Sensitivity analysis in non-inferiority trials with residual inconstancy after covariate adjustment

Journal of the
Royal Statistical Society

SERIES C
Applied
Statistics

Appl. Statist. (2014)

Sensitivity analysis in non-inferiority trials with residual inconstancy after covariate adjustment

Zhiwei Zhang, Lei Nie and Guoxing Soon

Food and Drug Administration, Silver Spring, USA

and Bo Zhang

Oregon State University, Corvallis, USA

[Received March 2012. Final revision August 2013]

Summary. A major issue in non-inferiority trials is the controversial assumption of constancy, namely that the active control has the same effect relative to placebo as in previous studies comparing the active control with placebo. The constancy assumption is often in doubt, which has motivated various methods that 'discount' the control effect estimate from historical data as well as methods that adjust for imbalances in observed covariates. We develop a new approach to deal with residual inconstancy, i.e. possible violations of the constancy assumption due to imbalances in unmeasured covariates after adjusting for the measured covariates. We characterize the extent of residual inconstancy under a generalized linear model framework and use the results to obtain fully adjusted estimates of the control effect in the current study based on plausible assumptions about an unmeasured covariate. Because such assumptions may be difficult to justify, we propose a sensitivity analysis approach that covers a range of situations. This approach is developed for indirect comparison with placebo and effect retention, and implemented through additive and multiplicative adjustments. The approach proposed is applied to two examples concerning benign prostate hyperplasia and human immunodeficiency virus infection, and evaluated in simulation studies.

Keywords: Active control; Conditional effect; Constancy; Discounting; Effect retention; Putative placebo

1. Introduction

The use of an active control is becoming increasingly common in randomized clinical trials, mostly for ethical reasons. The placebo control, which has traditionally been considered the gold standard for treatment evaluation, may be unethical to use when effective treatments are available and delaying treatment has irreversible consequences (Rothman and Michels, 1994; Temple and Ellenberg, 2000; Ellenberg and Temple, 2000; Schumi and Wittes, 2011; Witte *et al.*, 2011). Although a placebo control is usually not included in an active-controlled study, it remains relevant in important scientific and regulatory questions concerning the new treatment. For example, the efficacy or effectiveness of a treatment is often defined in comparison with a placebo in regulatory settings. Even for comparing the new and control treatments, it makes sense to consider not only the absolute difference but also the relative effect (i.e. the ratio of their effects relative to placebo), with interest in showing that the new treatment retains a certain

Address for correspondence: Zhiwei Zhang, Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA.
E-mail: zhiwei.zhang@fda.hhs.gov

0035–9254/14/63000

fraction of the control effect. Thus, appropriate analysis of an active-controlled trial, which is also known as a non-inferiority trial, usually requires external information about the control effect, which is often available in one or more historical studies comparing the active control with a placebo.

Among the many issues arising from non-inferiority trials is the controversial assumption of constancy, namely that the control effect remains constant between the historical study or studies and the current trial comparing the new treatment with the active control. The constancy assumption is often in doubt, which has motivated various 'discounting' methods (Snapinn, 2004). Under the fixed margin approach, a conservative estimate of the control effect based on historical data is used to define a non-inferiority margin, which is then treated as a constant in testing non-inferiority hypotheses (Hauck and Anderson, 1999; Wiens, 2002; Hung *et al.*, 2003; Rothman *et al.*, 2003). Under the synthesis approach, one seeks to demonstrate the effectiveness of the experimental treatment by showing that it preserves a specified fraction of the control effect (Holmgren, 1999; Hung *et al.*, 2003). There are also hybrid methods that combine the aforementioned two approaches in various ways (Food and Drug Administration, 1999; Wang and Huang, 2003; Gao and Ware, 2008; Witte *et al.*, 2011). Although attempting to address the lack of constancy, these discounting methods raise new issues. Scientifically, it is generally helpful to distinguish different objectives and different sources of uncertainty. For example, retaining a fraction of the control effect is an interesting hypothesis in its own right and should (ideally) be addressed separately from possible violations of the constancy assumption. Likewise, a lower confidence bound for the control effect is designed to account for uncertainty in the historical data and not that about the constancy assumption. When used as a fixed margin, it leads to an amount of discounting that is driven by the amount of historical data that are available and not by scientific considerations about between-trial differences. Practically, it can be difficult to determine the appropriate amount of discounting, and too much discounting can lead to unrealistically large sample sizes. Furthermore, even if the type I error rate is effectively controlled at or below the nominal level by means of discounting, it is not always clear how to estimate the effect of the new treatment relative to placebo in accordance with the discounting methods.

In light of these issues, covariate adjustment methods have been proposed to relax the constancy assumption (Zhang, 2009; Nie and Soon, 2010; Nie *et al.*, 2010). These methods allow the control effect to vary across studies, as long as the variation can be explained by relevant covariates (e.g. patient demographics and baseline characteristics) that are measured and differentially distributed in the current and historical studies. The latter assumption, which was formulated and termed 'conditional constancy' by Zhang (2009), is similar in spirit to the missingness at random assumption concerning missing data (Rubin, 1976) and the assumption of strongly ignorable treatment assignment in causal inference (Rosenbaum and Rubin, 1983). A practical limitation of covariate adjustment methods is the requirement for patient level data, unless the adjustment is focused on a single discrete covariate (Nie *et al.*, 2010). More importantly, the conditional constancy assumption, although less stringent than constancy, cannot be taken for granted because some relevant covariates may be unmeasured (e.g. drug resistance or concomitant drugs) or simply unknown to the investigator (e.g. unidentified genotype). Therefore, covariate adjustment does not eliminate all possible concerns about the lack of constancy, although it does represent a step forward.

In this paper, we propose a sensitivity analysis approach to deal with residual inconstancy, i.e. possible violations of the conditional constancy assumption due to unmeasured or unknown covariates. Under a generalized linear model framework, we show that the extent of residual inconstancy can be quantified under appropriate assumptions about the relationship of an unmeasured covariate with the outcome and its distributions in the studies. The results can be used to obtain fully adjusted estimates of the control effect in the current study by modifying the results of a partially adjusted analysis based on the observed covariates alone. Under the approach proposed, a variety of clinically plausible scenarios will be explored in preliminary calculations, a range will be specified for an additive or multiplicative adjustment that accounts for residual inconstancy, formal inference will be made for each value of the adjuster in the specified range and the results will be presented graphically and can be summarized numerically for specific purposes. This approach allows us to separate different sources of uncertainty, to crystallize the scientific question that requires clinical judgement and to deal with other types of uncertainty by using appropriate statistical techniques without unnecessary conservatism. It also provides some ability to work with summary statistics when patient level data are not available.

The approach proposed is developed for two distinct research questions:

(a) indirect comparison of the new treatment with placebo and
(b) the new treatment retaining a fraction of the control effect, both in the current patient population.

Both items have been discussed before, although their precise roles and interpretations vary across researchers (e.g. Food and Drug Administration (2010) and Huitfeldt *et al.* (2011)). Item (a) is also known as a putative placebo analysis (Hauck and Anderson, 1999; Fisher *et al.*, 2001; Hasselblad and Kong, 2001; D'Agostino *et al.*, 2003; Durrleman and Chaikin, 2003; Zhang, 2009). It should be noted that item (b), which has been used as a discounting mechanism, is regarded as a standalone research question in this paper.

The rest of the paper is organized as follows. The next section formulates the problem and explains the rationale for covariate adjustment. Section 3 describes residual inconstancy, and Section 4 presents a sensitivity analysis approach. The approach proposed is applied to two examples in Section 5 and evaluated in simulation studies in Section 6. The paper ends with a discussion in Section 7. Some additional information is provided as Web-based supplementary materials.

The programs that were used to analyse the data can be obtained from

```
http://wileyonlinelibary.com/journal/rss-datasets
```

2. Preliminaries

2.1. *Basic set-up*

For conceptual clarity, we formulate the statistical problem in terms of potential outcomes (Rubin, 1974). For a patient in the target population, let $Y(t)$ denote the (potential) clinical outcome that will realize if the patient receives a placebo ($t=0$), a standard treatment ($t=1$) or an experimental treatment ($t=2$). To fix ideas, suppose that the treatments are to be compared in terms of mean values of the corresponding outcomes: $\mu_t = E\{Y(t)\}$, $t=0,1,2$. The treatment differences will be denoted by $\Delta_{tt'} = \mu_{t'} - \mu_t$ for distinct treatments t and t'. The effectiveness of the experimental treatment, in a regulatory sense, may be defined by Δ_{02}, which compares the experimental treatment with a placebo. One may also be interested in the difference Δ_{12}, which compares the experimental treatment with the standard treatment, or the ratio $\lambda = \Delta_{02}/\Delta_{01}$, which measures the relative clinical utility of the experimental treatment to the standard of care. We are primarily interested in estimating Δ_{02} and λ in this paper.

Suppose that the experimental treatment is evaluated in a randomized study which also includes the standard treatment as an active control but not placebo. Write T for the actual treatment and $Y = Y(T)$ for the observed outcome. The study design implies that T is either 1 or 2 and never 0, so μ_1 and μ_2 are identifiable but μ_0 is not. The only treatment difference that is identified in this study is Δ_{12}. We note that

$$\Delta_{02} = \Delta_{01} + \Delta_{12}, \tag{1}$$

$$\lambda = (\Delta_{01} + \Delta_{12})/\Delta_{01}. \tag{2}$$

Thus, estimation of the quantities of interest could be helped by incorporating external information about Δ_{01}, which we assume is available from a historical study comparing the active control with a placebo with respect to the same clinical outcome as in the present study. We use a parallel notation system for the historical study, with an added asterisk to distinguish quantities from their counterparts in the present, active-controlled study. Thus $Y^*(t)$, T^* and Y^* denote respectively the potential outcomes, actual treatment and observed outcome in the historical study. Unlike T, T^* is either 0 or 1 and never takes the value 2. The historical data are directly informative about $\Delta^*_{01} = E\{Y^*(1) - Y^*(0)\}$, but not about Δ_{01} without additional assumptions.

The two studies could be connected through the so-called constancy assumption, namely that $\Delta_{01} = \Delta^*_{01}$. Under this assumption, Δ_{01} in equations (1) and (2) can be replaced by Δ^*_{01}, which shows that both Δ_{02} and λ can be identified from the two studies combined and estimated by substituting estimates of Δ_{12} and Δ^*_{01} from the present and historical studies respectively. Furthermore, variance formulae can be easily derived because the two studies are typically independent.

2.2. Covariate adjustment

The constancy assumption can be relaxed into the so-called conditional constancy assumption (Zhang, 2009), which essentially requires that any difference between Δ_{01} and Δ^*_{01} can be explained by imbalances across studies in a vector of baseline covariates that are measured in both studies. Denote this covariate vector by $\mathbf{X}$ for the current study and by $\mathbf{X}^*$ for the historical study. Define the conditional effects $\delta_{01}(\mathbf{x}) = E\{Y(1) - Y(0)|\mathbf{X} = \mathbf{x}\}$ and $\delta^*_{01}(\mathbf{x}) = E\{Y^*(1) - Y^*(0)|\mathbf{X}^* = \mathbf{x}\}$; then $\Delta_{01} = E\{\delta_{01}(\mathbf{X})\}$ and $\Delta^*_{01} = E\{\delta^*_{01}(\mathbf{X}^*)\}$. The conditional constancy assumption can be formulated as

$$\delta_{01}(\mathbf{x}) = \delta^*_{01}(\mathbf{x}) \qquad \text{for all } \mathbf{x}. \tag{3}$$

Assumption (3) may be more realistic than the constancy assumption in that the marginal effects Δ_{01} and Δ^*_{01} are allowed to differ. It allows Δ_{01} to be identified as

$$\Delta_{01} = E\{\delta^*_{01}(\mathbf{X})\}, \tag{4}$$

where the conditional effect $\delta^*_{01}(\cdot)$ is identifiable from the historical study, and the expectation is taken with respect to the covariate distribution in the current study. To fix ideas, consider the generalized linear model

$$E(Y^*|T^* = t, \mathbf{X}^* = \mathbf{x}) = g(\alpha_{t1} + \boldsymbol{\alpha}'_{tX}\mathbf{x}), \tag{5}$$

where g is an inverse link function. Write $\boldsymbol{\alpha}_t = (\alpha_{t1}, \boldsymbol{\alpha}'_{tX})'$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_0, \boldsymbol{\alpha}'_1)'$. This parameterization is chosen for notational convenience, as will become clear later. Because of randomization, the left-hand side of equation (5) is just $E\{Y^*(t)|\mathbf{X}^* = \mathbf{X}\}$, so the model implies that

$$\delta^*_{01}(\mathbf{x}) = g\{(1, \mathbf{x}')\boldsymbol{\alpha}_1\} - g\{(1, \mathbf{x}')\boldsymbol{\alpha}_0\}. \tag{6}$$

Now assumption (3) further implies that

$$\Delta_{01} = E[g\{(1, \mathbf{X}')\boldsymbol{\alpha}_1\} - g\{(1, \mathbf{X}')\boldsymbol{\alpha}_0\}], \tag{7}$$

which can be estimated by substituting regression parameter estimates based on the historical data and then averaging over X in the current study.

3. Residual inconstancy

In reality, even assumption (3) can be violated because the current and historical studies may differ in covariates that are unmeasured or simply unknown. We therefore relax assumption (3) by including an unmeasured baseline variable, which is denoted by U for the present study and by U^* for the historical study. Let $d_{01}(\mathbf{x}, u) = E\{Y(1) - Y(0)|\mathbf{X} = \mathbf{x}, U = u\}$ and $d_{01}^*(\mathbf{x}, u) = E\{Y^*(1) - Y^*(0)|\mathbf{X}^* = \mathbf{x}, U^* = u\}$; then the relaxed assumption can be written as

$$d_{01}(\mathbf{x}, u) = d_{01}^*(\mathbf{x}, u) \qquad \text{for all } (\mathbf{x}, u). \tag{8}$$

$\delta_{01}(\mathbf{x}) = E\{d_{01}(\mathbf{x}, U)|\mathbf{X} = \mathbf{x}\}$ and $\delta_{01}^*(\mathbf{x}) = E\{d_{01}^*(\mathbf{x}, U^*)|\mathbf{X}^* = \mathbf{x}\}$. Because $(U|\mathbf{X} = \mathbf{x})$ and $(U^*|\mathbf{X}^* = \mathbf{x})$ may follow different distributions, assumption (8) does allow the conditional effects $\delta_{01}(\mathbf{x})$ and $\delta_{01}^*(\mathbf{x})$ to differ. Thus assumption (3) no long holds, and equation (4) is generally invalid. We denote the right-hand side of equation (4) by Δ_{01}° and call it the partially adjusted control effect. Analogously to model (5), we assume that Y^* is related to $(T^*, \mathbf{X}^*, U^*)$ through the model

$$E(Y^*|T^* = t, \mathbf{X}^* = \mathbf{x}, U^* = u) = g(\beta_{t1} + \boldsymbol{\beta}_{tX}'\mathbf{x} + \beta_{tU}u). \tag{9}$$

Write $\boldsymbol{\beta}_t = (\beta_{t1}, \boldsymbol{\beta}_{tX}', \beta_{tU})'$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_0', \boldsymbol{\beta}_1')'$. Then

$$d_{01}^*(\mathbf{x}, u) = g\{(1, \mathbf{x}', u)\boldsymbol{\beta}_1\} - g\{(1, \mathbf{x}', u)\boldsymbol{\beta}_0\},$$

and, by assumption (8),

$$\Delta_{01} = E[g\{(1, \mathbf{X}', U)\boldsymbol{\beta}_1\} - g\{(1, \mathbf{X}', U)\boldsymbol{\beta}_0\}]. \tag{10}$$

As we shall see in Section 5.2, U and U^* may be vector valued and partially measured in one or both studies. In this section, however, we assume for simplicity that U and U^* are completely unmeasured scalars. Without additional information, it is impossible to fit model (9) directly or to estimate Δ_{01} by using equation (10). Therefore, we work with the induced model

$$E(Y^*|T^* = t, \mathbf{X}^* = \mathbf{x}) = \int g\{(1, \mathbf{x}', u)\boldsymbol{\beta}_t\}\,\mathrm{d}F^*(u|\mathbf{x}), \tag{11}$$

where $F^*(\cdot|\mathbf{x})$ denotes the conditional distribution of U^* given $\mathbf{X}^* = \mathbf{x}$. With (β_{0U}, β_{1U}) given and F^* fully specified, one could estimate the rest of $\boldsymbol{\beta}$ by fitting model (11). If we further specify F, the current study counterpart of F^*, then Δ_{01} could be estimated by using the relationship

$$\Delta_{01} = E\left(\int [g\{(1, \mathbf{X}', u)\boldsymbol{\beta}_1\} - g\{(1, \mathbf{X}', u)\boldsymbol{\beta}_0\}]\,\mathrm{d}F(u|\mathbf{X})\right). \tag{12}$$

This approach is available for any inverse link g and arbitrary distributions (F^*, F), but its implementation may require considerable computational effort for evaluating the integrals and fitting model (11) separately for each specification of $(\beta_{0U}, \beta_{1U}, F^*)$. In the rest of this section, we show that expressions (11) and (12) simplify for some common link functions combined with suitable assumptions about U^* and U. Specifically, we assume that U^* and U are independent of $\mathbf{X}^*$ and $\mathbf{X}$ respectively, that we are given plausible values of β_{0U}, β_{1U}, $\gamma^* = E(U^*)$ and $\gamma = E(U)$ (or some functions of these parameters), and in some cases that the (now marginal) distributions

F^* and F are fully specified (as Bernoulli or normal). The subsections to follow are independent in notation.

3.1. The identity link

For the identity link, without specifying F^* and F, model (5) is correct with

$$\alpha_{t1} = \beta_{t1} + \beta_{tU}\gamma^*,$$
$$\boldsymbol{\alpha}_{tX} = \boldsymbol{\beta}_{tX}.$$

Substituting these into equation (7) shows that the partially adjusted control effect is

$$\Delta_{01}^{\circ} = \alpha_{11} - \alpha_{01} + (\boldsymbol{\alpha}_{1X} - \boldsymbol{\alpha}_{0X})' E(\mathbf{X}) = \beta_{11} - \beta_{01} + (\beta_{1U} - \beta_{0U})\gamma^* + (\boldsymbol{\beta}_{1X} - \boldsymbol{\beta}_{0X})' E(\mathbf{X}). \quad (13)$$

From equation (10) and the assumptions stated, the true control effect is easily seen to be

$$\Delta_{01} = \beta_{11} - \beta_{01} + (\boldsymbol{\beta}_{1X} - \boldsymbol{\beta}_{0X})' E(\mathbf{X}) + (\beta_{1U} - \beta_{0U})\gamma. \quad (14)$$

Comparing equations (13) and (14), the bias in the partially adjusted estimate of Δ_{01} based on equation (7) is seen to be

$$\Delta_{01}^{\circ} - \Delta_{01} = (\beta_{1U} - \beta_{0U})(\gamma^* - \gamma). \quad (15)$$

The first difference on the right-hand side, $\beta_{1U} - \beta_{0U}$, represents the strength of the unmeasured covariate as an effect modifier, whereas the second, $\gamma^* - \gamma$, measures the discrepancy between the two studies with respect to the unmeasured effect modifier. The bias vanishes if either difference is 0. Once the two differences have been specified, expression (15) can be used to correct for bias in a sensitivity analysis.

3.2. The log-link

Now consider the log-link (i.e. $g \equiv \exp$). Without specifying F^* and F yet, it is easy to see that model (5) is correctly specified with

$$\alpha_{t1} = \beta_{t1} + \log[E\{\exp(\beta_{tU} U^*)\}],$$

$$\boldsymbol{\alpha}_{tX} = \boldsymbol{\beta}_{tX},$$

and the partially adjusted control effect is given by

$$\Delta_{01}^{\circ} = E[\exp\{(1, \mathbf{X}')\boldsymbol{\alpha}_1\} - \exp\{(1, \mathbf{X}')\boldsymbol{\alpha}_0\}].$$

In contrast, the true control effect is

$$\Delta_{01} = E[\exp\{(1, \mathbf{X}')\boldsymbol{\alpha}_1^{\diamond}\} - \exp\{(1, \mathbf{X}')\boldsymbol{\alpha}_0^{\diamond}\}], \quad (16)$$

where $\alpha_{t1}^{\diamond} = \alpha_{t1} + \log[E\{\exp(\beta_{tU} U)\}/E\{\exp(\beta_{tU} U^*)\}]$ and $\boldsymbol{\alpha}_{tX}^{\diamond} = \boldsymbol{\alpha}_{tX}$, $t = 0, 1$. With $\boldsymbol{\alpha}_t$ directly estimable from the historical data, expression (16) can be estimated as long as we can evaluate the ratio $E\{\exp(\beta_{tU} U)\}/E\{\exp(\beta_{tU} U^*)\}$, which requires assumptions on F and F^*. For example, if U and U^* are binary (0 or 1), then

$$\frac{E\{\exp(\beta_{tU} U)\}}{E\{\exp(\beta_{tU} U^*)\}} = \frac{1 - \gamma + \gamma \exp(\beta_{tU})}{1 - \gamma^* + \gamma^* \exp(\beta_{tU})}.$$

Alternatively, if we assume that

$$U \sim N(\gamma, \sigma^2) \quad \text{and} \quad U^* \sim N(\gamma^*, 1), \quad (17)$$

we then have

$$\frac{E\{\exp(\beta_{tU}U)\}}{E\{\exp(\beta_{tU}U^*)\}}=\frac{\exp(\beta_{tU}\gamma+\beta_{tU}^2\sigma^2/2)}{\exp(\beta_{tU}\gamma^*+\beta_{tU}^2/2)}.$$

In expression (17), there is no loss of generality in assuming that $\mathrm{var}(U^*)=1$ because the β_{tU} need to be specified anyway.

3.3. The probit link

Now let $g=\Phi$, the standard normal distribution function, and assume that expression (17) holds. Using the argument of Carroll *et al.* (1984), it can be shown that model (5) holds with regression coefficients $\boldsymbol{\alpha}_t=(1+\beta_{tU}^2)^{-1/2}(\beta_{t1}+\beta_{tU}\gamma^*,\boldsymbol{\beta}_{tX}')'$, and the partially adjusted control effect is

$$\Delta_{01}^{\circ}=E[\Phi\{(1,\mathbf{X}')\boldsymbol{\alpha}_1\}-\Phi\{(1,\mathbf{X}')\boldsymbol{\alpha}_0\}].$$

The same argument can be used to show that the true control effect is

$$\Delta_{01}=E[\Phi\{(1,\mathbf{X}')\boldsymbol{\alpha}_1^{\diamond}\}-\Phi\{(1,\mathbf{X}')\boldsymbol{\alpha}_0^{\diamond}\}],$$

where

$$\begin{aligned}\boldsymbol{\alpha}_t^{\diamond}&=(1+\beta_{tU}^2\sigma^2)^{-1/2}(\beta_{t1}+\beta_{tU}\gamma,\boldsymbol{\beta}_{tX}')'\\&=(1+\beta_{tU}^2\sigma^2)^{-1/2}((1+\beta_{tU}^2)^{1/2}\alpha_{t1}+\beta_{tU}(\gamma-\gamma^*),(1+\beta_{tU}^2)^{1/2}\boldsymbol{\alpha}_{tX}')'.\end{aligned}$$

Thus, as soon as $\boldsymbol{\alpha}_t$ is estimated and $(\beta_{tU},\gamma-\gamma^*,\sigma^2)$ specified, the above expressions can be used to estimate $\boldsymbol{\alpha}_t^{\diamond}$ and eventually the true control effect.

3.4. The logit link

Unfortunately, for the logit link with $g(z)=\mathrm{expit}(z)=\exp(z)/\{1+\exp(z)\}$, the induced model (11) does not seem to take a simple form, even for fully specified F^*. We therefore consider an approximation of the expit function by $\Phi(\cdot/c)$ with $c=15\pi/(16\sqrt{3})\approx1.70$ (e.g. Johnson and Kotz (1970), Zeger *et al.* (1988) and Liang and Liu (1991)). This allows us to write, under assumption (17),

$$\begin{aligned}E(T^*|T^*=t,\mathbf{X}^*=\mathbf{x})&=\int\mathrm{expit}\{(1,\mathbf{x}',u)\boldsymbol{\beta}_t\}\,\mathrm{d}F^*(u)\\&\approx\int\Phi\{c^{-1}(1,\mathbf{x}',u)\boldsymbol{\beta}_t\}\,\mathrm{d}F^*(u)\\&=\Phi\left\{\frac{\beta_{t1}+\boldsymbol{\beta}_{tX}'\mathbf{x}+\beta_{tU}\gamma^*}{c\sqrt{(1+\beta_{tU}^2/c^2)}}\right\}\\&\approx\mathrm{expit}\left\{\frac{\beta_{t1}+\boldsymbol{\beta}_{tX}'\mathbf{x}+\beta_{tU}\gamma^*}{\sqrt{(1+\beta_{tU}^2/c^2)}}\right\},\end{aligned}$$

where the second-to-last step follows from the same argument as used for the probit link. Thus, under assumption (17), model (5) holds approximately with parameters $\boldsymbol{\alpha}_t=(1+\beta_{tU}^2/c^2)^{-1/2}\times(\beta_{t1}+\beta_{tU}\gamma^*,\boldsymbol{\beta}_{tX}')'$. A similar argument leads to

$$\Delta_{01}\approx E[\mathrm{expit}\{(1,\mathbf{X}')\boldsymbol{\alpha}_1^{\diamond}\}-\mathrm{expit}\{(1,\mathbf{X}')\boldsymbol{\alpha}_0^{\diamond}\}],$$

where

$$\begin{aligned}\boldsymbol{\alpha}_t^{\diamond} &= (1+\beta_{tU}^2\sigma^2/c^2)^{-1/2}(\beta_{t1}+\beta_{tU}\gamma, \boldsymbol{\beta}_{tX}') \\ &= (1+\beta_{tU}^2\sigma^2/c^2)^{-1/2}((1+\beta_{tU}^2/c^2)^{1/2}\alpha_{t1}+\beta_{tU}(\gamma-\gamma^*), (1+\beta_{tU}^2/c^2)^{1/2}\boldsymbol{\alpha}_{tX}')'.\end{aligned}$$

Once again, these expressions allow us to perform a sensitivity analysis using estimates of α_t and specified values of $(\beta_{0U}, \beta_{1U}, \gamma-\gamma^*, \sigma^2)$.

4. Sensitivity analysis

4.1. Outline

The foregoing discussion suggests a sensitivity analysis approach that can be outlined as follows.

Initially, obtain a partially adjusted estimate of Δ_{01} based on model (5) and equation (7).

Step 1: obtain fully adjusted estimates of Δ_{01} based on clinically plausible specifications of $(\beta_{0U}, \beta_{1U}, F, F^*)$ or some functionals of these parameters, using the results of Section 3. Formal inference on the quantities of interest (Δ_{02} and λ) could be made for each specification of the unidentifiable parameters. However, this can be cumbersome because some parameters may need to be specified, giving rise to a potentially large number of combinations.

Step 2: for dimension reduction, one might focus instead on the difference $a=\Delta_{01}-\Delta_{01}^{\circ}$ or the ratio $r=\Delta_{01}/\Delta_{01}^{\circ}$ if both effects are positive. A range of plausible values, $\mathcal{A}$ for a or $\mathcal{R}$ for r, can be specified by using the results from step 1 and other sources of information (to be discussed shortly).

Step 3: modify the partially adjusted estimate of Δ_{01} additively or multiplicatively, and make formal inference on Δ_{02} and λ accordingly. This will be done for each $a\in\mathcal{A}$ or $r\in\mathcal{R}$ separately, assuming that $\Delta_{01}=\Delta_{01}^{\circ}+a$ or $\Delta_{01}=r\Delta_{01}^{\circ}$ respectively. The resulting point estimates and confidence intervals can then be plotted against a or r.

Step 4: the results of step 3 can be summarized succinctly for specific purposes. For testing hypotheses, say $H_0:\Delta_{02}\leqslant 0$ *versus* $H_1:\Delta_{02}>0$, one could report the smallest value of a or r at which the test is significant. To obtain a single confidence interval, one could take the union of the 'pointwise' confidence intervals over $a\in\mathcal{A}$ or $r\in\mathcal{R}$, or a subset of values that are deemed most likely.

4.2. Implementation

Let $(T_i, \mathbf{X}_i, Y_i)$, $i=1,\ldots,n$, denote the data from the current active-controlled study, which are assumed to be independent copies of $(T, \mathbf{X}, Y)$. Similar notation (with an asterisk) is used for the historical data.

For the initialization, we give a brief description of the partially adjusted estimates based on model (5) and equation (7). The working model (5) is estimated by solving a system of estimating equations:

$$\sum_{i=1}^{n^*}[Y_i^*-g\{(1,\mathbf{X}_i^{*\prime})\alpha_{T_i^*}\}]\mathbf{h}(T_i^*,\mathbf{X}_i^*)=0, \tag{18}$$

where $\mathbf{h}$ is a vector-valued function of the same dimension as $\boldsymbol{\alpha}$. For efficiency, $\mathbf{h}(T_i^*,\mathbf{X}_i^*)$ is usually taken to be an estimate of $\mathrm{var}(Y_i^*|T_i^*,\mathbf{X}_i^*)^{-1}g_\alpha(T_i^*,\mathbf{X}_i^*)$, where $g_\alpha(T_i^*,\mathbf{X}_i^*)=\partial g\{(1,\mathbf{X}_i^{*\prime})\boldsymbol{\alpha}_{T_i^*}\}/\partial\boldsymbol{\alpha}$. For the identity link, a unique solution to equation (18) exists in closed form. In general, equation (18) can be solved by using an iterative algorithm such as the iteratively reweighted least squares algorithm. Denote by $\hat{\boldsymbol{\alpha}}=(\hat{\boldsymbol{\alpha}}_0', \hat{\boldsymbol{\alpha}}_1')'$ the resulting estimate of $\boldsymbol{\alpha}$; then equation (7) suggests that Δ_{01} be estimated by

$$\hat{\Delta}_{01}=\frac{1}{n}\sum_{i=1}^{n}[g\{(1,\mathbf{X}_i')\hat{\boldsymbol{\alpha}}_1\}-g\{(1,\mathbf{X}_i')\hat{\boldsymbol{\alpha}}_0\}]. \tag{19}$$

Suppose that Δ_{12} is estimated by the observed treatment difference

$$\hat{\Delta}_{12}=\hat{\mu}_2-\hat{\mu}_1=\frac{1}{N_2}\sum_{i=1}^{n}I(T_i=2)Y_i-\frac{1}{N_1}\sum_{i=1}^{n}I(T_i=1)Y_i,$$

where $N_t=\Sigma_{i=1}^{n}I(T_i=t)$ $(t=1,2)$ and $I(\cdot)$ is the indicator function. Then partially adjusted estimates of Δ_{02} and λ are given by

$$\hat{\Delta}_{02}=\hat{\Delta}_{01}+\hat{\Delta}_{12},$$
$$\hat{\lambda}=\hat{\Delta}_{02}/\hat{\Delta}_{01}.$$

In step 1, we simply substitute these estimates together with specified values of unidentifiable parameters. With the identity link, for example, a fully adjusted estimate of Δ_{01} based on equation (15) is given by

$$\tilde{\Delta}_{01}=\hat{\Delta}_{01}+(\beta_{1U}-\beta_{0U})(\gamma-\gamma^*).$$

Suppose that the unmeasured effect modifier is a genotype whose presence results in an increase of 10 units in the effect of the active treatment relative to placebo. Assuming independence of this genotype and the measured effect modifiers, the above equation then indicates that an adjustment of $a=2$ would be appropriate if the genotype is 20% more prevalent in the current patient population than in the historical population. In the other special cases that were considered in Section 3, the fully adjusted estimate of Δ_{01} is given by

$$\tilde{\Delta}_{01}=\frac{1}{n}\sum_{i=1}^{n}[g\{(1,\mathbf{X}_i')\tilde{\boldsymbol{\alpha}}_1^{\diamond}\}-g\{(1,\mathbf{X}_i')\tilde{\boldsymbol{\alpha}}_0^{\diamond}\}],$$

where $\tilde{\boldsymbol{\alpha}}^{\diamond}$ is obtained from $\hat{\boldsymbol{\alpha}}$ through a linear transformation determined by unidentifiable parameters.

In step 2, we specify a range ($\mathcal{A}$ or $\mathcal{R}$) by using statistical techniques and clinical knowledge. This can be helped by comparing $\hat{\Delta}_{01}$ with $\tilde{\Delta}_{01}$ on the basis of a variety of assumptions. If $\dim(X)\geqslant 2$, a jackknife-type approach can be used to yield additional information. Denote by $\hat{\Delta}_{01}^{(-j)}$ the analogue of $\hat{\Delta}_{01}$ based on a reduced set of covariates excluding the jth element of $\mathbf{X}$ (and $\mathbf{X}^*$). Then the difference $a_j=\hat{\Delta}_{01}-\hat{\Delta}_{01}^{(-j)}$ measures the effect of omitting a covariate, and a candidate for $\mathcal{A}$ is given by the interval $[\min_j a_j, \max_j a_j]$. Similarly, a candidate for $\mathcal{R}$ may be obtained as $[\min_j r_j, \max_j r_j]$, where $r_j=\hat{\Delta}_{01}/\hat{\Delta}_{01}^{(-j)}$, $j=1,\ldots,\dim(X)$. These choices could be sharpened by incorporating clinical information, if available. For example, one could draw on the relationships

$$|a|=|E\{\delta_{01}(\mathbf{X})-\delta_{01}^*(\mathbf{X})\}|\leqslant\sup_{\mathbf{x}}|\delta_{01}(\mathbf{x})-\delta_{01}^*(\mathbf{x})|,$$
$$\inf_{\mathbf{x}}\{\delta_{01}(\mathbf{x})/\delta_{01}^*(\mathbf{x})\}\leqslant r\leqslant\sup_{\mathbf{x}}\{\delta_{01}(\mathbf{x})/\delta_{01}^*(\mathbf{x})\},$$

the latter assuming that both $\delta(\cdot)$ and $\delta^*(\cdot)$ are positive valued. To approximate the bounds, one might ask a clinician ‘Knowing what we know about the studies, are you concerned that an unmeasured factor alone could alter the control effect by a factor of k or more?’. The question could be repeated with different values of $k>1$ until the smallest value of k is found for which the answer is negative; then it seems reasonable to use the interval $(1/k,k)$ to truncate a tentative choice of $\mathcal{R}$ (obtained by using other methods). The question can be reworded to yield information on $\mathcal{A}$.

In step 3, we make simple (additive or multiplicative) adjustments to the point estimates from the initialization and adjust the inferential procedure accordingly. For $a \in \mathcal{A}$, the additive adjustment yields

$$\begin{aligned}\tilde{\Delta}_{01}^{+}(a) &= \hat{\Delta}_{01} + a,\\ \tilde{\Delta}_{02}^{+}(a) &= \tilde{\Delta}_{01}^{+}(a) + \hat{\Delta}_{12}\\ &= \hat{\Delta}_{02} + a,\\ \tilde{\lambda}^{+}(a) &= \tilde{\Delta}_{02}^{+}(a)/\tilde{\Delta}_{01}^{+}(a)\\ &= (\hat{\Delta}_{02} + a)/(\hat{\Delta}_{01} + a).\end{aligned}$$

For $r \in \mathcal{R}$, the multiplicative adjustment leads to

$$\begin{aligned}\tilde{\Delta}_{01}^{\times}(r) &= r\hat{\Delta}_{01},\\ \tilde{\Delta}_{02}^{\times}(r) &= \tilde{\Delta}_{01}^{\times}(r) + \hat{\Delta}_{12}\\ &= r\hat{\Delta}_{01} + \hat{\Delta}_{12},\\ \tilde{\lambda}^{\times}(r) &= \tilde{\Delta}_{02}^{\times}(r)/\tilde{\Delta}_{01}^{\times}(r)\\ &= (r\hat{\Delta}_{01} + \hat{\Delta}_{12})/(r\hat{\Delta}_{01}).\end{aligned}$$

Assuming that the historical and current samples grow proportionally, all these estimates are asymptotically normal though not necessarily consistent. The associated formulae for asymptotic inference are given in the on-line appendix A. Because the formulae for λ are somewhat cumbersome, one might prefer a bootstrap procedure for inference on λ.

5. Applications

5.1. A urological example

We now illustrate the proposed approach with a urological example analysed previously by Zhang (2009) using covariate adjustment. Trans-urethral microwave therapy (TUMT) is a noninvasive treatment of symptoms due to benign prostate hyperplasia. A randomized, sham-controlled, multicentre clinical study has been conducted to evaluate a TUMT device, say TUMT1. The study enrolled 300 male subjects over the age of 50 years who had been diagnosed with benign prostate hyperplasia and had not been treated for it, with prostate size 20–50 cm^3 and American Urology Association Symptom index AUASI at least 12. The AUASI-score ranges between 0 and 35 with higher values indicating more severe symptoms. The subjects in the study were assigned randomly, at a 2:1 ratio, to either treatment TUMT1 or a sham control, which 'treated' the patient with the same device in the off mode. Thus, the sham in a medical device trial is largely equivalent to a placebo in a drug trial. The subjects in this study were blinded to the assigned treatment. The primary effectiveness end point was the decrease in AUASI from baseline to 6 months post treatment. A summary of the primary analysis is given in Table 1, which is reproduced from Zhang (2009), Table 3, under the heading 'Historical study'. The results show a statistically significant treatment difference as well as a remarkable sham effect.

Of interest to us is a newly developed TUMT device which might be called TUMT2. Given the existence of proven effective treatments including TUMT1, another sham-controlled study for TUMT2 would have been difficult to implement because of ethical as well as practical issues such as patient enrolment. Thus, the evaluation of TUMT2 was based primarily on a non-inferiority study comparing TUMT2 with TUMT1. Like the original study for TUMT1, this latter study was randomized and blinded, involved multiple centres, used essentially the same inclusion–exclusion criteria and had the same primary end point. The key summary statistics,

Table 1. Summary of primary effectiveness analyses for the two benign prostate hyperplasia studies in Section 5.1

	Historical study results		*Current study results*	
	Sham	*TUMT1*	*TUMT1*	*TUMT2*
Number of evaluable patients	95	198	98	97
Mean decrease in AUASI	7.0	10.8	13.6	12.1
Standard deviation	6.9	7.4	7.9	7.2
Observed difference (left − right)	3.8		−1.5	
95% confidence interval	(2.1, 5.5)		(−3.6, 0.6)	
p-value for superiority (1 sided)	<0.0001		0.9171	

which are presented in Table 1 under 'Current study', show that the mean decrease in AUASI observed in the TUMT2 group was smaller than that seen in the TUMT1 group, although the difference was not statistically significant.

It is necessary to combine information from the two studies to answer the aforementioned research questions about TUMT2 (i.e. indirect comparison with sham and effect retention with respect to TUMT1). This seems promising given the apparent similarities between the two studies, with one notable exception: the mean baseline AUASI-score was higher in the second study than in the first (26.6 *versus* 23.5), even though the same entry criterion (baseline AUASI $\geqslant$ 12) was used. Fig. 1 shows histograms of baseline AUASI-scores in both studies as well as a non-parametric regression analysis of the historical data suggesting that the effect of TUMT1 relative to sham decreases with the patient's baseline AUASI-score. Thus, baseline AUASI-score appears to be an effect modifier that requires adjustment, and the studies may differ in other important aspects that are unmeasured or simply unknown. The latter source of uncertainty can be addressed by using the proposed sensitivity analysis approach, as we now demonstrate.

A linear regression model with both linear and quadratic terms was used to adjust for baseline AUASI-score, leading to a partially adjusted estimate of 3.1 for the control effect (i.e. the effect of TUMT1 relative to placebo in the current population), which is lower than the unadjusted estimate (3.8) assuming constancy. Under this linear model, the bias due to omitting an unmeasured effect modifier is given by equation (15), which involves the strength of the unmeasured effect modifier and the difference of its means in the two populations. Using this relationship (with the unmeasured covariate assumed to be a genotype), a wide range of scenarios was explored with the help of a clinical expert, leading to $\mathcal{A}=(-1.6, 1.6)$ and $\mathcal{R}=(0.5, 2)$ as plausible ranges for a and r respectively. Fig. 2 shows point estimates and confidence intervals for Δ_{02} and λ, obtained from an additive or multiplicative adjustment and plotted against a or r over the specified range. The confidence intervals for Δ_{02} are based on the formulae in the on-line appendix A, whereas those for λ are based on 1000 bootstrap samples. The results suggest that TUMT2 may have a positive effect relative to sham, but the evidence for that is not conclusive.

5.2. A human immunodeficiency virus example

Our second example concerns treatment of human immunodeficiency virus type 1 (HIV-1) in treatment-experienced patients. Raltegravir is an inhibitor of HIV-1 integrase active against HIV-1 susceptible or resistant to older antiretroviral drugs. The drug was developed by Merck & Co. and approved for marketing in the USA on the basis of two identically designed, random-

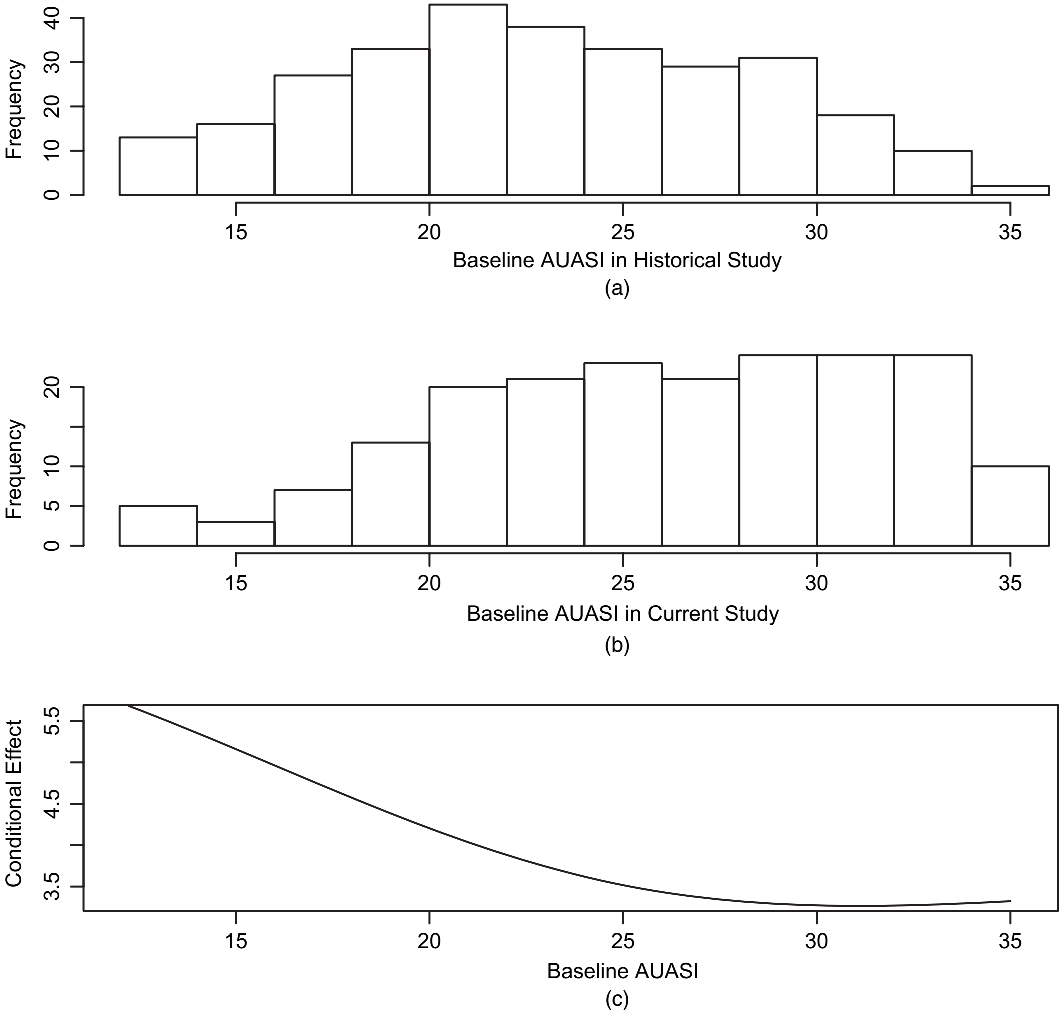

Fig. 1. Histograms of baseline AUASI-scores in (a) the historical and (b) the current studies, and (c) a kernel estimate of the conditional effect of TUMT1 relative to a sham

ized, placebo-controlled, double-blind, multicentre trials, which are combined for our purpose (because of consistent results) and collectively referred to as the 'BENCHMRK' study (Steigbigel *et al.*, 2008; Cooper *et al.*, 2008). The study randomized 703 HIV-1 patients to either Raltegravir or placebo, at a 2:1 ratio, both in combination with optimized background therapy. In our analysis, the primary end point is taken to be the virologic response rate (i.e. the proportion of patients with HIV ribonucleic acid (RNA) levels below 50 copies per millilitre) at week 48 of treatment, which is more informative of long-term effects than the original primary end point (virologic response rate at week 16). The key summary statistics for this primary end point, which are shown in Table 2, indicate clearly that the use of Raltegravir in combination with optimized background therapy improves the virologic response rate at week 48.

Our research question pertains to the efficacy of Elvitegravir, which is another HIV-1 integrase inhibitor under investigation. Elvitegravir was compared with Raltegravir in a randomized, double-blind, multicentre trial known as 'Study 145' in a population of treatment-experienced

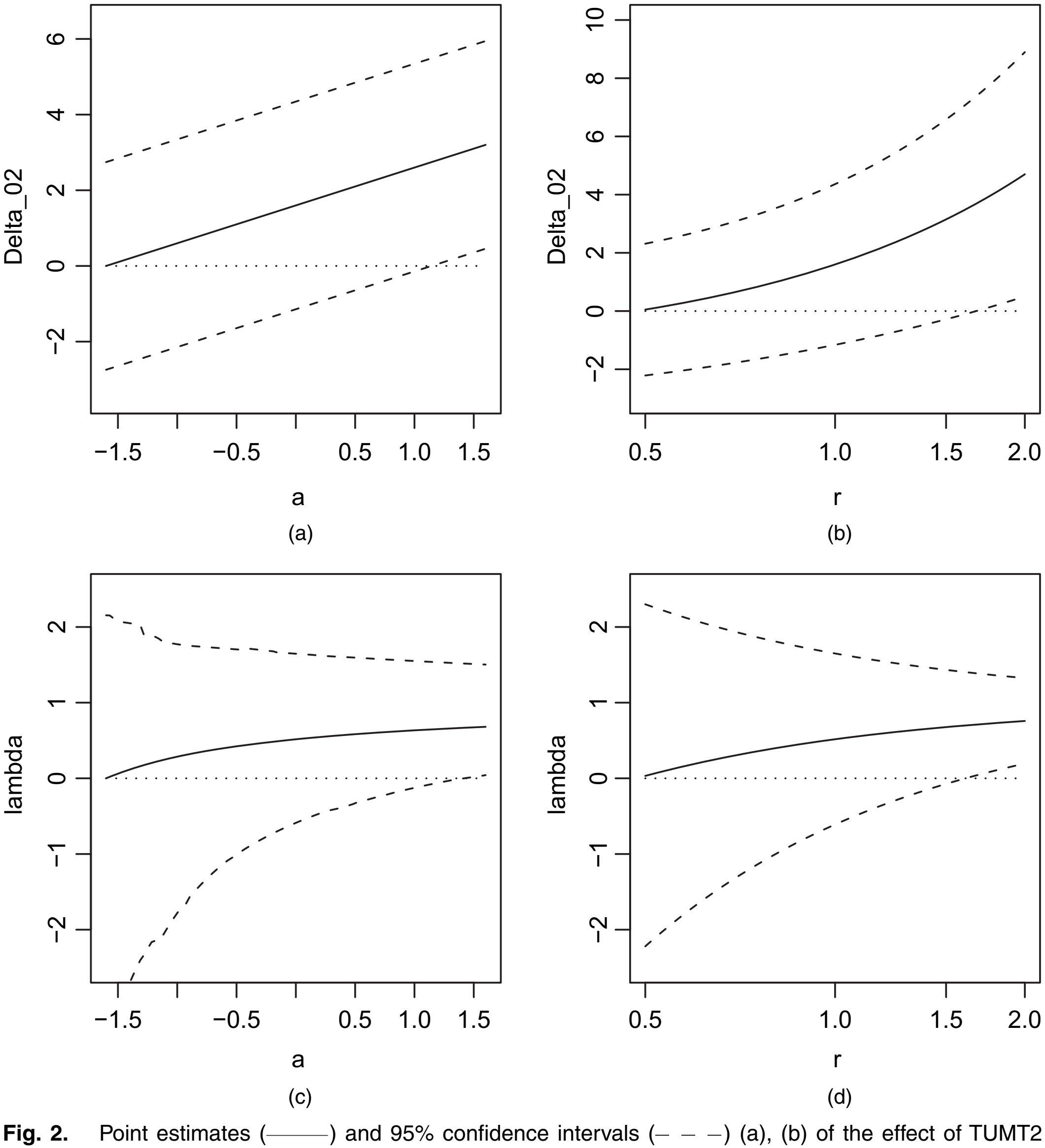

Fig. 2. Point estimates (———) and 95% confidence intervals (– – –) (a), (b) of the effect of TUMT2 relative to a sham (Δ_{02}) and (c), (d) the fraction of the control (TUMT1) effect retained by TUMT2 (λ), both in the current patient population, obtained by using the proposed sensitivity analysis approach with an additive ((a), (c)) or multiplicative ((b), (d)) adjustment following standard covariate adjustment (see Section 5.1 for details)

patients (Molina *et al.*, 2012). The study randomized 724 patients to Elvitegravir or Raltegravir with equal probability, both with a background regimen of a fully active, ritonavir-boosted protease inhibitor and a second agent. The virologic response rate at week 48 was the prespecified primary end point in Study 145, for which a summary is also presented in Table 2. The results show that Elvitegravir is associated with a response rate that is similar to, if not slightly higher than, that of Raltegravir. Though not sufficiently strong for superiority, these data appear to meet a non-inferiority criterion with a 10% margin (Molina *et al.*, 2012).

Table 2. Summary of primary efficacy analyses for the two HIV studies in Section 5.2

	BENCHMRK study results		*Study 145 results*	
	Placebo	*Raltegravir*	*Raltegravir*	*Elvitegravir*
Number of evaluable patients	228	443	351	351
Response proportion (%)	34	64	58	59
Observed difference (right – left)	30		1.1	
95% confidence interval	(22, 38)		$(-6.2, 8.4)$	
p-value for superiority (1 sided)	< 0.0001		0.3102	

Table 3. Relevant subgroup information for the two HIV studies in Section 5.2

Subgroup		*Study 145 proportion (%)*	*BENCHMRK study results*			
Variable	*Level*		*Proportion (%)*	*Response rate (%)*		
				Raltegravir	*Placebo*	*Difference*
RNA	$\leqslant 10^5$	74	65	73	43	30
	$> 10^5$	26	35	48	16	32
API	No	0	39	54	14	40
	Yes	100	61	71	49	22
GSS	0	?	26	45	3	42
	1	$\geqslant 15$	40	67	37	30
	2	$\geqslant 82$	23	77	62	15
	$\geqslant 3$	?	11	71	52	19
PSS	0	0	17	51	2	49
	1	?	32	61	29	32
	2	?	31	71	39	32
	$\geqslant 3$	0	20	71	61	10

To compare Elvitegravir with placebo, we need to adjust for important covariates such as RNA (baseline plasma HIV-1 RNA level, dichotomized at 100000 copies per millilitre), API (concomitant use of active protease inhibitors), 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). These four covariates were measured in the BENCHMRK study and reported in separate subgroup analyses (Cooper *et al.*, 2008). The right-hand half of Table 3 gives the proportion of each subgroup as well as the response rates and the treatment difference in each subgroup, calculated after excluding small amounts of missing data (less than 5%). Table 3 also shows the proportions of subgroups in Study 145, with partial information on GSS and PSS. Note that the study design implies that API is always positive in Study 145. It can be determined that at least 103 patients (14.7%) had GSS = 1 and 575 (81.9%) had GSS = 2 in Study 145 (Molina *et al.* (2012), Table 3), but it is unclear whether the other 24 patients (3.4%) belonged to the other categories or had missing GSS-information. PSS-information is not reported for Study 145; however, the study design implies that 1 and 2 are the only possible values.

In this example, the main motivation for a sensitivity analysis is the unavailability of patient level data (which are necessary for simultaneous adjustment for all four covariates). So we adjust for RNA only in a binary regression model with the identity link, which is convenient for our purpose, and consider the other three covariates in a sensitivity analysis. API is obviously a strong effect modifier, and Table 3 suggests that the control effect shrinks by 8 percentage points (from 30% overall to 22% in API positive patients) because of changes in API. Alternatively, this can be calculated by using equation (15) together with the relevant information in Table 3. Similar calculations for GSS show that the control effect is expected to shrink by 10–12 percentage points owing to changes in the GSS distribution, with the ambiguity arising from the 3.4% GSS-ambiguous patients. The lower end (10%) corresponds to all of these patients having $\mathrm{GSS}=0$, and the upper end to $\mathrm{GSS}=2$. Lastly, despite the lack of complete information about the PSS-distribution in Study 145, the available information in Table 3 suggests an increase of 2 percentage points because the observed treatment difference in the BENCHMRK study is 32% in both subgroups ($\mathrm{PSS}=1,2$) that are relevant in Study 145. Note that these estimates of effect modifications account for the different factors separately and not simultaneously. Because the three covariates in our sensitivity analysis are closely related and presumably have common pathways, it is likely that some of these effect modifications will be attenuated in a model that includes all three covariates as well as RNA level. Together, these considerations lead to an additive sensitivity analysis with $\mathcal{A}=(-20,0)$. The results, which are shown in Fig. 3, indicate that Elvitegravir would be superior to placebo and retain at least half of the control effect in the current population, assuming that the aforementioned covariates are sufficient for conditional constancy. If the latter assumption is violated, another sensitivity analysis could be conducted to incorporate additional covariates.

6. Simulation studies

In this section, we report simulation results concerning the accuracy of the fully adjusted estimate of Δ_{01} derived in Section 3, under correct and incorrect assumptions, as well as the operating characteristics of the sensitivity analysis approach of Section 4 for making inference on Δ_{02}.

6.1. Data generation

In general, our simulation of trial data $(T,\mathbf{X},Y)$ starts with an initial covariate vector $\mathbf{W}$, generates T independently of $\mathbf{W}$, and then generates Y according to a model for $(Y|T,\mathbf{W})$. One element of $\mathbf{W}$ (chosen *a priori* or randomly) will then be designated as U, and the rest as $\mathbf{X}$. The same approach is used to generate historical trial data, under the conditions that $\dim(\mathbf{W}^*)=\dim(\mathbf{W})=:$ m and that U^* relates to $\mathbf{W}^*$ in the same way that U relates to $\mathbf{W}$. The dimension m will be specified in each experiment. For a given m, we generate $\mathbf{W}^*\sim N_m(\mathbf{0},\mathbf{I}_m)$ and $\mathbf{W}\sim N_m(\boldsymbol{\mu},\mathbf{I}_m)$, where N_m is the m-variate normal distribution, $\mathbf{I}_m$ the $m\times m$ identity matrix and $\boldsymbol{\mu}$ a mean vector to be specified later. The vector $\boldsymbol{\mu}$ describes the discrepancies in baseline characteristics between the current and historical study populations. Independently of baseline covariates, T^* is generated as a Bernoulli variable with probability 0.5, and T takes the values 1 and 2 with equal probabilities. Given $\mathbf{W}^*$ and T^*, Y^* is generated according to model (9), which in the present notation may be written as

$$E(Y^*|T^*=t,\mathbf{W}^*=\mathbf{w})=g(\beta_{t1}+\beta_{tW}'\mathbf{w}),\qquad t=0,1.$$

Given $\mathbf{W}$ and T, Y is generated according to a similar model:

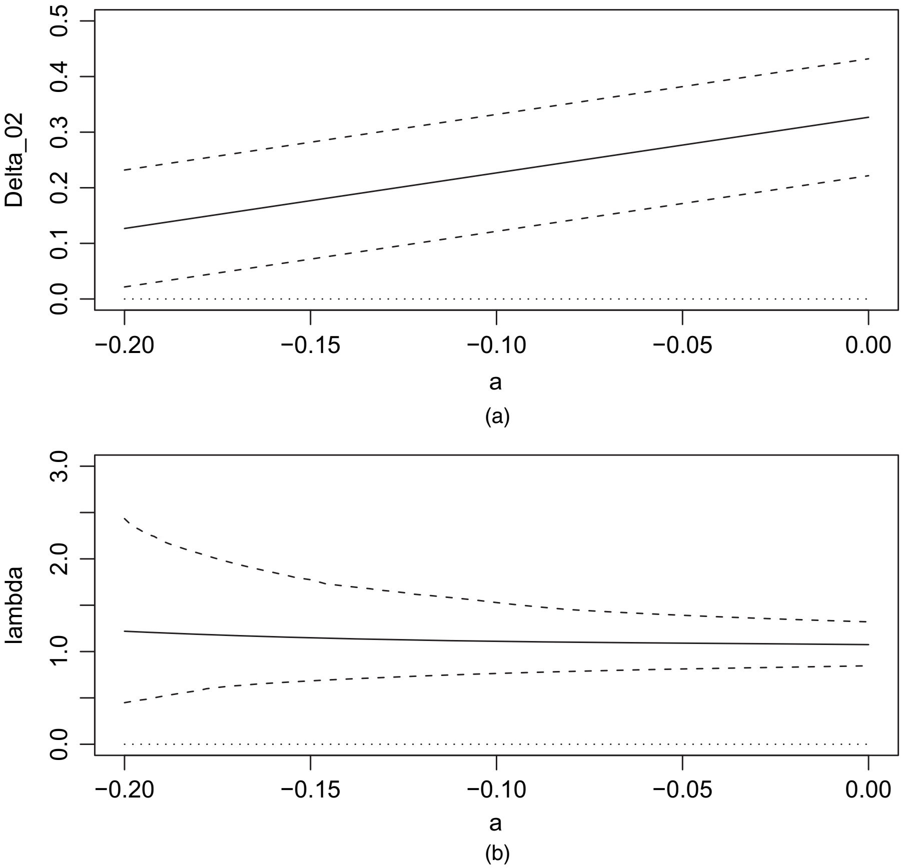

Fig. 3. Point estimates (———) and 95% confidence intervals (– – –) of (a) the effect of Elvitegravir relative to placebo (Δ_{02}) and (b) the fraction of the control (Raltegravir) effect retained by Elvitegravir (λ), both in the current patient population, obtained by using the proposed sensitivity analysis approach with an additive adjustment following standard covariate adjustment (see Section 5.2 for details)

$$E(Y|T=t,\mathbf{W}=\mathbf{w})=g(\beta_{11}+\boldsymbol{\beta}'_{1W}\mathbf{w}), \qquad t=1,2.$$

Note that the latter expression does not distinguish between $t=1$ and $t=2$, so treatments 1 and 2 have the same mean response in the current population. In our simulation studies, we consider two types of outcome: binary outcomes following logistic models (with $g\equiv\text{expit}$, defined in Section 3.4) and continuous outcomes following normal linear models (with an identity link). In the latter case, the error standard deviation is fixed at 0.5 to produce comparable variability with the binary case. In both cases, we fix $\beta_{01}=0$ and $\beta_{11}=1$, and consider the following scenarios concerning the values of $\boldsymbol{\mu}$, $\boldsymbol{\beta}_{0W}$ and $\boldsymbol{\beta}_{1W}$.

(a) $\boldsymbol{\mu}=(0,\ldots,m-1)'/(m-1)=:\mathbf{b}_m$ and $\boldsymbol{\beta}_{0W}=-\boldsymbol{\beta}_{1W}=\mathbf{s}_m/2$, where $\mathbf{s}_m$ is an m-vector of the form $(1,-1,1,-1,\ldots)$.

(b) $\boldsymbol{\mu}=\mathbf{1}$, $\boldsymbol{\beta}_{0W}=\mathbf{0}$ and $\boldsymbol{\beta}_{1W}=\mathbf{b}_m*\mathbf{s}_m$, where '$*$' denotes elementwise multiplication.

(c) $\boldsymbol{\mu}=\mathbf{1}$ and $\boldsymbol{\beta}_{0W}=-\boldsymbol{\beta}_{1W}=\mathbf{b}_m*\mathbf{s}_m/2$.

(d) $\mu_j\sim U(0,1)$, $\beta_{0Wj}\sim U(-0.5,0.5)$ and $\beta_{1Wj}\sim U(-0.5,0.5)$, independently of each other and across $j\in\{1,\ldots,m\}$, where the subscript j denotes the jth element of a vector and U denotes a uniform distribution.

(e) $\mu_j \sim N(0, 0.5^2)$, $\beta_{0Wj} \sim N(0, 0.25^2)$ and $\beta_{1Wj} \sim N(0, 0.25^2)$, independently of each other and across $j \in \{1, \ldots, m\}$.

The numerical values in these specifications are chosen to cover realistic situations in terms of population discrepancy and effect modification. For instance, it would be uncommon to have mean shifts that are greater than 1 standard deviation (i.e. $|\mu_j| > 1$) between the current and historical populations, which are only allowed in scenario (e) with a small probability. The scenarios with fixed parameter values ((a)–(c)) are designed to separate the consequences of population discrepancy and effect modification. In scenario (a), all covariates are equal in strength as effect modifiers, but they differ in the extent of population discrepancy. In scenarios (b) and (c), the extent of population discrepancy is fixed, but the covariates differ in their interactions with treatment (and also main effects, in scenario (b)). The scenarios with random parameter values ((d) and (e)) are designed to provide an overall evaluation in a range of situations comparable with the fixed value scenarios.

10000 data sets are simulated in each scenario for each given m (with the exception of Table 4, for which 1000 replications seem adequate). In this section, we focus on a common sample size of 500 for the non-inferiority trial and the historical trial, which represents a compromise between the two examples in Section 5 and a typical situation in practice. We have also experimented with different sample sizes (200 and 1000), which together cover a wide range of realistic situations, and the results are similar to those presented here (see the on-line appendix B).

Note that $\Delta_{01} = \Delta_{02}$ (because $\Delta_{12} = 0$) in all our simulation experiments. The true value of the effect is straightforward to calculate in the case of a continuous outcome. For a binary outcome, we approximate the true value of $\Delta_{01} = \Delta_{02}$ by using simulated data sets. Specifically, in a fixed value scenario ((a), (b) or (c)), we calculate

$$\frac{1}{n} \sum_{i=1}^{n} \{g(\beta_{11} + \boldsymbol{\beta}'_{1W}\mathbf{W}_i) - g(\beta_{01} + \boldsymbol{\beta}'_{0W}\mathbf{W}_i)\} \tag{20}$$

for each simulated data set and then take the average across the 1000 replicates. The resulting average is unbiased by construction, and its variability is negligible for all practical purposes. In a random-value scenario ((d) or (e)), each data set is associated with a different value of $\Delta_{01} = \Delta_{02}$, which we approximate by evaluating expression (20) on 10^4 hypothetical patients (in the current population), simulated by using the same parameter values that generate the 'observed' data. Larger numbers (such as 5×10^5) of hypothetical patients have been attempted without producing materially different results, and we therefore work with 10^4 hypothetical patients to reduce the computational burden.

6.2. *Point estimation*

Our evaluation of point estimates is focused on Δ_{01}, because estimation of Δ_{12} is straightforward and not subject to bias. Because theoretical results are readily available for a continuous outcome with the identity link, this simulation study is limited to a binary outcome with the logit link. Using the logit link allows us to assess also the quality of the approximation that is used in Section 3.4. To evaluate bias and variability in the usual sense, we restrict attention to the fixed value scenarios ((a)–(c)). In each scenario, we generate data with five covariates (i.e. $m = 5$) and apply the method of Section 3.4 (for a logit link) to a reduced data set (with one covariate omitted as U) under correct or incorrect assumptions (i.e. with values of γ, β_{0U} and β_{1U} for one of the original five covariates, which may or may not be the same as the covariate designated as U). All possible combinations (of true *versus* working choices for U) are included in our simulation study. For comparison, the simulation study also includes a naive estimate which is simply

Table 4. Point estimation of Δ_{01} with a binary outcome: comparison of several estimates (described in Section 6) in terms of sampling means and standard deviations (in parentheses)†

Adjust for	*Missing covariate*				
	W_1	W_2	W_3	W_4	W_5
Scenario (a)‡					
Proposed adjusted estimate (standard deviation, 0.05–0.07)					
W_1	0.09	0.05	0.19	−0.05	0.28
W_2	0.14	0.09	0.23	0.00	0.32
W_3	0.00	−0.05	0.09	−0.14	0.19
W_4	0.23	0.19	0.32	0.09	0.41
W_5	−0.09	−0.14	0.00	−0.23	0.10
Scenario (b)§					
Proposed adjusted estimate (standard deviation, 0.08–0.10)					
W_1	0.25	0.28	0.18	0.34	0.09
W_2	0.21	0.25	0.13	0.31	0.04
W_3	0.31	0.34	0.25	0.39	0.18
W_4	0.14	0.19	0.05	0.25	−0.06
W_5	0.35	0.38	0.30	0.42	0.25
Scenario (c)§§					
Proposed adjusted estimate (standard deviation, 0.08–0.10)					
W_1	0.10	0.05	0.21	−0.05	0.31
W_2	0.15	0.10	0.26	0.00	0.36
W_3	0.00	−0.05	0.11	−0.16	0.21
W_4	0.25	0.20	0.35	0.11	0.45
W_5	−0.10	−0.15	0.01	−0.25	0.11

†Each entry is based on 1000 replicates.

‡Gold standard, 0.10 (0.02); naive estimate, 0.19 (0.04); full data estimate, 0.09 (0.07).

§Gold standard, 0.25 (0.01); naive estimate, 0.18 (0.04); full data estimate, 0.25 (0.10).

§§Gold standard, 0.11 (0.01); naive estimate, 0.21 (0.04); full data estimate, 0.10 (0.11).

the observed treatment difference in the historical study, and a full data estimate that results from applying the standard covariate adjustment method of Section 2.2 to the full set of (five) covariates. Applying the method of Section 2.2 to the set of observed covariates (without any further adjustment) is equivalent to adjusting for W_1 (the first element of $\mathbf{W}$) as U by using the proposed method in the present context, where $U = W_1$ implies either $\gamma = \gamma^* = 0$ or $\beta_{0U} = \beta_{1U}$ and therefore is not shown as a separate method. The true value of Δ_{01} is approximated by averaging expression (20) across replicates, and the results are labelleded as ‘gold standard’ in Table 4.

Table 4 presents the sample mean and standard deviation of each method over 1000 replicates. Although Table 4 is based solely on $m = 5$, other values of m have been attempted and the results are qualitatively similar (and therefore have been omitted). Considering the observed variability and the number of replicates, the gold standard mean, as an approximation to the true value of Δ_{01}, is accurate to two decimal places with high confidence. As expected, the naive estimate is generally biased, and the full data estimate is nearly unbiased though more variable than the naive estimate. Under correct assumptions (i.e. adjusting for the ‘right’ covariate), the adjusted estimate proposed is also nearly unbiased, and its variability is similar to or slightly less than that of the full data estimate. It is not surprising that the estimate proposed can be more efficient than the full data estimate, because the former essentially compensates for unobserved data with true parameter values (under correct assumptions). Under incorrect assumptions, the estimate proposed can be seriously biased—in some cases more biased than the naive estimate. Thus, there is no guarantee that the proposed adjustment based on an arbitrary set of assumptions will necessarily outperform the naive estimate. This suggests that one should not rely on a single set of assumptions for inference unless there are good reasons for doing so. One may, however, explore a variety of assumptions in developing a range of plausible values for a sensitivity analysis.

6.3. Interval estimation

For inference, our proposal is to conduct a sensitivity analysis that summarizes results from multiple analyses based on different assumptions about U (see Section 4.1). We now evaluate this approach with Δ_{02} being the inferential target. This evaluation is focused on confidence intervals, which are closely related to hypothesis tests. Specifically, we consider the union confidence interval that is described in step 4 of Section 4.1, with an additive adjustment (a multiplicative adjustment would behave similarly, except for the complication that all effects involved must be positive). In practice, the range $\mathcal{A}$ for an additive adjustment should be chosen on the basis of all available information, including expert opinions as well as empirical evidence. However, in a simulation study it is convenient to use a data-driven mechanism for specifying $\mathcal{A}$, and we use the jackknife approach that was described in Section 4.2 in this particular study. In addition to the sensitivity analysis approach proposed, the study also includes a naive approach that assumes constancy and makes no use of covariate data, and a partially adjusted approach described in Section 2.2, which adjusts for observed covariates (and nothing more) under the conditional constancy assumption (3). This simulation study involves all five scenarios described in Section 6.1 and many different values of m.

Table 5 presents, for a continuous outcome with the identity link, simulation results (empirical coverage and average length) for the aforementioned three confidence intervals with a nominal level of 95%. Also shown in Table 5 are the dimension of $\mathbf{X}$, which is $m-1$ by design, and the true value of Δ_{02}, calculated by using equation (15) and knowledge of true distributions and parameter values. There is clearly an undercoverage problem for the naive method and, to a lesser extent, the partially adjusted method. The sensitivity analysis approach proposed has higher coverage than both of the other two methods, and its coverage probability increases with the dimension of $\mathbf{X}$. This should be expected because the jackknife version of $\mathcal{A}$ tends to become larger with more covariates available. In all five scenarios that are considered here, the sensitivity analysis approach has coverage probability close to the nominal level when $\dim(\mathbf{X}) \approx 8$. This approach suffers from an overcoverage problem when there are too many covariates, in which case one might want to shrink the jackknife version of $\mathcal{A}$ in some way. It might be possible to develop a formally justified procedure for shrinking $\mathcal{A}$ (or, rather, the union confidence interval) by treating the (observed and missing) covariates as a random sample from some universe of covariates, although the details of such an approach are not yet available. The higher coverage probability of the sensitivity analysis approach comes at a price: the resulting confidence intervals are generally longer than the partially adjusted intervals, with a typical ratio of 2–3 in average length.

Parallel results for a binary outcome with the logit link are shown in Table 6. The results in Table 6 are qualitatively similar to those in Table 5, with a few notable differences. First, a smaller number of covariates ($\dim(\mathbf{X}) \approx 5$) is required here for the sensitivity analysis approach to grow

Table 5. Interval estimation of Δ_{02} with a continuous outcome: empirical coverage and average length of (intended) 95% confidence intervals obtained by using a naive method that assumes constancy, a partially adjusted method that assumes conditional constancy (given observed covariates) and the proposed sensitivity analysis approach with a jackknife procedure specifying the range for additive adjustments (see Section 6 for details)†

dim($\mathbf{X}$) (*m* − 1)	Δ_{02}	*Empirical coverage for the following methods:*			*Average length for the following methods:*		
		Naive	*Partially adjusted*	*Sensitivity analysis*	*Naive*	*Partially adjusted*	*Sensitivity analysis*
Scenario (a)							
4	0.50	0.11	0.36	0.68	0.61	0.68	2.23
5	1.60	0.06	0.38	0.81	0.66	0.74	2.41
6	0.50	0.21	0.41	0.88	0.70	0.79	2.53
7	1.57	0.15	0.43	0.93	0.74	0.85	2.63
8	0.50	0.30	0.45	0.95	0.78	0.89	2.72
10	0.50	0.37	0.49	0.97	0.86	0.98	2.86
15	1.53	0.47	0.57	0.99	1.02	1.18	3.14
20	0.50	0.60	0.63	0.99	1.16	1.35	3.38
Scenario (b)							
4	1.50	0.14	0.40	0.75	0.64	0.76	2.31
5	0.40	0.07	0.42	0.85	0.68	0.83	2.49
6	1.50	0.23	0.45	0.91	0.73	0.88	2.63
7	0.43	0.17	0.48	0.95	0.77	0.94	2.74
8	1.50	0.33	0.51	0.96	0.81	1.00	2.84
10	1.50	0.40	0.56	0.98	0.88	1.10	3.02
15	0.47	0.48	0.66	0.99	1.04	1.32	3.41
20	1.50	0.62	0.73	0.99	1.18	1.52	3.74
Scenario (c)							
4	0.50	0.00	0.32	0.67	0.42	0.59	2.14
5	1.60	0.00	0.33	0.80	0.44	0.64	2.31
6	0.50	0.01	0.35	0.87	0.47	0.69	2.43
7	1.57	0.00	0.38	0.92	0.49	0.73	2.52
8	0.50	0.03	0.39	0.94	0.51	0.78	2.60
10	0.50	0.05	0.42	0.97	0.55	0.86	2.75
15	1.53	0.09	0.52	0.99	0.63	1.03	3.04
20	0.50	0.20	0.60	0.99	0.71	1.19	3.30
Scenario (d)							
4	1.00	0.30	0.66	0.85	0.40	0.45	0.93
5	1.00	0.28	0.68	0.89	0.43	0.48	1.03
6	1.00	0.28	0.70	0.91	0.45	0.51	1.12
7	1.00	0.28	0.71	0.93	0.47	0.54	1.20
8	1.00	0.27	0.72	0.94	0.49	0.56	1.27
10	1.00	0.26	0.75	0.95	0.53	0.62	1.40
15	1.00	0.26	0.79	0.97	0.62	0.73	1.64
20	1.00	0.25	0.82	0.99	0.70	0.83	1.84
Scenario (e)							
4	1.00	0.40	0.75	0.88	0.37	0.40	0.74
5	1.00	0.37	0.77	0.91	0.39	0.43	0.82
6	1.00	0.36	0.78	0.93	0.41	0.45	0.90
7	1.00	0.34	0.79	0.93	0.43	0.47	0.96
8	1.00	0.33	0.80	0.94	0.44	0.50	1.01
10	1.00	0.33	0.82	0.96	0.48	0.54	1.11
15	1.00	0.32	0.84	0.97	0.55	0.64	1.34
20	1.00	0.30	0.86	0.98	0.62	0.72	1.51

†Each entry is based on 10000 replicates.

Table 6. Interval estimation of Δ_{02} with a binary outcome: empirical coverage and average length of (intended) 95% confidence intervals obtained by using a naive method that assumes constancy, a partially adjusted method that assumes conditional constancy (given observed covariates) and the proposed sensitivity analysis approach with a jackknife procedure specifying the range for additive adjustments (see Section 6 for details)†

dim($\mathbf{X}$) ($m-1$)	Δ_{02}	*Empirical coverage for the following methods:*			*Average length for the following methods:*		
		Naive	*Partially adjusted*	*Sensitivity analysis*	*Naive*	*Partially adjusted*	*Sensitivity analysis*
Scenario (a)							
2	0.10	0.61	0.59	0.62	0.23	0.21	0.28
3	0.32	0.42	0.63	0.77	0.22	0.26	0.47
4	0.10	0.65	0.69	0.88	0.23	0.23	0.38
5	0.29	0.56	0.72	0.93	0.23	0.28	0.52
7	0.27	0.63	0.76	0.96	0.23	0.30	0.53
10	0.08	0.73	0.82	0.97	0.23	0.25	0.42
15	0.22	0.76	0.85	0.98	0.22	0.34	0.54
20	0.07	0.81	0.89	0.98	0.22	0.27	0.43
Scenario (b)							
2	0.27	0.72	0.76	0.75	0.22	0.28	0.41
3	0.06	0.51	0.80	0.88	0.24	0.26	0.37
4	0.25	0.74	0.85	0.93	0.22	0.34	0.51
5	0.07	0.63	0.86	0.95	0.24	0.29	0.43
7	0.07	0.68	0.89	0.97	0.23	0.32	0.46
10	0.22	0.80	0.90	0.97	0.22	0.42	0.59
15	0.06	0.79	0.92	0.98	0.23	0.38	0.53
20	0.18	0.84	0.91	0.97	0.22	0.48	0.65
Scenario (c)							
2	0.11	0.56	0.66	0.68	0.23	0.25	0.33
3	0.35	0.34	0.72	0.82	0.22	0.32	0.54
4	0.11	0.58	0.77	0.91	0.23	0.29	0.45
5	0.33	0.45	0.79	0.94	0.22	0.37	0.62
7	0.32	0.52	0.83	0.96	0.22	0.41	0.66
10	0.10	0.63	0.87	0.97	0.23	0.36	0.55
15	0.28	0.63	0.88	0.97	0.23	0.49	0.73
20	0.09	0.69	0.91	0.97	0.23	0.43	0.62
Scenario (d)							
2	0.22	0.73	0.88	0.89	0.23	0.25	0.29
3	0.21	0.69	0.87	0.91	0.23	0.26	0.32
4	0.21	0.65	0.87	0.92	0.23	0.26	0.34
5	0.21	0.62	0.86	0.93	0.23	0.27	0.35
7	0.20	0.59	0.86	0.91	0.23	0.28	0.37
10	0.20	0.53	0.84	0.88	0.23	0.29	0.40
15	0.18	0.49	0.81	0.86	0.23	0.31	0.42
20	0.17	0.45	0.78	0.82	0.23	0.32	0.44
Scenario (e)							
2	0.22	0.82	0.91	0.92	0.23	0.25	0.28
3	0.22	0.78	0.91	0.94	0.23	0.25	0.30
4	0.22	0.75	0.90	0.94	0.23	0.26	0.32
5	0.22	0.74	0.91	0.94	0.23	0.26	0.33
7	0.21	0.69	0.90	0.94	0.23	0.28	0.35
10	0.20	0.64	0.89	0.93	0.23	0.29	0.37
15	0.19	0.58	0.87	0.91	0.23	0.31	0.40
20	0.18	0.55	0.84	0.88	0.23	0.32	0.42

†Each entry is based on 10000 replicates.

close to the nominal confidence level. Second, the increase in average length for the sensitivity analysis approach *versus* the partially adjusted approach tends to be smaller here, in both absolute and relative terms, than in Table 5. Finally, in scenarios (d) and (e) in Section 6.1, the coverage probability appears to decline for all three methods when there are 'too many' covariates. This may be due to an increased chance of having 'difficult' parameter configurations that tend to produce erratic estimates (e.g. when there are few successes or failures in a treatment group).

7. Discussion

It is well recognized that possible violations of the constancy assumption present a serious challenge to the analysis and interpretation of non-inferiority trial data. Covariate adjustment, which accounts for inconstancy due to imbalances in observed covariates, represents a partial solution. This paper attempts to address the remaining part of the problem, namely residual inconstancy due to unmeasured covariates. We characterize residual inconstancy under a generalized linear model framework and derive fully adjusted estimates of the control effect in the current study based on plausible assumptions about an unmeasured covariate. It is admittedly difficult to specify and justify such assumptions, and we therefore propose a sensitivity analysis approach that covers a range of situations. The range for the sensitivity analysis may be based on clinical judgement about an unmeasured covariate as well as statistical analysis of observed data. The need for judgement is not a drawback of the approach proposed; rather, it reflects the nature of the problem and highlights the inherent limitations of active-controlled trials. Those limitations are best addressed by asking relevant scientific questions. In addition to clinical judgement, one could also use a jackknife procedure for gauging the effect of omitting a relevant covariate. Simulation results show that the jackknife-based sensitivity analysis has higher coverage probabilities than the partially adjusted approach and sometimes attains or approaches the nominal confidence level, although it may have an overcoverage problem. For a given application, further simulation experiments could be performed to assess the appropriateness of the approach.

Another possible approach to this problem would be a Bayesian approach, which has not been considered in this paper. Under a Bayesian approach, one could express the uncertainty about (conditional) constancy in the form of a prior distribution for some parameter that represents the extent of residual inconstancy. Such a prior distribution can provide some extra flexibility (relative to the range for a sensitivity analysis), although it may not be easy to specify. It may be worthwhile to explore such a Bayesian approach in the context of a suitable application.

The approach proposed is designed for interpretation of data with respect to a specific efficacy or effectiveness end point: not for decision making in a regulatory setting. The latter objective would require a different framework (decision theory rather than estimation) as well as additional information on safety, other efficacy end points and utility functions that quantify the consequences of different possible actions. Furthermore, it may be difficult to base a decision rule on a sensitivity analysis, which tends to be exploratory in nature. A Bayesian approach may once again prove helpful for this purpose.

It is for ease of presentation that we have focused on the simple case of a single historical study. The approach proposed extends easily to the case of multiple historical studies if assumption (8) continues to hold between the different historical studies and the same collection of covariates is measured in each study. If different collections of covariates are measured in different studies, as is often the case, this in principle could be addressed within the sensitivity analysis approach proposed, although the implementation will be more complicated. With multiple historical studies available, one may be able to estimate the variability between studies due to unobserved covariates without having to guess what it might be. An interesting possibility in that regard is the random-effects meta-analysis approach of Brittain *et al.* (2012).

This paper is focused on the mean difference as the effect measure of interest. This may be appropriate for binary data if the event rates are sufficiently high, as in the example of Section 5.2, and our discussion includes various link functions. For rare events, however, it would be more natural to consider the relative risk or the odds ratio as the effect measure. It will be of interest to extend the proposed approach to those effect measures with appropriate adjustment for possible non-collapsibility (Gail *et al.*, 1984; Greenland *et al.*, 1999).

Acknowledgements

We thank the Joint Editor, the Associate Editor and a referee for insightful comments that have improved the manuscript greatly. The views expressed in this paper do not necessarily represent the official position of the US Food and Drug Administration.

References

Brittain, E. H., Fay, M. P. and Follmann, D. A. (2012) A valid formulation of the analysis of noninferiority trials under random effects meta-analysis. *Biostatistics*, **13**, 637–649.

Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Kent, K. T. and Abbott, R. D. (1984) On error-in-variables for binary regression models. *Biometrika*, **71**, 19–25.

Cooper, D. A., Steigbigel, R. T., Gatell, J. M., Rockstroh, J. K., Katlama, C., Yeni, P., Lazzarin, A., Clotet, B., Kumar, P. N., Eron, J. E., Schechter, M., Markowitz, M., Loutfy, M. R., Lennox, J. L., Zhao, J., Chen, J., Ryan, D. M., Rhodes, R. R., Killar, J. A., Gilde, L. R., Strohmaier, K. M., Meibohm, A. R., Miller, M. D., Hazuda, D. J., Nessly, M. L., DiNubile, M. J., Isaacs, R. D., Teppler, H. and Nguyen, B. Y. for the BENCHMRK Study Teams (2008) Subgroup and resistant analyses of Raltegravir for resistant HIV-1 infection. *New Engl. J. Med.*, **359**, 355–365.

D'Agostino, R. B., Massaro, J. M. and Sullivan, L. M. (2003) Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Statist. Med.*, **22**, 169–186.

Durrleman, S. and Chaikin, P. (2003) The use of putative placebo in active control trials: two applications in a regulatory setting. *Statist. Med.*, **22**, 941–952.

Ellenberg, S. S. and Temple, R. (2000) Placebo-controlled trials and active-control trials in the evaluation of new treatments—Part 2: practical issues and specific cases. *Ann. Intern. Med.*, **133**, 464–470.

Fisher, L. D., Gent, M. and Büller, H. R. (2001) Active-control trials: how would a new agent compare with placebo?; a method illustrated with clopidogrel, aspirin, and placebo. *Am. Hrt J.*, **141**, 26–32.

Food and Drug Administration (1999) *Summary of CBER Considerations on Selected Aspects of Active Controlled Trial Design and Analysis for the Evaluation of Thrombolytics in Acute MI*. Washington DC: Department of Health and Human Services.

Food and Drug Administration (2010) *Guidance for Industry: Non-inferiority Clinical Trials*. Washington DC: Department of Health and Human Services.

Gail, M. H., Wieand, S. and Piantadosi, S. (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, **71**, 431–444.

Gao, P. and Ware, J. H. (2008) Assessing non-inferiority: a combination approach. *Statist. Med.*, **27**, 392–406.

Greenland, S., Robins, J. M. and Pearl, J. (1999) Confounding and collapsibility in causal inference. *Statist. Sci.*, **14**, 29–46.

Hasselblad, V. and Kong, D. F. (2001) Statistical methods for comparison to placebo in active-control trials. *Drug Inform. J.*, **35**, 435–449.

Hauck, W. W. and Anderson, S. (1999) Some issues in the design and analysis of equivalence trials. *Drug Inform. J.*, **33**, 109–118.

Holmgren, E. B. (1999) Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. *J. Biopharm. Statist.*, **9**, 651–659.

Huitfeldt, B. and Hummel, J. on behalf of European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) (2011) The draft FDA guidance on non-inferiority clinical trials: a critical review from European pharmaceutical industry statisticians. *Pharm. Statist.*, **10**, 414–419.

Hung, H. M. J., Wang, S. J., Tsong, Y., Lawrence, J. and O'Neil, R. T. (2003) Some fundamental issues with non-inferiority testing in active controlled trials. *Statist. Med.*, **22**, 213–225.

Johnson, N. L. and Kotz, S. (1970) *Distributions in Statistics*, vol. 2. Boston: Houghton-Mifflin.

Liang, K. Y. and Liu, X. H. (1991) Estimating equations in generalized linear models with measurement error. In *Estimating Functions* (ed. A. P. Godambe). Oxford: Clarendon.

Molina, J. M., Lamarca, A., Andrade-Villanueva, J., Clotet, B., Clumeck, N., Liu, Y. P., Zhong, L., Margot, N., Cheng, A. K. and Chuck, S. L. for the Study 145 Team (2012) Efficacy and safety of once daily elvitegravir versus twice daily raltegravir in treatment-experienced patients with HIV-1 receiving a ritonavir-boosted protease inhibitor: randomised, double-blind, phase 3, non-inferiority study. *Lancet Infect. Dis.*, **12**, 27–35.

Nie, L. and Soon, G. (2010) A covariate-adjustment regression model approach to noninferiority margin definition. *Statist. Med.*, **29**, 1107–1113.

Nie, L., Soon, G., Tauber, W. and Huque, M. (2010) An adaptive noninferiority margin and sample size adjustment in covariate-adjustment regression model approach to noninferiority clinical trials. *Mod. Assistd Statist. Appl.*, **5**, 169–177.

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

Rothman, M., Li, N., Chen, G., Chi, G. Y. H., Temple, R. and Tsou, H. H. (2003) Design and analysis of non-inferiority mortality trials in oncology. *Statist. Med.*, **22**, 239–264.

Rothman, K. J. and Michels, K. B. (1994) The continuing unethical use of placebo controls. *New Engl. J. Med.*, **331**, 394–398.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

Schumi, J. and Wittes, J. T. (2011) Through the looking glass: understanding non-inferiority. *Trials*, **12**, 106–117.

Snapinn, S. M. (2004) Alternatives for discounting in the analysis of noninferiority trials. *J. Biopharm. Statist.*, **14**, 263–273.

Steigbigel, R. T., Cooper, D. A., Kumar, P. N., Eron, J. E., Schechter, M., Markowitz, M., Loutfy, M. R., Lennox, J. L., Gatell, J. M., Rockstroh, J. K., Katlama, C., Yeni, P., Lazzarin, A., Clotet, B., Zhao, J., Chen, J., Ryan, D. M., Rhodes, R. R., Killar, J. A., Gilde, L. R., Strohmaier, K. M., Meibohm, A. R., Miller, M. D., Hazuda, D. J., Nessly, M. L., DiNubile, M. J., Isaacs, R. D., Nguyen, B. Y. and Teppler, H. for the BENCHMRK Study Teams (2008) Raltegravir with optimized background therapy for resistant HIV-1 infection. *New Engl. J. Med.*, **359**, 339–354.

Temple, R. and Ellenberg, S. S. (2000) Placebo-controlled trials and active-control trials in the evaluation of new treatments—Part 2: ethical and scientific issues. *Ann. Intern. Med.*, **133**, 455–463.

Wang, S. J. and Hung, H. M. J. (2003) TACT method for non-inferiority testing in active controlled trials. *Statist. Med.*, **22**, 227–238.

Wiens, B. L. (2002) Choosing an equivalence limit for noninferiority or equivalence studies. *Contr. Clin. Trials*, **23**, 2–14.

Witte, S., Schmidli, H., O'Hagan, A. and Racine, A. (2011) Designing a non-inferiority study in kidney transplantation: a case study. *Pharm. Statist.*, **10**, 427–432.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

Zhang, Z. (2009) Covariate-adjusted putative placebo analysis in active-controlled clinical trials. *Statist. Biopharm. Res.*, **1**, 279–290.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Web-based supplementary materials for "Sensitivity analysis in non-inferiority trials with residual inconstancy after covariate adjustment"'.