Randomization does not prevent bias when we assess treatment by covariate interaction *

Lei Nie[1],Zhiwei Zhang[2], Jialu Zang[3]
[1] Division of Biometrics IV, OB/OTS/CDER/FDA
Silver Spring, MD,20993-0002, USA.
[2] Division of Biostatistics, OSE/CDRH/FDA,
Silver Spring, MD,20993-0002, USA.
[3] Division of Biometrics I, OB/OTS/CDER/FDA
Silver Spring, MD,20993-0002, USA.

September 1, 2016

Abstract. Inference for the overall treatment effect resulted from a randomized clinical trial has the best credibility. In a traditional paradigm, where a clinical trial is primarily designed to answer a single question regarding the average (overall) treatment effect, randomization effectively prevents systematic confounding and bias when we assess the overall treatment effect. In an increasingly important new paradigm, inference of treatment heterogeneity is key to comparative effectiveness. In this new paradigm, randomization does not prevent systematic bias when we assess treatment by covariate interaction. Through a case study, this paper explains how this phenomenon occurs and examines whether and how this problem could be solved.

KEY WORDS: Bias; Treatment by covariate interaction;

* This article reflects the views of the author and should not be construed to represent FDA's views or policies.

1 Introduction

Inference resulted from a randomized clinical trial has the best credibility when the trial is primarily designed to answer a single question regarding the overall (average) treatment effect. An increasingly important new paradigm emerges in clinical trials in which the inference of treatment heterogeneity is key to comparative effectiveness. Please refer to Simon (2012a) and references therein for further discussion of the importance of new paradigm in the era of personalized medication. However, a comprehensive definition of this new paradigm has yet to come. As acknowledged in Simon (2012a), "there is uncertainty on what this new paradigm should be". In this paper, we are content to focus on the inference of treatment heterogeneity, a feature that will be included in the new paradigm.

This paper illustrates one major limitation of the randomization when we make inference regarding treatment by covariate interaction in the new paradigm. A limitation of randomization in estimating treatment effect was discussed in the traditional paradigm (see Gail et al. 1984, Neuhaus and Jewell, 1993); but it is not a major limitation. See Section 2 for detailed discussion.

A major limitation of randomization, in the new paradigm, stems from its lack of ability to prevent bias when we assess treatment by covariate interaction, which is the appropriate statistical method for assessing the heterogeneity(Pocock et al., 2002, Wang et al., 2007). The paper illustrates that the treatment by covariate interaction estimated in a model can be biased by interaction of treatment by unknown or unmeasured predictive variables even though treatments are randomized. This is a fundamental and intrinsic limitation of randomization: This limitation could be interpreted as inability to make correct inference for treatment heterogeneity. In addition, as this limitation also persists in linear models, it is more profound than the limitation described previously by Gail et al. (1984) and Neuhaus and Jewell (1993). However, we note that, our discussion of this limitation should not be considered as a criticism against randomization, as the limitation remains in nonrandomized trials. In fact, because there are more layers of sources of bias or confounding, problems are usually more serious in the nonrandomized trials.

This paper is organized as follows. In Section 2, we present a recent registrational trial of Brilinta, indicated for treatment of acute coronary syndromes, to describe the background and motivation of this research. The questions raised from this example were assessed in Sections 3 and 4 through a parametric and a nonparametric approach. In the parametric approach, we focus on the generalized linear model and assess the treatment effect expressed in odds ratio; in the nonparametric approach, we focus on assessing treatment effect expressed in rate difference. The results provide clear answers to the questions raised in clinical trial of Brilinta, which highlight that randomization does not prevent bias when we assess treatment by covariate interaction in Section 5. In Section 6, we discuss some limited strategies to solve the problem through trial designs and through statistical analyses.

2 Motivation and examples

In a series of works, Gail et al. (1984), Neuhaus and Jewell (1993), among others, some questions were raised about inferences of the treatment effect when some prognostic variables were omitted.

Consider, for example, a generalized linear model with density function:

$$\exp[\{y\theta - b(\theta)\}/a(\phi) + c(u, \phi)]$$

for some known functions $a(\cdot), b(\cdot), c(\cdot)$, and a scalar ϕ. The mean is

$$E(Y|T, X) = h(\alpha + \beta T + \gamma X).$$

Here Y is a response variable, X is a vector of a set of covariates, and T is a treatment indicator variable that takes values 1 or -1 with probabilities p and 1 -p respectively. The treatment assignment is independent of the levels of the covariates. If the variable X is ignored from the model, one would fit the following model

$$E(Y|T, X) = h(\alpha^* + \beta^* T).$$

Following Gail et al. (1984), we have

$$\beta^* = \frac{1}{2}\left(h^{-1}\left[E_X\{h(\alpha + \beta + \gamma X)\}\right] - h^{-1}\left[E_X\{h(\alpha - \beta + \gamma X)\}\right]\right).$$

The difference between $\beta^* - \beta$ was considered as bias by Gail et al. (1984), Neuhaus and Jewell (1993), while it was considered as a difference between two different metrics of the treatment effect by others, such as Greenland et al. (1999).

Below, we explain why the difference observed between β^* and β is not a major problem for randomized clinical trials.

First, the conclusion about the efficacy of a new treatment (T=1) vs. the control (T=-1) based on β and β^* is consistent when the sample size is large enough and when $h(.)$ is a monotonic function. It is easy to find that $\beta^* = 0$ when $\beta = 0$. Furthermore, it is also easy to show that β^* has the same sign as β. That is, $\beta > 0 (< 0)$ implies $\beta^* > 0 (< 0)$. To see this, without of any loss of generality, we assume $\beta > 0$ and $h(.)$ is an increasing function. Under these assumptions, $h(\alpha + \beta + \gamma X) - h(\alpha - \beta + \gamma X) > 0$. Thus, $E_X\{h(\alpha + \beta + \gamma X)\} > E_X\{h(\alpha - \beta + \gamma X)\}$. As h^{-1} is also an increasing function, thus $h^{-1}\left[E_X\{h(\alpha + \beta + \gamma X)\}\right] > h^{-1}\left[E_X\{h(\alpha - \beta + \gamma X)\}\right]$.
The last inequality implies that $\beta^* > 0$. Readers please also refer to Gail et al. (1984), Neuhaus and Jewell (1993), for a more thorough discussion. The results highlight that, although β and β^* can be different, they both make the same conclusion about the treatment effect in a randomized clinical trial when the sample size is large. Second, the difference between β and β^* is due to the nonlinearity of metrics, therefore it may not be a fundamental difference in our opinion. To see this, let $h(t) = t$. It is easy to see that $\beta^* = \beta$. That is, the difference only occurs in some nonlinear models due to nonlinearity, similar to difference between relative risk and odds ratio.

This traditional paradigm focuses more on the average treatment effect and less on potential heterogeneity in treatment effects, which does not necessarily provide a scientifically sound basis because treatment heterogeneity is often present (Simon, 2012 a,b). It is known that treatment heterogeneity has been well documented in many disease areas. In oncology, as noted in Simon (2004) ,“a large body of evidence indicates that cancers of most primary sites are heterogeneous with regard to molecular pathogenesis, genomic signatures, and phenotypic properties. Consequently, it is not necessarily reasonable to expect such tumors to have equal sensitivities to a drug that inhibits a particular protein target. The protein target may be driving tumor growth in only a subset of the tumors". Treatment heterogeneities are also present in many other areas, including respiratory disease (such as The IMpact-RSV Study Group, 1998), Cardiovascular disease (such as Sacks et al., 1996, Wallentin et al. 2009), and HIV (such as Cooper et al., 2008).

Unsurprisingly, assessing treatment heterogeneity has become a common practice for many trialists and is often recommended, see such as Assmann et al. (2000), Pocock et al. (2002), Lagakos (2006), Wang et al. (2007), and references therein. In this new paradigm, we usually need to answer questions such as whether the treatment effect was different across different levels of a factor, as exemplified below. When the goal is to make inference for treatment heterogeneity in the new paradigm, we encounter a bigger problem in the randomized trials than the problem previously discussed. Different from the setup in Gail et al. (1984) and Neuhaus and Jewell (1993), one of the goals in the new paradigm is to make inference about the heterogeneity. The problem we discuss here is possible false inference of treatment by covariate interaction as illustrated below in the case study. Randomization does not necessarily prevent it from happening.

The clinical evidence for the effectiveness of Brilinta is derived from PLATO trial, a randomized double-blind trial comparing Brilinta to Clopidogrel. Both treatments are given in combination with Aspirin and other standard therapies, in patients with acute coronary syndromes (ACS), see Wallentin et al. (2009). Patients were treated for at least 6 months and up to 12 months. The study's primary endpoint was the composite of first occurrence of cardiovascular death, non-fatal Myocardial Infarction (MI) (excluding silent MI). The benefit of Brilinta to Clopidogrel was statistically significant overall. However, in the United States (US), Brilinta was numerically inferior to Clopidogrel. The test for treatment by region (US vs. non-US) interaction is statistically significant, and the same trend is seen in both CV death and non-fatal MI. As treatment by region (US versus non-US) interaction receives special attention in recent years, experience from large international trials suggests that regional difference may not occur simply due to chance, Lawrence et al. (2012). What does the significant interaction in PLATO trial tell? Suppose a patient travels between Europe and US frequently, shall we recommend that the patient take Brilinta in Europe and take Clopidogrel in US? We doubt that anyone would consider the recommendation sensible. But how do we interpret the observed heterogeneity?

We know that randomization prevent systematic confounding and bias in the traditional paradigm when we make inference about the overall treatment effect. Note that, randomization is typically applied to treatment allocation but not to other variables. Consequently, bias could occur when we make inference on treatment by covariate interaction. In this paper, we use examples and theoretical results to show how the bias occurs. Before doing that, we provide more details of the PLATO trial.

The PLATO protocol left the choice of Aspirin maintenance dose up to investigators. The pattern of Aspirin maintenance dose usage is very different in the US compared to elsewhere: about 10% of patients outside US received Aspirin doses above 100 mg, while more than 50% of patients in US received doses above 100 mg.. Consequently, the usage of Aspirin maintenance dose is highly imbalanced between US and other countries.

Table 1 provides details of the usage of Aspirin maintenance dose. In the US and non-US groups, the difference of event rates between Brilinta and Clopidogrel are 1.4%, respectively and -2.0%.The treatment by region interaction in terms of event rate is 3.4% with a 95% confidence interval of (0.04%, 6.7%). The odds ratio of Brilinta to Clopidogrel in US and Non-US patients are 1.2 and 0.78, respectively. The treatment by region interaction in odds ratio is significant with a p-value of 0.04.

However, when we look into the analysis stratified by Aspirin dose, the regional heterogeneity disappears. Let us consider subjects who had high Aspirin dose. The difference of event rate of Brilinta over Clopidogrel in the US and non-US groups are 4.7% and 3.6%. The treatment by region interaction in terms of event rate is 1.1% with a 95% confidence interval of (-9.0%, 11.2%). Now consider subjects who had low Aspirin dose. The difference of event rate of Brilinta over Clopidogrel in the US and non-US groups are -2.4% and -2.1%.The treatment by region interaction in terms of event rate is -0.4% with a 95% confidence interval of (-5.0%, 4.2%). Neither the interaction measured by difference in the event rate nor the interaction measured by odds ratio are significant after taking Aspirin into account.

The analysis seems to support that Aspirin dose is the covariate that cause the treatment by region interaction. But it remains possible that there are other factors and Aspirin dose is a surrogate of these factors. The fact is that we are not able to answer the question even in a randomized controlled trial.

3 Bias in the treatment by covariate interaction in generalized linear model

In this section we assume that the response variable Y follows a generalized linear model, McCullagh and Nelder (1983), with density function:

$$\exp[\{y\theta - b(\theta)\}/a(\phi) + c(u, \phi)]$$

for some known functions $a(\cdot), b(\cdot), c(\cdot)$, and a scalar ϕ. The mean of Y is $\mu = h(\eta)$

$$\eta = \alpha + T\beta + X\beta_x + Z\beta_z + TX\gamma + TZ\gamma_{tz} + XZ\gamma_{xz} + TXZ\gamma_{txz} + C'\xi. \tag{1}$$

Here $h^{-1}(\cdot)$ is the link function and ϕ is the dispersion parameter. X is a set of covariates that are independent from T. Z is a different set of covariates: some components are dependent on X and therefore could interact with treatment; some interact with treatment but do not necessarily depend on X. C is a vector of covariates that do not interact with treatment and are independent of T and X. The variance, $b''(\theta)a(\phi)$, can be defined as a function of the mean

$$var(Y) = V_\phi(\mu).$$

The above generalized linear model (1) examines the treatment effect and treatment-by-covariate interaction with covariate adjustment. When $\gamma = 0$ (i.e. there is no treatment by covariate interaction), this model reduces to the model considered by Gail et al (1984), among others. This model can be more accurately defined as a conditional model for the given covariates

$$\eta|X, Z, C = \alpha + T\beta + X\beta_x + Z\beta_z + TX\gamma + TZ\gamma_{tz} + XZ\gamma_{xz} + TXZ\gamma_{txz} + C'\xi. \tag{2}$$

Typically, when assessing the overall treatment effect, the following alternative model may also be used to assess the overall treatment effect,

$$\mu = h(\eta^*), \quad \eta^* = \alpha^* + T\beta^*. \tag{3}$$

This model is often called the covariate-unadjusted generalized linear model for treatment effect. The resulting marginal treatment effect β may be restricted to the trial population.

Let

$$\zeta_1 = E(Y|T=1) = E_{Z,C,X}\Big\{h\Big(\alpha + \beta + X\beta_x + Z\beta_z + X\gamma + Z\gamma_{tz} + XZ\gamma_{xz} + XZ\gamma_{txz} + C'\xi\Big)\Big\},$$
$$\zeta_0 = E(Y|T=-1) = E_{Z,C,X}\Big\{h\Big(\alpha - \beta + X\beta_x + Z\beta_z - X\gamma - Z\gamma_{tz} + XZ\gamma_{xz} - XZ\gamma_{txz} + C'\xi\Big)\Big\}.$$

With some regularity assumptions similar to what have been assumed in Gail et al. (1984), the marginal (crude) treatment effect is obtained from model (3) as

$$\beta^* = \frac{1}{2}\{h^{-1}(\zeta_1) - h^{-1}(\zeta_0)\}.$$

When $h(.)$ is the identical link, such as in linear regression,

$$\beta^* = \beta + E(X)\gamma + E(Z)\gamma_{tz} + E_{X,Z}(XZ)\gamma_{txz}.$$

If we further assume that there is no three way interaction and centralize X and Z so that $E(X) = 0$ and $E(Z) = 0$, then $\beta^* = \beta$.

Model (3) allows us to estimate the treatment effect. However, without an interaction term in the model, it does not allow an estiamtion of the treatment by X interaction. Similar to model (3), the following model could be used to assess the overall treatment effect and treatment by X interaction.

$$\mu = h(\eta^*), \quad \eta^* = \alpha^* + T\beta^* + X\beta_x^* + TX\gamma^*. \tag{4}$$

The notations α^* and β^* used here are similar to notations used in model (3). We do not (or marginal) treatment by X interaction estimated from the covariate-adjusted model (1) with γ^*, obtained from the covariate-unadjusted model (4).

Following Gail et al.(1984), we let

$$\begin{aligned}
\zeta_{11} = E(Y|T=1,X=1) &= E_{Z,C|X}\Big\{h\Big(\alpha+\beta+\beta_x+Z\beta_z+\gamma+Z\gamma_{tz}+Z\gamma_{xz}+Z\gamma_{txz}+C'\xi\Big)\Big\},\\
\zeta_{10} = E(Y|T=1,X=-1) &= E_{Z,C|X}\Big\{h\Big(\alpha+\beta-\beta_x+Z\beta_z-\gamma+Z\gamma_{tz}-Z\gamma_{xz}-Z\gamma_{txz}+C'\xi\Big)\Big\},\\
\zeta_{01} = E(Y|T=-1,X=1) &= E_{Z,C|X}\Big\{h\Big(\alpha-\beta+\beta_x+Z\beta_z-\gamma-Z\gamma_{tz}+Z\gamma_{xz}-Z\gamma_{txz}+C'\xi\Big)\Big\},\\
\zeta_{00} = E(Y|T=-1,X=-1) &= E_{Z,C|X}\Big\{h\Big(\alpha-\beta-\beta_x+Z\beta_z+\gamma-Z\gamma_{tz}-Z\gamma_{xz}+Z\gamma_{txz}+C'\xi\Big)\Big\}.
\end{aligned}$$

Here $E_{Z,C}$ is the expectation with respect to Z and C. Let $\kappa(\eta)=\frac{\partial\mu}{\partial\eta}\cdot\frac{1}{V_\phi(\mu)}$ Similar to Gail et al. (1984), the maximum likelihood equations divided by the sample size converge to

$$\begin{aligned}
&E\big[\kappa(\eta)\cdot\big\{Y-h(\eta)\big\}\big]=0,\\
&E\big[\kappa(\eta)\cdot T\cdot\big\{Y-h(\eta)\big\}\big]=0,\\
&E\big[\kappa(\eta)\cdot X'\cdot T\cdot\big\{Y-h(\eta)\big\}\big]=0,\\
&E\big[\kappa(\eta)\cdot X'\cdot\big\{Y-h(\eta)\big\}\big]=0,
\end{aligned}$$

Throughout this paper, we use $\dot{h}(\cdot)$ and $\ddot{h}(\cdot)$ to denote the first and second derivatives, respectively, of a function $h(\cdot)$. The relationship between the parameters in model (4) and the parameters in model (1) is given in the following result.

Under the following assumptions:
(a) T is independent of all covariate variables;
(b) $E\big[\kappa(\eta)\times\big\{Y-h(\eta)\big\}\big]$, $E\big[\kappa(\eta)\times T\times\big\{Y-h(\eta)\big\}\big]$, $E\big[\kappa(\eta)\times X'\times T\cdot\big\{Y-h(\eta)\big\}\big]$, $E\big[\kappa(\eta)\times X'\times\big\{Y-h(\eta)\big\}\big]$, ζ_{11}, ζ_{10}, ζ_{01}, and ζ_{00} exist;
(c) $h(\cdot)$ has a unique inverse $h^{-1}(\cdot)$ which is well defined at ζ_1 and ζ_2; $\dot{h}$ and $\ddot{h}$ exist;
(d) $h^{-1}(\cdot)$ is nonsingular at $h\big(\zeta_{kl}\big)$, $k,l=0,1$;
(e) $\kappa(\zeta_{kl})$, $k,l=0,1$ do not vanish;
we have the following result:

$$\gamma^*=\frac{1}{4}\{h^{-1}(\zeta_{11})-h^{-1}(\zeta_{10})-h^{-1}(\zeta_{01})+h^{-1}(\zeta_{00})\}$$

Under the same conditions described above, if $h(\eta)=\eta$ (e.g. the linear model and the Poisson model with identity link),

$$\gamma^* \;=\; \gamma+\frac{E(Z|X=1)-E(Z|X=-1)}{2}\gamma_{tz}+\frac{E(Z|X=1)+E(Z|X=-1)}{2}\gamma_{txz}.$$

Note that, the difference between γ and γ^* has multiple sources: 1) the bias noted in Gail et al. (1984); 2) the effect of γ_{tz}; 3) the interaction γ_{txz}. If $h(\eta)=\eta$, $\gamma^*=\gamma$ when $\gamma_{tz}=0$ and $\gamma_{txz}=0$. However, if $h(\eta)\neq\eta$, γ^* is generally different from γ. In this circumstance, when $\gamma_{tz}=0$ and $\gamma_{txz}=0$, the difference between γ^* and γ is the same as the bias characterized in Gail et al. (1984).

4 Bias in the nonparametric treatment by covariate interaction in additive scale

In this section, we generally do not make any parametric assumption about the random variable Y. However, we do have a similar interest in estimating the overall treatment effect and treatment by X interaction in additive scale, i.e.

$$\gamma^* \;=\; \frac{E_{T=1,X=1}(Y) - E_{T=1,X=-1}(Y) - E_{T=-1,X=1}(Y) + E_{T=-1,X=-1}(Y)}{4}.$$

This happens in e.g. a HIV trial, where estimating treatment by subgroup interaction in terms of success rate is one of the most important goals because physicians need to choose appropriate treatments from more than 25 approved treatments for their patients.
Let $E_{C|Z}E_{Y|T,X,Z,C}(Y) = \mu(T,X,Z)$ and $\nu(t,x) = E_{Z|X=x}\mu(t,x,Z)$. As C is independent of X, T, we have

$$\begin{aligned}
E_{Y|T,X}(Y) &= E_{Z|X,T}E_{C|X,T,Z}E_{Y|T,X,Z,C}(Y)\\
&= E_{Z|X}E_{C|Z}E_{Y|T,X,Z,C}(Y)\\
&= E_{Z|X}\{\mu(T,X,Z)\}\\
&= \nu(T,X).
\end{aligned}$$

Thus

$$\begin{aligned}
4\gamma^* &= E_{T=1,X=1}(Y) - E_{T=1,X=-1}(Y) - E_{T=-1,X=1}(Y) + E_{T=-1,X=-1}(Y)\\
&= \nu(1,1) - \nu(1,-1) - \nu(-1,1) + \nu(-1,-1).
\end{aligned}$$

Here γ^* measures the overall (marginal) effect modification by X. Considering a prognostic/predictive variable Z and other variables C, we also take the average treatment by X interaction into account:

$$\begin{aligned}
4\gamma &= E_{Z,C}\{E_{T=1,X=1,Z,C}(Y) - E_{T=1,X=-1,Z,C}(Y) - E_{T=-1,X=1,Z,C}(Y) + E_{T=-1,X=-1,Z,C}(Y)\}\\
&= E_Z\{\mu(1,1,Z) - \mu(-1,1,Z) - \mu(1,-1,Z) + \mu(-1,-1,Z)\}.
\end{aligned}$$

Here γ is the averaged conditional interaction effect.

Using the relationship

$$E_{Z,C}E_{T,X,Z,C}(Y) \;=\; E_Z\{E_{C|Z}E_{T,X,Z,C}(Y)\} = E_Z\mu(T,X,Z),$$

we obtain

$$\begin{aligned}
& 4(\gamma^* - \gamma)\\
= \;& E_{Z|X=1}\{\mu(1,1,Z)\} - E_{Z|X=1}\{\mu(-1,1,Z)\} - E_{Z|X=-1}\{\mu(1,-1,Z)\} + E_{Z|X=-1}\{\mu(-1,-1,Z)\}\\
- \;& E_Z\{\mu(1,1,Z) - \mu(-1,1,Z) - \mu(1,-1,Z) + \mu(-1,-1,Z)\}.
\end{aligned}$$

When Z is a binary variable with values 1 and -1, we let $r = P(X=1)$, $q_{1|x} = P(Z=1|X=x)$

and $q = P(Z = 1)$ and obtain

$$\begin{aligned}
& 4(\gamma^* - \gamma) \\
= \; & \mu(1,1,1)q_{1|1} + \mu(1,1,-1)(1 - q_{1|1}) - \mu(-1,1,1)q_{1|1} - \mu(-1,1,-1)(1 - q_{1|1}) \\
- \; & \mu(1,-1,1)q_{1|-1} - \mu(1,-1,-1)(1 - q_{1|-1}) + \mu(-1,-1,1)q_{1|-1} + \mu(-1,-1,-1)(1 - q_{1|-1}) \\
- \; & \mu(1,1,1)q - \mu(1,1,-1)(1 - q) + \mu(-1,1,1)q + \mu(-1,1,-1)(1 - q) \\
+ \; & \mu(1,-1,1)q + \mu(1,-1,-1)(1 - q) - \mu(-1,-1,1)q - \mu(-1,-1,-1)(1 - q) \\
= \; & \mu(1,1,1)(q_{1|1} - q_{1|-1})(1 - r) - \mu(1,1,-1)(q_{1|1} - q_{1|-1})(1 - r) \\
- \; & \mu(-1,1,1)(q_{1|1} - q_{1|-1})(1 - r) + \mu(-1,1,-1)(q_{1|1} - q_{1|-1})(1 - r) \\
+ \; & \mu(1,-1,1)(q_{1|1} - q_{1|-1})r - \mu(1,-1,-1)(q_{1|1} - q_{1|-1})r \\
- \; & \mu(-1,-1,1)(q_{1|1} - q_{1|-1})r + \mu(-1,1,-1)(q_{1|1} - q_{1|-1})r \\
= \; & \sum_{T,Z=1,-1} \{\mu(T,1,Z\} \times T \times Z\}(q_{1|1} - q_{1|-1})(1 - r) \\
+ \; & \sum_{T,Z=1,-1} \{\mu(T,-1,Z\} \times T \times Z\}(q_{1|1} - q_{1|-1})r.
\end{aligned}$$

As $a_1b_1 + a_2b_2 = \{ra_1 + (1 - r)a_2\}\{b_1/r + b_2/(1 - r)\}/2 + \{ra_1 - (1 - r)a_2\}\{b_1/r - b_2/(1 - r)\}/2$, we have,

$$\begin{aligned}
& 4(\gamma - \gamma^*) \\
= \; & \sum_{T,Z=1,-1} [\{r\mu(T,1,Z) + (1 - r)\mu(T,-1,Z)\} \times T \times Z]\frac{c_0 + c_1}{2} \\
+ \; & \sum_{T,Z=1,-1} [\{r\mu(T,1,Z) - (1 - r)\mu(T,-1,Z)\} \times T \times Z]\frac{c_0 - c_1}{2},
\end{aligned}$$

where $c_0 = (q_{1|1} - q_{1|-1})(1 - r)/r$ and $c_1 = (q_{1|1} - q_{1|-1})r(1 - r)$. Note that,

$$\sum_{T,Z=1,-1} \{r\mu(T,1,Z) + (1 - r)\mu(T,-1,Z)\} \times T \times Z$$

can be viewed as a weighted interaction between treatment T and covariate Z. On the other hand, the term

$$\sum_{T,Z=1,-1} \{r\mu(T,1,Z) - (1 - r)\mu(T,-1,Z)\} \times T \times Z$$

can be viewed as a weighted three way interaction among treatment T, X, and covariate Z. Thus, the differences between γ and γ^* in the second approach has two sources: 1) the effect of a related treatment by covariate interaction; 2) the three way interaction.

5 Understanding the problem through the Brilinta example

In PLATO, Brilinta is superior to Clopidogrel in overall population. However, heterogeneous treatment effects were observed across different regions. The PLATO protocol left the choice of Aspirin maintenance dose up to the investigator and patterns were very different in US and elsewhere.

Below we compare estimations of treatment by region interaction obtained from models that include Aspirin maintenance dose and models that does not. The comparison illustrates the potential problems in making inference of treatment by region interaction.

We first fit a logistic regression model to p, the rate of first occurrence of cardiovascular death and non-fatal MI. The model includes treatment indicator T, region indicator X, the indicator of high ($\geq$ 300 mg) Aspirin maintenance dose Z, and all interaction terms. That is,

$$\text{logit}(p) = \alpha + T\beta + X\beta_x + Z\beta_z + TX\gamma + TZ\gamma_{tz} + XZ\gamma_{xz} + TXZ\gamma_{txz}. \tag{5}$$

The statistical test for the treatment by region interaction is not statistically significant (p=0.67). The point estimate of the treatment by region interaction is $\hat{\gamma} = -0.03$ with a standard error of 0.07. See the Appendix for details.

We now theoretically calculate the treatment by region interaction through the following model,

$$\text{logit}(p) = \alpha^* + T\beta^* + X\beta_x^* + TX\gamma^*, \tag{6}$$

in which Aspirin maintenance dose is omitted.

Using the results stated in previous section and replacing the parameters in model (5) by their point estimations, we derive the treatment by region interaction of $\gamma^* = -0.108$ without conducting actual data analysis using model (6).

Now, let's compare this prediction with results obtained from actual data analysis. The statistical test for treatment by region interaction is statistically significant (p=0.04). The point estimation of the treatment by region interaction is $\hat{\gamma}^* = -0.106$ with a standard error of 0.05. That is, the theoretically predicted treatment by region interaction of -0.108 is very similar to the actual observed treatment by region interaction of -0.106.

Using the theoretical results we can further explore the relationship between γ^* and γ. In the Appendix, we described a scenario in which there is no treatment by covariate X interaction. However, because the presence of treatment by covariate Z interaction, we may mistakenly conclude the presence of treatment by covariate X interaction through model (6). That is $\gamma^* \neq 0$ even $\gamma = 0$. Please refer to Appendix to details. This is different from the result we discussed in Section 2, in which $\beta = 0$ implies $\beta^* = 0$.

Next, we also discuss the treatment effect measured in risk difference through a nonparametric approach, VanderWeele, Mukherjeec, and Chen (2012). Let Y be the indicator of first occurrence of cardiovascular death and non-fatal MI and $\mu(T, X, Z) = E(Y|T, X, Z)$. The overall proportions of subjects taking high dose and low dose of Aspirin are 5.8% and 94.2%, respectively. The treatment by region interaction in event rate difference is 1.1% in the high dose of Aspirin subjects and -0.4% in the low dose Aspirin subjects. So the averaged conditional treatment by region interaction in the trial population is

$$4\gamma \;=\; E_Z\{\mu(1,1,Z) - \mu(-1,1,Z) - \mu(1,-1,Z) + \mu(-1,-1,Z)\} = -0.0031$$

A 95% C.I. for this estimate is (-4.7%, 4.1%), suggesting a non significant treatment by region interaction. Note that, the average is obtained for trial population.

If we ignore the Aspirin maintainable dose, we would obtain the following estimate of treatment by region interaction,

$$4\gamma^* \;=\; E_{T=1,X=1}(Y) - E_{T=1,X=-1}(Y) - E_{T=-1,X=1}(Y) + E_{T=-1,X=-1}(Y) = 0.034$$

with a 95% C.I. of (0.04%, 6.7%), suggesting a significant treatment by region interaction. The difference between $4\gamma^*$ and 4γ is 0.037.

The difference between 4γ and $4\gamma^*$ is shown to be

$$\begin{aligned} &\sum_{T,Z=1,-1} [\{r\mu(T,1,Z) + (1-r)\mu(T,-1,Z)\} \times T \times Z]\frac{c_0 + c_1}{2} \\ +\; &\sum_{T,Z=1,-1} [\{r\mu(T,1,Z) - (1-r)\mu(T,-1,Z)\} \times T \times Z]\frac{c_0 - c_1}{2}, \end{aligned} \tag{7}$$

Please refer to the Appendix for the derivation and definitions of r, c_0, and c_1. Note that, the first term can be viewed as weighted treatment by variable Z interaction and the second term can be viewed as the weighted three way interaction among treatment T, X, and covariate Z. Thus, the differences between γ and γ^* in the second approach has two sources: 1) effect of a related treatment by covariate interaction; 2) three way interaction.

Using the theoretical results in materials, the differences between γ and γ^*, two typically used metrics of interaction, can be clearly explained. The results also highlight that conclusion based on γ^* could be misleading because of the bias, even in randomized trials.

6 Closing remarks

In this paper, we demonstrated that inference for treatment by subgroup (covariate) interaction has major limitations even in randomized clinical trials: Some factors, perhaps related to unknown/unmeasured predictive variables, can be systematically biased by the treatment by covariate interaction. Unlike inference for the overall treatment effect interaction, precisely in the setup of this work, the bias factors may lead to red herring finding.

If the factors that led to the bias are known measured covariates, many existing methods could be used to make valid inferences. If, however, the factors are unknown or unmeasured, the inference for the treatment heterogeneity may end up with a misleading conclusion. For example, in the Brilinta example, if the Aspirin dose was not measured, one may conclude the treatment by region interaction. Furthermore, even though the Aspirin dose measurement explains the regional heterogeneity from a statistical point of view, we still cannot conclude that Aspirin changed (modified) the treatment effect as there is no persuasive biological explanation so far, and we cannot rule out the possibility there might be some unmeasured factors that bias the results.

In the Brilinta example, there is one way to confirm whether treatment by Aspirin interaction is a cause of treatment by region interaction. An additional confirmatory trial can be conducted through a factorial design, in which units are randomized by levels of treatment (factor A) and Aspirin dose (factor B). Through this design, randomization can prevent bias when we assess treatment by Aspirin dose interaction. In such design, we learned that both c_0 and c_1 in equation (7) are 0. Therefore

$$4(\gamma^* - \gamma) = 0$$

Unfortunately, such factorial design is not always feasible. For example, we would like to study the interaction of treatment and baseline disease status (such as baseline HIV viral loads $>$ or $\leq$ 100,0000 copies/ml) in a HIV trial. The baseline disease status is highly correlated with many other baseline characteristics such as baseline CD4 counts and HIV symptotics; therefore randomization on treatment and baseline viral load is not feasible to separate factors such as baseline CD4 counts and HIV symptotics. In the Brilinta example, treatment and Aspirin dose can be considered as two add-on treatments and the interaction can be considered as treatment by treatment interaction, Senn (2004). However, in the example of the treatment of HIV, the interaction of treatment by baseline viral load is interaction of treatment by block, thus a factorial design is not possible.

REFERENCES

Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, **355**, 1064-1069.

Cooper, D. A., Steigbigel, R. T., Gatell, J. M., et al. (2008). Subgroup and resistance analyses of raltegravir for resistant HIV-1 infection. *N Engl J Med.* **359**, 355-365.

Gail, M., Wieand, S., Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted covariates. *Biometrika*, **71**, 431-44.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, *14*, 29-46.

Greenland, S. (2009). Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology*, *20*, 14-17.

IMpact-RSV Study Group (1998). Palivizumab, a humanized respiratory syncytial virus monoclonal antibody, reduces hospitalization from respiratory syncytial virus infection in high-risk infants. *Pediatrics*, **102**, 531-537.

Lagakos, S. W. (2006). The challenge of subgroup analyses–reporting without distorting. *N Engl J Med*, **354**, 1667-1669.

Lawrence, J., Bai, S., Hung, J., O'Neil, R. (2012). Regional Treatment Effects in Studies of Cardiorenal Drugs: A Summary of Recent Clinical Trials. *Journal of the American College of Cardiology*, **63**, 1117-1118.

Peto, R., Pike, M. C., Armitage, P., Cox, D.R., Howard,S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer*, **34**, 585-612.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med*, **21**, 2917-2930.

Sack, A.M., Pfffer, M.A., et al (1996). The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels, *N Engl J Med*, **335**, 1001-1009.

Simon, R. (2004) An agenda for Clinical Trials: clinical trials in the genomic era, *Clinical trials*, **1**, 468-470.

Simon, R. (2012a). How to Develop Treatments for Biologically Heterogeneous "Diseases". *Clinical Cancer Research*, **18**, 4001-4003.

Simon, R. (2012b). Clinical trials for predictive medicine. *Statistics in Medicine*, **31**, 3031-3040.

Neuhaus, J. M., Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, **80**, 807-815.

Rimsky, L., Vingerhoets, J., Van Eygen, V., et al. Genotypic and phenotypic characterization of HIV-1 isolates obtained from patients on rilpivirine therapy experiencing virologic failure in the phase 3 ECHO and THRIVE studies: 48-week analysis (2012). *J Acquir Immune Defic Syndr*, **59**, 39-46.

Senn, S. (2004). Added values - Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine*, **23**, 3729-3753.

VanderWeele, T. J. (2009). On the distinction between interaction and effect modification. *Epidemiology*, **20**, 863-871.

VanderWeele, T. J., Mukherjeec, B. and Chen, J. (2012). Sensitivity analysis for interactions under unmeasured confounding. *Statistics in Medicine*, **31**, 25522564.

Wallentin L, Becker RC, Budaj A, et al. (2009) Ticagrelor versus Clopidogrel in patients with acute coronary syndromes. *N Engl J Med*;**361**, 1045-57.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007). Statistics in medicine–reporting of subgroup analyses in clinical trials. *N Engl J Med*, **357**, 2189-2194.

Table 1: Events rate by region and by Aspirin dose (# of events/sample size)

US	Aspirin dose	Brilinta	Clopidogrel	diff (95% C.I.)
	$\geq$ 300 mg	12.4% (40/324)	7.7% (27/352)	4.7% (0.1%, 9.2%)
	$\leq$ 100 mg	6.7% (19/284)	9.1% (24/263)	-2.4% (-7.0%, 2.1%)
	overall	9.7%	8.3%	1.4% (-1.8%, 4.6%)
non US	Aspirin dose	Brilinta	Clopidogrel	diff (95% C.I.)
	$\geq$ 300 mg	20.0% (28/140)	16.4% (23/140)	3.6% (-5.5%, 12.6%)
	$\leq$ 100 mg	7.3% (546/7449)	9.4% (699/7443)	-2.1% (-3.0%, 1.2%)
	overall	7.6%	9.5%	-2.0% (-2.9%, 1.1%)