Harvard University
Harvard University Biostatistics Working Paper Series

Year 2016 *Paper* 208

Moving beyond the conventional stratified analysis to estimate an overall treatment efficacy with the data from a comparative randomized clinical study

Lu Tian* Fei Jiang† Takahiro Hasegawa‡
Hajime Uno** Marc Alan Pfeffer†† L.J. Wei‡‡

*Stanford University, lutian@stanford.edu
†The University of Hong Kong, feijiang@hku.hk
‡Shionogi & Co., Ltd, duo@mars.dti.ne.jp
**Dana-Farber Cancer Institute, huno@jimmy.harvard.edu
††Brigham and Women's Hospital, mpfeffer@rics.bwh.harvard.edu
‡‡Harvard University, wei@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper208

Moving beyond the conventional stratified analysis to estimate an overall treatment efficacy with the data from a comparative randomized clinical study

Lu Tian*

Department of Biomedical Data Science,

Stanford University,

Stanford, California 94305, U.S.A.

Fei Jiang

Department of Statistics & Actuarial Science,

The University of Hong Kong,

Hong Kong

Takahiro Hasegawa

Shionogi & Co., Ltd, Osaka, Japan

Hajime Uno

Division of Biostatistics and Computational Biology,

Dana-Farber Cancer Institute,

Boston, Massachusetts 02215, U.S.A.

Marc Alan Pfeffer

Department of Medicine,

Brigham and Women's Hospital,

Boston, Massachusetts 02115, U.S.A.

L.J. Wei

Department of Biostatistics,

Harvard University,

Boston, Massachusetts 02115, U.S.A.

*lutian@stanford.edu

September 22, 2016

Abstract

For a two-group comparative study, a stratified inference procedure is routinely used to estimate an overall group contrast to increase the precision of the simple two-sample estimator. Unfortunately most commonly used methods including the Cochran-Mantel-Haenszel statistic for a binary

outcome and the stratified Cox procedure for the event time endpoint do not serve this purpose well. In fact, these procedures may be worse than their two-sample counterparts even when the observed treatment allocations are imbalanced across strata. Various procedures beyond the conventional stratified methods have been proposed to increase the precision of estimation when the naive estimator is consistent. In this paper, we are interested in the case when the treatment allocation proportions vary markedly across strata. We study the stochastic properties of the two-sample naive estimator conditional on the ancillary statistics: the observed treatment allocation proportions and/or the stratum sizes, and present a biased-adjusted estimator. This adjusted estimator is asymptotically equivalent to the augmentation estimators proposed under the unconditional setting (Zhang *and others*, 2008). Moreover, this consistent estimation procedure is also equivalent to a rather simple procedure, which estimates the mean response of each treatment group first via a stratum-size weighted average and then constructs the group contrast estimate. This simple procedure is flexible and readily applicable to any target patient population by choosing appropriate weights for each treatment effect estimate. All the proposals are illustrated with the data from a cardiovascular clinical trial, whose treatment allocations are imbalanced. Ancillary statistic; Augmentation estimation procedure; Conditional inference; CMH statistic; Mixture population

1 Introduction

In a comparative, randomized study, suppose that we are interested in estimating an overall group difference, θ, between a treatment and a control via their individual population parameters, τ_2 and τ_1, respectively. For example, τ is the mean

value of the study subject's outcome variable and $\theta = \tau_2 - \tau_1$. Assume that the study utilizes a $M : 1$ random treatment allocation rule for assigning patients to the treatment and control groups. Although the primary analysis of the study is generally based on a two-sample empirical counterpart $\hat{\theta}$ of θ, a stratified inference procedure is often utilized to increase the estimation precision for $\hat{\theta}$ (Valliant, 1993; Miratrix *and others*, 2013). Moreover, when the observed allocation proportions of patients assigned to either arm vary substantially across strata, a stratified estimator $\hat{\theta}_S$, has been generally perceived as less biased than $\hat{\theta}$.

Unfortunately a number of routinely used stratified procedures cannot be guaranteed to increase the precision nor reduce the bias for estimating θ, especially when θ is non-linearly related to τ_1 and τ_2. As an example, consider a simple case where the outcome is a binary variable and θ is the odds ratio (OR) of the event rates τ's. That is,

$$\theta = \frac{\tau_2/(1-\tau_2)}{\tau_1/(1-\tau_1)}.$$

The two-sample empirical counterpart is

$$\hat{\theta} = \frac{\hat{\tau}_2/(1-\hat{\tau}_2)}{\hat{\tau}_1/(1-\hat{\tau}_1)},$$

where $\hat{\tau}_j$ is the observed empirical counterpart of $\tau_j, j = 1, 2$. For a large randomized study, $\hat{\theta}$ is consistent for θ. Now, suppose that one considers a stratified inference procedure with K strata to estimate θ. Let the observed stratum sizes be $n_1, n_2, \cdots, n_K$ and $n = \sum_{k=1}^{K} n_k$. For the jth group in the kth stratum, let the group size be n_{jk}, the true event rate be $\tau_{jk}, \hat{\pi}_k = n_{2k}/n_k$ and $\hat{w}_k = n_k/n, j = 1, 2; k = 1, \cdots, K$. A commonly used stratified estimate for θ is based on the Cochran-Mantel-Haenszel (CMH) statistics (Mantel and Haenszel, 1959), which results in

$$\hat{\theta}_S = \frac{\sum_{k=1}^{K} \hat{\tau}_{2k}(1-\hat{\tau}_{1k})\hat{\pi}_k(1-\hat{\pi}_k)\hat{w}_k}{\sum_{k=1}^{K}(1-\hat{\tau}_{2k})\hat{\tau}_{1k}\hat{\pi}_k(1-\hat{\pi}_k)\hat{w}_k},$$

where $\hat{\tau}_{jk}$ is the observed empirical counterpart of τ_{jk}. Note that even when $n_k \to \infty$ and $\hat{\pi}_k \to M/(M+1), k = 1, \cdots, K$, $\hat{\theta}_S$ converges to

$$\theta_S = \frac{\sum_{k=1}^{K} \tau_{2k}(1-\tau_{1k})w_k}{\sum_{k=1}^{K}(1-\tau_{2k})\tau_{1k}w_k},$$

where w_k, the limit of $\hat{w}_k$, is the mixing proportion of the kth stratum in the study population. In general, $\theta_S \neq \theta$, and therefore, $\hat{\theta}_S$ is an inconsistent estimator. Moreover, θ_S is not a simple weighted average of the stratum-specific OR's as usually interpreted by the practitioners. Here, the kth weight is proportional to $(1-\tau_{2k})\tau_{1k}w_k$, a rather complex form of the underlying stratum-specific event rates.

In survival analysis, a stratified inference procedure, which is routinely used for comparing two groups, is to estimate the hazard ratio (HR) θ under a two-sample proportional hazards model (Cox, 1972). Here, the outcome variable is the time to a specific event. The corresponding stratified method is the stratified Cox procedure (Cox, 1972; Mehrotra *and others*, 2012), which suffers from the same limitation of the CMH method for the binary outcome. More details will be given in Section 4 of this paper.

Alternatives to aforementioned conventional stratified procedures have been studied extensively when the simple estimator $\hat{\theta}$ is consistent. For this case, the goal is mainly to increase the estimation precision of $\hat{\theta}$ with the baseline covariate information (Koch *and others*, 1998; Zhang *and others*, 2008; Moore and van der Laan, 2009; Rosenblum and van der Laan, 2010; Tian *and others*, 2012). The performance of such alternative estimation procedures are assessed under an "unconditional" setting by considering all possible realizations of the estimator generated under the random treatment allocation rule utilized in the study. Now, suppose that (i) $\hat{\pi}_k'$s, the observed proportions of study patients assigned to the treatment group,

vary substantially across strata, and/or (ii) $\hat{w}_k'$s, the observed proportions of the patients in each stratum, are substantially different from the underlying w_k's. Then generally, for the case with a practical sample size, the observed naive estimate $\hat{\theta}$ would not be close to θ. It is not clear, however, how to empirically quantify the bias of $\hat{\theta}$. One possible approach to handle this problem would be based on a conditional inference principle via ancillary statistics for the treatment difference θ (Kalbfleisch, 1975). For the present case, both the empirical proportions of patients assigned to the treatment $\{\hat{\pi}_k, k = 1, \cdots, K\}$ and the empirical proportions of the strata $\{\hat{w}_k, k = 1, \cdots, K\}$ are ancillary statistics. The distribution of $(\hat{\theta} - \theta)$ conditional on such ancillary statistics would be more "relevant" and informative than its unconditional counterpart to study the stochastic behavior of $\hat{\theta}$ (Senn, 1989; Pocock *and others*, 2002). Specifically, we consider all realizations of $\hat{\theta}$ generated from the random allocation rule utilized in the study, but $\{(\hat{\pi}_k, \hat{w}_k), k = 1, \cdots, K\}$ of each realized sample would be the same as its observed counterparts. By doing so, the individual realizations in the above conditional sample space are obtained under the experimental condition similar to the observed one. In this paper, using this procedure, one can empirically quantify the bias of $\hat{\theta}$ and then obtain a consistent estimator for θ by modifying $\hat{\theta}$. It is interesting to note that the above modified estimator is identical or asymptotically equivalent to the augmentation estimators proposed by Zhang *and others* (2008); Tsiatis *and others* (2008) and Tian *and others* (2012) under the unconditional setting. Moreover, the above modified estimator is also equivalent to a rather simple estimator obtained via a mixture estimation procedure across strata. Specifically, for the above example for θ being the OR, we first estimate the overall event rate for the control arm using a weighted average of stratum-specific event rate es-

timates, where the weight for the kth stratum-specific estimate is w_k, the target proportion of patients from the kth stratum. Similarly, we estimate the overall event rate for the treated arm. Then, we construct the OR estimate with these two overall event rate estimates. The details are given in Sections 2 and 3. In Section 4, we apply the proposal to the case with the censored event time observations. All the procedures are illustrated with the data from a comparative, randomized cardiovascular clinical trial.

2 Modifying the naive estimator $\hat{\theta}$ in the presence of treatment allocation imbalance

Using the notations in the Introduction for the general case, that is, let τ_j and $\hat{\tau}_j$ be the population mean and its empirical counterpart of the subject's outcome variable for the jth group, respectively, and let τ_{jk} and $\hat{\tau}_{jk}$ be the corresponding quantities in the kth stratum, $k = 1, \cdots, K$; $j = 1, 2$. Let $\theta = g(\tau_1, \tau_2)$ and $\hat{\theta} = g(\hat{\tau}_1, \hat{\tau}_2)$, where $g(\cdot, \cdot)$ is a given smooth function. With a random assignment allocation rule and the random sampling assumption of subjects in each group and stratum, for large $n_k, k = 1, \cdots, K$, it is straightforward to show that the joint distribution of $(\hat{\theta} - \theta)$ and $(\hat{w}_1 - w_1, \cdots, \hat{w}_K - w_K, \hat{\pi}_1 - \bar{\pi}, \cdots, \hat{\pi}_K - \bar{\pi})'$ is approximately normal with mean 0 and covariance matrix $\hat{\Sigma}$ given in the Appendix A, where $\bar{\pi} = \sum_{k=1}^{K} \hat{w}_k \hat{\pi}_k$. Heuristically, it follows that the conditional distribution of

$$(\hat{\theta} - \theta) \Big| \{(\hat{w}_k - w_k, \hat{\pi}_k - \bar{\pi}), k = 1, \cdots, K\}$$

is approximately normal with a mean of

$$\hat{b}_{\pi w} = \dot{g}_1(\tau_1^\dagger, \tau_2^\dagger) \sum_{k=1}^{K} \hat{\tau}_{1k} \left(\frac{\hat{w}_k(1-\hat{\pi}_k)}{1-\bar{\pi}} - w_k \right) + \dot{g}_2(\tau_1^\dagger, \tau_2^\dagger) \sum_{k=1}^{K} \hat{\tau}_{2k} \left(\frac{\hat{w}_k \hat{\pi}_k}{\bar{\pi}} - w_k \right)$$

and a variance of

$$\hat{\sigma}^2_{\pi w} = \sum_{k=1}^{K} \left\{ \dot{g}_2(\tau_1^\dagger, \tau_2^\dagger)^2 \hat{w}_k^2 \hat{\sigma}^2_{2k} + \dot{g}_1(\tau_1^\dagger, \tau_2^\dagger)^2 \hat{w}_k^2 \hat{\sigma}^2_{1k} \right\},$$

where

$$\tau_j^\dagger = \sum_{k=1}^{K} w_k \hat{\tau}_{jk}, \, j = 1, 2,$$

$\dot{g}_j(\tau_1, \tau_2)$ is the partial derivative of $g(\tau_1, \tau_2)$ with respect to τ_j and $\hat{\sigma}^2_{jk}$ is a consistent estimator for the variance of $\hat{\tau}_{jk}$, $j = 1, 2$. Note that the above large sample normal approximation is not straightforward. The theoretical justification of this approximation to the conditional distribution is given in Appendix A. The above conditional distribution shows that $\hat{\theta}$ is in general inconsistent for θ when there are treatment allocation imbalances within individual strata or $\hat{w}_k'$s are different from w_k's. An obvious consistent estimator for θ by directly adjusting $\hat{\theta}$ is

$$\hat{\theta}_{\pi w} = \hat{\theta} - \hat{b}_{\pi w}.$$

Note that $\hat{\theta}_{\pi w}$ is a sum of $\hat{\theta}$ and a linear combination of $(\hat{w}_k - w_k)$ and $(\hat{\pi}_k - \bar{\pi}), k = 1, \cdots, K$ and is asymptotically equivalent to the augmentation estimators proposed and discussed in Zhang *and others* (2008); Moore and van der Laan (2009) and Tian *and others* (2012). Note also that the augmentation procedures in the literature do not have $\hat{w}_k - w_k, k = 1, \cdots, K$, as one of the augmented terms. On the other hand, from the semi-parametric efficiency argument, it is a trivial extension to do so, when $w_k, k = 1, \cdots, K$ are known. As a result, unconditionally, $\hat{\theta}_{\pi w}$ minimizes the asymptotical variance of the sum of $\hat{\theta}$ and any linear combination

of $\{(\hat{w}_k - w_k), (\hat{\pi}_k - \bar{\pi}), k = 1, \cdots, K\}$. However, all the augmented estimators were developed to improve efficiency over the naive estimator $\hat{\theta}$ under an unconditional setting for which the naive estimator $\hat{\theta}$ is consistent.

When w_k for the study population is unknown, one may consider an adjusted estimator $\hat{\theta}_\pi$ based on the conditional distribution of $\hat{\theta} - \theta$ given $\{\hat{\pi}_k - \bar{\pi}, k = 1, \cdots, K\}$ only. Specifically, $\hat{\theta}_\pi = \hat{\theta} - \hat{b}_\pi$, where

$$\hat{b}_\pi = \dot{g}_1(\tau^\dagger_{1\pi}, \tau^\dagger_{2\pi}) \sum_{k=1}^{K} \hat{w}_k \hat{\tau}_{1k} \left(\frac{1 - \hat{\pi}_k}{1 - \bar{\pi}} - 1 \right) + \dot{g}_2(\tau^\dagger_{1\pi}, \tau^\dagger_{2\pi}) \sum_{k=1}^{K} \hat{w}_k \hat{\tau}_{2k} \left(\frac{\hat{\pi}_k}{\bar{\pi}} - 1 \right).$$

The variance of $\hat{\theta}_\pi$ can be estimated consistently by

$$\hat{\sigma}^2_\pi = \hat{\sigma}^2_{\pi w} + n^{-1} \sum_{k=1}^{K} \hat{w}_k \left\{ \dot{g}_2(\tau^\dagger_{1\pi}, \tau^\dagger_{2\pi})(\hat{\tau}_{2k} - \hat{\tau}_2) + \dot{g}_1(\tau^\dagger_{1\pi}, \tau^\dagger_{2\pi})(\hat{\tau}_{1k} - \hat{\tau}_1) \right\}^2,$$

where

$$\tau^\dagger_{j\pi} = \sum_{k=1}^{K} \hat{w}_k \hat{\tau}_{jk}.$$

It is interesting to note $\hat{\sigma}^2_\pi$ is greater than $\hat{\sigma}^2_{\pi w}$ due to the the sampling variation from $\{\hat{w}_k, k = 1, \cdots, K\}$. This adjusted estimator would not be unbiased under the conditional setting with the additional conditioning event, $\{(\hat{w}_k - w_k), k = 1, \cdots, K\}$. It is important to note that there is no general rule on the choice of ancillary statistics. This topic has been discussed extensively in the literature. In practice, the choice of the ancillary statistics would be made on a case by case basis. For our present case, the choice ancillary statistics, $\{\hat{\pi}_k\}$ and $\{\hat{w}_k\}$ was similar to those of the inference for analyzing multiple 2×2 tables (Fraser, 2004). The choice of the ancillary statistic $\{\hat{\pi}_k\}$ only was also suggested by Senn (1989) and Pocock *and others* (2002).

As an example, consider the data from a cardiovascular trial "Valsartan in acute myocardial infarction (VALIANT) study" (Pfeffer *and others*, 2003) to il-

lustrate the above estimation procedures. There were three arms in this study, the patients in the first group were treated by ARB valsartan, the second group was with ACE inhibitor captopril and the third one was a combination of these two drugs. For illustration, we consider the time to the first hospitalization or death as the endpoint and compare the monotherapy (combing two treatment groups) with the combo therapy. The study enrolled a total of 14,703 patients were equally assigned to three arms. The median follow-up time was 24.7 months after randomization. For the entire study, there is no difference with respect to the endpoint considered here. In Fig. 1, we show the Kaplan-Meier curves for this endpoint for two comparators. On the other hand, with the data from 302 patients in Australia, the monotherapy somehow appears to be statistically significantly better than its combo counterpart (see Fig. 2). Note that Australia was the only country among 24 countries participated in the VALIANT study, whose patients tend to have better outcomes for the monotherapy than those for the combo therapy.

We will discuss the case with an event time as the endpoint in Section 4. Here, let us consider the outcome from the above study to be a binary, either the patient had event by or at month 18. Note that there were no censored observations in the study before month 18. There are two important patient's covariates, which are related to this binary endpoint: BMI and history of diabetes. For simplicity, we dichotomize BMI with a cutoff value of 25 and create four strata (i.e, 1: low BMI and no diabetic history; 2: low BMI and with diabetic history; 3: high BMI and no diabetic history; 4: high BMI and with diabetic history). Table 1 shows the number of 302 Australian patients assigned to each treatment group with respect to the four strata, whose sizes $(\hat{w}_1, \hat{w}_2, \hat{w}_3, \hat{w}_4) = (0.24, 0.06, 0.54, 0.16)$. Note that if the randomization scheme worked well, the treatment allocation ratio (mono

vs. combo) would be around 2:1. From Table 1, there is a non-trivial treatment allocation imbalance with respect to these two factors.

The naive estimate $\hat{\theta}$ for the OR of two event rates (combo vs. mono) at month 18 is 1.99 with a 0.95 confidence interval of (1.12, 3.51) in favor of monotherapy. The corresponding CMH estimate is 1.83 with a 0.95 confidence interval of (1.03, 3.25). Using the empirical strata weights $\{\hat{w}_i\}$, the bias adjusted estimate $\hat{\theta}_\pi$ is 1.73 with a 0.95 confidence interval of (0.96, 3.12). The confidence interval includes one, suggesting that the group difference is not statistically significant based on the bias adjusted estimator. If we assume that the true mixing proportions for the Australia substudy are identical to the observed proportions in the entire VALIANT study, i.e., $(w_1, w_2, w_3, w_4) = (0.24, 0.04, 0.53, 0.19)$, then $\hat{\theta}_{\pi w}$ for OR becomes 1.75 with a slightly different 0.95 confidence interval of (0.97, 3.13), also indicating an insignificant treatment effect.

3 A simple consistent stratified estimator for θ with two group-specific weighted averages of the stratum mean outcomes

Under a stratification setting with K strata, the mean of the outcome τ_j can be rewritten as $\sum_{k=1}^{K} w_k \tau_{jk}$, and can be estimated consistently with $\tau_j^\dagger = \sum_{k=1}^{K} w_k \hat{\tau}_{jk}$, $j = 1, 2$. It follows that

$$\hat{\theta}_{new} = g(\tau_1^\dagger, \tau_2^\dagger)$$

is a consistent estimator unconditionally or conditionally on $\{\hat{\pi}_k, k = 1, \cdots, K\}$ and/or $\{\hat{w}_k, k = 1, \cdots, K\}$. It is interesting to note that this new estimator is asymp-

totically equivalent to the bias-adjusted estimator $\hat{\theta}_{\pi w}$. This equivalence is due to the fact that

$$\tau_1^\dagger - \hat{\tau}_1 = -\sum_{k=1}^{K} \hat{\tau}_{1k}\left(\frac{\hat{w}_k(1-\hat{\pi}_k)}{1-\bar{\pi}} - w_k\right),$$

$$\tau_2^\dagger - \hat{\tau}_2 = -\sum_{k=1}^{K} \hat{\tau}_{2k}\left(\frac{\hat{w}_k\hat{\pi}_k}{\bar{\pi}} - w_k\right),$$

and

$$g(\hat{\tau}_1, \hat{\tau}_2) - g(\tau_1^\dagger, \tau_2^\dagger) \approx \dot{g}_1(\tau_1^\dagger, \tau_2^\dagger)(\hat{\tau}_1 - \tau_1^\dagger) + \dot{g}_2(\tau_1^\dagger, \tau_2^\dagger)(\hat{\tau}_2 - \tau_2^\dagger) = \hat{b}_{\pi w}.$$

Note also that when $\{w_k, k = 1, \cdots, K\}$ is not known, $\hat{\theta}_{new}$ can be replaced by

$$\hat{\theta}_{new\pi} = g(\tau_{1\pi}^\dagger, \tau_{2\pi}^\dagger),$$

with $\tau_{j\pi}^\dagger = \sum_{k=1}^{K} \hat{w}_k\hat{\tau}_{jk}$, $j = 1, 2$. For this case, the resulting $\hat{\theta}_{new\pi}$ is asymptotically equivalent to $\hat{\theta}_\pi$ discussed in Section 2.

Considering the estimator $\hat{\theta}_{new\pi}$ with $\hat{w}_k$ as the weights, first $\tau_{j\pi}^\dagger$ estimates the mean response of the subjects in arm j by a weighted average with the observed proportion of the stratum size as the weight. One then constructs a group contrast measure with these two resulting treatment group-specific estimators. Note that this estimation procedure for the between-group difference is constructed in a rather different way from the conventional stratified counterparts. For the conventional stratification methods, we first estimate the stratum specific group contrasts and then empirically combine them across strata, whose weights may not have clinical or physical interpretation. Moreover, the conventional stratified methods for the case with a non-linear $g(\tau_1, \tau_2)$ may not increase the estimation efficiency and introduce nontrivial bias as discussed in the previous section. Since $\hat{\theta}_{new}$ or $\hat{\theta}_{new\pi}$ is asymptotically equivalent to its augmentation estimation procedure, this estimator is always more efficient than $\hat{\theta}$ asymptotically.

As an estimator for the marginal mean response, $\tau_j^\dagger$ not only serves as the building block for adjusting the bias of the naive estimator but also provides important reference level in interpreting the group contrast measure θ. For example, although both $(\tau_1^\dagger, \tau_2^\dagger) = (0.050, 0.095)$ and $(0.500, 0.667)$ yield the same OR of 2.0, they may have quite different implications in practice. The conventional stratified estimator and the estimators $\hat{\theta}_{\pi w}$ and its corresponding equivalent augmented counterparts do not have the benchmark value from the control arm for clinical decision makings.

In the VALIANT example, the estimated event rate is $\tau_{1\pi}^\dagger = 0.67$ with a 0.95 confidence interval of (0.61, 0.74) for monotherapy and $\tau_{2\pi}^\dagger = 0.77$ with a 0.95 confidence interval interval of (0.72, 0.83) for combo therapy based on the observed stratum sizes. Note that if we are interested in a difference group contrast measure, for instance, the event rate difference as θ, these values are readily available from this simple new procedure to make inferences.

4 APPLICATION TO THE CASE WITH THE EVENT TIME OUTCOME VARIABLE

The most commonly used stratification estimation procedure is based on the stratified Cox model (Mehrotra *and others*, 2012). Here, the parameter of interest is the overall HR θ by assuming that the two hazard functions are proportional of each other over the entire study time. As discussed extensively in the statistical and medical literature, the HR estimate is difficult to interpret especially when the PH assumption is violated (Uno *and others*, 2014, 2015). The stratified Cox procedure follows the same approach as other conventional stratified methods for

analyzing non-censored outcomes. That is, for each stratum, we assume that the PH assumption is plausible. We then obtain the HR estimate for each stratum, and combine those estimates. The resulting estimator is asymptotically equivalent to linearly combining the log-transformed stratum-specific HR estimators over K strata. The weight of the combination is proportional to the inverse of the variance estimate for the log-transformed stratum-specific HR estimate. However, even when PH assumption holds within each stratum, the PH assumption for entire study population is almost always violated (see Appendix B). Consequently, the combined HR estimator from the stratified Cox procedure cannot be interpreted as the HR for the entire population. Now, the question is whether we can apply the marginal treatment effect method discussed in Section 3 to deal with the HR. Unfortunately, since the hazard function is not a probability measure, the overall hazard function cannot be expressed as a weighted average of stratum-specific hazard functions. The simple estimation approach discussed in Section 3 cannot be applied to the case using HR as the group contrast.

There are several alternative summary measures to quantify survivorship for each treatment group. For example, one may consider the median survival time. However, it is in general not a weighted average of stratum-specific median failure times either. Two other alternatives are the event rate and restricted survival time at a specific time point (Uno *and others*, 2014, 2015).

For the $t-$year event rate τ_j, $j = 1, 2$, we can use the same approach discussed in Section 3. That is, we estimate τ_j by

$$\tau_j^\dagger = \sum_{k=1}^{K} w_k \hat{\tau}_{jk} \quad \text{or} \quad \tau_{j\pi}^\dagger = \sum_{k=1}^{K} \hat{w}_k \hat{\tau}_{jk},$$

where $\hat{\tau}_{jk} = \hat{S}_{jk}(t)$ and $\hat{S}_{jk}(\cdot)$ is the Kaplan-Meier estimator for the survival func-

tion of the group j in the kth stratum. Then $\hat{\theta}_{new} = g(\tau_1^\dagger, \tau_2^\dagger)$ and $\hat{\theta}_{new\pi} = g(\tau_{1\pi}^\dagger, \tau_{2\pi}^\dagger)$ are unbiased for estimating θ even when there are markedly observed treatment allocation imbalance.

As an example, we also consider the substudy analysis for Australia in VALIANT study. The estimated event rate (death or hospitalization) at $t = 1000$ days is 0.88 in the combo therapy group and 0.77 in the monotherapy group. The naive estimate of the OR of two event rates is 2.33 with a 95% confidence interval of [1.39, 3.93], suggesting that the event rate by $t = 1000$ days for patients receiving monotherapy is significantly lower than that for combo therapy. Recall that there are four strata defined by BMI and diabetic history with the stratum size, $(\hat{w}_1, \hat{w}_2, \hat{w}_3, \hat{w}_4) = (0.24, 0.06, 0.54, 0.16)$. With the above simple procedure, the adjusted event rate is 0.87 in the combo therapy group and 0.77 in the monotherapy group. The resulting estimator $\hat{\theta}_{new}$ of the OR is 1.92 (95% confidence interval: 0.88 to 4.22). This new OR estimator is smaller than the unadjusted counterpart with a p-value of 0.103.

The second alternative is the $t-$year mean survival time (t-MST) up to time point t. That is, $E\{\min(T, t)\}$ for survival time T, which is the area under the survival function up to the time point t (Uno *and others*, 2014, 2015). The corresponding parameter summarizing the treatment effect is $g(\tau_1, \tau_2)$, a contrast between t-MSTs from two groups, where τ_j is the t-MST in group j. For instance, $g(\tau_1, \tau_2)$ is the ratio of two t-MSTs. For the above VALIANT study, the area under the Kaplan-Meier curves for the monotherapy and combo therapy arms are $\hat{\tau}_1 = 395$ days and $\hat{\tau}_2 = 260$ days up to $t = 1000$ days, respectively. That means in the future, if we treat the patient with the new treatment up to 1000 days, we expect on average, the patient receiving monotherapy will have 395 event-free days,

and 260 event-free days for patients receiving combo therapy. If we let θ be the ratio of two t-MSTs at 1000 days, $\hat{\theta} = 1.58$ with a 95% confidence interval of (1.15, 2.18) and a p-value less than 0.01.

Now, for a stratified analysis with respect to the t-MST measure, we first estimate this parameter with a weighted sum of stratum-specific t-MSTs, that is,

$$\tau_j^{\dagger} = \sum_{k=1}^{K} w_k \hat{\tau}_{jk} \quad \text{or} \quad \tau_{j\pi}^{\dagger} = \sum_{k=1}^{K} \hat{w}_k \hat{\tau}_{jk},$$

where $\hat{\tau}_{jk} = \int_0^t \hat{S}_{jk}(u)du$. For the above numerical example, $\tau_{1\pi}^{\dagger} = 394$ days and $\tau_{2\pi}^{\dagger} = 280$ days based on observed $\{\hat{w}_k, k = 1, 2, 3, 4\}$. The ratio of these two, $\hat{\theta}_{new} = 1.41$ with a 95% confidence interval of (1.03, 1.92). Although statistically the monotherapy is significantly better than the combo, the confidence interval of the ratio of two t-MSTs is shifted toward the null value of one.

5 REMARKS

The marginal treatment effect approach discussed in Section 3 is rather flexible, which can handle the case when the target patient population is different from the study population. For example, in a cardiovascular clinical study, the majority of study patients is male. The target future population may be evenly divided with respect to gender. With the conventional stratified analysis, it is difficult to figure out the overall treatment difference for the target population.

The choice of the stratification factors is important. If we overly stratify the study, that is, some factors are not related or mildly related to the outcome, the precision of the stratified inference procedure can be worse than the naive two-sample estimate. Recently Tian *and others* (2012) and Bloniarz *and others* (2015) gener-

alized the augmentation method originally proposed by Zhang *and others* (2008) to efficiently select relevant baseline covariates among a set of pre-specified candidates under an unconditional setting for adjusting the consistent two-sample estimator. It is not clear how to apply this idea to handle the case when there is a potential imbalance with respect to a large set of stratification factors to avoid over-stratification. Further research on appropriate selection of stratification factors is warranted.

APPENDIX

Appendix A: Asymptotical Properties of the Naive Estimator Conditional on Observed Allocation Imbalances.

Firstly, we assume that the observed data consist of n i.i.d observations (Y_i, R_i, X_i), $i = 1, \cdots, n$, where Y_i is the response, $R_i = 1$ or 2 is the treatment indicator for group 1 and 2, respectively, and X_i takes values $1, 2, \cdots, K$, representing the stratum of the ith subject. Without loss of generality, let $\hat{\theta} = g(\hat{\tau}_1, \hat{\tau}_2)$, where $\hat{\tau}_j$ is the observed mean response in arm j and $\hat{\tau}_{jk}$ is the observed mean response in arm j of stratum k, $j = 1, 2; k = 1, \cdots, K$. Our goal is to derive the limiting distribution of

$$n^{1/2}(\hat{\theta} - \theta)$$

conditional on

$$\hat{w}_k - w_k, \hat{\pi}_k - \bar{\pi}, k = 1, \cdots, K,$$

where $n = \sum_{k=1}^{K} n_k$. To this end, we first have the expansion

$$\begin{pmatrix} \hat{\theta} - \theta \\ \hat{w}_1 - w_1 \\ \vdots \\ \hat{w}_K - w_K \\ \hat{\pi}_1 - \bar{\pi} \\ \vdots \\ \hat{\pi}_K - \bar{\pi} \end{pmatrix} = n^{-1} \sum_{i=1}^{n} \begin{pmatrix} \frac{\dot{g}_2(\tau_1,\tau_2) I(R_i=2)(Y_i-\tau_2)}{\pi} + \frac{\dot{g}_1(\tau_1,\tau_2) I(R_i=1)(Y_i-\tau_1)}{1-\pi} \\ I(X_i = 1) - w_1 \\ \vdots \\ I(X_i = K) - w_K \\ \frac{\{I(X_i=1)-w_1\}\{I(R_i=2)-\pi\}}{w_1} \\ \vdots \\ \frac{\{I(X_i=K)-w_K\}\{I(R_i=2)-\pi\}}{w_K} \end{pmatrix} + \begin{pmatrix} \xi_\theta \\ 0 \\ \xi_\pi \end{pmatrix},$$

where $\pi = \text{pr}(R_i = 2)$ and $|\xi_\theta| + |\xi_\pi| = o(n^{-1/2})$ almost surely. Let

$$U_n = n^{-1/2} \sum_{i=1}^{n} U_i \text{ and } V_n = n^{-1/2} \sum_{i=1}^{n} V_i,$$

where

$$U_i = \frac{\dot{g}_2(\tau_1, \tau_2) I(R_i = 2)(Y_i - \tau_2)}{\pi} + \frac{\dot{g}_1(\tau_1, \tau_2) I(R_i = 1)(Y_i - \tau_1)}{1 - \pi}$$

and

$$V_i = \begin{pmatrix} I(X_i = 1) - w_1 \\ \vdots \\ I(X_i = K) - w_K \\ \frac{\{I(X_i=1)-w_1\}\{I(R_i=2)-\pi\}}{w_1} \\ \vdots \\ \frac{\{I(X_i=K)-w_K\}\{I(R_i=2)-\pi\}}{w_K} \end{pmatrix}$$

By central limit theorem, $(U_n, V_n')'$ converges weakly to $(U_0, V_0')'$, a multivariate normal with mean 0 and a variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_\theta^2 & D_{\theta w}' & D_{\theta\pi}' \\ D_{\theta w} & \Sigma_{ww} & 0 \\ D_{\theta\pi} & 0 & \Sigma_{\pi\pi} \end{pmatrix},$$

where

$$\sigma_\theta^2 = \frac{\dot{g}_2(\tau_1,\tau_2)\sigma_{Y2}^2}{\pi} + \frac{\dot{g}_1(\tau_1,\tau_2)\sigma_{Y1}^2}{1-\pi}$$

$\sigma_{Yj}^2 = \text{var}(Y_i|R_i = j)$, $D_{\theta w} = (d_{w1}, \cdots, d_{wK})'$, $D_{\theta\pi} = (d_{\pi 1}, \cdots, d_{\pi K})'$,

$$d_{wk} = \dot{g}_2(\tau_1,\tau_2)w_k(\tau_{2k} - \tau_2) + \dot{g}_1(\tau_1,\tau_2)w_k(\tau_{1k} - \tau_1),$$

$$d_{\pi k} = \dot{g}_2(\tau_1,\tau_2)(1-\pi)(\tau_{2k} - \tau_2) - \dot{g}_1(\tau_1,\tau_2)\pi(\tau_{1k} - \tau_1),$$

$$\Sigma_{ww} = \begin{pmatrix} w_1(1-w_1) & \cdots & -w_1 w_K \\ \cdots & \cdots & \cdots \\ -w_K w_1 & \cdots & w_K(1-w_K) \end{pmatrix}$$

and

$$\Sigma_{\pi\pi} = \pi(1-\pi)\begin{pmatrix} (1-w_1)/w_1 & \cdots & -1 \\ \cdots & \cdots & \cdots \\ -1 & \cdots & (1-w_K)/w_K \end{pmatrix}.$$

It follows from Steck (1957), as $n \to \infty$, $U_n|V_n = v$ converges weakly to a limiting distribution in the sense that for any sequence $\{v_n\}$ such that $v_n \in A_n$, the range of V_n, and $\lim_{n\to\infty} v_n = v$,

$$\sup_u |F_n^{v_n}(u) - F^v(u)| = o(1),$$

where $F_n^{v_n}(u)$ is the cumulative distribution function of U_n conditional on $V_n = v_n$ and $F^v(u)$ is the cumulative distribution function of U_0 conditional on $V_0 = v$.

Furthermore, since $|\xi_\theta| + |\xi_\pi| = o(n^{-1/2})$ almost surely, for any $\delta_n = (\delta'_{nw}, \delta'_{n\pi})' \to \delta$,

$$\begin{aligned}
&P\left\{n^{1/2}(\hat{\theta} - \theta) \le u \mid n^{1/2}(\hat{w}_k - w_k) = \delta_{nw}, n^{1/2}(\hat{\pi}_k - \bar{\pi}) = \delta_{n\pi}\right\} \\
=&P\left\{U_n \le u - n^{1/2}\xi_\theta \mid V_n = (\delta'_{nw}, \delta'_{n\pi} - n^{1/2}\xi'_\pi)'\right\} \\
=&F_n^{\tilde{v}_n}(u - n^{1/2}\xi_\theta) \\
=&F^\delta(u - n^{1/2}\xi_\theta) + o(1) \\
=&F^\delta(u) + o(1),
\end{aligned}$$

where $\tilde{v}_n = (\delta'_{nw}, \delta'_{n\pi} - n^{1/2}\xi'_\pi)' \to \delta$. Therefore, the conditional distribution $n^{1/2}(\hat{\theta} - \theta) \mid n^{1/2}(\hat{w}_k - w_k), n^{1/2}(\hat{\pi}_k - \bar{\pi})$ converges to the normal distribution with a mean b_θ and variance σ^2_{adj}. Next, we will derive the explicit expression for mean b_θ and variance σ^2_{adj}. We first note that the variance-covariance matrices $\Sigma_{\pi\pi}$ and Σ_{ww} are singular, the conditional distribution $U_0|V_0$ is the same as that only conditioning on the components corresponding to $(w_1, \cdots, w_{K-1}, \pi_1, \cdots, \pi_{K-1})'$. Let $\tilde{D}_{\theta\pi}$ $\tilde{D}_{\theta w}$, $\tilde{\Sigma}_{ww}$ and $\tilde{\Sigma}_{\pi\pi}$ be the associated covariance vectors and variance matrices.

$$\begin{aligned}
b_\theta =&n^{1/2}(\hat{w}_1 - w_1, \cdots, \hat{w}_{K-1} - w_{K-1})\tilde{\Sigma}^{-1}_{\theta w}\tilde{D}_{\theta\pi} + n^{1/2}(\hat{\pi}_1 - \bar{\pi}, \cdots, \hat{\pi}_{K-1} - \bar{\pi})\tilde{\Sigma}^{-1}_{\pi\pi}\tilde{D}_{\theta\pi} \\
=&n^{1/2}\begin{pmatrix} \hat{w}_1 - w_1 \\ \vdots \\ \hat{w}_{K-1} - w_{K-1} \end{pmatrix}'\left\{\dot{g}_2(\tau_1, \tau_2)\begin{pmatrix} \tau_{21} - \tau_{2K} \\ \vdots \\ \tau_{2(K-1)} - \tau_{2K} \end{pmatrix} + \dot{g}_1(\tau_1, \tau_2)\begin{pmatrix} \tau_{11} - \tau_{1K} \\ \vdots \\ \tau_{1(K-1)} - \tau_{1K} \end{pmatrix}\right\} \\
&+ n^{1/2}\begin{pmatrix} \hat{\pi}_1 - \bar{\pi} \\ \vdots \\ \hat{\pi}_{K-1} - \bar{\pi} \end{pmatrix}'\left\{\frac{\dot{g}_2(\tau_1, \tau_2)}{\pi}\begin{pmatrix} w_1(\tau_{21} - \tau_{2K}) \\ \vdots \\ w_{K-1}(\tau_{2(K-1)} - \tau_{2K}) \end{pmatrix} - \frac{\dot{g}_1(\tau_1, \tau_2)}{1 - \pi}\begin{pmatrix} w_1(\tau_{11} - \tau_{1K}) \\ \vdots \\ w_{K-1}(\tau_{1(K-1)} - \tau_{1K}) \end{pmatrix}\right\} \\
=&n^{1/2}\left[\dot{g}_2(\tau_1, \tau_2)\sum_{k=1}^{K}\left\{\hat{w}_k - w_k + \frac{\hat{w}_k(\hat{\pi}_k - \bar{\pi})}{\bar{\pi}}\right\}\tau_{2k} + \dot{g}_1(\tau_1, \tau_2)\sum_{k=1}^{K}\left\{\hat{w}_k - w_k - \frac{\hat{w}_k(\hat{\pi}_k - \bar{\pi})}{1 - \bar{\pi}}\right\}\tau_{1k}\right] + o_p(1)
\end{aligned}$$

which is asymptotically equivalent to $n^{1/2}(\hat{\theta}_{\pi w} - \hat{\theta})$. It is clear that the bias of $\hat{\theta}$ can then be consistently estimated by $\hat{b}_{\pi w}$.

For the conditional variance, we first have

$$\sigma^2_{\pi w} = \sigma^2_{\theta} - \tilde{D}'_{\theta w}\tilde{\Sigma}^{-1}_{ww}\tilde{D}_{\theta w} - \tilde{D}'_{\theta \pi}\tilde{\Sigma}^{-1}_{\pi\pi}\tilde{D}_{\theta \pi}.$$

Since

$$\begin{aligned}
\tilde{D}'_{\theta w}\tilde{\Sigma}^{-1}_{ww}\tilde{D}_{\theta w} =& \tilde{D}'_{\theta w}\left\{\dot{g}_2(\tau_1,\tau_2)\begin{pmatrix}\tau_{21}-\tau_{2K}\\ \vdots \\ \tau_{2(K-1)}-\tau_{2K}\end{pmatrix} + \dot{g}_1(\tau_1,\tau_2)\begin{pmatrix}\tau_{11}-\tau_{1K}\\ \vdots \\ \tau_{1(K-1)}-\tau_{1K}\end{pmatrix}\right\} \\
=& \sum_{k=1}^{K} w_k \left\{\dot{g}_2(\tau_1,\tau_2)(\tau_{2k}-\tau_2) + \dot{g}_1(\tau_1,\tau_2)(\tau_{1k}-\tau_1)\right\}^2,
\end{aligned}$$

$$\begin{aligned}
\tilde{D}'_{\theta w}\tilde{\Sigma}^{-1}_{ww}\tilde{D}_{\theta w} =& \tilde{D}'_{\theta w}\left\{\frac{\dot{g}_2(\tau_1,\tau_2)}{\pi}\begin{pmatrix}w_1(\tau_{21}-\tau_{2K})\\ \vdots \\ w_{K-1}(\tau_{2(K-1)}-\tau_{2K})\end{pmatrix} - \frac{\dot{g}_1(\tau_1,\tau_2)}{1-\pi}\begin{pmatrix}w_1(\tau_{11}-\tau_{1K})\\ \vdots \\ w_{K-1}(\tau_{1(K-1)}-\tau_{1K})\end{pmatrix}\right\} \\
=& \frac{\dot{g}_2(\tau_1,\tau_2)^2}{\pi}\sum_{k=1}^{K} w_k(\tau^2_{2k}-\tau^2_2) + \frac{\dot{g}_1(\tau_1,\tau_2)^2}{1-\pi}\sum_{k=1}^{K} w_k(\tau^2_{1k}-\tau^2_1) \\
& - \sum_{k=1}^{K} w_k \left\{\dot{g}_2(\tau_1,\tau_2)(\tau_{2k}-\tau_2) + \dot{g}_1(\tau_1,\tau_2)(\tau_{1k}-\tau_1)\right\}^2,
\end{aligned}$$

and $\sigma^2_{Yj} = \sum_{k=1}^{K} w_k(\sigma^2_{Yjk} + \tau^2_{jk}) - \tau^2_j$, we have

$$\sigma^2_{adj} = \frac{\dot{g}_2(\tau_1,\tau_2)^2}{\pi}\sum_{k=1}^{K} w_k\sigma^2_{Y2k} + \frac{\dot{g}_1(\tau_1,\tau_2)^2}{1-\pi}\sum_{k=1}^{K} w_k\sigma^2_{Y1k},$$

where $\sigma^2_{Yjk} = \mathrm{var}(Y_i|R_i = j, X_i = k)$. Furthermore, the conditional variance of $\hat{\theta} - \theta$ can be consistently estimated by $\hat{\sigma}^2_{\pi w} = \sum_{k=1}^{K}\left\{\dot{g}_2(\tau^{\dagger}_1,\tau^{\dagger}_2)^2\hat{w}^2_k\hat{\sigma}^2_{2k} + \dot{g}_1(\tau^{\dagger}_1,\tau^{\dagger}_2)^2\hat{w}^2_k\hat{\sigma}^2_{1k}\right\}$.

In the aforementioned derivation, we used the fact that

$$\tilde{\Sigma}_{ww}\begin{pmatrix} \tau_{j1}-\tau_{jK} \\ \vdots \\ \tau_{j(K-1)}-\tau_{jK} \end{pmatrix} = \begin{pmatrix} w_1(\tau_{j1}-\tau_j) \\ \vdots \\ w_{K-1}(\tau_{j(K-1)}-\tau_j) \end{pmatrix}$$

and

$$\tilde{\Sigma}_{\pi\pi}\begin{pmatrix} w_1(\tau_{j1}-\tau_{jK}) \\ \vdots \\ w_{K-1}(\tau_{j(K-1)}-\tau_{jK}) \end{pmatrix} = \pi(1-\pi)\begin{pmatrix} \tau_{j1}-\tau_j \\ \vdots \\ \tau_{j(K-1)}-\tau_j \end{pmatrix}, j = 1, 2.$$

Appendix B: Incompatibility of PH assumption in the entire study and within stratum

Assume that the PH assumption holds within each of the K strata. Let $S_{jk}(t)$ denote the survival function in the kth stratum of arm j, $j = 1, 2$. Then we have

$$S_{2k}(t) = S_{1k}(t)^{r_0}$$

for $k = 1, \cdots, K$ and a common HR r_0. Furthermore, the marginal survival function for the entire study population is

$$S_j(t) = \sum_{k=1}^{K} w_k S_{jk}(t),$$

for arm j, where $w_1 + \cdots + w_K = 1, w_k \in (0, 1), k = 1, \cdots, K$. If the PH assumption holds for the marginal survival function, then $S_2(t) = S_1(t)^r$ for a constant r, i.e.,

$$\left\{\sum_{k=1}^{K} w_k S_{1k}(t)\right\}^r = \sum_{k=1}^{K} w_k S_{1k}(t)^{r_0}.$$

Taking derivative with respect to t for both sides at $t = 0$, we obtain that

$$r = \frac{r_0 \sum_{k=1}^{K} w_k S_{1k}(0)^{r_0-1} f_{1k}(0)}{\sum_{k=1}^{K} w_k f_{1k}(0) \left(\sum_{k=1}^{K} w_k S_{1k}(0)\right)^{r-1}} = r_0.$$

Without loss of generality, we assume that $r_0 \geq 1$, otherwise we always can switch two groups of interest to ensure that the HR is greater than or equal to 1. Under this assumption

$$S_2(t) = \sum_{k=1}^{K} w_k S_{1k}(t)^{r_0} \geq \left(\sum_{k=1}^{K} w_k S_{1k}(t) \right)^{r_0} = S_1(t)^{r_0},$$

for $0 < S_{1k}(t) < 1$ due to the convexity of the function x^{r_0}. The equality holds only when $S_{11}(t) = \cdots = S_{1K}(t)$ or $r_0 = 1$. Thus the PH model with a HR of r_0 is not true for the entire study population unless $S_{11}(t) = \cdots = S_{1K}(t)$, i.e., the survival distributions are the same across strata or $r_0 = 1$, i.e., there is no difference in survivorship in any stratum.

References

BLONIARZ, A., LIU, H., ZHANG, C., SEKHON, J. AND YU, B. (2015). Lasso adjustments of treatment effect estimates in randomized experiments. *PNAS*.

COX, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2), 187–220.

FRASER, D A S. (2004). Ancillaries and conditional inference. *Statistical Science* **19**(2), 333–269.

KALBFLEISCH, J. (1975). Sufficiency and conditionality. *Biometrika* **62**(2), 251–259.

KOCH, G., TANGEN, C., JUNG, J. AND AMARA, I. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized

clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine* **17**, 1863–1892.

MANTEL, N. AND HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**(4), 719–748.

MEHROTRA, D., SU, S. AND LI, X. (2012). An efficient alternative to the stratified cox model analysis. *Statistics in Medicine* **31**(17), 1849–1856.

MIRATRIX, L W., SEKHON, J S. AND YU, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B* **75**, 369–396.

MOORE, K. AND VAN DER LAAN, M. (2009). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine* **28**(1), 39–64.

PFEFFER, M., MCMURRAY, J., VELAZQUEZ, E., ROULEAU, J., KOBER, L., MAGGIONI, A., SOLOMON, S., SWEDBERG, K., DE WERF F., VAN, WHITE, H., LEIMBERGER, J., HENIS, M., EDWARDS, S., ZELENKOFSKE, S., SELLERS, M., CALIFF, R. *and others*. (2003). Valsartan, captopril, or both in myocardial infarction complicated by heart failure, left ventricular dysfunction, or both. *New England Journal of Medicine* **349**(20), 1893–1906.

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. *Statistics in Medicine* **21**(19), 2917–2930.

ROSENBLUM, M. AND VAN DER LAAN, M. (2010). Targeted maximum likelihood estimation of the parameter of a marginal structural model. *International Journal of Biostatistics* **6**(1), 13.

SENN, S J. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* **8**(4), 467–475.

STECK, G.P. (1957). Limit theorem for conditional distribution. *Univ. California Publ. Statst.* **2**, 237–284.

TIAN, L., CAI, T., ZHAO, L. AND WEI, LJ. (2012). On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial. *Biostatistics* **13**(2), 256–273.

TSIATIS, A., DAVIDIAN, M., ZHANG, M. AND LU, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine* **27**, 4658–4677.

UNO, H., CLAGGETT, B., TIAN, L., INOUE, E., GALLO, P., MIYATA, T., SCHRAG, D., TAKEUCHI, M., UYAMA, Y., ZHAO, L., SKALI, H., SOLOMON, S., JACOBUS, S., HUGHES, M., PACKER, M. *and others*. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology* **32**(22), 2380–2385.

UNO, H., WITTES, J., FU, H., SOLOMON, S., CLAGGETT, B., TIAN, L., CAI, T., PFEFFER, M., EVANS, S. AND WEI, LJ. (2015). Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. *Annals of Internal Medicine* **163**(2), 127–134.

VALLIANT, R. (1993). Post-stratification and conditional variance estimation. *Journal of the American Statistical Association* **88**.

ZHANG, M., TSIATIS, A. AND DAVIDIAN, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64**(3), 707–715.

Table 1: Stratified event rate of death or hospitalization up to 18 months in Australia region of VALIANT Study

stratification factors		Australia Data (# of events / # of patients)		$\hat{\pi}_k$	OR
BMI	History of Diabetes	Monotherapy	Combo therapy		
< 25	No	43/60	8/13	0.82	0.63
< 25	Yes	9/10	6/8	0.56	0.33
≥ 25	No	65/108	44/54	0.67	2.91
≥ 25	Yes	18/24	22/25	0.49	2.44

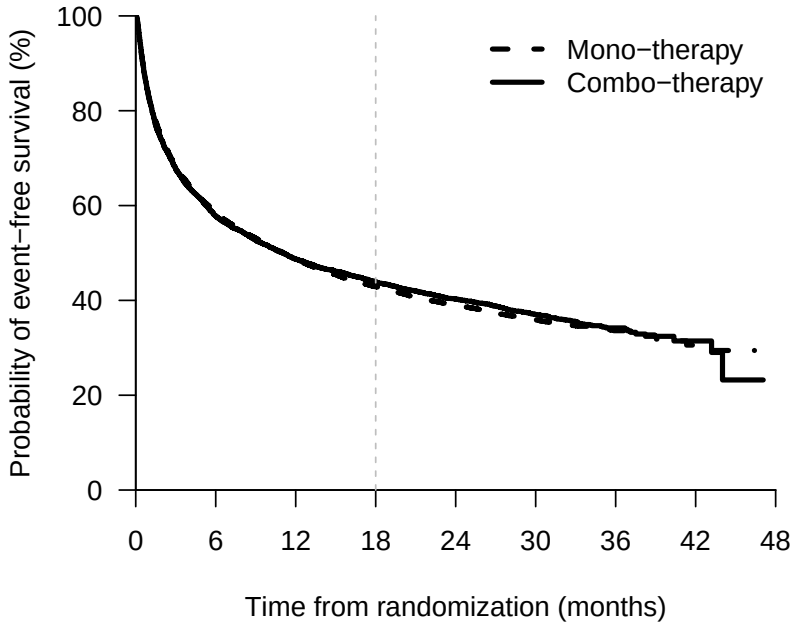

Figure 1: The survival curves for entire VALIANT study by arms: monotherapy and combo therapy arms.

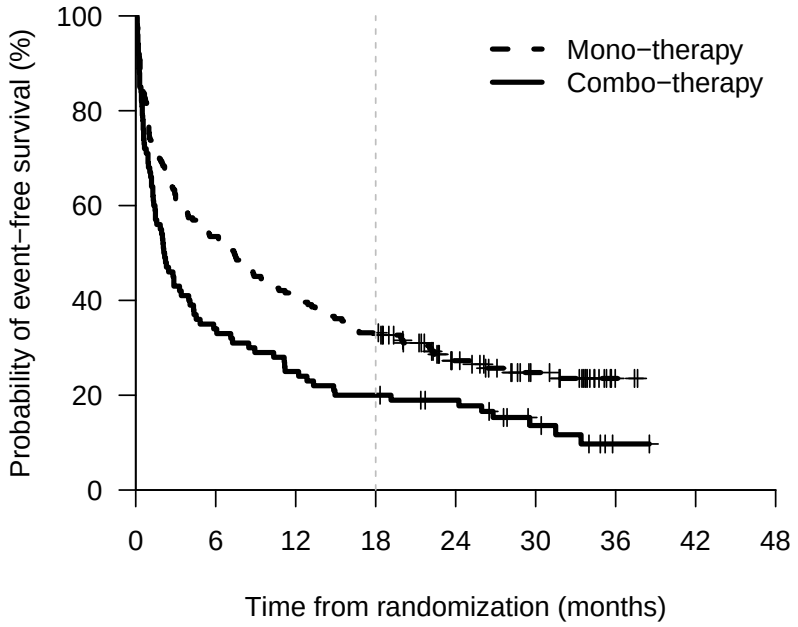

Figure 2: The survival curves for Australian patients in VALIANT study by arms: monotherapy and combo therapy arms.