In the rest of this subsection, we consider estimation of $\boldsymbol{\theta}^*$ and π_{jk} with σ_U^2 fixed, discuss how the resulting inference depends on σ_U^2 and then suggest ways to extract information about σ_U^2 from longitudinal data.

For a given value of σ_U^2, $\boldsymbol{\theta}^*$ can be estimated by maximizing the likelihood

$$\boldsymbol{\theta}^* \mapsto \prod_{i=1}^{n} \int_{-\infty}^{\infty} p^*(Y_i|T_i, \mathbf{X}_i, u; \boldsymbol{\theta}^*)\, \phi(u;0,\sigma_U^2)\mathrm{d}u. \quad (13)$$

Although this may appear complicated, it can be simplified for common link functions. In Appendix A, we show that, under the GLMM (11), the GLM (8) is correct with $\boldsymbol{\theta}=(1+\sigma_U^2)^{-1/2}\boldsymbol{\theta}*$ for the probit link, and approximately correct with $\boldsymbol{\theta}=(1+\sigma_U^2/c^2)^{-1/2}\boldsymbol{\theta}*$, where c ≈ 1.70, for the logit link. This suggests that we can fit model (8) by using standard GLM software and then recover $\boldsymbol{\theta}$* as $(1+\sigma_U^2)^{1/2}\boldsymbol{\theta}$ for the probit link or as $(1+\sigma_U^2/c^2)^{1/2}\boldsymbol{\theta}$ (approximately) for the logit link. The resulting estimate of $\boldsymbol{\theta}$* can then be substituted into expression (12) together with the given value of σ_U^2.

Given a GLM estimate of $\boldsymbol{\theta}$, $\boldsymbol{\theta}$* in expression (12) varies as a function of $\sigma_U^2 \in (0,\infty)$. When σ_U^2 approaches 0, $\boldsymbol{\theta}$* approaches $\boldsymbol{\theta}$ and expression (12) approaches the estimate (9) based on conditional independence. When $\sigma_U^2 \to \infty$, expression (12) approaches a limit that corresponds to maximal positive dependence between *Y*(0) and *Y*(1) given **X**, as shown in Appendix B. For the case of π_{10}, Appendix B also shows that the limit of expression (12) when *n* and σ_U^2 both approach ∞ is just the lower limit in inequality (3), which is analogous to but generally higher than the lower limit in inequality (1). Thus, a sensitivity analysis based on $\sigma_U^2 \in (0,\infty)$ can account for any and all positive dependence between *Y*(0) and *Y*(1) given **X**.

Such a sensitivity analysis can be sharpened by using reliable information about σ_U^2, which may be available from longitudinal data. Suppose that we have longitudinal measurements of the same outcome following an expanded GLMM:

$$P(Y_{im}=1|T_i,\mathbf{X}_i,U_{im})=\psi\{\theta^*_{1m}+\theta^*_{Tm}\,T_i+\boldsymbol{\theta}^{*'}_{Xm}\,\mathbf{X}_i+\boldsymbol{\theta}^{*'}_{TXm}\,(T_i\mathbf{X}_i)+U_{im}\}, \quad (14)$$

where the subscript *m* denotes the *m*th measurement. We let *m* = 1,…, *M* with *M* corresponding to the outcome of primary interest. The original Y_i, U_i and $\boldsymbol{\theta}$* are now known as Y_{im}, U_{im} and $\boldsymbol{\theta}^*_M$ respectively. Although $\boldsymbol{\theta}^*_m=(\theta^*_{1m},\theta^*_{Tm},\boldsymbol{\theta}^{*'}_{Xm},\boldsymbol{\theta}^{*'}_{TXm})'$ is allowed to depend on *m* in an arbitrary fashion, some components may be assumed constant in *m* if this is scientifically plausible. Suppose that $U_{im} = V_{\mathrm{i}} + W_{im}$, where V_i is characteristic of a subject and W_{im} represents random fluctuation within a subject. The essence of this assumption is that, within a subject, the contemporaneous correlation (between potential outcomes at the same time point) is stronger than the non-contemporaneous correlation (between two outcomes at different time points). Suppose that the V_i and the W_{im} are independent of each other and of the (T_i, $\mathbf{X}_i$). If $V_i \sim N(0,\sigma_V^2)$ and $W_{im} \sim N(0,\sigma_W^2)$ (*i* = 1,…,*n; m* = 1,…, *M*), then $\sigma_U^2=\sigma_V^2+\sigma_W^2$. This requires var($U_{im}$) to be constant over time, at least within a suitable time window containing the primary end point (to be discussed later). Using the arguments of Appendix A, we can integrate the W_{im}-component of U_i out of GLMM (14) and obtain

$$P(Y_{im}=1|T_i, \mathbf{X}_i, V_i)=\psi\{\theta^{**}_{1m}+\theta^{**}_{Tm}\,T_i+\boldsymbol{\theta}^{**\prime}_{Xm}\,\mathbf{X}_i+\boldsymbol{\theta}^{**\prime}_{TXm}\,(T_i\mathbf{X}_i)+rV_i\}, \quad (15)$$

exactly for the probit link and approximately for the logit link. Here,

$$\boldsymbol{\theta}^{**}_{m}=(\theta^{**}_{1m},\theta^{**}_{Tm},\boldsymbol{\theta}^{**\prime}_{Xm},\boldsymbol{\theta}^{**\prime}_{TXm})^{\prime}=r\boldsymbol{\theta}^{*}_{m},$$

$r=(1+\sigma^2_W)^{-1/2}$ for the probit link and $r=(1+\sigma^2_W/c^2)^{-1/2}$ for the logit link. Note that r ≤ 1 with equality holding if and only if $\sigma^2_W=0$. Because model (15) is a standard GLMM, $\mathrm{var}(\mathrm{rV}_i)=r^2\sigma^2_V$ is directly estimable by using standard GLMM software. If we assume that $\sigma^2_W=0$, so that $W_{im}\equiv 0$, then $\sigma^2_U=\sigma^2_V=\mathrm{var}(rV_i)$ becomes identifiable and estimable. Otherwise, noting that $\mathrm{var}(rV_i)\leq\sigma^2_V\leq\sigma^2_U$, var($rV_i$) may serve as a lower bound for σ^2_U in a sensitivity analysis. To implement this analysis based on model (15), one should choose a set of repeated measurements containing the outcome of primary interest as well as some adjacent measurements within a suitable time window. The choice of the time window represents a bias–variance trade-off, with wider windows affording better precision at the expense of potential bias due to model misspecification.

In addition to a lower bound for σ^2_U, one could recover $\boldsymbol{\theta}^*$ from model (15) as $\boldsymbol{\theta}^{*}_{M}=(1+\sigma^2_W)^{1/2}\boldsymbol{\theta}^{**}_{M}$ for the probit link or as $\boldsymbol{\theta}^{*}_{M}=(1+\sigma^2_W/c^2)^{1/2}\boldsymbol{\theta}^{**}_{M}$ for the logit link, if one is confident about model (15). The resulting estimate of $\boldsymbol{\theta}^*$ (based on a given value of σ^2_W) may be more efficient than the estimate that maximizes likelihood (13) (based on a given value of σ^2_U), especially when the $\boldsymbol{\theta}^{*}_{m}$ (m = 1,…, M) are closely related to each other. However, this approach relies heavily on correct specification of model (15) and is thus more susceptible to misspecification bias.

3.3. Sensitivity analysis based on an odds ratio

To accommodate alternative dependence structures (e.g. negative dependence) for $Y(0)$ and $Y(1)$ given $\mathbf{X}$, we now propose a sensitivity analysis approach based on the odds ratio given by equation (6). Note that assumption (4) corresponds to $\rho(\mathbf{X})\equiv 1$. In general, we could specify a model for the odds ratio such as $\rho(\mathbf{X};\beta)=\exp(\beta_1+\beta^{\prime}_X\mathbf{X})$, where $\beta=(\beta_1,\beta^{\prime}_X)^{\prime}$. Specification of the model and plausible parameter values must be based on substantive knowledge, because the observed data provide no information about $\rho(\mathbf{X})$. When the dimension of β is high, it can be difficult to cover a wide range of β-values in obtaining and presenting estimates of the π_{jk} as functions of β. Therefore, without reliable and concrete information about $\rho(\mathbf{X})$, it does not seem advantageous to perform a sensitivity analysis based on a covariate-dependent odds ratio. In the rest of this subsection, we shall work with a fixed odds ratio, $\rho(\mathbf{X})\equiv\rho\in(0,\infty)$, for ease of interpretation, implementation and presentation, although the methodology extends easily to a more general model for $\rho(\mathbf{X})$.

In a 2×2 contingency table with cell probabilities $\mathbf{q} = (q_{00}, q_{01}, q_{10}, q_{11})'$, q_{11} can be determined from the odds ratio ρ and the marginal probabilities $q_{1+} = q_{10} + q_{11}$ and $q_{+1} = q_{01} + q_{11}$ as

$$q_{11}=\frac{1+(q_{1+}+q_{+1})(\rho-1)-\omega}{2(\rho-1)},$$

where

$$\omega=\surd[\{1+(q_{1+}+q_{+1})(\rho-1)\}^2+4\rho(1-\rho)q_{1+}q_{+1}].$$

The rest of $\mathbf{q}$ can then be obtained as $q_{10} = q_{1+} - q_{11}$, $q_{01} = q_{+1} - q_{11}$ and $q_{00} = 1 - q_{11} - q_{10} - q_{01}$. Let $\mathbf{Q}$ denote the map $(q_{1+}, q_{+1}, \rho) \mapsto \mathbf{q}$ for recovering cell probabilities. Then we can write

$$\begin{array}{rcl} \pi & = & (\pi_{00},\pi_{01},\pi_{10},\pi_{11})' \\ & = & E\{(\pi_{00|X}(\mathbf{X}),\pi_{01|X}(\mathbf{X}),\pi_{10|X}(\mathbf{X}),\pi_{11|X}(\mathbf{X}))'\} \\ & = & E[\mathbf{Q}\{\pi_{1+|X}(\mathbf{X}),\pi_{+1|X}(\mathbf{X}),\rho\}]. \end{array}$$

Now let $p(\hat{y}|t, \mathbf{X})$ be an estimate of $p(y|t, \mathbf{x}) = P(Y = y|T = t, \mathbf{X} = \mathbf{x})$ from Section 3.1, which may be obtained non-parametrically through stratification or parametrically under a regression model. Because of randomization, $p(\hat{1}|0, \mathbf{x})$ estimates $\pi_{1+|X}(\mathbf{x})$ and $p(\hat{1}|t, \mathbf{x})$ estimates $\pi_{+1|X}(\mathbf{x})$. It follows that π, the vector of joint probabilities, is estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{Q}\{\hat{p}(1\,|0,\mathbf{X}_i),\hat{p}(1|\,1,\mathbf{X}_i),\rho\}.$$

It is easy to see that, when n and ρ both approach ∞, the estimate of π_{10} given above converges to the lower bound in inequality (3) corresponding to maximal positive dependence (conditional on $\mathbf{X}$). Analogous results hold for $\rho \to -\infty$ and for the other cell probabilities. Thus, stringent as it may seem, the assumption of a constant odds ratio does not exclude any possible value of the π_{jk}.

3.4. Summarizing remarks

We have presented three methods that are roughly increasing in generality, except for some modelling assumptions (GLMMs in Section 3.2; a constant odds ratio in Section 3.3). The method of Section 3.1 assumes conditional independence of $Y(0)$ and $Y(1)$ given $\mathbf{X}$, the method of Section 3.2 assumes positive conditional dependence (of a particular structure), and the method of Section 3.3 accommodates both positive and negative conditional dependence. The first method is simpler though perhaps less credible than the other two.

The second method is probably the most demanding in terms of data, modelling assumptions and possibly computation, but it also can be the most informative, because it provides a data-driven lower bound on the extent of positive dependence. The last method can be restricted to positive dependence by forcing $\rho > 1$, but further restriction would require substantive knowledge.

4. Application

We now apply the methods of Section 3 to the MOTIVATE trial that was introduced in Section 1, a randomized, double-blinded, placebo-controlled, confirmatory clinical trial comparing maraviroc plus OBT with placebo plus OBT for treating HIV-1. The MOTIVATE study consists of two substudies that were identically designed and conducted (albeit in different countries), produced similar results and are therefore combined in our analysis. The study enrolled a total of 1049 patients with R5 HIV-1 who had been treated with or had resistance to three antiretroviral drug classes and had HIV-1 RNA levels of more than 5000 copies ml^{-1}. The patients were randomized in a 2:2:1 ratio to receive one of three antiretroviral regimens (maraviroc once daily, maraviroc twice daily and placebo), each of which also included OBT based on treatment history and drug resistance testing. Our analysis is focused on comparing the maraviroc twice daily group with the placebo group, as the two maraviroc groups produced similar results.

The outcome of interest to us is virologic response (defined as HIV RNA level below 400 copies ml^{-1}) at week 48 of treatment. As mentioned earlier, the observed virologic response rates are 57.5% and 22.5% in the maraviroc twice daily and placebo groups respectively, and the difference between the two groups (35.0%; 95% confidence interval 27.7−42.4%) is highly significant ($p < 0.0001$). Fig. 1 shows that the cell probabilities π_{jk} (j, k = 0, 1) vary widely as functions of the marginal log-odds ratio, $\log\{\pi_{00}\pi_{11}/(\pi_{01}\pi_{10})\}$, which is not identifiable from the data. Without prior information about the marginal log-odds ratio, a sensitivity analysis based on Fig. 1 would not be very informative. Using the methods proposed, we now show that the ranges of the π_{jk} can be narrowed by borrowing information from relevant covariates and repeated measurements.

The covariates that were included in our analysis are age, RNA (the logarithm of the baseline level of HIV RNA), CD4 (the logarithm of the baseline count of CD4 cells), and GSS and PSS (genotypic and phenotypic sensitivity scores, defined as the number of antiretroviral drugs used concomitantly to which a patient's HIV was fully susceptible, as determined by genotypic and phenotypic resistance testing at baseline). GSS and PSS take integer values from 0 to 3. The regression model for covariate adjustment is a logistic regression model given by equation (8), where **X** consists of the five covariates just defined. An estimate of (π_{00}, π_{01}, π_{10}, π_{11}) is obtained as (0.359, 0.432, 0.062, 0.147) by using the method of Section 3.1, with respective standard errors (0.0092, 0.0061, 0.0021, 0.0069). Throughout this section, we base inference on 400 non-parametric bootstrap samples obtained by sampling with replacement from the original subjects, without changing the observed data within a subject. Standard errors are obtained as empirical standard deviations, and confidence intervals as 2.5th and 97.5th empirical percentiles, across the 400 bootstrap samples.

The data are also analysed by using the method of Section 3.2 under the random-effect model given by equation (11), with a logit link and the same covariate vector **X** as in the previous analysis. For a given value of σ_U^2, we estimate $\boldsymbol{\theta}^*$ as $(1+\sigma_U^2/c^2)^{1/2}\boldsymbol{\theta}$ and the π_{jk} by using expression (12). The results (point estimates and confidence intervals of cell probabilities) are plotted in Fig. 2 as functions of σ_U^2. As $\sigma_U^2 \to 0$, the results converge to those reported in the preceding paragraph (based on conditional independence), as expected. With increasing σ_U^2, the estimated probabilities tend to increase for concordant pairs (π_{00}, π_{11}) and to decrease for discordant pairs (π_{01}, π_{10}). To gauge the magnitude of σ_U^2, a GLMM analysis based on model (15) is performed on repeated measurements at 24, 32, 40 and 48 weeks. Although earlier measurements (from baseline to 20 weeks) are also available, we restrict our analysis to the later measurements in an attempt to reduce misspecification bias. Indeed, Fig. 2A of Gulick *et al.* (2008) shows that virologic response rates first rise in a non-linear fashion and then decline in a slow, steady and apparently linear fashion. An analysis of all repeated measurements would require a complex model for the mean structure and possibly additional random effects for the correlation structure. The resulting inference would be more susceptible to bias and of questionable relevance to our goal of understanding σ_U^2. Our restricted analysis of later measurements is carried out with θ_{1m}^{**} depending on *m* freely and the rest of $\boldsymbol{\theta}_m^{**}$ independent of *m*; this amounts to adding an indicator for the follow-up visit as a categorical covariate. From this analysis, var(rV_i) is estimated to be 20.2 (95% confidence interval 12.2−30.2). Because $\sigma_U^2 \geq \mathrm{var}(rV_i)$, a conservative 97.5% lower confidence bound for σ_U^2 is given by 12.2, the lower confidence bound for var(rV_i). This suggests that interpretation of Fig. 2 should be focused on later portions of the curves, which happen to be flatter than the earlier portions and have narrower ranges. In particular, the range of π_{10}, the proportion of patients who would be harmed by maraviroc, is much narrower in Fig. 2 than in Fig. 1. The argument of Berger and Boos (1994) can be used to conclude that, with σ_U^2 treated as an unknown nuisance parameter, a 95% upper confidence bound for π_{10} is below 4%; this is obtained by evaluating the upper confidence bound curve for π_{10} at the lower confidence bound for σ_U^2. In contrast, a 97.5% lower confidence bound for π_{01}, the proportion of patients who would benefit from maraviroc, is clearly above 25%; the Berger−Boos argument is not needed here because we are taking the infimum of the lower confidence bound curve.

Fig. 3 shows another sensitivity analysis based on $\rho(\mathbf{X}) \equiv \rho$, as described in Section 3.3. It includes negative conditional dependence (i.e. $\rho < 1$) for completeness, although there is no reason to expect a negative dependence between *Y*(0) and *Y*(1) given **X** in this situation. Even without prior information about ρ, the curves in Fig. 3 are notably more constrained than those in Fig. 1. The extra information obviously comes from the covariates as well as the regression model (8). The results at $\rho = 1$ correspond to those based on conditional independence as well as those in Fig. 2 with $\sigma_U^2 \to 0$. If we restrict attention to positive conditional dependence (i.e. $\rho > 1$), then this analysis yields the same range of point estimates for each joint probability π_{jk} as in Fig. 2.

The above results should be weighed according to the plausibility of the underlying assumptions for the different methods. Because the GLMM analysis indicates a high level of within-subject correlation (after adjusting for **X**), the conditional independence assumption (4) is seriously in doubt, and the resulting analysis is not as credible as the two sensitivity analyses. Although based on different assumptions, both sensitivity analyses can cover the full range between conditional independence and maximal positive conditional dependence of $Y(0)$ and $Y(1)$ given **X**. In addition, the random-effect-based approach, under the assumptions of Section 3.2, also provides a lower bound on the extent of positive conditional dependence, which can be used to narrow the spectrum of the sensitivity analysis. The key assumption for the lower bound, that the contemporaneous correlation is stronger than the non-contemporaneous correlation, seems quite plausible in the present situation, and we therefore place more emphasis on the random-effect-based analysis than on the odds-ratio-based analysis.

The random-effect-based analysis suggests that it is fairly unlikely for maraviroc to affect adversely the virologic response (at week 48) of a patient who already receives OBT. This finding adds considerable assurance to the current knowledge of maraviroc (Gulick *et al.*, 2008; Fatkenheuer *et al.*, 2008). If this were not so (i.e. if the analysis showed that a large proportion of patients could be harmed by the addition of maraviroc to OBT), it might be necessary to consider restricting the use of maraviroc to a subpopulation of patients with a more favourable benefit–risk profile. The search for such a subpopulation could be facilitated by intermediate results in the preceding analyses. For example, potential effect modifiers could be identified by examining the regression coefficients in models (8) and (15), and the various estimates of $\pi_{jk|X}(\mathbf{X})$ could be used to develop a candidate subpopulation. To contain the adverse effect of maraviroc, it makes sense to look for patients with small estimates of $\pi_{10|X}(\mathbf{X})$. Once a candidate subpopulation has been identified, the methods proposed can then be applied to the target subpopulation to re-estimate the π_{jk}, presumably by using a cross-validation approach to avoid overfitting. If no subpopulation based on **X** can be found with a satisfactory benefit–risk profile, the scientific community may then be motivated to search for new and more informative biomarkers.

5. Discussion

This paper adds to the literature on HTE assessment, which is currently dominated by subgroup analyses and multiplicity adjustments, with a different approach based on potential outcomes of individual patients. Gadbury and colleagues have previously studied HTE in terms of potential outcomes. This paper builds on their work by developing new and practical methods that incorporate relevant information in covariates and repeated measurements in assessing HTE. To address the inherent identifiability issue, we propose a covariate adjustment method based on conditional independence as well as two sensitivity analysis methods, one of which extends the work of Gadbury *et al.* (2001). The HIV example in Section 4 shows that relevant covariates and repeated measurements can indeed help to reduce the uncertainty about HTE. Originally developed for a binary outcome, the methods are extended to a continuous outcome in Appendix C.

In practice, the choice between the three methods proposed should be based on the available information and the plausibility of the underlying assumptions. The first method (Section 3.1), which is conceptually simple and easy to implement, could serve as a starting point. It would not be the final analysis, though, unless one is fairly confident about the conditional independence assumption. In situations like the MOTIVATE study, where longitudinal data are available, the second method (Section 3.2) could be used to assess the effect of positive conditional dependence and also to estimate a lower bound for that dependence, under a suitable GLMM framework. The last method (Section 3.3), which is even more general though perhaps less informative than the second method, could nonetheless be used as a safety net when reliable information is unavailable about the dependence structure of $Y(0)$ and $Y(1)$ given $\mathbf{X}$.

As alluded to earlier, a promising approach to understanding HTE may be the crossover design, where potential outcomes for different treatments are actually observed on the same subject, albeit in different time periods. However, the crossover design has its own issues, as noted by Poulson *et al.* (2012). Depending on the disease being studied and the treatments being compared, a crossover design may be infeasible or seriously compromised by a substantial carry-over effect. Even without these problems, evaluation of HTE in a crossover study could be complicated by an individual period effect. Poulson *et al.* (2012) show in a two-period two-treatment setting that the sample variance of the observed individual treatment difference overestimates the true variance of the individual treatment effect. Nonetheless, the crossover design can still be a valuable tool for studying HTE. It will be of interest to see whether the methods that are developed here can be extended to the crossover design and combined with the results of Poulson *et al.* (2012) to yield further insights into HTE.

Although developed for randomized clinical trials, our methods can certainly be extended to observational studies. The main issue in such extensions is to ensure adequate control for confounding. For this, all potential confounders need to be measured and included in $\mathbf{X}$ to meet the condition of strongly ignorable treatment assignment (Rosenbaum and Rubin, 1983).

Acknowledgments

We thank the Joint Editor and two reviewers for their insightful and constructive comments, which have greatly improved the manuscript. The views expressed in this paper do not represent the official position of the US Food and Drug Administration.

Appendix A: Relationship between θ and (θ, σU2) in Section 3.2

For the probit link, model (11) implies that model (8) holds with $\boldsymbol{\theta}=(1+\sigma_U^2)^{-1/2}\boldsymbol{\theta}^*$. This observation, which was noted previously by Carroll *et al.* (1984), can be argued as follows. With $\psi = \Phi$, the standard normal distribution function, we have

$$\begin{aligned} p(1|t,\mathbf{x}) &= \int p(1|t,\mathbf{x},u)\mathrm{d}F_{U|T,X}(u|t,\mathbf{x}) \\ &= \int \Phi\{(1,t,\mathbf{x}',t\mathbf{x}')\boldsymbol{\theta}^{*}+u\}\phi(u;0,\sigma_U^2)\mathrm{d}u \\ &= \int P\{Z<(1,t,\mathbf{x}',t\mathbf{x}')\boldsymbol{\theta}^{*}+u\}\phi(u;0,\sigma_U^2)\mathrm{d}u, \end{aligned}$$

where Z is a standard normal variable independent of *U*. The second step in this expression makes use of the assumed independence between *U* and (*T*, **X**), and the last step follows from the definition of Φ. Now we can write *p*(1|*t*, **x**) as the marginal probability

$$\begin{aligned} P\{Z<(1,t,\mathbf{x}',t\mathbf{x}')\boldsymbol{\theta}^{*}+U\} &= P\{Z-U<(1,t,\mathbf{x}',t\mathbf{x}')\boldsymbol{\theta}^{*}\} \\ &= P\left\{\frac{Z-U}{\surd(1+\sigma_U^2)}<\frac{(1,t,\mathbf{x}',t\mathbf{x}')\boldsymbol{\theta}^{*}}{\surd(1+\sigma_U^2)}\right\} \\ &= \Phi\left\{\frac{(1,t,\mathbf{x}',t\mathbf{x}')\boldsymbol{\theta}^{*}}{\surd(1+\sigma_U^2)}\right\}, \end{aligned}$$

which leads to the result claimed.

For the logit link, we can approximate ψ(*a*) = exp(*a*)/{1 + exp(*a*)} by Φ(*a*/*c*), where *c*= 15*π*/(16√3)≈ 1.70 (e.g. Johnson and Kotz (1970), Zeger *et al.* (1988) and Liang and Liu (1991)). This allows us to write

$$\begin{aligned} p(1|t,\mathbf{x}) &= \int \psi\{(1,t,\mathbf{x}',t\mathbf{x}')\boldsymbol{\theta}^{*}+u\}\phi(u;0,\sigma_U^2)\mathrm{d}u \\ &\approx \int \Phi\left\{\frac{(1,t,\mathbf{x}',t\mathbf{x}')\boldsymbol{\theta}^{*}+u}{c}\right\}\phi(u;0,\sigma_U^2)\mathrm{d}u \\ &= \Phi\left\{\frac{(1,t,\mathbf{x}',t\mathbf{x}')\boldsymbol{\theta}^{*}}{c\surd(1+\sigma_U^2/c^2)}\right\} \\ &\approx \psi\left\{\frac{(1,t,\mathbf{x}',t\mathbf{x}')\boldsymbol{\theta}^{*}}{\surd(1+\sigma_U^2/c^2)}\right\}. \end{aligned}$$

In this case, model (11) implies that model (8) holds approximately with $\boldsymbol{\theta}=(1+\sigma_U^2/c^2)^{-1/2}\boldsymbol{\theta}^{*}$.

Appendix B: Limit of expression (12) as σU2→∞

We fix *θ* and consider *θ** as a function of $\sigma_U^2{:}\boldsymbol{\theta}^{*}(\sigma_U^2)=(1+\sigma_U^2)^{1/2}\boldsymbol{\theta}$ for the probit link and $\boldsymbol{\theta}^{*}(\sigma_U^2)=(1+\sigma_U^2/c^2)^{1/2}\boldsymbol{\theta}$ for the logit link. To fix ideas, we focus on estimating π_{10}, i.e. the case that *j* = 1 and *k* = 0 in expression (12); the other cases can be treated with the same argument. Let us define

$$\begin{aligned} h(\mathbf{x},\sigma_U^2) &= P\{Y(0)=1,Y(1)=0|\mathbf{X}=\mathbf{x};\boldsymbol{\theta}^{*}(\sigma_U^2),\sigma_U^2\} \\ &= \int_{-\infty}^{\infty} p^{*}\{1|0,\mathbf{x},u;\boldsymbol{\theta}^{*}(\sigma_U^2)\}p^{*}\{0|1,\mathbf{x},u;\boldsymbol{\theta}^{*}(\sigma_U^2)\}\phi(u;0,\sigma_U^2)\mathrm{d}u \\ &= \int_0^1 \psi\{\boldsymbol{\theta}^{*}(\sigma_U^2)'\mathbf{b}(0,\mathbf{x})+\sigma_U z_q\}[1-\psi\{\boldsymbol{\theta}^{*}(\sigma_U^2)'b(1,x)+\sigma_U z_q\}]\mathrm{d}q, \end{aligned} \quad (16)$$

where **b**(t, **x**) = (1, *t*, **x**′, *t***x**′)′, z_q is the *q*th quantile of the standard normal distribution and a change of variables is used in the last step. Now expression (12) can be rewritten as $n^{-1}\sum_{i=1}^{n} h(\mathbf{X}_i, \sigma_U^2)$.

Let us focus on the probit link for the moment. As $\sigma_U^2 \to \infty$, we have, for *t* = 0, 1,

$$\psi\{\boldsymbol{\theta}^*(\sigma_U^2)'\mathbf{b}(t,\mathbf{x})+\sigma_U z_q\}=\psi\{(1+\sigma_U^2)^{1/2}\boldsymbol{\theta}'\mathbf{b}(t,\mathbf{x})+\sigma_U z_q\} \to \begin{cases} 1 & \text{if } \boldsymbol{\theta}'\mathbf{b}(t,\mathbf{x})+z_q>0, \\ 0 & \text{if } \boldsymbol{\theta}'\mathbf{b}(t,\mathbf{x})+z_q<0. \end{cases}$$

The case that θ' **b**(*t*, **x**) + z_q = 0 is irrelevant for our purpose. The above expression implies that the integrand in equation (16) converges to

$$I\{\boldsymbol{\theta}'\mathbf{b}(0,\mathbf{x})+z_q>0, \boldsymbol{\theta}'\mathbf{b}(1,\mathbf{x})+z_q<0\}=I\{-\boldsymbol{\theta}'\mathbf{b}(0,\mathbf{x})<z_q<-\boldsymbol{\theta}'\mathbf{b}(1,\mathbf{x})\},$$

where *I* is the indicator function, for almost every *q* (write respect to Lebesgue measure). By the dominated convergence theorem,

$$h(\mathbf{x},\sigma_U^2) \to \max[0,\Phi\{\boldsymbol{\theta}'\mathbf{b}(0,\mathbf{x})\}-\Phi\{\boldsymbol{\theta}'\mathbf{b}(1,\mathbf{x})]=\max\{0,\pi_{1+|X}(\mathbf{x})-\pi_{+1|X}(\mathbf{x})\}=:h(\mathbf{x},\infty).$$

Note that the above limit is simply the lower limit for $\pi_{10|X(\mathbf{x})}$ given in inequality (2), which is attained under maximal positive dependence between *Y*(0) and *Y*(1) given **X** = **x**. Applying the dominated convergence theorem once again (with respect to **x**), we see that expression (12) converges to $n^{-1}\sum_{i=1}^{n} h(\mathbf{X}_i,\infty)$ as $\sigma_U^2 \to \infty$. Furthermore, the population counterpart of expression (12), with the sample average replaced by expectation (with respect to **X**), converges to *E*{*h*(**X**, ∞)}. Note that

$$\begin{aligned} E\{h(\mathbf{X},\infty)\} &= \int \max\{0,\pi_{1+|X}(\mathbf{x})-\pi_{+1|X}(\mathbf{x})\}\mathrm{d}F_X(\mathbf{x}) \\ &\geq \max\left[0,\int\{\pi_{1+|X}(\mathbf{x})-\pi_{+1|X}(\mathbf{x})\}\mathrm{d}F_X(\mathbf{x})\right] \\ &= \max(0,\pi_{1+}-\pi_{+1}). \end{aligned}$$

Thus, the limit *E*{h(**X**, ∞)}, which is also the lower limit in inequality (3), is indeed higher (i.e. sharper) than the lower limit in inequality (1), which does not involve any covariate information.

With minimal modifications, the result in the above paragraph remains valid (approximately) for the logit link.

Appendix C: Extension to continuous outcomes

The joint distribution, $F(y_0, y_1) = P\{\mathrm{Y}(0) \leq y_0, \mathrm{Y}(1) \leq y_1\}$, contains all relevant information about HTE for an arbitrary outcome variable. This joint distribution is itself a succinct summary of HTE for a binary outcome and categorical outcomes with more than two levels, to which the methods of Section 3 extend readily. This appendix is therefore focused on a continuous or quantitative outcome, for which there may be more succinct summaries of HTE than the joint distribution of potential outcomes. For example, the difference $D = Y(1) - Y(0)$ can be used to represent the effect of the experimental treatment relative to the control on an individual patient, and one might be interested in estimating $\sigma_D^2 = \mathrm{var}(D)$, which measures the overall extent of HTE, or more generally the distribution function F_D. From F_D one could further derive important quantities such as selected quantiles of D or the proportion of patients who would benefit from, or be harmed by, the new treatment. These quantities are functionals of the joint distribution F which are not determined by the marginal distributions $F_t(y) = P\{Y(t) \leq y\}$, $t = 0, 1$.

Under assumption (4), $F(y_0, y_1) = E\{F_{Y|T,X}(y_0|0, \mathbf{X})\, F_{Y|T,X}(y_1|1, \mathbf{X})\}$ can be estimated by substituting an estimate of $F_{Y|T,X}$, which may be parametric or non-parametric, and replacing expectation with sample average (over the $\mathbf{X}_i$). From this it is straightforward to derive estimates of HTE-related quantities. For example, suppose that $F_{Y|T,X}$ follows a normal linear model given by

$$Y = \theta_1 + \theta_T T + \boldsymbol{\theta}_X^{\prime}\mathbf{X} + \boldsymbol{\theta}_{TX}^{\prime}(T\mathbf{X}) + \varepsilon, \quad (17)$$

where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, independently of $(T, \mathbf{X})$. Then assumption (4) implies that

$$F_D(d) = E\left\{\Phi\left(\frac{d - \theta_T - \boldsymbol{\theta}_{TX}^{\prime}\mathbf{X}}{\sigma_\varepsilon\sqrt{2}}\right)\right\}.$$

When assumption (4) is in doubt, the random-effect-based approach of Section 3.2 extends easily to any outcome for which a GLMM is appropriate. Suppose, for example, that $F_{Y|T,X,U}$ follows the linear mixed model

$$Y = \theta_1^* + \theta_T^* T + \theta_X^{*\prime}\mathbf{X} + \theta_{TX}^{*\prime}(T\mathbf{X}) + U + \varepsilon*, \quad (18)$$

where $U \sim N(0, \sigma_U^2)$ and $\varepsilon^* \sim N(0, \sigma_{\varepsilon*}^2)$, independently of each other and of $(T, \mathbf{X})$. Then assumption (5) implies that

$$F_D(d) = E\left\{\Phi\left(\frac{d - \theta_T^* - \boldsymbol{\theta}_{TX}^{*\prime}\mathbf{X}}{\sigma_{\varepsilon*}\sqrt{2}}\right)\right\}. \quad (19)$$

The parameters in model (18) may be (partially) estimable from longitudinal data under suitable assumptions (see Section 3.2). In any case, we note that model (18) implies model (17) with $\theta = \theta^*$ and $\sigma_\varepsilon^2 = \sigma_U^2 + \sigma_{\varepsilon *}^2$, so that θ^* is directly estimable from a linear regression analysis based on model (17), and $\sigma_{\varepsilon *}^2$ can be recovered as $\sigma_\varepsilon^2 - \sigma_U^2$ for a specified value of σ_U^2.

The odds-ratio-based approach of Section 3.3 can be extended to general outcomes by using copulas (Nelsen, 1999). A copula is a multivariate distribution function with uniform (0, 1) marginals. It captures the dependence structure in a multivariate distribution without being constrained by the marginal distributions. In the present context, it allows us to represent the joint distribution of $Y(0)$ and $Y(1)$ given $\mathbf{X}$ as

$$P\{Y(0) \le y_0, Y(1) \le y_1 | \mathbf{X} = \mathbf{x}\} = C\{F_{Y|T,X}(y_0|0, \mathbf{x}), F_{Y|T,X}(y_1|1, \mathbf{x})\},$$

where C is the copula. Instead of conditional independence, we now assume that, given $\mathbf{X}$, the conditional dependence of $Y(0)$ and $Y(1)$ is described by the copula $C(\cdot, \cdot; \rho)$, where C is taken from a parametric family of copulas and ρ is a parameter that represents the strength of the dependence. Note that the copula could be allowed to depend on $\mathbf{X}$ if necessary. Now $F(y_0, y_1)$ can be identified and estimated as $E[C\{F_{Y|T, X}(y_0|0, \mathbf{X}), F_{Y|T, X}(y_1|1, \mathbf{X}); \rho\}]$. In particular, if $F_{Y|T, X}$ follows model (17) and $C(\cdot, \cdot; \rho)$ is a normal copula with correlation coefficient ρ, we then have

$$F_D(d) = E\left(\Phi\left[\frac{d - \theta_T - \theta'_{TX}\mathbf{X}}{\sigma_\varepsilon \sqrt{\{2(1-\rho)\}}}\right]\right). \quad (20)$$

A sensitivity analysis based on equation (20) is similar to an analysis based on equation (19) but perhaps more general in that it accommodates negative dependence between $Y(0)$ and $Y(1)$ given $\mathbf{X}$.

References

Berger RL, Boos DD. P values maximised over a confidence set for the nuisance parameter. J Am Statist Ass. 1994; 89:1012–1016.

Carroll RJ, Spiegelman CH, Lan KKG, Kent KT, Abbott RD. On error-in-variables for binary regression models. Biometrika. 1984; 71:19–25.

Fatkenheuer G, Nelson M, Lazzarin A, Konourina I, Hoepelman AI, Lampiris H, Hirschel B, Tebas P, Raffi F, Trottier B, Bellos N, Saag M, Cooper DA, Westby M, Tawadrous M, Sullivan JF, Ridgway C, Dunne MW, Felstead S, Mayer H, van der Ryst E, MOTIVATE Study Teams. Subgroup analyses of maraviroc in previously treated R5 HIV-1 infection. New Engl J Med. 2008; 359:1442–1455. [PubMed: 18832245]

Gadbury GL, Iyer HK. Unit-treatment interaction and its practical consequences. Biometrics. 2000; 56:882–885. [PubMed: 10985231]

Gadbury GL, Iyer HK, Albert JM. Individual treatment effects in randomized trials with binary outcomes. J Statist Planng Inf. 2004; 121:163–174.

Gadbury GL, Iyer HK, Allison DB. Evaluating subject-treatment interaction when comparing two treatments. J Biopharm Statist. 2001; 11:313–333.

Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics. 1985; 41:361–372. [PubMed: 4027319]

Gulick RM, Lalezari J, Goodrich J, Clumeck N, DeJesus E, Horban A, Nadler J, Clotet B, Karlsson A, Wohlfeiler M, Montana JB, McHale M, Sullivan J, Ridgway C, Felstead S, Dunne MW, van der Ryst E, Mayer H, MOTIVATE Study Teams. maraviroc for previously treated patients with R5 HIV-1 infection. New Engl J Med. 2008; 359:1429–1441. [PubMed: 18832244]

Holland PW. Statistics and causal inference (with discussion). J Am Statist Ass. 1986; 81:945–970.

Huang Y, Gilbert PB, Janes H. Assessing treatment-selection markers using a potential outcomes framework. Biometrics. 2012; 68:687–696. [PubMed: 22299708]

Johnson, NL.; Kotz, S. Distributions in Statistics. Vol. 2. Boston: Houghton-Mifflin; 1970.

Liang, KY.; Liu, XH. Estimating equations in generalized linear models with measurement error. In: Godambe, AP., editor. Estimating Functions. Oxford: Clarendon; 1991.

Nelsen, RB. An Introduction to Copulas. New York: Springer; 1999.

Peto, R. Statistical Aspects of Cancer Trials. London: Chapman and Hall; 1982.

Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statist Med. 2002; 21:2917–2930.

Poulson RS, Gadbury GL, Allison DB. Treatment heterogeneity and individual qualitative interaction. Am Statistn. 2012; 66:16–24.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55.

Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974; 66:688–701.

Rubin DB. Inference and missing data. Biometrika. 1976; 63:581–592.

Russek-Cohen E, Simon RM. Evaluating treatments when a gender by treatment interaction may exist. Statist Med. 1997; 16:455–464.

Stefanski LA, Boos DD. The calculus of M-estimation. Am Statistn. 2002; 56:29–38.

van der Vaart, AW. Asymptotic Statistics. Cambridge: Cambridge University Press; 1998.

Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics. 1988; 44:1049–1060. [PubMed: 3233245]

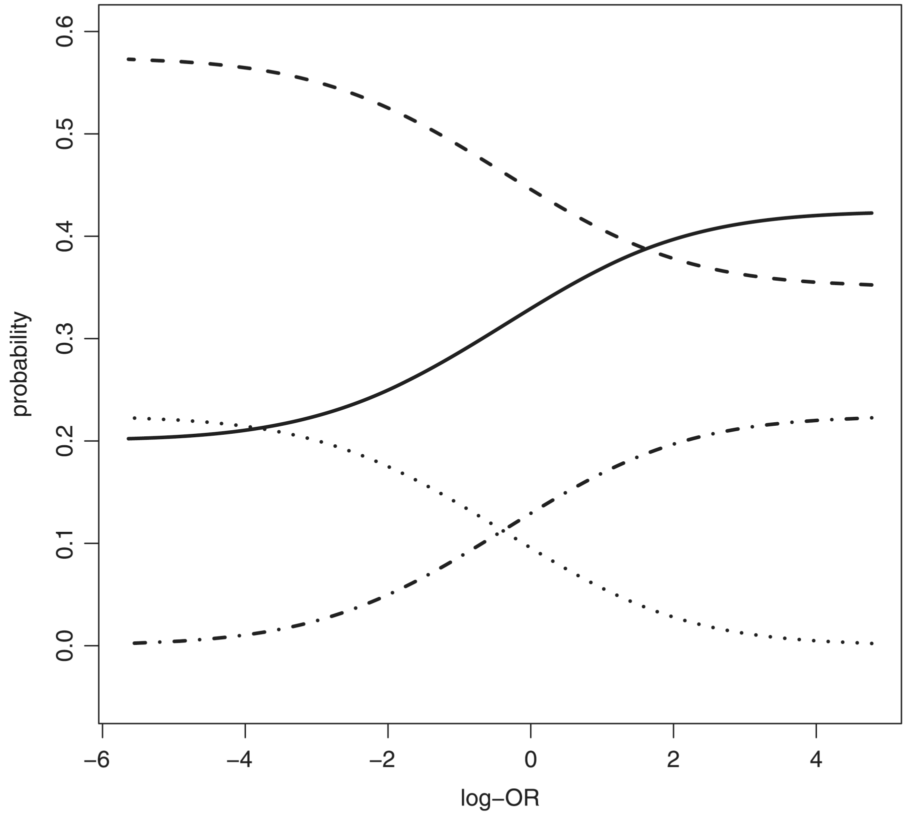

Fig. 1.

Possible values of the cell probabilities in Table 1 as functions of the log-odds ratio, $\log\{\pi_{11}\pi_{00}/(\pi_{01}\pi_{10})\}$ for fixed marginal probabilities $\pi_{1+} = 0.225$ and $\pi_{+1} = 0.575$ as estimated from the MOTIVATE study (see Section 4 for details): ——, π_{00};— — —, π_{01}; · · · · · ·, π_{10}; · — · —, π_{11}

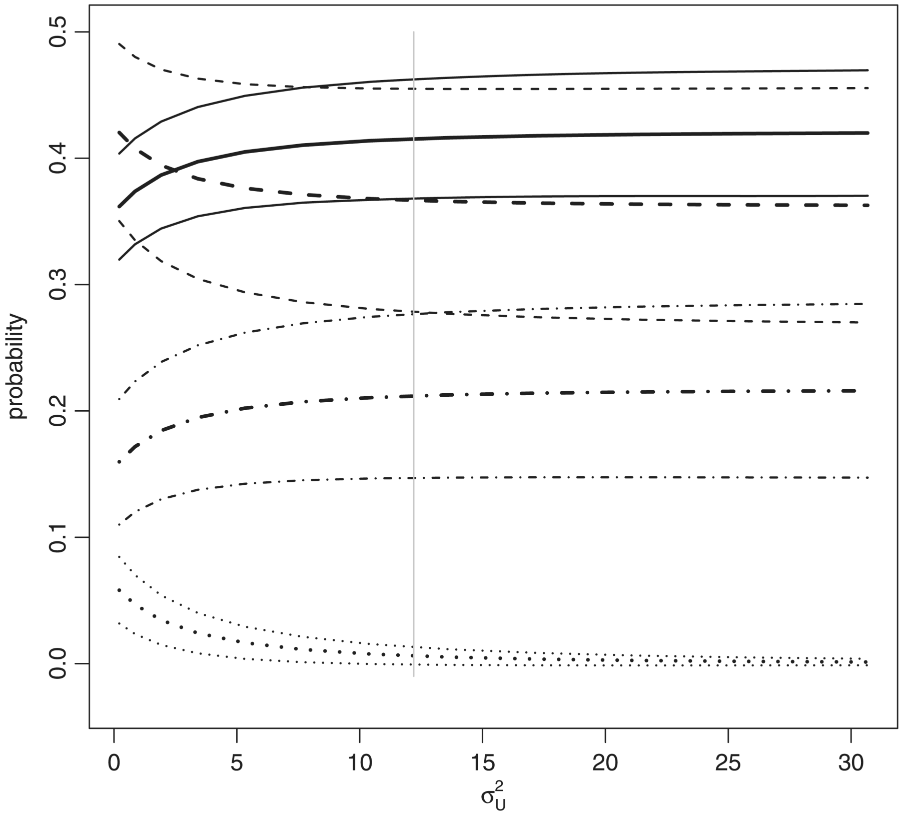

Fig. 2.

Sensitivity analysis for the MOTIVATE study of Section 4 based on a random-effect model (|, conservative 97.5% lower confidence bound for σ_U^2, which is obtained as 12.2 from a GLMM analysis (see Section 4 for details)): point estimates (▬▬, — — —, · — · —, · · · · · · ·) and 95% confidence intervals (——, — — —, · — · —, · · · · · · ·) for π_{00} (▬▬), π_{01}(— — —), π_{10}(· · · · · · ·) and π_{11} (· — · —)

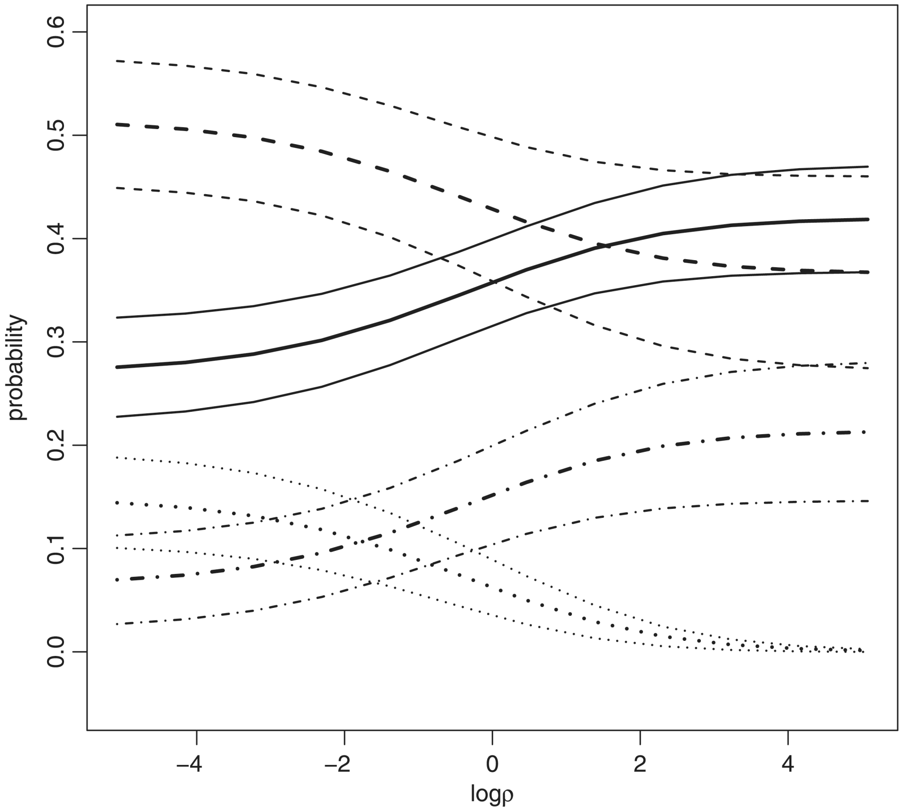

Fig. 3.

Sensitivity analysis for the MOTIVATE study of Section 4 based on the conditional odds ratio $\rho(\mathbf{X})$ defined by equation (6): point estimates (——, — — —, · — · —, · · · · · · ·) and 95% confidence intervals (——, — — —, ·— · —, ·· · · · ·) for π_{00} (——), π_{01} (— — —), π_{10} (· · · · · · ·) and π_{11} (· — · —), as functions of $\rho(\mathbf{X}) \equiv \rho$

Table 1

Contingency table for binary potential outcomes in HIV-1 patients treated with placebo or maraviroc (in addition to optimized background therapy)

Placebo	**Maraviroc**		**Row sum**
	Failure	**Success**	
Failure	π_{00}	π_{01}	π_{0+}
Success	π_{10}	π_{11}	π_{1+}
Column sum	π_{+0}	π_{+1}	1