Biometrical Journal **0** (0000) 0, zzz–zzz / DOI: 10.1002/bimj.200100000

Confidence regions for treatment effects in subgroups in biomarker stratified designs

Fang Wan[*,1], **Cornelia U. Kunz** [1], and **Thomas F. Jaki** [1]

[1] Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK

Received zzz, revised zzz, accepted zzz

Subgroup analysis has important applications in the analysis of controlled clinical trials. Sometimes the result of the overall group fails to demonstrate that the new treatment is better than the control therapy, but for a subgroup of patients the treatment benefit may exist; or sometimes the new treatment is better for the overall group but not for a subgroup. Hence we are interested in constructing a simultaneous confidence interval for the difference of the treatment effects in a subgroup and the overall group. Subgroups are usually formed on the basis of a predictive biomarker such as age, sex or some genetic marker. While, for example age can be detected precisely, it is often only possible to detect the biomarker status with a certain probability. Because patients detected with a positive or negative biomarker may not be truly biomarker positive or negative, responses in the subgroups depend on the treatment therapy as well as on the sensitivity and specificity of the assay used in detecting the biomarkers. In this work we show how (approximate) simultaneous confidence intervals and confidence ellipsoid for the treatment effects in subgroups can be found for biomarker stratified clinical trials using a normal framework with normally distributed or binary data. We show that these intervals maintain the nominal confidence level via simulations.

Key words: Biomarker stratified design; Confidence region; Subgroup; Treatment effects.

1 INTRODUCTION

Biomarkers are measurable indicators of biological conditions. This could be gene expression, the presence of a certain antibody or some proteins. With the development and use of biomarkers, modern medicine shifts from empirical to precision medicine (Kelloff and Sigman, 2012; Henry and Hayes, 2012). While for empirical medicine drugs under study target the whole population, precision medicine allows the drugs to target some specific subgroups of the population. Since a different status of a biomarker indicates a different biological condition, it is not unusual that patients from different (predictive) biomarker subgroups respond differently to the same drug in clinical trials. Sometimes the response from the overall group fails to demonstrate that the new treatment is better than the control or existing therapy, but for a subgroup of patients the treatment benefit may exist; or the new treatment is significantly better than control for the overall group but not for the subgroups. Hence subgroup analysis is becoming increasingly popular (Friede *et al*., 2012; Spienssens and Debois, 2010).

Biomarker stratified designs are often used to determine the subgroups of patients that benefit (most) from the new drug (see, for example, Freidlin *et al*., 2010; Baker *et al*., 2012; Kaplan *et al*., 2013; Mandrekar and Sargent, 2009; Simon, 2010) . The patients are classified into subgroups according to the biomarker status before being randomized within each subgroup into treatment and control arms.

As the patients in different biomarker subgroups may respond very differently to the same drug, it is important to know which subgroup does in fact benefit and how much it benefits from the drug in order to optimize the treatment outcome and minimize adverse events. Point estimation of the treatment effect

*Corresponding author: e-mail: f.wan@lancaster.c.uk, Phone: +00-999-999-999

(i.e. the difference of mean responses between treatment and control arms) has often been performed in analyzing the results in some predefined biomarker subgroups. However, the point estimate is based on random observations of the individual response and hence is almost never the true effect. For some samples of the population, the point estimate could be reasonably close to the true effect, but in some unlucky circumstances the estimates may indicate effects entirely by chance. With the randomness of response of individual patients and the uncertainty of the biomarker detection, interval estimation should be employed to evaluate the treatment effects in true biomarker subgroups. Confidence ellipsoids and simultaneous confidence intervals for the treatment effects, which give a direct indication (simultaneously) whether subgroups are likely to have treatment effects, are the ideal statistics to be used in addressing this issue.

Furthermore, while some biomarkers can be detected precisely, it is often only possible to detect the biomarker status with a certain probability. Therefore the estimates of the treatment effects in the true biomarker subgroups depend on the response of patients in the subgroups as well as on the sensitivity and specificity of the assay used. Without taking into consideration the sensitivity and specificity, the resulting estimates are in fact the treatment effects in the observed biomarker subgroups rather than in the true biomarker subgroups (see, for example, Freidlin *et al*., 2012; Brusselle *et al*., 2013; Jubb *et al*., 2011). Liu *et al*. (2014) proposed a method for correcting the bias in the estimated effects caused by the biomarker misclassification in a normal framework. In this paper, we build upon their method and we derive a variance-covariance matrix for the treatment effects for two different cases: (1) numbers in each biomarker subgroup are random and (2) when the they are fixed through stratification. We then propose a simultaneous confidence ellipse and simultaneous confidence intervals for the treatment effects in different populations when the membership to the population is imperfect. We consider normally distributed and binary data in our evaluations and consider the simple case of only a single biomarker that is either positive or negative, although the method developed can be extended to the situation where more than two (sub)groups are used.

2 METHOD

Suppose there is a new therapy for treating a given disease and we want to determine whether it works better than the control therapy (active control or placebo). Some prior knowledge suggests that the existence of treatment effects may be related to the positive or negative status of a certain biomarker. A clinical trial is designed for this purpose: a fixed number of patients are recruited and classified into two subgroups with respect to their test results of the biomarker status. In each of the subgroups, patients are randomized to treatment and control arms.

Let N be the number of patients recruited into the study and γ be the prevalence of the positive biomarker among the population under study, that is, the probability of each individual having a positive biomarker. Let λ_1 and λ_2 be the sensitivity and specificity of testing the biomarker. In order to be usable, both the sensitivity and specificity should be strictly greater than 0.5. We assume γ, λ_1 and λ_2 are either given by medical experts or estimated from previous studies. Let $\oplus$ and $\ominus$ denote the test results of the biomarker status and $+$ and $-$ the true biomarker status. Note that the test results of the biomarker status may not be the true biomarker status, depending on the sensitivity and specificity of the test.

By using Bayesian Theory, the probabilities that an individual classified into the biomarker positive subgroup or the negative subgroup using such test are given by

$$p_{\oplus} = \gamma\lambda_1 + (1-\gamma)(1-\lambda_2).$$

We consider two different designs: in the first design the total number of patients is fixed but the number of patients in each subgroup is random following a binomial distribution, $Binom(N, p_{\oplus})$, while in the second design we fix the number of subjects in each biomarker subgroup through stratification. We assume the sample size of patients in the true positive biomarker subgroup follows a binomial distribution $Binom(N, \gamma)$. The responses are denoted by Y_{jk}^i, $i \in \{T, C\}$, $j \in \{\oplus, \ominus, +, -\}$ and $k = 1, 2, \cdots, N_j^i$

where T and C denote the treatment and control arms and N_j^i denote the corresponding sample size. For example, $Y_{\oplus 2}^C$ denotes the response of the second subject in the control arm in the detected biomarker positive subgroup. Note that responses in the true biomarker positive/negative subgroups are not observable unless the sensitivity/specificity equals one. The sample means are denoted by $\bar{Y}_j^i$, and are again observable only for $j \in \{\oplus, \ominus\}$. Next, we consider the construction of confidence regions for the treatment effects in the two subgroups when the response distributions are normal and Bernoulli, and then extend the method for comparisons of one subgroup versus the full population.

2.1 Confidence region with Normally distributed response

In this section, we assume $Y_{jk}^i \sim N(\mu_{i,j}, \sigma_{i,j}^2)$ with unknown mean $\mu_{i,j}$ and unknown variance $\sigma_{i,j}^2$ for $j \in \{+,-\}$ and all the responses, either within or between the four arms, are independently distributed.

Let $\boldsymbol{\mu} = (\mu_{+,T} - \mu_{+,C}, \mu_{-,T} - \mu_{-,C})'$ be the vector of treatment effects that we are interested in; that is the differences in the mean treatment effects amongst all truly biomarker positive and negative patients, respectively. Note that simple estimates of $\boldsymbol{\mu}$ based on sample means are not available as we only know the status based on the diagnostic test. We want to construct confidence ellipses and simultaneous confidence intervals for $\boldsymbol{\mu}$. The mean treatment effects in the observed subgroups are expressed as

$$E(\bar{Y}_{\oplus}^i) = w_1 \mu_{+,i} + (1 - w_1)\mu_{-,i} \quad (1)$$

$$E(\bar{Y}_{\ominus}^i) = (1 - w_2)\mu_{+,i} + w_2 \mu_{-,i} \quad (2)$$

where $i \in \{T, C\}$ and w_1 and w_2 denote the true positive and true negative (predictive) rate, i.e.,

$$w_1 = \frac{\lambda_1 \gamma}{\lambda_1 \gamma + (1 - \lambda_2)(1 - \gamma)}$$

$$w_2 = \frac{\lambda_2 (1 - \gamma)}{\lambda_2 (1 - \gamma) + (1 - \lambda_1)\gamma}.$$

Note that $\mu_{+,i}$ and $\mu_{-,i}$ can be found from Equations (1) and (2):

$$\mu_{+,i} = m_1 E(\bar{Y}_{\oplus}^i) + (1 - m_1) E(\bar{Y}_{\ominus}^i)$$

$$\mu_{-,i} = m_2 E(\bar{Y}_{\ominus}^i) + (1 - m_2) E(\bar{Y}_{\oplus}^i)$$

with $m_1 = w_2/(w_1 + w_2 - 1)$ and $m_2 = w_1/(w_1 + w_2 - 1)$. Hence if we define

$$\boldsymbol{V} = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} m_1(\bar{Y}_{\oplus}^T - \bar{Y}_{\oplus}^C) + (1 - m_1)(\bar{Y}_{\ominus}^T - \bar{Y}_{\ominus}^C) \\ m_2(\bar{Y}_{\ominus}^T - \bar{Y}_{\ominus}^C) + (1 - m_2)(\bar{Y}_{\oplus}^T - \bar{Y}_{\oplus}^C) \end{pmatrix},$$

it is clear that $E(\boldsymbol{V}) = \boldsymbol{\mu}$, that is, $\boldsymbol{V}$ is an unbiased estimate of the vector of treatment effects in the true positive biomarker subgroup and true negative biomarker subgroup. The derivation of $\boldsymbol{V}$ can also be found in Liu *et al*. (2014).

The variance-covariance matrix of $\boldsymbol{V}$, Σ, is usually not known. Liu *et al*. (2014) provided the estimates of the variances of V_1 and V_2 for the case that the subgroup sample sizes are fixed through stratification. We estimate the covariance between V_1 and V_2 and extend the estimates to case (1) where the subgroup sample sizes are treated as random variables. Our variance-covariance matrix is given by

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \cdot & \hat{\Sigma}_{22} \end{pmatrix} := \begin{pmatrix} m_1^2 S_{\oplus}^2 + (1 - m_1)^2 S_{\ominus}^2 & m_1(1 - m_2)S_{\oplus}^2 + m_2(1 - m_1)S_{\ominus}^2 \\ \cdot & (1 - m_2)^2 S_{\oplus}^2 + m_2^2 S_{\ominus}^2 \end{pmatrix}$$

where $S_{\oplus}^2 = E(\frac{1}{N_{\oplus}^T})\left(\hat{\sigma}_{\oplus,T}^2 + \hat{\sigma}_{\oplus,C}^2\right)$ and $S_{\ominus}^2 = E(\frac{1}{N_{\ominus}^T})\left(\hat{\sigma}_{\ominus,T}^2 + \hat{\sigma}_{\ominus,C}^2\right)$ are the estimated variances of $\bar{Y}_{\oplus}^T - \bar{Y}_{\oplus}^C$ and $\bar{Y}_{\ominus}^T - \bar{Y}_{\ominus}^C$ and $\hat{\sigma}_{j,i}^2$ is the usual estimate of $\sigma_{j,i}^2$ with $j \in \{\oplus, \ominus\}$ and $i \in \{T, C\}$. It can be

shown that if the observed biomarker status is exactly the true biomarker status, that is, when $\lambda_1 = \lambda_2 = 1$, we have $w_1 = w_2 = 1$ and $m_1 = m_2 = 1$. Note that in practice, we need at least 2 patients in each arm in order to estimate the variances. The computation of the expectation in the first case therefore becomes

$$\begin{aligned} E(\frac{1}{N^i_{\oplus}}) &= \frac{\Sigma_{n=2}^{0.5N-2}\frac{1}{n}\binom{0.5N-2}{n}p_{\oplus}^n(1-p_{\oplus})^{0.5N-2-n}}{\Sigma_{n=2}^{0.5N-2}\binom{0.5N-2}{n}p_{\oplus}^n(1-p_{\oplus})^{0.5N-2-n}} \\ E(\frac{1}{N^i_{\ominus}}) &= \frac{\Sigma_{n=2}^{0.5N-2}\frac{1}{n}\binom{0.5N-2}{n}(1-p_{\oplus})^n p_{\oplus}^{0.5N-2-n}}{\Sigma_{n=2}^{0.5N-2}\binom{0.5N-2}{n}(1-p_{\oplus})^n p_{\oplus}^{0.5N-2-n}} \end{aligned}$$

in order to incorporate the practical consideration. By applying Jensen's inequality, it can be shown that $\frac{1}{E(N^T_{\oplus})} \leq E(\frac{1}{N^i_{\oplus}})$ and $\frac{1}{E(N^T_{\ominus})} \leq E(\frac{1}{N^i_{\ominus}})$. In the case of the second design, that is, the sample size of the positive biomarker subgroup, $N_{\oplus}$, is fixed through stratification, the expectations $E(\frac{1}{N^T_{\oplus}})$ and $E(\frac{1}{N^T_{\ominus}})$ simply become $\frac{1}{N^T_{\oplus}}$ and $\frac{1}{N^T_{\ominus}}$. The variances of treatment effects proposed in Liu *et al.* (2014) can be seen as a special case of $\hat{\Sigma}_{11}$ and $\hat{\Sigma}_{22}$ in our variance-covariance matrix.

The sensitivity and specificity of the biomarkers are assumed known in the above inference. However, often there will be situations where the sensitivity and specificity are not known with certainty but come from clinical opinion or are estimated from previous data. In this case, we can assume that the observed or estimated sensitivity, $\hat{\lambda}_1$, and specificity, $\hat{\lambda}_2$, are random variables with unknown mean λ_i for $i = 1, 2$. Since $\hat{\lambda}_i$ is essentially a proportion, it is natural to assume $logit(\hat{\lambda}_i)$ follows a normal distribution centered at $logit(\lambda_i)$. Then asymptotically the resulting estimate of (m_1, m_2), $(\hat{m}_1, \hat{m}_2)$, follows a normal distribution where the covariance matrix is found by using the Delta method. The test statistic becomes

$$\boldsymbol{V} = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} \hat{m}_1(\bar{Y}^T_{\oplus} - \bar{Y}^C_{\oplus}) + (1-\hat{m}_1)(\bar{Y}^T_{\ominus} - \bar{Y}^C_{\ominus}) \\ \hat{m}_2(\bar{Y}^T_{\ominus} - \bar{Y}^C_{\ominus}) + (1-\hat{m}_2)(\bar{Y}^T_{\oplus} - \bar{Y}^C_{\oplus}) \end{pmatrix}.$$

The covariance matrix of the statistic $\boldsymbol{V}$ is then computed following the law of total variance/covariance. See the supplementary material for details.

If the sample size is large, according to the Central Limit Theorem, asymptotically we have

$$(\boldsymbol{V} - \boldsymbol{\mu})'\hat{\Sigma}^{-1/2} \sim N(\boldsymbol{0}, I). \tag{3}$$

Let $(z_1, z_2)' = (\boldsymbol{V} - \boldsymbol{\mu})'\hat{\Sigma}^{-1/2}$, then z_1 and z_2 are i.i.d following a standard normal distribution. Hence

$$\left((\boldsymbol{V} - \boldsymbol{\mu})'\hat{\Sigma}^{-1/2}\right)\left((\boldsymbol{V} - \boldsymbol{\mu})'\hat{\Sigma}^{-1/2}\right)' = z_1^2 + z_2^2 \sim \chi^2_2.$$

Based on the above asymptotic results, confidence regions and simultaneous confidence intervals can be constructed straightforwardly:
A natural $1-\alpha$ confidence ellipse for $\boldsymbol{\mu}$ is given by

$$\left\{\boldsymbol{\theta}:(\boldsymbol{V} - \boldsymbol{\theta})'\hat{\Sigma}^{-1}(\boldsymbol{V} - \boldsymbol{\theta}) < \chi^2_{2,\alpha}\right\} \tag{4}$$

and an exact simultaneous confidence interval is given by

$$\left\{V_1 - r\sqrt{\hat{\Sigma}_{11}} < \mu_{+,T} - \mu_{+,C} < V_1 + r\sqrt{\hat{\Sigma}_{11}}, V_2 - r\sqrt{\hat{\Sigma}_{22}} < \mu_{-,T} - \mu_{-,C} < V_2 + r\sqrt{\hat{\Sigma}_{22}}\right\} \tag{5}$$

where the critical constant r is such that

$$P\left\{V_1 - r\sqrt{\hat{\Sigma}_{11}} < \mu_{+,T} - \mu_{+,C} < V_1 + r\sqrt{\hat{\Sigma}_{11}}, V_2 - r\sqrt{\hat{\Sigma}_{22}} < \mu_{-,T} - \mu_{-,C} < V_2 + r\sqrt{\hat{\Sigma}_{22}}\right\} = 1-\alpha.$$

The exact simultaneous confidence interval is 'exact' in the sense that it gives an exact $1-\alpha$ confidence level if the distribution in (3) is exact. The critical constant r can be found by using the method given by Genz and Bretz (2009) and can be computed directly using, for example, the R package mvtnorm (Genz and Bretz, 2009; Genz *et al*., 2016). If we choose $r = Z_{\alpha/4}$ where Z_α is the upper α-quantile number of a standard normal distribution, (5) becomes a Bonferroni confidence interval (Dunn, 1961). The Bonferroni confidence intervals are slightly conservative given that the approximation in (3) is accurate. An alternative conservative simultaneous confidence interval is given by simply projecting the confidence ellipse on the axes (referred to as SCIs without shrinkage henceforth), but this is often very conservative compared with the other confidence intervals we considered above.

2.2 Extension to binary response data

The most commonly used confidence intervals for the proportion, p, of a Bernoulli distribution Bern(p) is based on a Normal approximation confidence interval, also called the Wald confidence interval, which has the form

$$\left\{\hat{p} - z_{\alpha/2}\sqrt{\frac{1}{N}\hat{p}(1-\hat{p})}, \hat{p} + z_{\alpha/2}\sqrt{\frac{1}{N}\hat{p}(1-\hat{p})}\right\}$$

where N is the sample size and $\hat{p}$ is the sample proportion. For a difference in proportions, let $\hat{p}_1$ and $\hat{p}_2$ be the sample proportions and N_1 and N_2 be the relevant sample sizes, the Wald confidence interval for $p_1 - p_2$ is given by

$$\left\{\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{N_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{N_2}}, \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{N_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{N_2}}\right\}.$$

However, it has been shown that the coverage of the Wald confidence interval is not satisfactory, especially when the value of the proportion parameter is close to either 0 or 1 (Brown *et al*., 2001; Agresti and Caffo, 2000). Various alternative confidence intervals have been proposed to deal with this 'boundary effect' (see, for example, Anbar, 1983; Wilson, 1927; Newcombe, 1998), among which the method given by Agresti and Coull (1998) seems to be both simple and effective (Brown *et al*., 2001). The basic idea of their method is to adjust the Wald interval by adding some pseudo data to the response. If the confidence interval is for the proportion parameter, then add 2 successes and 2 failures; if it is for the difference between two proportions, then add 1 success and 1 failure to each sample group (Agresti and Caffo, 2000). In this section, we adapt their method in the construction of confidence ellipse and simultaneous confidence intervals for the difference between proportions in the presence of misclassification.

Assume that the response of patient k in treatment arm i with true biomarker status j follows a Bernoulli distribution $Y_{jk}^i \sim Bern(\pi_j^i)$ for $i \in \{T, C\}$ and $j \in \{+, -\}$ where π_j^i is the unknown proportion parameter of the Bernoulli distribution. For $j \in \{\oplus, \ominus\}$, π_j^i represents the response rate in the corresponding observed biomarker subgroup. All the responses, either within or between the four arms, are assumed to be independently distributed. Note that when the sensitivity and specificity are not equal to 1, the numbers of positive responses in the observed subgroups, for example $\Sigma_k Y_{\oplus,k}^T$, are the weighted sum of two Binomial distributions, $Binom(N_+^T, \pi_+^T)$ and $Binom(N_-^T, \pi_-^T)$. It can be shown that $\Sigma_{k=1}^n Y_{\oplus,k}^T$ conditional on a fixed total sample size n also follows a Binomial distribution. Therefore, each realization of $Y_{\oplus,k}^T,\ k = 1, \cdots, n$ can be seen as n samples from a Bernoulli distribution. This holds for other observed subgroups as well. Hence if we let $\boldsymbol{\mu} = (\pi_+^T - \pi_+^C, \pi_-^T - \pi_-^C)'$ be the vector of treatment effects in the two biomarker subgroups, following Section 2.1, an estimate of $\boldsymbol{\mu}$ is given by

$$\boldsymbol{V}_b = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} = \begin{pmatrix} m_1(\pi_\oplus^T - \pi_\oplus^C) + (1-m_1)(\pi_\ominus^T - \pi_\ominus^C) \\ m_2(\pi_\ominus^T - \pi_\ominus^C) + (1-m_2)(\pi_\oplus^T - \pi_\oplus^C) \end{pmatrix}$$

where $\pi_j^i = \frac{\Sigma_k Y_{jk}^i + t}{N_j^i + 2t}$ for $i \in \{T, C\}$ and $j \in \{\oplus, \ominus\}$ with t being the added number of successes and failures in that arm. The variance-covariance matrix $\hat{\Sigma}_b$ is given by $\hat{\Sigma}$ in Section 2.1 but with $\hat{\sigma}^2_{j,i} = \pi_j^i(1-\pi_j^i)$ for $i \in \{T, C\}$ and $j \in \{\oplus, \ominus\}$. Then the confidence ellipse and SCIs can be constructed by substituting $\boldsymbol{V}_b$ and $\hat{\Sigma}_b$ for $\boldsymbol{V}$ and $\hat{\Sigma}$ in (4) and (5) in Section 2.1.

2.3 Extension to the entire group

If the interest is on the treatment effects of one subgroup, for example the biomarker positive subgroup $\mu_{+,T} - \mu_{+,C}$, and the full sample population $\mu_T - \mu_C$, the test statistic $\boldsymbol{V}^*$ can be easily obtained by using simple matrix transformation

$$\boldsymbol{V}^* = \begin{pmatrix} 1 & 0 \\ \gamma & 1-\gamma \end{pmatrix} \times \boldsymbol{V}$$

and the estimated variance-covariance matrix is given by

$$\hat{\Sigma}^* = \begin{pmatrix} 1 & 0 \\ \gamma & 1-\gamma \end{pmatrix} \times \hat{\Sigma} \times \begin{pmatrix} 1 & 0 \\ \gamma & 1-\gamma \end{pmatrix}'.$$

Then the confidence ellipse and simultaneous confidence intervals constructed above can be directly applied by substituting $\boldsymbol{V}^*$ and $\hat{\Sigma}^*$ for $\boldsymbol{V}$ and $\hat{\Sigma}$.

If the interest is on the treatment effects of both the subgroups and the full population, then the test statistic is given by

$$\boldsymbol{V}^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \gamma & 1-\gamma \end{pmatrix} \times \boldsymbol{V}$$

where any two elements in $\boldsymbol{V}^*$ determine the third one given γ. Hence the SCIs for the treatment effects in the two subgroups and the whole group are simply the SCIs for the treatment effects in the two subgroups and the weighted combination of these SCIs, with weights γ and $1-\gamma$, in the whole group. For example, by substituting $\boldsymbol{V}^*$ for $\boldsymbol{V}$ the limits of the SCIs in expression (5) becomes

$$\left\{ \begin{pmatrix} V_1 - r\sqrt{\hat{\Sigma}_{11}} \\ V_2 - r\sqrt{\hat{\Sigma}_{22}} \\ \gamma(V_1 - r\sqrt{\hat{\Sigma}_{11}}) + (1-\gamma)(V_2 - r\sqrt{\hat{\Sigma}_{22}}) \end{pmatrix}, \begin{pmatrix} V_1 + r\sqrt{\hat{\Sigma}_{11}} \\ V_2 + r\sqrt{\hat{\Sigma}_{22}} \\ \gamma(V_1 + r\sqrt{\hat{\Sigma}_{11}}) + (1-\gamma)(V_2 + r\sqrt{\hat{\Sigma}_{22}}) \end{pmatrix} \right\}.$$

3 EXAMPLE

In Corren *et al.* (2011), a clinical trial was designed and conducted to compare the efficacy (the relative change in prebronchodilator forced expiratory volume in 1 second (FEV_1) from baseline to week 12) of lebrikizumab in adults with Asthma. Patient subgroups were pre-specified according to baseline type 2 helper T-cell status (Th2-high or Th2-low). The Th2 status were assessed on the basis of total IgE level and blood eosinophil count. Patient with a total IgE level of more than 100 IU per milliliter and an eosinophil count of 0.14×10^9 cells per liter or more was classified to the Th2-high subgroup, otherwise was classified to the Th2-low subgroup. The gene expression microarrays of a group of 42 patients were analyzed to estimate the sensitivity($\lambda_1 = 0.86$), specificity($\lambda_2 = 0.65$), true positive rate ($w_1 = 0.73$) and true negative rate ($w_2 = 0.83$) of this classification (see Table S2 in Supplementary Appendix in

Table 1 Mean relative change from baseline FEV_1 in patients at 12 weeks with known sample sizes

	Placebo		Lebrikizumab	
	mean (SD)	sample size	mean (SD)	sample size
all patients	4.27 (15.36)	112	9.78 (19.71)	106
Th2-high	3.12 (14.92)	54	9.54 (18.18)	58
Th2-low	5.35 (15.81)	58	10.08 (21.62)	48

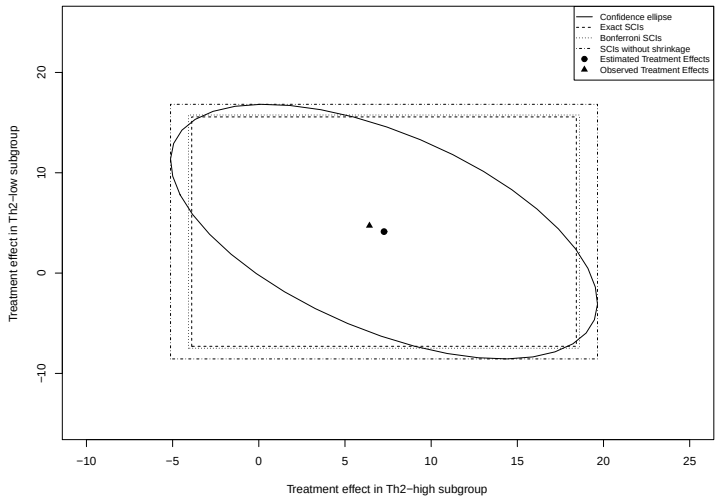

Figure 1 The confidence ellipse and the simultaneous confidence intervals for the treatment effects in the example in Section 3.

Corren *et al*. (2011)). In total, 112 of the eligible patients were classified into the Th2-high subgroup and the remaining 106 were classified into the Th2-low subgroup. Patients in each subgroup were randomly assigned to receive lebrikizumab or placebo. The outcomes were recorded in Table 1.

In order to establish whether the new therapy is effective, we want to construct a confidence ellipse and simultaneous confidence intervals for the treatment effects ($\%$ change in FEV_1 in eligible patients) in the Th2-high and Th2-low subgroups. The confidence ellipse and simultaneous confidence intervals are constructed by using the methods introduced in Section 2 and plotted in Figure 1. Note that the 'estimated treatment effects' denoted by a solid dot is the estimate of the treatment effects in the true biomarker subgroups, while the 'observed treatment effects' denoted by a triangle represents the treatment effects in the observed biomarker subgroups; they will overlap with each other if the sensitivity and specificity both equal to 1. The confidence ellipse is given by the solid ellipse and the 3 simultaneous confidences intervals are given by the projection of the 3 rectangular regions onto the two axes. It is shown that the 'exact' SCIs lie within the Bonferroni SCIs and the 'SCIs without shrinkage' and the latter contains also the confidence ellipse, as expected by construction. Although the observed treatment effects and the estimated treatment effects are larger than 0 in both axes, it is shown that both the confidence ellipse and simultaneous confidence intervals contain the origin 0. Hence in this example, we cannot conclude that the new therapy is significantly better than the placebo in either of the subgroups.

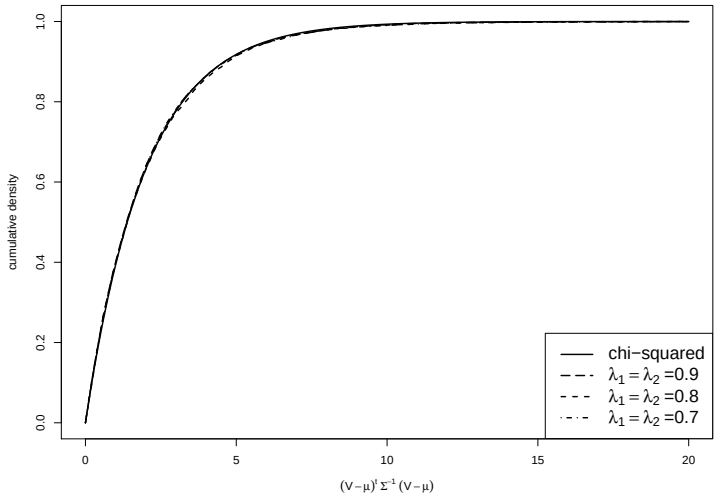

Figure 2 Cumulative density of χ^2_2 and $(\boldsymbol{V}-\boldsymbol{\mu})'\hat{\Sigma}^{-1}(\boldsymbol{V}-\boldsymbol{\mu})$ with different values of λ_1 and λ_2. Assume $\mu_{+,T}=2$, $\mu_{+,C}=\mu_{-,T}=\mu_{-,C}=1$, $\sigma_{+,T}=\sigma_{+,C}=\sigma_{-,T}=\sigma_{-,C}=1$, $N=100$, $\gamma=0.3$. The result is computed by $1,000,000$ simulation.

4 SIMULATION

In this section, simulations are used to assess finite sample properties of the asymptotic normal distribution in (3). Note that if normality in (3) holds, the statistic $(\boldsymbol{V}-\boldsymbol{\mu})'\hat{\Sigma}^{-1}(\boldsymbol{V}-\boldsymbol{\mu})$ should follow a chi-squared distribution with 2 degrees of freedom, i.e., the density of the statistic $(\boldsymbol{V}-\boldsymbol{\mu})'\hat{\Sigma}^{-1}(\boldsymbol{V}-\boldsymbol{\mu})$ should be close to that of a chi-squared distribution when the sample size is large. Assume $\mu_{+,T}=2$, $\mu_{+,C}=\mu_{-,T}=\mu_{-,C}=1$, $\sigma_{+,T}=\sigma_{+,C}=\sigma_{-,T}=\sigma_{-,C}=1$, $N=100$, $\gamma=0.3$. Firstly, we assume the sensitvity and specificity are constants. For λ_1 and λ_2 equal to $0.7, 0.8$ and 0.9, we randomly draw $1,000,000$ samples and compute $(\boldsymbol{V}-\boldsymbol{\mu})'\hat{\Sigma}^{-1}(\boldsymbol{V}-\boldsymbol{\mu})$ and the results are plotted in Figure 2. It is shown that the cumulative distribution curves of the χ^2_2 approximation almost overlap with the true χ^2_2 cumulative distribution curve for all 3 sensitivity and specificity settings, indicating the asymptotic distribution in (3) is valid.

Next, we plot the confidence ellipses for the treatment effects with respect to random sample size ($N_\oplus \sim Binom(N, p_\oplus)$ with $\gamma=0.3$) and fixed sample size ($N_\oplus=30$) using the simulation setting above. The reason we choose to use $N_\oplus=30$ for the fixed sample size design is that for $\gamma=0.3$, the expectation of $N_\oplus \sim Binom(N, p_\oplus)$ with perfect biomarker is 30. Different levels of λ_1 and λ_2 are used to illustrate the changes of confidence region with respect to sensitivity and specificity of biomarkers (Figure 3). It is clear that when the sensitivity and specificity are close to 1, the confidence ellipses are similar to those with perfect biomarker (i.e. $\lambda_1=\lambda_2=1$), and the confidence ellipse expand when the sensitivity and specificity decrease. This coincides with our expectation, as when the uncertainty increases the confidence region should expand to meet the same coverage requirement.

The coverage probability and area of the 95% confidence ellipse and simultaneous confidence intervals with total sample size 100 and 1000 for normally distributed data are provided in Table 2. It is shown that the coverage probability is positively related to the sample size while the area in the confidence region is negatively associated with both the sample size and the sensitivity and specificity of the biomarker. When the sample size increases, the precision improves and hence the coverage probability of the confidence region and the exact simultaneous confidence intervals are closer to the nominal 95% level. Furthermore, from Table 2 it is shown that the areas of the confidence regions change faster if we vary the specificity

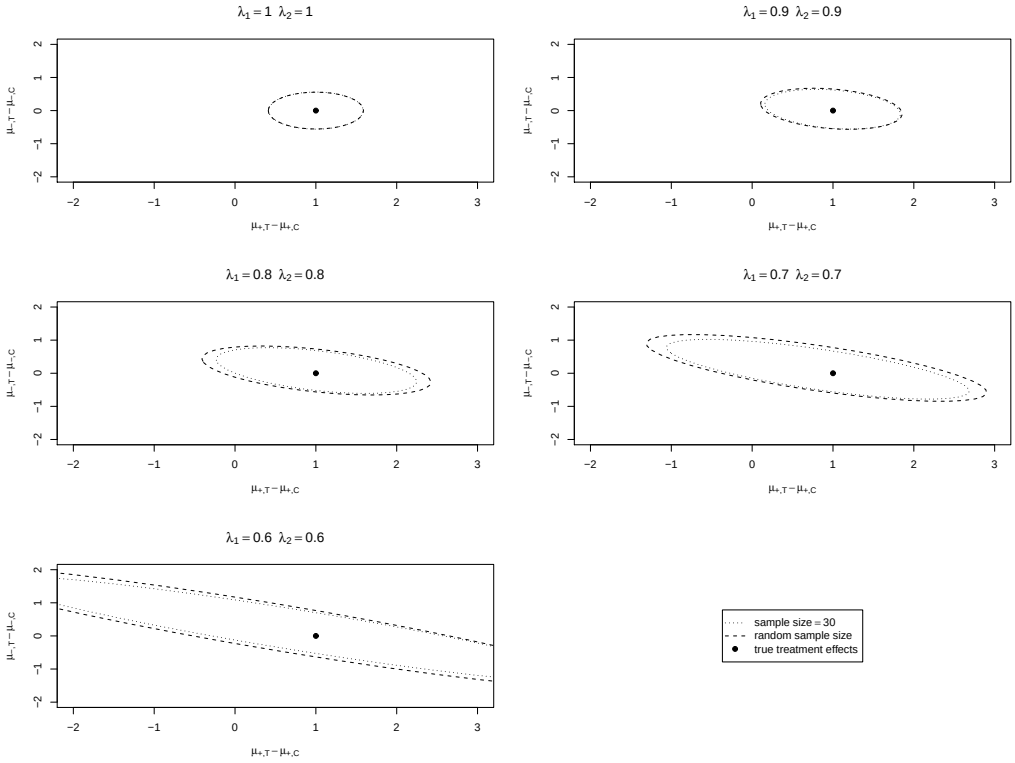

Figure 3 Plots of confidence regions of treatment effects in the combination of random sample size ($N_{\oplus} \sim Binom(N, p_{\oplus})$) or fixed sample size ($N_{\oplus} = Np_{\oplus}$) and different levels of sensitivity and specificity of the biomarker. The simulation setting is $N = 100$, $\gamma = 0.3$, $\mu_{+,T} = 2$, $\mu_{+,C} = \mu_{-,T} = \mu_{-,C} = 1$ and $\sigma_{+,T} = \sigma_{+,C} = \sigma_{-,T} = \sigma_{-,C} = 1$.

while fixing the sensitivity, than if we vary the sensitivity and fix the specificity, i.e., specificity has a larger impact on area size compared to sensitivity. This is because in the simulation setting, we assume the prevalence of the target population with true positive biomarker, γ, is equal to 0.3, hence misclassification is more sensitive with specificity than sensitivity. For example, if the sensitivity decrease by 10% then $0.3 \times 10\% = 3\%$ more patients will be misclassified. But if the specificity decreases by 10%, there will be $(1 - 0.3) \times 10\% = 7\%$ more patients misclassified. On the other hand, if we assume $\gamma < 0.5$, the impact of specificity and sensitivity on the confidence region will be the other way round; if $\gamma = 0.5$, the sensitivity and specificity will have equal impact.

The coverage probability of the confidence region versus the sample size is depicted in Figures 4. It is shown that the coverage probability increases quickly when the total sample size increases from 50 to 200 and then slowly approaches the true coverage. The coverage probabilities of both the Bonferroni SCIs and the SCIs without shrinkage tend to go above 95% as they are conservative confidence intervals, that is, in theory the true coverage should be greater than 95%. However, when the sample size is small (< 200), it seems the Bonferroni SCIs is better than the exact SCIs in the sense that it provides a better coverage probability. If one would like a region with guaranteed confidence level, then the SCIs without shrinkage might be the choice as they are almost always over-coverage. The area of the confidence regions increase when the sensitivity and specificity or the sample size decrease because, as is pointed out earlier, when the uncertainty increases the confidence region should expand to meet the same coverage requirement.

More simulation scenarios with constant sensitivity and specificity are included in the supplementary materials: Table S1 and Figure S1 in Section C show that the coverage probabilities of the confidence regions changed very slightly when varying the prevalence of the positive biomarker among the population, and the area of the confidence regions decreases significantly when the prevalence increases towards 50%; Table S2 illustrates the influence of variance on the area of the confidence regions while the confidence level is less affected; Table S3 compares the coverage level and area of the regions in these situations: (1) There are treatment effects in both subgroups, in one of the subgroups and in none of the subgroups. It seems that the coverage probability is similar among the three situations, and the area is larger in the case where treatment effect only exist in one of the subgroups.

Next, we assume the sensitivity and specificity are random variables rather than constants. Table S4 in Section C in the supplementary materials shows the resulting coverage and area of the confidence regions with different combinations of λ_1 and λ_2. Note that here λ_1 and λ_2 represent the expected values of the sensitivity and specificity. The relative performance of the four confidence regions is similar to when we assume constant sensitivity and specificity: The SCI without shrinkage is the largest in both the coverage and the area and the only one that guarantees the nominal confidence level (95%). Comparing with Table 4, it shows that the coverages in the random sensitivity and specificity case are smaller than if those are constants. For example, if we assume the sensitivity and specificity are both 0.9 then the coverage of the confidence ellipse is 94.8 (Table 4) comparing with 94.5 if we assume the sensitivity and specificity are random variables with both expectations equal to 0.9 (Table S4). The drop in coverage is caused by approximating the distribution of m_1 and m_2 when deriving the variance-covariance matrix of the test statistic $\boldsymbol{V}$.

For the case of binary response data, we assume the true binary response rates for the relevant treatment arm and biomarker subgroup are $\pi_+^T = 0.9$, $\pi_+^C = \pi_-^T = \pi_-^C = 0.2$ in order to explore the 'boundary effect'. The sensitivity and specificity are assumed equal for simplicity and the added number of success(es) and failure(s) in each arm, t, varies from 0.1 to 2. Note that it is not possible to consider the case where t=0 since, for small sample sizes, there is a non-negligible chance the confidence region will be undefined due to a zero count in one of the observed groups. The prevalence of true positive biomarker is 0.3, the same as in the normal case. The coverage probability and areas of the confidence regions, based on 100,000 simulations for each setting, are given in Table 3. The results show a similar trend as in the normal case if we change the sensitivity/specificity or the sample size for a fixed t. When t increases, the coverage probability increases and so is the area in the confidence region. Further more, the increment is smaller for a total sample size of 1000 than that of 100. This is nature due to the fact that the proportion of $4 \times t$ in 1000 is smaller than if it is in 100 and so is its influence on the confidence region. For a sample size of 100, it seems $t = 1$ is a reasonable choice as the nominal confidence level is guaranteed in all four confidence regions, at least for a reasonable sensitivity and specificity. As for $N = 1000$, $t = 1$ gives good coverage for the case when the sample sizes in the subgroups are random. However, if the subgroup sample size is prefixed, the Bonferroni SCIs is recommended as its coverage is the closest to the nominal confidence level while only the SCI without shrinkage guarantees that level. The coverage probability versus t for a total sample size 100 is also plotted in Figure 5.

In general, we find throughout our simulations that the confidence ellipse and the exact SCIs are generally close to the nominal 95% confidence level but sometimes under cover, and the SCIs without shrinkage is almost always substantially conservative, especially when the Normal approximation is accurate. The Bonferroni SCIs is generally be close to nominal coverage without systematic bias and hence the best choice for two-subgroup problems. The main factor that affects the coverage probability is the sample size – a greater sample size is associated with better coverage. The area of the region is affected by many factors: the prevalence of the positive biomarker, the variance of the sample comparing to the mean and the sample size, which all contribute to the variability of the estimates. Hence the area of the region is negatively associated with the variability in order to satisfy the coverage requirement.

Table 2 Coverage probability%(area) of the 95% confidence region with prevalence $\gamma = 0.3$, true means $\mu_{+,T} = 2$, $\mu_{+,C} = \mu_{-,T} = \mu_{-,C} = 1$, communal variance $\sigma = 1$ using $1,000,000$ simulations

λ_1	λ_2	$N_\oplus = 30$, $N = 100$				$N_\oplus = 300$, $N = 1000$			
		Confidence Ellipse	Simultaneous Confidence Intervals			Confidence Ellipse	Simultaneous Confidence Intervals		
			exact	without shrinkage	Bonferroni		exact	without shrinkage	Bonferroni
1	1	93.9(1.62)	93.9(1.72)	96.3(2.07)	94.0(1.73)	94.9(0.16)	95.0(0.17)	97.1(0.21)	95.0(0.18)
0.95	0.95	93.9(1.92)	93.9(2.05)	96.3(2.46)	94.0(2.07)	94.9(0.19)	94.9(0.21)	97.1(0.25)	95.0(0.21)
0.9	0.9	93.9(2.28)	93.9(2.48)	96.3(2.98)	94.0(2.5)	94.9(0.23)	94.9(0.25)	97.1(0.3)	95.0(0.25)
0.8	0.8	93.9(3.27)	94.0(3.88)	96.5(4.74)	94.3(3.97)	94.9(0.33)	94.9(0.39)	97.2(0.48)	95.2(0.4)
0.9	1	93.9(1.71)	93.9(1.82)	96.3(2.18)	94.0(1.83)	94.9(0.17)	94.9(0.18)	97.1(0.22)	95.0(0.18)
0.8	1	94.0(1.79)	93.9(1.92)	96.3(2.3)	94.0(1.93)	94.9(0.18)	94.9(0.19)	97.1(0.23)	95.0(0.2)
1	0.9	93.9(2.08)	93.9(2.22)	96.3(2.66)	94.0(2.23)	94.9(0.21)	94.9(0.22)	97.1(0.27)	95.0(0.23)
1	0.8	93.9(2.51)	93.9(2.71)	96.3(3.25)	94.1(2.73)	94.9(0.25)	94.9(0.27)	97.1(0.33)	95.0(0.28)
λ_1	λ_2	$N_\oplus \sim Binom(N, p_\oplus)$, $N = 100$				$N_\oplus \sim Binom(N, p_\oplus)$, $N = 1000$			
		Confidence Ellipse	Simultaneous Confidence Intervals			Confidence Ellipse	Simultaneous Confidence Intervals		
			exact	without shrinkage	Bonferroni		exact	without shrinkage	Bonferroni
1	1	94.8(1.74)	94.7(1.85)	96.8(2.22)	94.7(1.86)	95.0(0.17)	95.0(0.18)	97.1(0.21)	95.0(0.18)
0.95	0.95	94.8(2.03)	94.7(2.17)	96.8(2.6)	94.7(2.18)	95.0(0.19)	95.0(0.21)	97.1(0.25)	95.1(0.21)
0.9	0.9	94.8(2.36)	94.7(2.57)	96.8(3.09)	94.8(2.59)	95.0(0.22)	95.0(0.24)	97.1(0.29)	95.1(0.25)
0.8	0.8	94.8(3.31)	94.8(3.93)	97.1(4.8)	95.2(4.03)	95.0(0.31)	95.0(0.37)	97.2(0.46)	95.3(0.38)
0.9	1	94.7(1.90)	94.6(2.03)	96.7(2.43)	94.6(2.03)	95.0(0.18)	94.9(0.19)	97.1(0.23)	95.0(0.19)
0.8	1	94.6(2.08)	94.5(2.23)	96.6(2.68)	94.6(2.25)	95.0(0.2)	95.0(0.21)	97.1(0.25)	95.0(0.21)
1	0.9	94.8(2.11)	94.7(2.26)	96.8(2.71)	94.8(2.27)	95.0(0.2)	95.0(0.21)	97.1(0.26)	95.1(0.22)
1	0.8	94.9(2.47)	94.8(2.71)	96.9(3.26)	95.0(2.73)	95.0(0.24)	95.0(0.26)	97.2(0.31)	95.1(0.26)

Table 3 Coverage probability%(area) of the 95% confidence region for treatment effects in biomarker subgroups with binary response. The simulation setting: (1) true binary response rates are $\pi_+^T = 0.9$, $\pi_+^C = \pi_-^T = \pi_-^C = 0.2$, (2) sensitivity and specificity are $\lambda_1 = \lambda_2 = \lambda$, (3) added number of success(es) and failure(s) is t, (4) the prevalence of true positive biomarker is 0.3 and (5) simulation number is $100,000$.

λ	t	$N_\oplus = 30$, $N = 100$				$N_\oplus = 300$, $N = 1000$			
		Confidence	Simultaneous Confidence Intervals			Confidence	Simultaneous Confidence Intervals		
		Ellipse	exact	without shrinkage	Bonferroni	Ellipse	exact	without shrinkage	Bonferroni
1	0.1	91.8(0.22)	92.0(0.23)	93.3(0.28)	92.2(0.24)	94.8(0.02)	95.0(0.02)	97.0(0.03)	95.1(0.02)
	0.5	95.8(0.24)	95.9(0.25)	97.6(0.3)	95.9(0.25)	94.8(0.02)	94.6(0.02)	96.9(0.03)	94.6(0.02)
	1	97.1(0.25)	97.1(0.27)	98.7(0.32)	97.1(0.27)	95.3(0.02)	95.3(0.02)	97.2(0.03)	95.3(0.03)
	2	96.8(0.28)	97.0(0.3)	98.6(0.35)	97.0(0.3)	95.2(0.02)	95.3(0.03)	97.3(0.03)	95.4(0.03)
0.9	0.1	92.3(0.34)	91.7(0.37)	95.0(0.45)	91.9(0.38)	94.1(0.04)	94.2(0.04)	96.5(0.05)	94.3(0.04)
	0.5	94.2(0.36)	94.4(0.39)	96.4(0.47)	94.5(0.39)	94.3(0.04)	94.3(0.04)	96.7(0.05)	94.4(0.04)
	1	96.3(0.37)	96.2(0.41)	97.9(0.49)	96.4(0.41)	94.7(0.04)	94.7(0.04)	96.9(0.05)	94.8(0.04)
	2	97.2(0.4)	97.0(0.43)	98.6(0.52)	97.1(0.43)	94.7(0.04)	94.8(0.04)	97.0(0.05)	94.9(0.04)
0.8	0.1	90.9(0.51)	91.7(0.61)	94.7(0.75)	92.3(0.63)	93.3(0.05)	93.6(0.06)	96.2(0.08)	94.0(0.07)
	0.5	94.2(0.53)	94.1(0.63)	96.5(0.77)	94.4(0.64)	93.6(0.05)	93.7(0.06)	96.3(0.08)	94.1(0.07)
	1	95.5(0.54)	95.6(0.65)	97.5(0.79)	95.8(0.66)	93.9(0.05)	94.0(0.06)	96.5(0.08)	94.3(0.07)
	2	96.8(0.57)	96.8(0.68)	98.4(0.83)	97.1(0.69)	94.1(0.05)	94.2(0.06)	96.7(0.08)	94.6(0.07)

λ	t	$N_\oplus \sim Binom(N, p_\oplus)$, $N = 100$				$N_\oplus \sim Binom(N, p_\oplus)$, $N = 1000$			
		Confidence	Simultaneous Confidence Intervals			Confidence	Simultaneous Confidence Intervals		
		Ellipse	exact	without shrinkage	Bonferroni	Ellipse	exact	without shrinkage	Bonferroni
1	0.1	92.1(0.24)	91.8(0.25)	93.6(0.30)	91.9(0.25)	94.8(0.02)	94.8(0.02)	96.9(0.03)	94.8(0.02)
	0.5	96.2(0.25)	96.1(0.27)	97.5(0.32)	96.1(0.27)	95.0(0.02)	95.0(0.02)	97.0(0.03)	95.0(0.03)
	1	97.6(0.27)	97.6(0.29)	98.8(0.35)	97.6(0.29)	95.4(0.02)	95.4(0.03)	97.4(0.03)	95.4(0.03)
	2	97.6(0.3)	97.8(0.32)	99.0(0.38)	97.8(0.32)	95.2(0.02)	95.4(0.03)	97.4(0.03)	95.5(0.03)
0.9	0.1	93.5(0.37)	93.5(0.40)	95.7(0.48)	93.6(0.41)	94.8(0.04)	94.8(0.04)	97.0(0.05)	94.9(0.04)
	0.5	95.8(0.38)	95.6(0.42)	97.4(0.5)	95.7(0.42)	95.1(0.04)	95.0(0.04)	97.1(0.05)	95.1(0.04)
	1	97.2(0.4)	97.1(0.43)	98.4(0.52)	97.2(0.44)	95.3(0.04)	95.3(0.04)	97.3(0.05)	95.4(0.04)
	2	98.0(0.42)	98.0(0.46)	99.1(0.55)	98.1(0.46)	95.3(0.04)	95.3(0.04)	97.4(0.05)	95.5(0.04)
0.8	0.1	94.1(0.55)	94.2(0.66)	96.5(0.80)	94.6(0.67)	94.8(0.05)	94.9(0.06)	97.1(0.08)	95.2(0.07)
	0.5	95.7(0.57)	95.6(0.67)	97.5(0.82)	95.9(0.69)	95.1(0.05)	95.0(0.06)	97.3(0.08)	95.3(0.07)
	1	97.0(0.58)	97.0(0.69)	98.4(0.84)	97.2(0.71)	95.2(0.05)	95.2(0.06)	97.3(0.08)	95.5(0.07)
	2	98.2(0.61)	98.1(0.72)	99.1(0.88)	98.3(0.74)	95.4(0.05)	95.4(0.06)	97.5(0.08)	95.7(0.07)

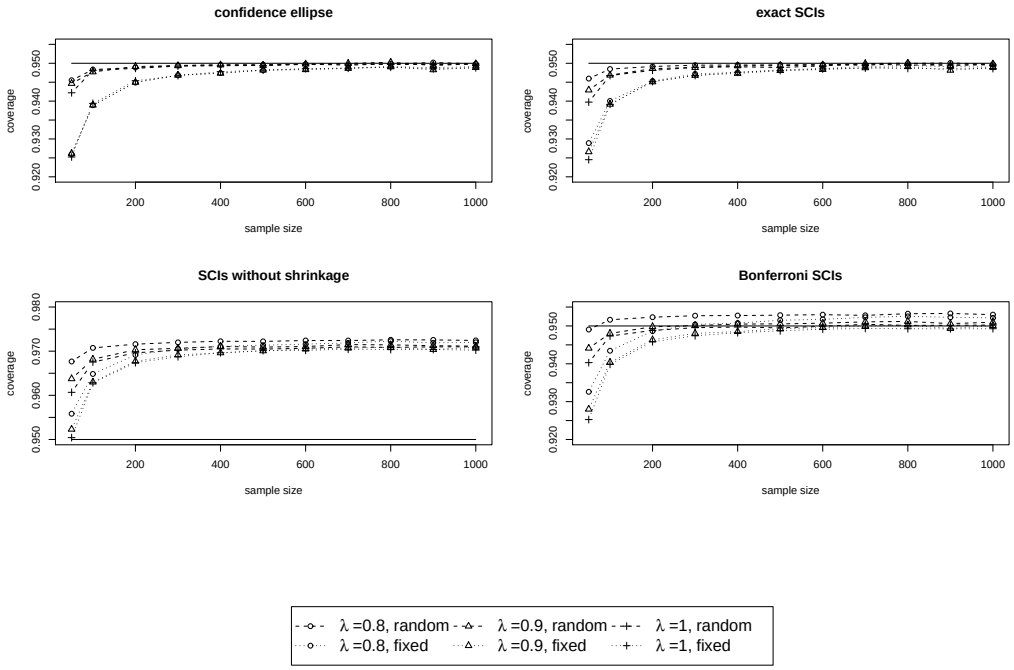

Figure 4 Plots of coverage probability of confidence regions with $\gamma = 0.3$, $\mu_{+,T} = 2$, $\mu_{+,C} = \mu_{-,T} = \mu_{-,C} = 1$, communal variance $\sigma = 1$ and $N_{\oplus} = N\gamma$.

5 CONCLUSION AND DISCUSSION

In subgroup analysis, the confidence regions of treatment effects are frequently used to evaluate the accuracy of the estimate. However, when constructing confidence regions the sensitivity and specificity of the biomarker is typically ignored. As is shown in the example in Section 3, the estimates of treatment effect in the observed biomarker subgroups are different from those in the true biomarker subgroups. If the interest is in constructing confidence regions for the treatment effects in the observed (detected) biomarker subgroups, then clearly there's no need to take into consideration the impact of the sensitivity and specificity. But if the interest is in the confidence regions in the true biomarker subgroup, then the sensitivity and specificity need to be considered in order to give an accurate estimate. Hence we recommend the use of the statistics $\boldsymbol{V}$ and $\hat{\Sigma}$ when constructing confidence regions for the treatment effects in true biomarker subgroups.

In this paper, the construction of confidence regions for treatment effects in biomarker subgroups with both perfect and imperfect biomarker in preplanned clinical trials is discussed. An asymptotic normal distribution for the treatment effects is given and a confidence ellipse and three simultaneous confidence intervals based on the asymptotic distribution are constructed. The method provided in this paper is straightforward and the relevant coverage probabilities seem to be satisfactory. From the simulation result in Section 4, it is shown that for normally distributed data the exact SCIs work well if the normal approximation is accurate, otherwise (e.g. when the sample size is small) the Bonferroni simultaneous confidence intervals and simultaneous confidence intervals without shrinkage provide better (greater) coverage. For

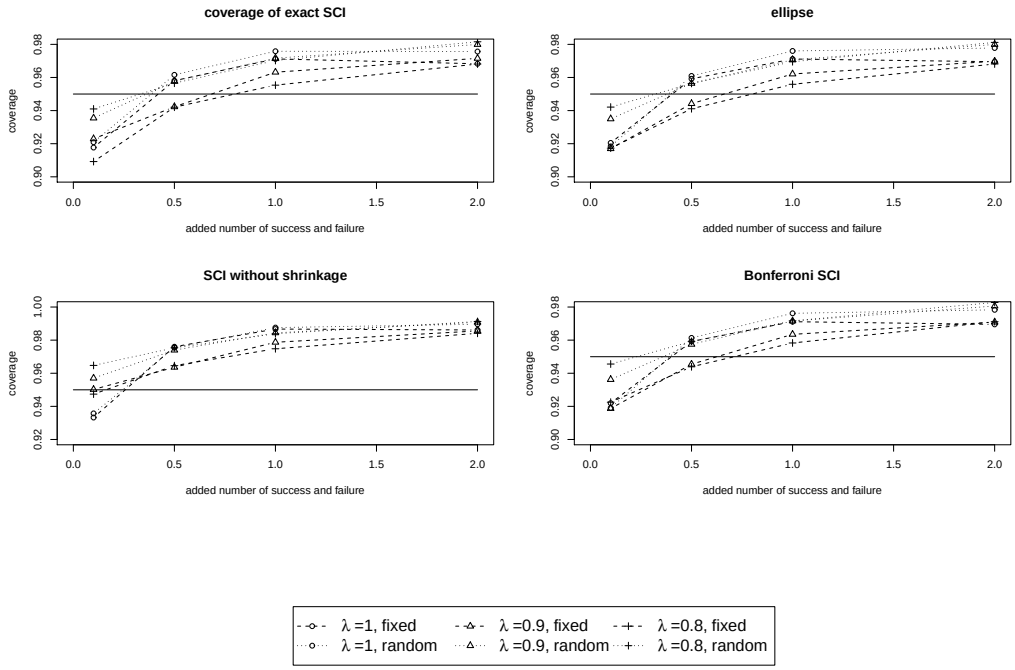

Figure 5 Plots of coverage probability of confidence regions with $\gamma = 0.3$, $N = 100$, $\pi_+^T = 0.9$, $\pi_+^C = \pi_-^T = \pi_-^C = 0.2$ using $100,000$ simulation.

binary data, adding 1 success and 1 failure to each treatment arm (that is, $t = 1$) seem to generally work well with a better balance between over-coverage and under-coverage compared to other choices of t.

Acknowledgements This work is independent research arising in part from Dr Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research and Dr Kunz's Medical Research Council fellowship (MR/M014525/1). Funding for this work was also provided by the Medical Research Council (MR/M005755/1 and MR/K025635/1). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

References

Agresti, A. and Caffo, B. (2000). Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures. *The American Statistician* **54**(4), 280-288.

Agresti, A. and Coull, B.A. (1998). Approximate Is Better than 'Exact' for Interval Estimation of Binomial Proportions. *The American Statistician* **52**(2), 119-126.

Anbar, D. (1983). On Estimating the Difference between Two Probabilities, with Special Reference to Clinical Trials. *Biometrics* **39**(1), 257-262.

Baker, S., Kramer, B., Sargent, D. and Bonetti, M. (2012). Biomarkers, subgroup evaluation, and clinical trial design. *Discovery Medicine* **13**(70), 187-192.

Brown, L.D., Cai, T. and DasGupta, A. (2001) Interval Estimation for a Binomial Proportion.Statistical Science. *Statistical Science* **16**(2), 101-117.

Brusselle, G.G., Vanderstichele, C., Jordens, P., Slabbynck, H., Ringoet, V., Verleden, G., Demedts, I.K., Verhamme, K., Delporte, A., Demeyere, B., Claeys, G., Boelens, J., Padalko, E., Verschakelen, J., Van Maele, G., Deschepper, E. and Joos, G.F. (2013). Azithromycin for prevention of exacerbations in severe asthma (AZISAST): a multicentre randomised double-blind placebo-controlled trial. *Thorax* **68**, 322-329.

Corren, J., Lemanske, R.F.Jr, Hananian, N.A., Korenblat, P.E., Parsey, M.V., Arron, J.R., Harris, J.M., Scheerens, H., Wu, L.C., Su, Z., Mosesova, S., Eisner, M.D., Bohen, S.P., Matthews, J.G. (2011). Lebrikizumab treatment in adults with Asthma. *The New England Journal of Medicine* **365**(12), 1088-98.

Dunn, J.O. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association* **56**(293), 5264.

Friede, T., Parsons, N. and Stallard, N. (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine* **31**(30), 43094320.

Freidlin, B., McShane, L.M. and Korn, E.L. (2010). Randomized clinical trials with biomarkers: Design issue. *Journal of the National Cancer Institute* **102**(3), 152-160.

Freidlin, B., McShane, L.M., Polley, M.C. and Korn, E.L. (2012). Randomized Phase II Trial Designs With Biomarkers. *Journal of Clinical Oncology* **30**(26), 3304-3309.

Genz, A. and Bretz, F. (2009). Computation of Multivariate Normal and t Probabilities. *Springer Science and Business Media* **195**.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2016). mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-5. http://CRAN.R-project.org/package=mvtnorm.

Henry, N.L. and Hayes, D.F. (2012). Cancer biomarkers. *Molecular Oncology* **6**(2), 140146.

Jubb, A.M., Miller, K.D., Rugo, H.S., Harris, A.L., Chen, D., Reimann, J.D., Cobleigh, M.A., Schmidt, M., Langmuir, V.K., Hillan, K.J., Chen, D.S. and Koeppen, H. (2011). Impact of Exploratory Biomarkers on the Treatment Effect of Bevacizumab in Metastatic Breast Cancer. *Clinical Cancer Research* **17**(2), 372-381.

Kaplan, R., Maughan, T., Crook, A., Fisher, D., Wilson, R., Brown, L. and Parmar, M. (2013). Evaluating Many Treatments and Biomarkers in Oncology:A New Design. *Journal of Clinical Oncology* **31**(36), 4562-8.

Kelloff, G.J and Sigman, C.C. (2012). Cancer biomarkers: selecting the right drug for the right patient. *Nature Reviews Drug Discovery* **11**(3), 201-214.

Liu, C., Liu, A., Hu, J., Yuan, V. and Halabi, S. (2014). Adjusting for misclassification in a stratified biomarker clinical trial. *Statistics in Medicine* **33**(18), 3100-3113.

Mandrekar, S.J. and Sargent, D.J. (2009). Clinical Trial Designs for Predictive Biomarker Validation: Theoretical Considerations and Practical Challenges. *Journal of Clinical Oncology* **27**(24), 4027-4034.

Newcombe, R.G. (1998). Statistics in Medicine. Improved Confidence Intervals for the Difference between Binomial Proportions Based on Paired Data. *Statistics in Medicine* **17**, 2635-2650.

Simon, R. (2010). Clinical trials for predictive medicine: new challenges and paradigms. *Clinical Trials* **7**(5): 516524.

Spienssens, B. and Debois, M. (2010). Adjusted significance levels for subgroup analysis in clinical trials. *Contemporary Clinical Trials* **31**(6), 647-656.

Wilson, E.B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association* **22**(158), 209-212.

