Extending multiple testing with unknown test dependency via the CoCo test with applications to cancer studies

Abstract

Multiple testing problems are ubiquitous in clinical and scientific investigations. Central to multiple testing is to control for the type I error. The behavior of multiple testing procedures for α-control when the tests are independent or dependent but with a known joint distribution is relatively well known. When the joint distribution of test statistics is unknown, one can still guarantee the α-control, if the *positive dependency through stochastic ordering* (PDS) condition is satisfied. Despite the frequent occurrence of unknown test dependency in multiple testing and the importance of the PDS condition in endorsing its validity, little do we know about how to verify the condition. Here, we develop a new nonparametric statistical test, called the *CoCo test*, that can validate the condition of PDS, through which one can control for α regardless of the prior knowledge of the dependency between test statistics. Simulation studies show that the CoCo test can faithfully detect the violation of the PDS condition or lack thereof. To further evaluate the efficacy of the CoCo test, we apply it to investigate two meta-analyses in oncology. An R package `cocotest` to implement the proposed methodology is available at CRAN.

Keywords. Clinical trials, Concordance, Dependence test, Hazard ratio, Hochberg procedure, Multiple testing procedure.

1 Introduction

Modern clinical trials consider multiple primary outcomes in their designs. This is because a single primary endpoint generally only evaluates and reveals a limited dimension of the effect of a medical intervention (Hamasaki et al., 2018). Two chief primary endpoints in clinical trials in general, and oncology trials in particular, are the overall survival (OS) and the progression-free survival (PFS). The individual importance of OS and PFS in clincal trails have been extensively studied. Recent years, however, have seen an increasing interest in investigating the association between the OS and PDF (Buyse et al., 2007; Halabi et al., 2009; Amir et al., 2012; Adunlin et al., 2015; Gyawali et al., 2018; Hess et al., 2019; Belin et al., 2020; Pasalic et al., 2020; Chase et al., 2023; Courtinard et al., 2023). The reasons are twofold. First, clinicians want to determine the circumstances under which PFS can act as a validated surrogate endpoint for OS, thus establishing the clinical efficacy of a treatment. Second, one wants to provide a validation of the dependence assumptions underlying widely applied multiple testing procedures in clinical trials with two or more endpoints.

While the first aforementioned reason is relatively straightforward, the second reason perhaps needs some explanation. Consider the Hochberg procedure (Hochberg, 1988), one of the most commonly used adjustment approaches in multiple testing to control for the overall type I error α. When there are multiple endpoints and the tests are independent, the Hochberg procedure provides adequate overall α control. When two or more primary endpoints in confirmatory clinical trials are not independent, however, a common misconception about the Hochberg procedure is that it gurantees adequate overall α control when the test statistics are positively correlated. Yet, this is only true when the distribution of the endpoints and their correlation structure satisfies certain positive dependence assumptions (FDA, 2022). More specifically, when there are multiple endpoints tests and the tests are either independent or positively correlated with a known joint distribution (e.g., the test statistics are jointly bivariate normal), the Hochberg procedure provides adequate overall α-control (FDA, 2022). When the joint distributions of test statistics are unknown, a valid control of overall type I error is only guaranteed if the *positive dependency through stochastic ordering* (PDS) condition is satisfied (Sarkar and Chang, 1997; Sarkar, 1998; Benjamini and Yekutieli, 2001; Sarkar, 2002).

Despite the frequent occurrence of unknown test dependency in multiple testing, and, therefore, the cardinal importance of the PDS condition in guaranteeing the overall control of the type I error, little do we know about how to statistically check the condition. To address this issue, here we design a nonparametric statistical test, called the CoCo test, based on ranked correlation coefficient and a simple, yet effective, algebraic arrangement of the Spearman's ρ and Kendall's τ, to evaluate and test the PDS condition in multiple testing problems. Operationally, we first introduce the test, and then examine and demonstrate its efficacy using simulation studies and apply it to investigate two representative primary endpoints (i.e., OS and PFS) in two meta-analyses with 84 cancer clinical trials. Our exploration suggest the utility of the CoCo test in validating the PDS condition in multiple testing. With this new, simple, and general statistical tool, one may witness expanded applications of the Hochberg procedure in a wide range of "non-standard" cases when the tests are either correlated under an uncommon distribution or with an unknown correlation structure; one may also extend other multiplicity correction methods (e.g., Benjamini and

Hochberg (1995)'s false discovery rate controlling procedure) by embedding the CoCo test into them.

The rest of this paper is structured as follows. Section 2 outlines the PDS condition and introduces the CoCo test to assess the PDS condition. Section 3 performs simulation studies to investigate the performance of the CoCo test for controlling the type I error rate and statistical power under various scenarios. Section 4 presents two case studies on 84 cancer clinical trials including 72 trials on patients with metastatic breast cancer treated with anthracyclines, taxanes, or targeted therapies, and 12 trials on patients with advanced solid tumors treated with programmed cell death protein 1 (PD-1) inhibitors, and examines the association between two primary outcomes, PFS and OS, in both studies. Section 5 concludes the paper with discussions and future directions.

2 The PDS condition and the CoCo Test

In this section, we introduce the definition of the *positive dependency through stochastic ordering* (PDS) condition and the methodological framework of the nonparametric CoCo test, based on ranked correlation coefficient, to check the PDS condition.

The general PDS condition. Let us begin with a random vector $\boldsymbol{W} = (W_1, W_2, \ldots, W_m)$. We say $\boldsymbol{W}$ is positively dependent through stochastic ordering (PDS) if, for any $i = 1, 2, \ldots, m$, the conditional expected value

$$\mathbb{E}\left[g(\boldsymbol{W}) \mid W_i = w\right]$$

is non-decreasing in w, for any non-decreasing function g.

The PDS condition with two variables. A special case of the PDS condition is regarding a bivariate random vector (X, Y). More specifically, we say X and Y are PDS if: $\mathbb{E}\left[g(Y) \mid X = x\right]$ and $\mathbb{E}\left[g(X) \mid Y = y\right]$ are non-decreasing in x and y, respectively, for any non-decreasing function g. This special PDS condition is also known as positive regression dependence (Tukey, 1958; Lehmann, 1966; Shaked, 1977; Block et al., 1985).

The PDS condition with a multivariate normal distribution. Another special case of the PDS distribution is the multivariate normal distribution with non-positive non-diagonal elements in the precision matrix $\boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Sigma}$ is the covariance matrix (Karlin and Rinott, 1980). A multivariate normal distribution may not satisfy the PDS condition when all the elements in the covariance matrix $\boldsymbol{\Sigma}$ are non-negative.

The PDS condition is related to several other positive dependence structures, including the positive quadrant dependence (PQD) and positive association dependence (PAD), but it is more strigent. See Tong (1990), Joe (1997) and Gou (2023) for a comprehensive summary of these positive dependence structures.

The PDS condition is critical to ensure that a multiple testing procedure provides valid control of the type I error. Here, we use Hochberg (1988) procedure as a demonstration, because it is a widely used multiplicity adjustment method in confirmatory clinical trials involving multiple endpoints. One, however, can relatively straightforwardly extend our arguments to other multiple testing procedures, including Hommel (1988)'s procedure, Benjamini and Hochberg (1995)'s FDR controlling procedure, etc.

Let us now proceed with the Hochberg procedure. Consider m hypotheses $H_1, \ldots, H_m$ and the associated p-values $p_1, \ldots, p_m$. Denote the ordered p-values by $p_{(1)} \leq \cdots \leq p_{(m)}$. The Hochberg procedure rejects $H_{(1)}, \ldots, H_{(k)}$, where k is the largest index such that $p_{(k)} \leq \alpha/(m-k+1)$. When test statistics are (a) independent, (b) dependent but under some known joint distribution that satisfies some specific condition (such as joint normal with a precision matrix where all non-diagonal elements are positive or zero), or (c) dependent with an unknown joint distribution but the dependency satisfies the PDS condition, the Hochberg procedure strongly controls the familywise error rate (Hochberg and Tamhane, 1987), since it is a shortcut of the closed testing procedure (Marcus et al., 1976) using the Simes (1986) method to test the intersection hypotheses.

Whereas (a) and (b) are clear and straightforward to verify, case (c) seems labyrinthine. Naturally, one would ask, could we relax the PDS condition to some other (less stringent) positive dependence assumption, such as the PDQ and PAD conditions, with known statistical properties? The short answer is no. More concretedly, Gou and Tamhane (2018) and Gou (2023) showed that the PDS condition may not be relaxed to some other positive dependence assumptions. In other words, a more relaxed positively correlated test statistics may not guarantee the validity of the Hochberg procedure.

The key objective of this paper, therefore, aims to design a test that can guarantee strong familywise error rate control during a multiple testing correction procedure, such as the Hochberg procedure, when simply studying positive correlation coefficients between test statistics do not guarantee the desired α-control that is the strong familywise error rate control. To solve this, we draw insights from nonparemetric statistics and particularly the ranked correlation coefficients (CoCo). Before introducing the CoCo test, let us first review the sample correlation coefficients and measures of variability, including Pearson's r, Spearman's ρ, and Kendall's τ (Spearman, 1904; Kendall, 1938, 1945). Although their forms are well-known, presenting them side-by-side would help to peek into the construction of the CoCo test, a simple, yet effective, algebraic combination of the Spearman's ρ and Kendall's τ.

Pearson's r. For a sample of n data pairs $\{X_i, Y_i\}_{i=1}^n$ that are independent and identically distributed, the sample Pearson correlation coefficient is:

$$\widehat{r} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}$$

where $\overline{X}$ and $\overline{Y}$ are the sample means.

Spearman's ρ. The sample Spearman's ρ is defined as the sample Pearson's r between the rank variables and can be computed using the formula

$$\widehat{\rho} = 1 - \frac{6\sum_{i=1}^n (R(X_i) - R(Y_i))^2}{(n-1)n(n+1)}$$

where $R(X_i)$ is the rank of X_i among $\{X_1, \ldots, X_n\}$, and analogously for $R(Y_i)$.

Kendall's τ. The sample Kendall's τ considers the numbers of concordant pairs and discordant pairs and is computed as

$$\widehat{\tau} = \frac{\sum_{i \neq j} \operatorname{sign}(X_i - X_j) \cdot \operatorname{sign}(Y_i - Y_j)}{n(n-1)}$$

where sign() stands for the signum function.

Prerequisite of the CoCo test. Spearman's ρ and Kendall's τ are nonparametric, meaning they are invariant under monotonic transformations. This invariant property makes them well-suited for constructing the CoCo tests. For positive ρ and τ, Daniels (1950) and Durbin and Stuart (1951) showed that they are constrained by the inequality

$$1 - \sqrt{2(1-\rho)} \leq \tau \leq \frac{1+2\rho}{3}, \tag{1}$$

as shown in Figure 1 by the dashed boundaries. When test statistics satisfy the PDS condition, Capéraà and Genest (1993) further showed that

$$\max\left\{\frac{\rho}{3}, 1 - \sqrt{2(1-\rho)}\right\} \leq \tau \leq \rho. \tag{2}$$

To find the bounds, Hutchinson and Lai (1990) and Hürlimann (2003) suggested a narrower region of possible values of ρ and τ when the PDS condition is satisfied:

$$1 - \sqrt{1-\rho} \leq \tau \leq \frac{\rho^2+2\rho}{3}, \tag{3}$$

as shown in Figure 1 by the solid boundaries. For the aforementioned special case involving two normally distributed variables, assuming a bivariate normal distribution, Moran (1948) showed that Spearman's $\rho = \frac{6}{\pi}\arcsin\frac{r}{2}$, and Esscher (1924) found that Kendall's $\tau = \frac{2}{\pi}\arcsin r$, where r is the bivariate normal correlation coefficient. Under the normality assumption, the relation between Spearman's ρ and Kendall's τ is

$$\tau = \frac{2}{\pi}\arcsin\left(2\sin\frac{\pi\rho}{6}\right). \tag{4}$$

as shown in Figure 1 by the dash-dotted line. When test statistics follow a bivariate normal distribution with positive r, we have $\tau \leq \rho$, and the equality holds only when $\rho = \tau = 0$ and $\rho = \tau = \pm 1$, since $\frac{d\tau}{d\rho} = \frac{2}{3}\sqrt{\frac{1-r^2/4}{1-r^2}}$ is monotonically increasing on $\rho \in [0,1]$.

The CoCo test. Using the lower boundary and the upper boundary on τ in terms of ρ, we have the type I and type II test statistics based on the estimated Spearman's ρ and Kendall's τ, which are

$$\text{C1} = \frac{(1-\widehat{\tau})^2}{1-\widehat{\rho}}$$

and

$$\text{C2} = \frac{1+3\widehat{\tau}}{(1+\widehat{\rho})^2}.$$

We call them CoCo–1 (C1) statistic and CoCo–2 (C2) statistic, respectively. Under the null hypothesis where the PDS condition holds, both C1 and C2 statistics are less than or equal to one.

Formally, the corresponding CoCo hypothesis tests are presented in Eqs. (5) and (6).

$$\text{CoCo–1 test}: \quad H_0: \frac{(1-\tau)^2}{1-\rho} \leq 1 \quad \text{versus} \quad H_a: \frac{(1-\tau)^2}{1-\rho} > 1 \tag{5}$$

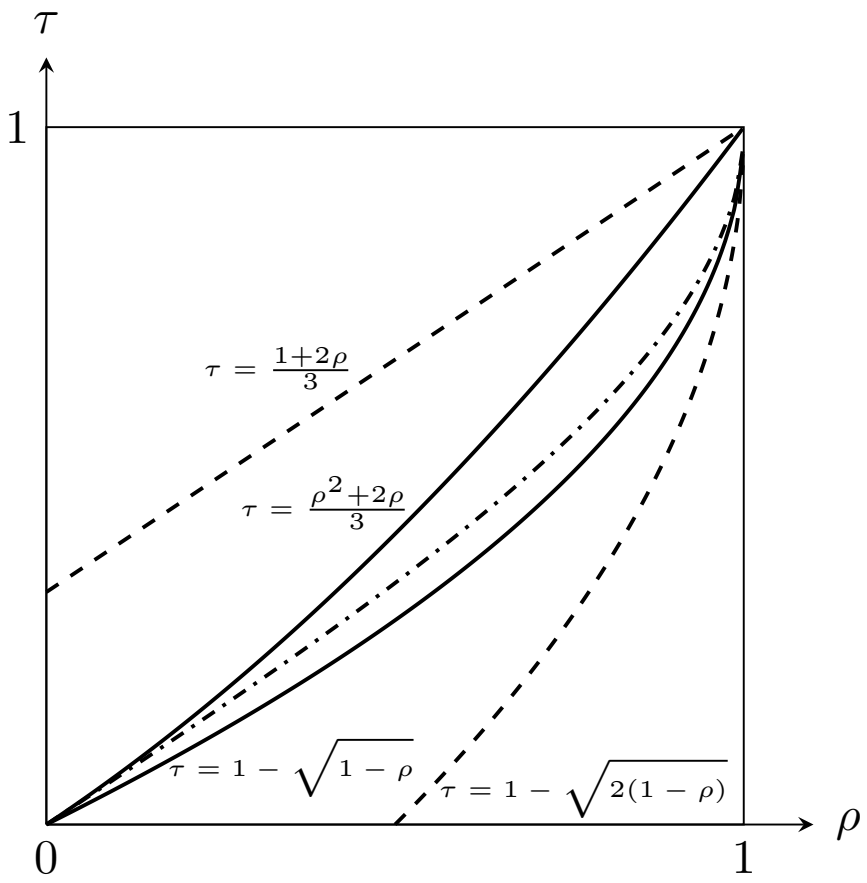

Figure 1: Boundaries of Kendall's τ vis-à-vis Spearman's ρ. The figure contains two bounded regions and one line: the region corresponding to the "no dependence condition" (i.e., the area between two dashed lines), the region corresponding to the "PDS condition" (i.e., the area between two solid lines), and the line corresponding to the bivariate normal distribution (i.e., the dash-dotted line).

$$\text{CoCo–2 test}: \quad H_0: \frac{1+3\tau}{(1+\rho)^2} \leq 1 \quad \text{versus} \quad H_a: \frac{1+3\tau}{(1+\rho)^2} > 1 \tag{6}$$

The test statistics of the CoCo test are constructed based on the boundaries in Eq. (3). One can test the null hypotheses in Eqs. (5) and (6) by computing the one-sided confidence intervals of C1 statistic $\frac{(1-\widehat{\tau})^2}{1-\widehat{\rho}}$ and C2 statistic $\frac{1+3\widehat{\tau}}{(1+\widehat{\rho})^2}$ using bootstrap. Here we apply the bias-corrected and accelerated bootstrap confidence intervals (Efron, 1987), since it provides accurate results across a wide range of settings and is second-order accurate (Hesterberg, 2015).

3 Simulation Studies

After presenting the foundation, construction, and intuition of the CoCo test, here we perform simulation studies to investigate how the proposed test controls the type I error and how it affects the results of statistical power analysis.

Following the logic in Section 2, we consider, without loss of generality, a simulation study for evaluating type I error control where the underlying true distribution is bivariate normal, since a bivariate normal distribution with positive correlation coefficient r satisfies the PDS condition.

Assuming the normality, given Pearson's correlation coefficient r, one can compute Spearman's ρ and Kendall's τ as:

$$\rho = \frac{6}{\pi} \arcsin\left(\frac{r}{2}\right) \quad \text{and} \quad \tau = \frac{2}{\pi} \arcsin\left(r\right). \tag{7}$$

We simulate bivariate normally distributed random numbers $(x_1, y_1), \ldots, (x_n, y_n)$ assuming the correlation coefficient $r = 0.2$, 0.5 and 0.8, and various sample sizes: $n = 6$, 8 and 10. For each sample within these scenarios, we test the hypotheses in Eqs. (5) and (6) at level $\alpha = 5\%$ based on the bootstrap one-sided confidence intervals. One rejects the corresponding hypothesis and concludes the violation of the PDS condition, if the confidence interval does not include C1 = 1 or C2 = 1.

We replicate the simulation 3×10^4 times and report, in Table 1, the simulated type I errors using the C1 and C2 tests, along with the true ρ, τ, C1, and C2 values, where $C1 = \frac{(1-\tau)^2}{1-\rho}$ and $C2 = \frac{1+3\tau}{(1+\rho)^2}$. Our result suggests that the simulated type I error rates of the C1 and C2 tests are smaller than the nominal significance level $\alpha = 5\%$, since the true C1 and C2 values are less than the boundary value 1 for the PDS condition. Another way to see this is via the dissimilarity plot (the difference between τ and ρ) in Figure 1, where the bivariate normal line lies between the two boundaries of PDS distributions.

Next, we evaluate the statistics power using a distribution that does not satisfy the PDS condition. Specifically, consider a bivariate random variable (X, Y) with joint cumulative distribution function:

$$F(x, y) = \min\left\{x, y, \frac{x^2 + y^2}{2}\right\}, \quad (x, y) \in [0, 1]^2.$$

Table 1: Simulated type I error rate of the CoCo tests ($\alpha = 5\%$)

r	ρ	τ	true C1	true C2	n	Type I error of C1	Type I error of C2
$r = 0.2$	0.191	0.128	0.940	0.976	6	0.39%	1.50%
					8	0.32%	0.26%
					10	0.10%	0.07%
$r = 0.5$	0.483	0.333	0.859	0.910	6	0.42%	1.43%
					8	0.26%	0.19%
					10	0.19%	0.07%
$r = 0.8$	0.786	0.590	0.784	0.869	6	0.34%	1.55%
					8	0.23%	0.21%
					10	0.15%	0.07%

It follows that (X, Y) meets the positive quadrant dependence (PQD) condition - a weaker positive dependence condition than the PDS condition, but does not satisfy the PDS condition (Nelsen, 2006). Using this distribution, we simulate random numbers $(x_1, y_1), \ldots, (x_n, y_n)$ with varying sample sizes of $n = 6$, 8, 10, 15, 20, 30 and 50. We reject the null hypothesis that the PDS condition holds, if the bootstrap confidence intervals do not contain $C1 = 1$ or $C2 = 1$, where the significance level $\alpha = 5\%$.

Again, we replicate this simulation 3×10^4 times, and includes, in Table 2, power from the simulated data with corresponding true ρ, τ, C1 and C2 values. Since the true C1 value is less than 1 and the true C2 value is greater than 1, the bivariate distribution of (X, Y) only violates the null hypothesis of the C2 test in Eq. (6). Therefore, we only report the simulated probability of rejecting the null hypothesis of the C2 test, which represents the probability of the C2 test favoring the alternative hypothesis. The simulation results suggest the effectiveness of the CoCo test: even with a moderate sample size, the test has a reasonable statistical power to detect the violation of PDS condition.

Table 2: Simulated statistical power of the PDS tests ($\alpha = 5\%$)

ρ	τ	true C1	true C2	Sample size n	Power of C2 test
0.288	0.333	0.624	1.206	6	50.0%
				8	37.4%
				10	37.2%
				15	45.1%
				20	54.9%
				30	71.4%
				50	90.4%

4 Application

After verifying the performance of the CoCo test using simulation studies, we apply it to investigate the association between the overall survival (OS) and the progression-free survival (PFS) in two trial-level meta-analyses on cancer. Specifically, the first study is a meta-analysis of 72 clinical trials between 1990 and 2015 on patients with metastatic breast cancer (MBC) treated with anthracyclines, taxanes, or targeted therapies (Adunlin et al., 2015). The second study is a meta-analysis of 12 trials between 2015 and 2017 on patients with advanced solid tumors treated with programmed cell death 1 (PD-1) inhibitors (e.g., nivolumab and pembrolizumab) (Gyawali et al., 2018). Figure 2 shows the relations between the log hazard ratios for OS and PFS in these two meta-analyses.

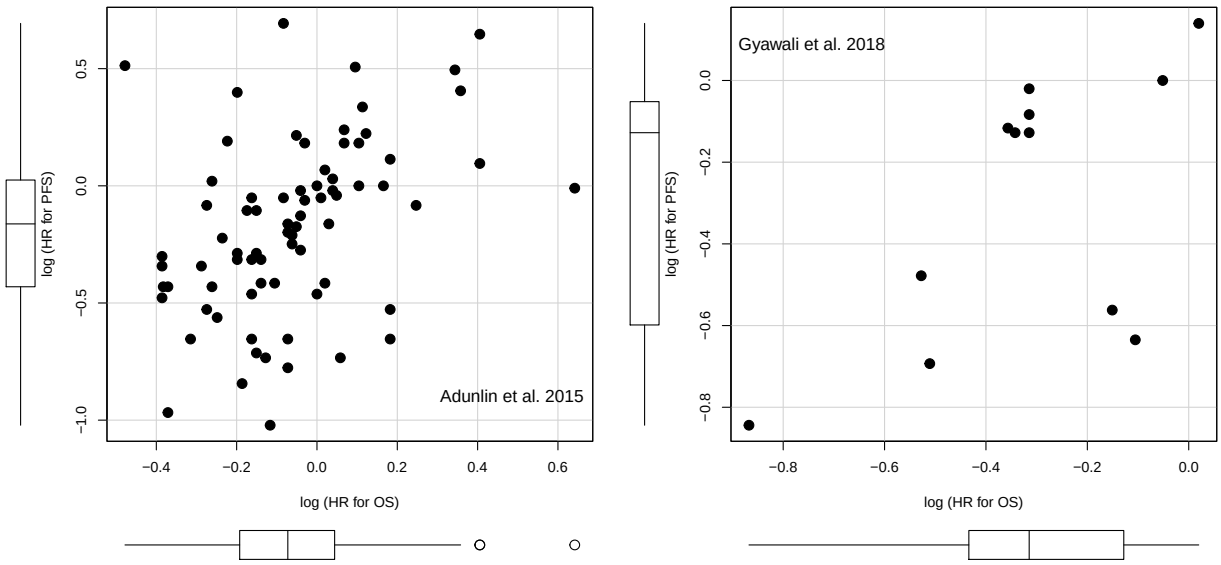

Figure 2: Log hazard ratios for OS and PFS from Adunlin et al. (2015) (left panel) and Gyawali et al. (2018) (right panel).

Study one. In the 72 trials on patients with metastatic breast cancer who underwent treatment of with anthracyclines, taxanes, or targeted therapies (Adunlin et al., 2015), Kendall's τ between PFS and OS is $\widehat{\tau} = 0.344$ with a 95% confidence interval $\tau \in (0.183, 0.492)$, and Spearman's ρ between PFS and OS is $\widehat{\rho} = 0.462$ with a 95% confidence interval $\rho \in (0.241, 0.682)$.

To determine whether a CoCo test is needed, we first assess the normality of the log hazard ratios for OS and PFS. The Anderson-Darling test for marginal normality gives p-values $p = 0.910$ for PFS and $p = 0.148$ for OS (Anderson and Darling, 1952). Next, we check whether the relation between the log hazard ratios for OS and PFS follows a bivariate normal distribution. The Henze-Zirkler test for multivariate normality indicates that it does not, with a p-value of $p = 0.0274$ (Henze and Zirkler, 1990). Since the bivariate normality seems not to hold, a simple test of correlation alone is insufficient to verify the satisfaction of the PDS condition.

Now we invoke the two CoCo tests to evaluate whether the PDS condition is met. The CoCo–1 test yields a test statistic $C1 = 0.7984$ and the corresponding 95% one-sided confidence interval is $C1 > 0.7160$. The p-value to reject the null hypothesis in Eq. (5) is greater than 0.999. The CoCo–2 test results in a test statistic $C2 = 0.9517$ and the 95% one-sided confidence interval is $C2 > 0.8979$. The p-value to reject the null hypothesis in Eq. (6) is 0.887. Taken together, both tests suggest a valid PDS relationship between PFS and OS.

Study two. In the 12 trials on patients with advance solid cancer treated with PD-1 inhibitors (Gyawali et al., 2018), Kendall's τ between PFS and OS is 0.469 with a 95% confidence interval $\tau \in (-0.145, 0.902)$, and Spearman's ρ is 0.593 with a 95% confidence interval $\rho \in (0.028, 0.870)$. Similarly, we conduct normality tests to see if a simple correlation test is sufficient to verify the PDS condition. Considering the the log hazard ratios for OS and PFS, the Anderson-Darling test for marginal normality gives $p = 0.081$ for PFS and $p = 0.342$ for OS, and the Henze-Zirkler test rejects the multivariate normality with $p = 0.0233$. Therefore, we need to apply the CoCo tests to evaluate the PDS condition. The CoCo–1 test gives a test statistic $C1 = 0.6926$, with a 95% one-sided confidence interval $C1 > 0.3574$ and p-value 0.856. The CoCo–2 test outputs a test statistic $C2 = 0.9488$, with a 95% one-sided confidence interval is $C2 > 0.7761$ and the p-value is 0.881. Together, both tests indicate a valid PDS relationship between PFS and OS.

5 Discussion

Assessing the conditions underpining the hypothesis testing is chief in guaranteeing the validity of the test. In multiple testing, when the dependency between tests are unknown, ensuring the satisfaction of the PDS condition is not only important for providing a suitable overall Type I error protection, but also, via the protection, for ensuring that the appraisal of multiple endpoints, and hence the evaluation of the drugs, is valid. In fact, in clinical studies, the PDS condition needs to be satisfied to demonstrate a strong familywise error rate control in many commonly used multiple testing procedures including the Hochberg (1988) procedure, the Hommel (1988) procedure, and the Benjamini and Hochberg (1995) false discovery rate (FDR) procedure.

In this paper, we develop a nonparametric statistical test, called the CoCo test, based on ranked correlation coefficients, to judge the validity of the PDS condition. While the CoCo test can be used when the tests are independent or dependent with a known standard joint distribution, it is especially useful when the multivariate distribution of test statistics is unknown or unclear. For example, consider the overall survival (OS) and progression-free survival (PFS) as the two endpoints measured in an oncology study. The log hazard ratios of OS and PFS are both normally distributed. As the joint distribution of OS and PFS is undetermined, the parametric statistical methods based on multivariate normal distribution may not be suitable. In this case, the CoCo test leverages the the nonparametric concordance measures of Spearman's ρ and Kendall's τ in Eqs. (5) and (6) and, via their simple, but effective, algebraic reformulation, can evaluate the PDS condition and may help addressing the concerns regarding the Hochberg procedure from the FDA (FDA, 2022).

One can extend the Cloe test to more generalized territories by utilizing Spearman's ρ, Kendall's τ, and other nonparametric concordance measures to evaluate various assumptions

of distribution. For example, under the normal assumption, using the relation that Pearson's correlation $r = 2\sin(\pi\rho/6) = \sin(\pi\tau/2)$, we construct a multivariate normality test:

$$\text{CoCo multivariate normality test:} \quad H_0 : \frac{\sin(\pi\tau/2)}{2\sin(\pi\rho/6)} = 1 \quad \text{versus} \quad H_a : \frac{\sin(\pi\tau/2)}{2\sin(\pi\rho/6)} \neq 1$$

where the decision of rejection can be made based on the bootstrap confidence interval. In addition, we can compute the sample variance of $\frac{\sin(\pi\tau/2)}{2\sin(\pi\rho/6)}$ via the delta method, using the variance estimation of Kendall's τ (Esscher, 1924), that of Spearman's ρ (Fieller et al., 1957; David and Mallows, 1961; Borkowf, 1999), and the covariance estimation between ρ and τ (David et al., 1951; Xu et al., 2013). Here, the CoCo multivariate normality test only evaluates whether the dependence structure is multivariate normal or not. If a complete test of multivariate normality is needed, we can combine the CoCo multivariate normality test with some univariate normality tests (Anderson and Darling, 1952; Shapiro and Wilk, 1965; Jarque and Bera, 1980) for the marginal distributions. The validity of this combined test is guaranteed by Sklar's theorem, since a copula describing the dependence structure and univariate marginal distribution functions can be fused to create any multivariate distribution (Sklar, 1959).

References

Adunlin, G., Cyrus, J. W. W., and Dranitsaris, G. (2015). Correlation between progression-free survival and overall survival in metastatic breast cancer patients receiving anthracyclines, taxanes, or targeted therapies: a trial-level meta-analysis. *Breast Cancer Research and Treatment* **154,** 591–608.

Amir, E., Seruga, B., Kwong, R., Tannock, I. F., and Ocaña, A. (2012). Poor correlation between progression-free and overall survival in modern clinical trials: Are composite endpoints the answer? *European Journal of Cancer* **48,** 385–388.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics* **23,** 193–212.

Belin, L., Tan, A., De Rycke, Y., et al. (2020). Progression-free survival as a surrogate for overall survival in oncology trials: a methodological systematic review. *British Journal of Cancer* **122,** 1707–1714.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57,** 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29,** 1165–1188.

Block, H. W., Savits, T. H., and Shaked, M. (1985). A concept of negative dependence using stochastic ordering. *Statistics and Probability Letters* **3,** 81–86.

Borkowf, C. B. (1999). A new method for approximating the asymptotic variance of spearman's rank correlation. *Statistica Sinica* **9,** 535–558.

Buyse, M., Burzykowski, T., Carroll, K., Michiels, S., Sargent, D. J., Miller, L. L., Elfring, G. L., Pignon, J.-P., and Piedbois, P. (2007). Progression-free survival is a surrogate for survival in advanced colorectal cancer. *Journal of Clinical Oncology* **25,** 5218–5224.

Capéraà, P. and Genest, C. (1993). Spearman's ρ is larger than Kendall's τ for positively dependent random variables. *Journal of Nonparametric Statistics* **2,** 183–194.

Chase, D. M., Mahajan, A., Scott, D. A., Hawkins, N., and Kalilani, L. (2023). Correlation between progression-free survival and overall survival in patients with ovarian cancer after cytoreductive surgery: a systematic literature review. *International Journal of Gynecologic Cancer* **33,** 1602–1611.

Courtinard, C., Gourgou, S., Jacot, W., et al. (2023). Association between progression-free survival and overall survival in women receiving first-line treatment for metastatic breast cancer: evidence from the ESME real-world database. *BMC Medicine* **21,** 87.

Daniels, H. E. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society. Series B (Methodological)* **12,** 171–191.

David, F. N. and Mallows, C. L. (1961). The variance of Spearman's rho in normal samples. *Biometrika* **48,** 19–28.

David, S. T., Kendall, M. G., and Stuart, A. (1951). Some questions of distribution in the theory of rank correlation. *Biometrika* **38,** 131–140.

Durbin, J. and Stuart, A. (1951). Inversions and Rank Correlation Coefficients. *Journal of the Royal Statistical Society: Series B (Methodological)* **13,** 303–309.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82,** 171–185.

Esscher, F. (1924). On a method of determining correlation from the ranks of the variates. *Scandinavian Actuarial Journal* **1924,** 201–219.

FDA (2022). *Multiple endpoints in clinical trials guidance for industry.* U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). FDA-2016-D-4460.

Fieller, E. C., Hartley, H. O., and Pearson, E. S. (1957). Tests for rank correlation coefficients. i. *Biometrika* **44,** 470–481.

Gou, J. (2023). On dependence assumption in p-value based multiple test procedures. *Journal of Biopharmaceutical Statistics* **33,** 596–610.

Gou, J. and Tamhane, A. C. (2018). Hochberg procedure under negative dependence. *Statistica Sinica* **28,** 339–362.

Gyawali, B., Hey, S. P., and Kesselheim, A. S. (2018). A comparison of response patterns for progression-free survival and overall survival following treatment for cancer with PD-1 inhibitors: A meta-analysis of correlation and differences in effect sizes. *JAMA Network Open* **1,** e180416.

Halabi, S., Vogelzang, N. J., Ou, S.-S., Owzar, K., Archer, L., and Small, E. J. (2009). Progression-free survival as a predictor of overall survival in men with castrate-resistant prostate cancer. *Journal of Clinical Oncology* **27,** 2766–2771.

Hamasaki, T., Evans, S. R., and Asakura, K. (2018). Design, data monitoring, and analysis of clinical trials with co-primary endpoints: A review. *Journal of Biopharmaceutical Statistics* **28,** 28–51.

Henze, N. and Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics – Theory and Methods* **19,** 3595–3617.

Hess, L., Brnabic, A., Mason, O., Lee, P., and Barker, S. (2019). Relationship between progression-free survival and overall survival in randomized clinical trials of targeted and biologic agents in oncology. *Journal of Cancer* **10,** 3717–3727.

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician* **69,** 371–386.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75,** 800–802.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley and Sons, New York, New York.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75,** 383–386.

Hutchinson, T. P. and Lai, C. D. (1990). *Continuous Bivariate Distributions, Emphasizing Applications*. Rumsby Scientific Publishing, Adelaide.

Hürlimann, W. (2003). Hutchinson-Lai's conjecture for bivariate extreme value copulas. *Statistics & Probability Letters* **61,** 191–198.

Jarque, C. M. and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* **6,** 255–259.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall, London.

Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. multivariate totally positive distributions. *Journal of Multivariate Analysis* **10,** 467–498.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* **30,** 81–93.

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika* **33,** 239–251.

Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics* **37,** 1137–1153.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63,** 655–660.

Moran, P. A. P. (1948). Rank correlation and product-moment correlation. *Biometrika* **35,** 203–206.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Series in Statistics. Springer Science+Business Media, Inc., New York, New York, 2 edition.

Pasalic, D., McGinnis, G. J., Fuller, C. D., Grossberg, A. J., Verma, V., Mainwaring, W., Miller, A. B., Lin, T. A., Jethanandani, A., Espinoza, A. F., Diefenhardt, M., Das, P., Subbiah, V., Subbiah, I. M., Jagsi, R., Garden, A. S., Fokas, E., Rödel, C., Thomas, C. R., Minsky, B. D., and Ludmir, E. B. (2020). Progression-free survival is a suboptimal predictor for overall survival among metastatic solid tumour clinical trials. *European Journal of Cancer* **136,** 176–185.

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture. *The Annals of Statistics* **26,** 494–504.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics* **30,** 239–257.

Sarkar, S. K. and Chang, C. K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Asociation* **92,** 1601–1608.

Shaked, M. (1977). A family of concepts of dependence for bivariate distributions. *Journal of the American Statistical Association* **72,** 642–650.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52,** 591–611.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73,** 751–754.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris* **8,** 229–231.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology* **15,** 72–101.

Tong, Y. L. (1990). *The Multivariate Normal Distribution*. Springer-Verlag, New York.

Tukey, J. W. (1958). A problem of berkson, and minimum variance orderly estimators. *The Annals of Mathematical Statistics* **29,** 588–592.

Xu, W., Hou, Y., Hung, Y., and Zou, Y. (2013). A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. *Signal Processing* **93,** 261–276.