applied sciences

MDPI

Article

Optimal Tests for Combining *p*-Values

Zhongxue Chen

Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, 1025 E. 7th Street, Bloomington, IN 47405, USA; zc3@indiana.edu; Tel.: +1-812-855-1163

Abstract: Combining information (*p*-values) obtained from individual studies to test whether there is an overall effect is an important task in statistical data analysis. Many classical statistical tests, such as chi-square tests, can be viewed as being a *p*-value combination approach. It remains challenging to find powerful methods to combine *p*-values obtained from various sources. In this paper, we study a class of *p*-value combination methods based on gamma distribution. We show that this class of tests is optimal under certain conditions and several existing popular methods are equivalent to its special cases. An asymptotically and uniformly most powerful *p*-value combination test based on constrained likelihood ratio test is then studied. Numeric results from simulation study and real data examples demonstrate that the proposed tests are robust and powerful under many conditions. They have potential broad applications in statistical inference.

Keywords: chi-square test; constrained likelihood ratio test; Fisher test; gamma distribution; uniformly most powerful test

1. Introduction

In statistical inference and decision making, it is critical but challenging to appropriately aggregate information from different sources. *p*-value combination approaches provide possible solutions. A *p*-value combination method usually combines the transformed statistics via the original individual *p*-values, and then, an overall *p*-value is obtained. The development of combining *p*-value has a long history. Many pioneer statisticians, including R. A. Fisher [1] and K. Pearson [2], had important contributions in this area. Their methods (e.g., Fisher test), along with others, such as the z-test [3] and the minimal *p*-test [4], are still widely used in today's statistical practice. Many studies have been conducted to compare the performances among those *p*-value combination tests [5–7]; it turned out that no test is uniformly most powerful although some methods may perform better than others under certain conditions. Combining dependent *p*-values is another research direction, as many robust and powerful methods have been proposed in the literature, including a recently proposed test based on Cauchy distribution (CCT) [8]. Although the CCT can be applied to both independent and dependent *p*-values, it has been shown that this test can never obtain a *p*-value less than the smallest *p*-value to be combined, and therefore, is not recommended for combining independent *p*-values [9]. In this paper, we focus on the situation where we have independent *p*-values to be combined.

It is well known that combining *p*-value methods have important applications in meta-analysis [10,11]. However, it is less recognized that combining *p*-value methods are more frequently but implicitly used in statistical testing. For instance, the commonly used chi-square tests, including the likelihood ratio test, the score test, and the Wald test, with degrees of freedom (df) greater than one can be viewed as *p*-value combination methods, which are special cases of our proposed gamma distribution-based tests (see Section 2). In other words, the popular chi-square tests only provide possible and special ways to combine *p*-values that are not necessarily optimal; more powerful methods for combining *p*-value may be found and used instead.

Citation: Chen, Z. Optimal Tests for Combining *p*-Values. *Appl. Sci.* **2022**, *12*, 322. https://doi.org/10.3390/app12010322

Academic Editor: Pentti Nieminen

Received: 16 November 2021
Accepted: 27 December 2021
Published: 29 December 2021

With recent technical developments, larger volume data, such as genome-wide genomic data, are generated more easily and rapidly. Consequently, advanced statistical methods, including p-value combination tests, are highly desirable [12–20]. For instance, meta-analyses, which combine information from different genome-wide association studies (GWASs), have identified many associated genetic variants that could not be identified from a single GWAS [21,22]. It is expected that with more powerful p-value combination tests being developed and available, more and more significant associated genetic variants will be discovered in cancer genomics.

Unfortunately, as Birnbaum [23] already noticed, there exists no uniformly most powerful (UMP) p-value combination test for all alternative hypotheses. However, it is possible that a method is UMP under a certain condition. Moreover, if the true condition is unknown, it is desirable to choose a robust p-value combination method in the sense that it has reasonable detection power under many conditions.

In this paper, we first propose a class of p-value combination methods based on gamma distribution. We show that several existing popular methods are equivalent to certain special cases of this class of tests. Then, we show that the proposed tests are UMP when the p-values to be combined are from certain distributions. When the p-values to be combined are from certain type of distributions whose parameters are partially or fully unknown, asymptotically UMP tests based on constrained likelihood ratio test (CLRT) are proposed and studied.

The rest of the manuscript is organized as follows: In Section 2, we first introduce some existing popular p-value combination tests, describe our proposed tests based on gamma distributions, and then study their connections to existing popular methods and their properties as of UMP tests. Finally, some asymptotically UMP tests based on CLRT are proposed and studied. In Section 3, we compare the performances of the proposed tests with some existing popular methods through a simulation study. In Section 4, two examples of real data applications are demonstrated to illustrate the desired performance of the proposed tests. This paper concludes in Section 5 with discussion and conclusions.

2. Methods

Suppose we have n independent p-values, denoted as $P_i (i = 1, \cdots, n)$, obtained from testing the individual null hypothesis H_{i0} versus the alternative hypothesis H_{i1}, respectively. In addition, throughout this paper, we assume $P_i \sim \mathrm{U}(0,1)$ under H_{i0}, where $\mathrm{U}(0,1)$ stands for the uniform distribution between 0 and 1. For a combining p-value test, in this paper, we consider testing the global null hypotheses, $H_0 = \cap H_{i0}$ vs. the global alternative hypothesis, $H_1 = \cup H_{i1}$. In statistical literature, several p-value combination tests were proposed long time ago but are still widely used today. We introduce some of the most popular ones as follows.

2.1. Some Existing Popular Tests

2.1.1. The Minimal p-Value (Min p) Test

This test is denoted as T_p, with the test statistic defined as [4]:

$$\min(P_1, P_2, \cdots, P_n) \tag{1}$$

whose null distribution is the beta distribution $Beta(1, n)$ and its p-value is defined as $1 - \left(1 - P_{(1)}\right)^n$, where $P_{(1)} = \min(P_1, P_2, \cdots, P_n)$. The Tippett's Min p test in (1) is closely related to the Bonferroni method [24]. When the minimal p-value $p_{(1)}$ is small, the two tests obtain similar results, and both are close to $np_{(1)}$.

2.1.2. The Chi-Square Test with n Degrees of Freedom

Denoted as χ^2_n, it has the test statistic [25,26]:

$$\sum_{i=1}^{n}\left(\Phi^{-1}(1-P_i)\right)^2 \tag{2}$$

where $\Phi^{-1}(\cdot)$ is the inverse function of the cumulative distribution function (CDF) of the standard normal distribution, $N(0,1)$. The null distribution of the test χ^2_n in (2) is the chi-square distribution with n df.

2.1.3. The Fisher Test

Denoted as F_p, it has the following test statistic [1]:

$$-2\sum_{i=1}^{n}\ln(P_i) \tag{3}$$

whose null distribution is χ^2_{2n}, the chi-square distribution with $2n$ df.

2.1.4. The z Test

Denoted as Z_p, it has the test statistic [3]:

$$\sum_{i=1}^{n}\Phi^{-1}(1-P_i)/\sqrt{n} \tag{4}$$

whose null distribution is $N(0,1)$.

Note that for all of the above tests, their overall one-sided p-values are calculated based on the right-tails, i.e., the areas beyond the test statistics from the right sides of their null distributions, to reflect the fact that smaller individual p-values provide stronger evidence to support the global alternative hypothesis.

2.2. New Tests Based on Gamma Distribution

We use $Gamma(\alpha,\beta)$ to denote a random variable that has a gamma distribution with shape parameter α and rate parameter β, where both parameters are positive. The probability density function (PDF) of $Gamma(\alpha,\beta)$ is:

$$f_{G(\alpha,\beta)}(x) = \beta^{\alpha}x^{\alpha-1}\exp(-\beta x)/\Gamma(\alpha) \tag{5}$$

for $x > 0$, where the gamma function $\Gamma(z) = \int_0^{\infty} x^{z-1}e^{-x}dx$. Denote the corresponding CDF as $F_{G(\alpha,\beta)}(x)$. We can combine n independent p-values using $Gamma(\alpha,\beta)$ and obtain an overall p-value accordingly.

2.2.1. Gamma Distribution-Base Test $T_{G(\alpha,\beta)}$

Define the following test statistic:

$$T_{G(\alpha,\beta)}(P_1,P_2,\cdots,P_n) = \sum_{i=1}^{n}F^{-1}_{G(\alpha,\beta)}(1-P_i), \tag{6}$$

where $F^{-1}_{G(\alpha,\beta)}(y)$ is the inverse function of the CDF $F_{G(\alpha,\beta)}(x)$.

We also define the following right-tailed p-value for $T_{G(\alpha,\beta)}$:

$$P = \Pr(Gamma(n\alpha,\beta) > t) = 1 - F_{G(n\alpha,\beta)}(t) = S_{G(n\alpha,\beta)}(t) \tag{7}$$

where t is the observed value for the test $T_{G(\alpha,\beta)}$, and $F_{G(n\alpha,\beta)}(\cdot)$ and $S_{G(n\alpha,\beta)}(\cdot)$ are the CDF and the survival function of $Gamma(n\alpha,\beta)$, respectively.

The above test determined by the test statistic $T_{G(\alpha,\beta)}$ in (6) and the p-value P in (7) is called $T_{G(\alpha,\beta)}$. The p-value for a specific test T is also denoted as P_T. Therefore, the p-value in (7) can also be written as $P_{T_{G(\alpha,\beta)}}$. Based on the properties of gamma distributions, we have the following result for the test $T_{G(\alpha,\beta)}$:

Proposition 1. *The test $T_{G(\alpha,\beta)}$ with statistic defined in (6) and p-value defined in (7) controls type I error rate exactly at given significance level.*

To study the properties of the new tests, we use the following definitions:

Definition 1. *Two tests T_1 and T_2 are called equivalent and denoted as $T_1 \equiv T_2$ if $p_{T_1}(x) = p_{T_2}(x)$ for any observed data x, where $p_{T_i}(x)$ is the p-value obtained by test $T_i (i = 1, 2)$ from given data x.*

Based on the properties of the gamma distributions, we can easily verify the following result:

Proposition 2. *For any $\alpha > 0$ and $\beta > 0$, $T_{G(\alpha,\beta)} \equiv T_{G(\alpha,1)}$.*

Therefore, the rate parameter of the gamma distribution has no effect on the test $T_{G(\alpha,\beta)}$. For convenience, in this paper, we set $\beta = 1$ and use $T_{G(\alpha)}$ to denote $T_{G(\alpha,1)}$ hereafter unless otherwise specified.

Definition 2. *A positively-valued function $c(\theta)$ is called the exact slope of the sequence of tests T_n if $\lim_{n\to\infty}[(-2/n)ln(1 - F_n(T_n))] = c(\theta)$ with probability 1, where $F_n(T_n)$ is the CDF of T_n. A test is called asymptotically Bahadur optimal (ABO) if its exact slope is maximal at every $\theta \in \Theta - \Theta_0$, where Θ is the parameter space, and Θ_0 is the parameter space under the null hypothesis. The exact slope is a measure of the rate at which the attained p-value of a test statistic tends to 0 and is a measure of asymptotic efficiency.*

It has been proven that $T_{G(\alpha,1/2)}$ is ABO for any $\alpha \in (0, \infty)$ [27]. Hence, from proposition 2, we have the following property:

Proposition 3. *The test $T_{G(\alpha)}$ is ABO for any $\alpha \in (0, \infty)$.*

Previously, we proposed a different p-value combination test based on gamma distribution with the test statistic $T = \sum_{i=1}^{n} F^{-1}_{G(1/P_i,1)}(1 - P_i)$, which uses the random shape parameter $a_i = 1/P_i$ for p-value P_i, and its null distribution is intractable, and a resampling method is used to estimate the p-value. While in the current proposed tests, the same shape parameter is used for all individual p-values.

2.2.2. Connections between $T_{G(\alpha)}$ and Existing Popular Tests

Although the existing popular tests described in Section 2.1 were proposed a long time ago, and their theoretical properties and empirical performances have been extensively studied and compared [6,7,25,28,29], surprisingly, their relationships have not been fully investigated, and the theoretical explanation on the differences of their performances is lacking. In this subsection, we show that they are connected to the aforementioned gamma distribution-based tests $T_{G(\alpha)}$ with different α values. In fact, we have the following results:

Theorem 1. *The class of gamma distribution-based tests $T_{G(\alpha)}$ include special cases that are equivalent to the aforementioned existing popular methods. More specifically,*

$$T_{G(0)} \triangleq \lim_{\alpha \to 0+} T_{G(\alpha)} \equiv T_p;$$
$$T_{G(0.5)} \equiv \chi^2_n;$$
$$T_{G(1)} \equiv F_p; \text{ and}$$
$$T_{G(\infty)} \triangleq \lim_{\alpha \to \infty} T_{G(\alpha)} \equiv Z_p.$$

The proof of Theorem 1 is given in the Appendix A.

2.2.3. $T_{(\alpha)}$ as the Uniformly Most Powerful Test

Besides the ABO property, there is another attractive property for the gamma distribution-based test $T_{G(\alpha)}$: under certain conditions, it is UMP. We thus have the following theorem:

Theorem 2. *Suppose $P_1, P_2, \cdots, P_n$ are iid from the following common density function with parameters $\alpha > 0$ and $0 < c < 1$*

$$f_{\alpha,c}(p) = (1-c)^{\alpha} exp\left[cF^{-1}_{G(\alpha)}(1-p)\right] \text{ for } p \in (0,1), \tag{8}$$

If both α and c are known, then test $T_{G(\alpha)}$ is UMP.
The proof is given in the Appendix A.

Remark 1. *(i) For $p \in (0,1)$, when $c = 0$, $f_{\alpha,0}(p) = 1$. Therefore, $f_{\alpha,0}(p)$ corresponds to the global null hypothesis. (ii) Under the condition $f_{\alpha,c}(p) = (1-c)^{\alpha} exp\left[cF^{-1}_{G(\alpha)}(1-p)\right]$ with $0 < c < 1$, from Theorem 1, the existing tests T_p, χ^2_n, F_p, and Z_p defined in (1)–(4) are UMP for given $\alpha = 0$, 0.5, 1, and ∞, respectively. (iii) Along with Theorem 1, Theorem 2 provides insightful explanations on when and why an existing popular test described in Section 2.1 is preferred. For instance, when the p-values are extremely heterogeneous (e.g., a very small α in (8)), the Min p test (i.e., $T_{G(0)}i$) s more powerful than the other gamma distribution-based tests. On the other hand, when the p-values are more homogeneous (e.g., a large α in (8)), the z test (i.e., $T_{G(\infty)}$) is preferred. (iv) The function $f_{\alpha,c}(p) = (1-c)^{\alpha} e^{cF^{-1}_{G(\alpha)}(1-p)}$ with two parameters α ($\alpha > 0$) and c ($0 < c < 1$) represents a large class of densities and can be used to approximate the true density functions under the alternative hypotheses. Based on simulation study, we found that under many situations, the true density functions under the alternative hypotheses can be closely approximated by $f_{\alpha,c}(p)$ with the two parameters being estimated from the data (see Figures S1–S19 in the Supplementary File).*

2.3. Constrained Likelihood Ratio Tests

In Section 2.2, we have shown that $T_{G(\alpha)}$ is UMP if the p-values to be combined are from the common density as described in (8) with known constants α and c. When c is unknown, or both α and c are unknown, constrained likelihood ratio tests (CLRTs) can be constructed accordingly. In this subsection, we study those CLRT-based tests when parameter c is unknown with α being known or unknown.

2.3.1. CLRT-Based Tests with Known α Values

Under (8), with known $\alpha = \alpha_0$, the CLRT-based tests can be constructed based on the constrained MLE of c obtained through maximizing $l(\alpha_0, c) = l(c) = n\alpha_0 ln(1-c) + c\sum_{i=1}^{n} F^{-1}_{G(\alpha_0)}(1-P_i)$. More specifically, we define the following test statistic:

$$T_{CLRT,\alpha_0}(P_1, \cdots, P_n) = 2\alpha_0 n ln\left(1 - \hat{c}_{CLRT,\alpha_0}\right) + 2\left(\hat{c}_{CLRT,\alpha_0}\right) \sum_{i=1}^{n} F^{-1}_{G(\alpha_0)}(1-P_i), \tag{9}$$

where $\hat{c}_{CLRT,\alpha_0}$ is the constrained MLE of c through maximizing the log-likelihood $l(\alpha_0, c)$ with the constrain $0 < c < 1$. We will reject the overall null hypothesis if the test statistic is large. In other words, a one-sided p-value will be calculated from the test.

For the above CLRT-based test T_{CLRT,α_0} in (9), we have the following result [30]:

Proposition 4. *The asymptotic distribution of the test T_{CLRT,α_0} is a mixture of chi-square distributions $\sum_{i=0}^{1} w_i\chi_i^2$, where χ_i^2 is the chi-square distribution with $df = i$, χ_0^2 is the random variable with probability 1 of being 0, and the weights w_0, w_1 are determined by the null and the alternative hypothesis.*

In practice, the p-values of the test T_{CLRT,α_0} can be estimated through resampling methods (e.g., see Section 2.3.2 for an example). However, the following result shows that this test is tightly connected with the gamma distribution-based test $T_{G(\alpha_0)}$, whose p-value can be calculated directly.

Theorem 3. *Let t_{CLRT,α_0} be the observed statistic of test T_{CLRT,α_0}; the p-value of T_{CLRT,α_0} is determined by the gamma distribution-based test $T_{G(\alpha_0)} = \sum_{i=1}^{n} F_{G(\alpha_0)}^{-1}(1 - P_i)$ as follows:*

$$Pr\left[T_{CLRT,\alpha_0} > t_{CLRT,\alpha_0}\right] = \begin{cases} P_{T_{G(\alpha_0)}} \text{ if } t_{CLRT,\alpha_0} > 0 \\ Pr\left[T_{G(\alpha_0)} < n\alpha_0\right] \text{ if } t_{CLRT,\alpha_0} = 0 \end{cases} \tag{10}$$

The proof is given in the Appendix A.

Proposition 5. *Under the conditions specified in (8), if the parameter $\alpha = \alpha_0$ is known, then asymptotically $T_{G(\alpha_0)} \equiv T_{CLRT,\alpha_0}$.*

Proof. From the proof of Theorem 3, we know that under (8), when $0 \le c < 1$, $Pr[T_{G(\alpha_0)} < n\alpha_0] = Pr[\hat{c}_{\alpha_0} \le 0] \to 0$ $(as\ n \to \infty)$, and hence, the two p-values from tests $T_{G(\alpha_0)}$ and T_{CLRT,α_0} are asymptotically equal with probability 1. □

Theorem 4. *Under the conditions specified in (8), if the parameter $\alpha = \alpha_0$ is known, then the gamma distribution-based test $T_{G(\alpha_0)}$ and the CLRT-based test T_{CLRT,α_0} are both asymptotically UMP.*

Proof. When $\alpha = \alpha_0$ is known, it can be shown that the constrained MLEs for c is consistent (see, for instance, Theorem 1 of Self and Liang [30]). Hence, from Theorem 2, $T_{G(\alpha_0)}$ is asymptotically UMP. From Proposition 5, T_{CLRT,α_0} is asymptotically UMP. □

2.3.2. The Optimal CLRT-Based Test When α Is Unknown

When both parameters α and c in (8) are unknown, they need to be estimated via the constrained MLEs, from which the following constrained likelihood ratio test is defined:

$$T_{CLRT}(P_1, \cdots, P_n) = 2\hat{\alpha}_{CLRT} nln(1 - \hat{c}_{CLRT}) + 2(\hat{c}_{CLRT}) \sum_{i=1}^{n} F_{G(\hat{\alpha}_{CLRT})}^{-1}(1 - P_i) \tag{11}$$

where $\hat{\alpha}_{CLRT}$ and $\hat{c}_{CLRT}$ are the constrained MLEs for parameters α and c, respectively, through maximizing the log-likelihood function $l(\alpha, c) = n\alpha ln(1 - c) + c\sum_{i=1}^{n} F_{G(\alpha)}^{-1}(1 - P_i)$ with the constrains $0 < c < 1$ and $\alpha > 0$. The R function "nlminb" can be applied to find the constrained MLEs and the corresponding test statistic. The proposed test was implemented using R; the R package "opt" (optimal p-value combination test) can be freely download from https://github.com/zchen2020/opt (accessed on 15 November 2021).

For the above CLRT-based test T_{CLRT}, similar to Proposition 4, we have the following result [30]:

Proposition 6. *The asymptotic distribution of the test T_{CLRT} is a mixture of chi-square distributions $\sum_{i=0}^{2} w_i\chi_i^2$, where χ_i^2 is the chi-square distribution with $df = i$, χ_0^2 is the random variable with probability 1 of being 0, and the weights w_0, w_1, w_2 are determined by the null and the alternative hypothesis.*

The above asymptotic result may not be directly applicable to estimate the p-value for this test, as the number of p-values n is usually small, and more seriously, the weights $w_i's$ are difficult to obtain. Instead, a simple resampling method can be used to approximate the null distribution and to estimate the p-value of T_{CLRT}. More specifically, for given sample size n, randomly sample n null p-values evenly distributed between 0 and 1, then calculate the test statistic using (10). Repeat this process many times (e.g., 10^5); then, the empirical distribution of the test statistic can be used to approximate the null distribution and therefore the p-value of T_{CLRT}.

Similar to Theorem 4, we have the following result for T_{CLRT}:

Theorem 5. *Under the conditions specified in (8), the CLRT-based test T_{CLRT} is asymptotically UMP.*

Proof. Under conditions (8), it can be shown that the constrained MLEs for α and c are consistent (see, for instance, Theorem 1 of Self and Liang [30]). Hence, T_{CLRT} is asymptotically equivalent to T_{CLRT,α_0} for known $\alpha = \alpha_0$ in (8). Then from Theorem 4, T_{CLRT} is asymptotically UMP. □

Remark 2. *(i) When $\alpha = \alpha_0$ is known, compared with test T_{CLRT,α_0}, the test $T_{G(\alpha_0)}$ is preferred because (a) its test statistic and p-value are easier to get and (b) the two tests in general have very similar performances. (ii) When both α and c are unknown, the test T_{CLRT} is asymptotically UMP, while, in general, neither T_{CLRT,α_0} nor $T_{G(\alpha_0)}$ for preset $\alpha = \alpha_0$ is UMP or asymptotically UMP. Therefore, it is expected that T_{CLRT} is more robust, and overall, it has better performance than each individual gamma distribution-based test $T_{G(\alpha)}$, including the existing popular ones described in Sections 2.1.1–2.1.4.*

3. Numeric Studies

In this section, we assess the performances of the proposed tests through a simulation study. In the simulation, we compare the optimal CLRT-based test T_{CLRT} with the popular and representative gamma distribution-based tests, $T_{G(0)}$, $T_{G(1)}$, and $T_{G(\infty)}$ (i.e., the Min p, Fisher, and z test, respectively).

In the simulation study, fifty ($n = 50$) independent p-values are simulated and combined. Among these 50 p-values, m ($m = 0$, 10, 20, 40, 50) are assumed from the true individual alternative hypotheses, and the rest ($n - m$) are from the true individual null hypotheses. When m = 0, all 50 individual null hypotheses are true, and the empirical power obtained under this condition is the empirical type I error rate. The p-values from the true null hypotheses are randomly sampled between 0 and 1 from the uniform distribution. For a true individual alternative hypothesis H_{i1} ($i = 1, \cdots, m$), we assume the p-value p_i is obtained via a random variable $z_i \sim N(\mu_i, 1)$. We randomly set k of the m $u_i's$ as positive or negative (those alternative hypotheses with the same direction of the effects are called concordant alternatives) and the rest of $m - k$ having the other direction. A two-sided p-value for each true individual alternative hypothesis are obtained via the standard z test by comparing the test statistic with the standard normal distribution $N(0, 1)$.

For the true alternative hypotheses, we consider three different scenarios for the effects of $\mu_i's$. Scenario 1: $|\mu_i| = \mu v_i / \sum_{i=1}^{m} v_i$, where $v_i = 10^{r_i}$; $r_i \sim N(0.3, 1)$; and $\mu = 0.8, 0.6$, 0.4, 0.3 when there are 10, 20, 40, and 50 true individual alternatives, respectively. Scenario 2: $|\mu_i| = \mu v_i / \sum_{i=1}^{m} v_i$, where $v_i \sim U(1, 100)$ and $\mu = 1.2$, 1.0, 0.6, 0.5 when the number of true individual alternatives is $m = 10$, 20, 40, and 50, respectively. Scenario 3: $|\mu_i| = \mu / m$, and $\mu = 1.5$, 1.2, 0.8, and 0.6 when $m = 10$, 20, 40, and 50, respectively. Note that (i) the constants (e.g., μ, the parameters in the normal distribution for r_i and the uniform distribution for v_i) are chosen in the way so that the empirical powers are appreciable for comparison. (ii) For all the three scenarios, the sum of the absolute effect sizes is equal to μ; and (iii) for given m, the degree of heterogeneity of the effect sizes among the true individual alternatives decreases from Scenario 1 to Scenario 3. More

specifically, in scenario 1, the effect sizes are extremely heterogenous when the number of the true individual alternatives is small. In Scenario 3, the effect sizes are more homogenous. The situations in Scenario 2 are between those in Scenarios 1 and 3. By considering those different conditions, we tried to conduct a reasonable and realistic simulation study to fairly compare our proposed tests with others.

The empirical power values of the tests are estimated using the rejection proportions based on 1000 replicates at the significance level of 0.05. For the new tests, T_{CLRT} 10^5 replicates are used to estimate their p-values from the resampling method described in Section 2.3.2. Under the overall null hypotheses (i.e., all individual null hypotheses are true), the empirical type I error rates for all methods with different significance levels were obtained. From the simulation study, all methods controlled type I error rate quite well (see Table S4 in the Supplementary File).

Figures 1–3 plot the empirical power values of the Min p (Min), Fisher (Fisher), z test (Z), and the proposed CLRT-based test T_{CLRT} (LRT_CS) when p-values are combined under Scenarios 1 to 3, respectively. We have the following observations: First, for Scenario 1 (Figure 1), where the effect sizes from the true individual alternative hypotheses are extremely heterogeneous, the Min p test (i.e., $T_{G(0)}$) usually performs better than the Fisher test ($T_{G(1)}$), which in turn performs better than the z test ($T_{G(\infty)}$). Second, when the degree of heterogeneity of the effect sizes among individual alternative hypotheses are less extreme, as in Scenarios 2 and 3 (Figures 2 and 3), Fisher test and the z test usually perform better or much better than the Min p test. Third, for the Min p, Fisher, and z test, one may perform very well for some conditions but very poorly under others. For instance, under Scenario 1 (extremely heterogeneous effect sizes among the alternatives), the Min p test is more powerful than the other two, while it is much less powerful under scenario 3 (homogeneous effect sizes among the alternatives). The opposite direction was observed for the z test. Fourth, under all conditions considered, the new test T_{CLRT} is either the best or very comparable to the best one. When the number of p-values to be combined is small, we observed similar patterns (see the simulation results in Tables S1–S3 in the Supplementary Materials when n = 10). This demonstrates that, as expected, T_{CLRT} is a robust test in the sense that under many conditions, it has reasonable detection power compared with other tests. We would like to point out that, as expected, the empirical power values from the CLRT-based test T_{CLRT,α_0} are very close to those from the corresponding gamma distribution-based tests $T_{G(\alpha)}$, with $\alpha = 0,\ 1,\ \infty$ being fixed (data not shown).

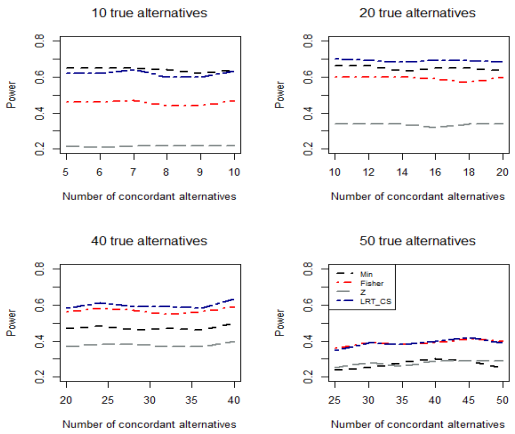

Figure 1. Empirical power values of the tests based on two-sided p-values under Scenario 1: $|\mu_i| = \mu v_i / \sum_{i=1}^{m} v_i$, where $v_i = 10^{r_i}$, $r_i \sim N(0.3, 1)$, and $\mu = 0.8,\ 0.6,\ 0.4,\ 0.3$ when there are 10, 20, 40, and 50 true individual alternatives, respectively.

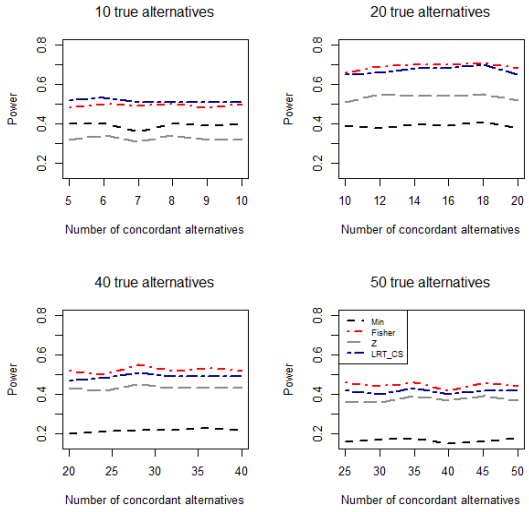

Figure 2. Empirical power values of the tests based on two-sided p-values under scenario 2: $|\mu_i| = \mu v_i / \sum_{i=1}^{m} v_i$, where $v_i \sim U(1, 100)$, and $\mu = 1.2,\ 1.0,\ 0.6,\ 0.5$ when the number of true individual alternatives is $m = 10,\ 20,\ 40,$ and 50, respectively.

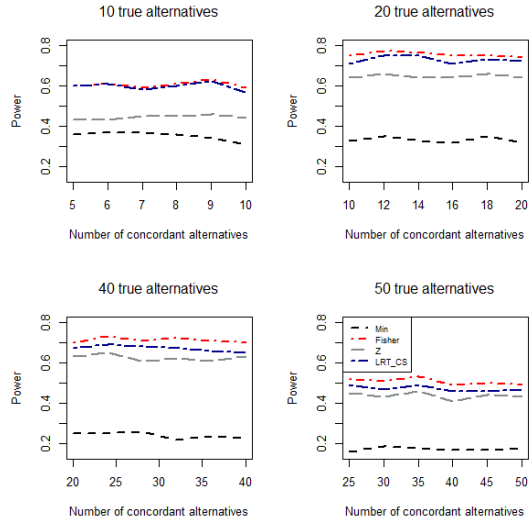

Figure 3. Empirical power values of the tests based on two-sided p-values under scenario 3: $|\mu_i| = \mu / m$, and $\mu = 1.5,\ 1.2,\ 0.8,$ and 0.6 when $m = 10,\ 20,\ 40,$ and 50, respectively.

4. Real Data Examples

In this section, we apply the proposed tests along with others to two real-world problems to demonstrate the usefulness of the proposed test.

4.1. Example 1: A Meta-Analysis

In a meta-analysis, 12 randomized trials examining the effect of patient rehabilitation designed for geriatric patients versus usual care on improving functional outcome at 3–12 month follow-up were used [31,32]. The estimated odds ratios (ORs) from the 12 trials are listed in Table 1.

The p-value from the Cochran's test for homogeneity was 0.021, indicating that the commonly used fixed effect model of meta-analysis was inadequate for this data set. Therefore, the authors ran the meta-analysis with a random effect model and estimated the overall OR as 1.36 with 95% CI (1.08, 1.71) [32]. However, the goodness-of-fit test for the random effect model obtained a p-value of 0.025 [33], indicating the lack of fit of the random effect model for this data set. Therefore, instead of using the problematic

fixed or random effect models to combine information from the 12 trials, we use *p*-value combination methods to test whether there is an overall effect.

Table 1. Estimated odds ratio and its 95% CI from each study in a meta-analysis with 12 trials. Data were taken from Bachmann et al. and Riley et al.

Study	OR	95% CI	Study	OR	95% CI	Study	OR	95% CI
1	1.11	0.51, 2.39	5	0.88	0.39, 1.95	9	1.06	0.63, 1.79
2	0.97	0.78, 1.21	6	1.28	0.71, 2.30	10	2.95	1.54, 5.63
3	1.13	0.73, 1.72	7	1.19	0.69, 2.08	11	2.36	1.18, 4.72
4	1.08	0.42, 2.75	8	3.82	1.37, 10.60	12	1.68	1.05, 2.70

In order to use the *p*-value combination methods, for each trial, we calculate its *p*-value based on its reported 95% CI. Denote U and L the upper and lower limits of the 95% CI; the test statistic can be approximated as $t = \ln(U \times L)/\sqrt{4\ln(U/L)/3.92}$, whose asymptotic null distribution is $N(0,1)$. The sample sizes of these 12 trials were relatively large, ranging from 108 and 1388; therefore, we can reasonably estimate their *p*-values using the asymptotic null distribution. We calculate the two-sided *p*-value for each trial and apply the gamma distribution-based tests. The *p*-values from the Min *p* (i.e., $T_{G(0)}$), Fisher ($T_{G(1)}$), z test ($T_{G(\infty)}$), and T_{CLRT} for combining those two-sided *p*-values are 0.013, 0.0068, 0.075, and 0.0047, respectively. Except for the z test, which is known to be less powerful for this heterogeneous situation, all methods obtained *p*-values less than 0.05, and the proposed test T_{CLRT} had the smallest *p*-value, indicating that the proposed test is more powerful than the other tests under this specific situation.

4.2. Example 2: A Survival Analysis from a Clinical Trial

The second data set is from the randomized, double-blinded Digoxin Intervention Trial [34]. In this trial, patients with left ventricular ejection fractions of 0.45 or less were randomly assigned to digoxin (3397 patients) or placebo (3403 patients) groups. A primary outcome was the mortality due to worsening heart failure (see Figure S20 in the Supplementary Materials). In the original study, the authors used the log-rank (LR) test and obtained a *p*-value of 0.061, indicating that the evidence of the effectiveness of digoxin, in terms of reducing the mortality due to worsening heart failure, is at most marginal.

However, it is well known that the LR test may fail to detect the difference between two survival functions if their hazard rate functions are crossing [35,36]. We apply the two-stage approach [35,36] to this data set and obtained two *p*-values of 0.06 and 0.04 for the two stages. Since, under the null hypothesis, the two *p*-values from the two stages are asymptotically independent [35,36], we can combine them using the proposed test T_{CLRT} and the gamma distribution-based tests. The *p*-values are 0.078, 0.017, 0.0099, and 0.011 from the Min *p*, Fisher, z test, and T_{CLRT}, respectively. The proposed test obtained the second smallest *p*-value, which is slightly larger than the smallest one obtained by the z test.

In addition, the drug effect may differ between males and females. To investigate the possible interaction between sex and treatment, we divide the data into four groups based on the combinations of sex and treatment: Male—Placebo (MP), Male—Drug (MD), Female—Placebo (FP), and Female—Drug (FD). The sample sizes for the four groups are 2639, 2642, 764, and 755, respectively. We then compare the survival functions in the following three pairs of groups: MP vs. MD, (MP + MD) vs. FP, and (MP + MD + FP) vs. FD, where (MP + MD) is a new group with pooled data from groups MP and MD, and (MP + MD + FP) includes all the subjects from groups MP, MD, and FP (see Figure S21 in the Supplementary Materials). For each comparison, the two-stage approach is applied. We obtain the following six *p*-values: 0.019, 0.026, 0.504, 0.092, 0.975, and 0.050. It has been shown that under the null hypothesis, the six *p*-values obtained from the above approach are asymptotically independent [36]. Applying the gamma distribution-based tests, along with T_{CLRT}, to the six asymptotically independent *p*-values, we obtain *p*-values

of 0.11, 0.0067, 0.020, and 0.013 from the Min p, Fisher, z test, and T_{CLRT}, respectively. It noticeable that the proposed test obtained the second smallest p-value, while Fisher test obtained the smallest one among those methods. These results show that except for the Min p test, all other tests obtain p-values less than 0.05. In addition, the p-values from T_{CLRT} in general are close to the smallest ones, while the popular tests, Fisher and z test, may get quite different p-values under different situations (two groups vs. four groups).

Compared with the original analysis, the results of the proposed optimal test T_{CLRT} using p-values from the two-stage approach applied to either two groups (placebo vs. drug) or four groups (combinations of sex and drug) provide stronger evidence against the null.

5. Discussion and Conclusions

In this paper, we studied a class of gamma distribution-based p-value combination methods, which include special cases that are equivalent to some existing popular methods. This class of tests provide unlimited choices for combining independent p-values. However, under a given situation, some of them may perform very poorly. Therefore, arbitrarily picking one of them may result in failing to detect true alternatives. On the other hand, if we try many different methods and report the smallest p-value, we need to adjust this p-value due to multiple comparison issue; otherwise, we will have more false findings than expected due to inflated type I error rate. Therefore, it is desirable to develop methods that can adaptively find the optimal approach from candidate tests. Our proposed CLRT-based test T_{CLRT} is one of such methods. We have shown that if the p-values to be combined are from a common density function $f_{\alpha,c}(p) = (1-c)^{\alpha} e^{cF^{-1}_{G(\alpha)}(1-p)}$ for $p \in (0,1)$, the gamma distribution-based test $T_{G(\alpha)}$ is UMP when both parameters α and c are known. When $\alpha = \alpha_0$ is known but c unknown, both $T_{G(\alpha)}$ and the CLRT-based test T_{CLRT,α_0} are asymptotically UMP. Furthermore, when both α and c are unknown, the proposed CLRT-based test T_{CLRT} is asymptotically UMP.

In a meta-analysis, it is natural to assign different weights to individual studies [5,7,37,38]. For instance, a larger weight can be assigned to a study with more subjects; hence, in the z test, a p-value from a larger study may receive a greater weight. Weights can also be assigned based on other quantities, such as the variances of the estimated effect sizes. However, there is no consensus on weight assignment. For our proposed tests, we can easily incorporate weights assigned to each individual p-value. For instance, the weighted gamma distribution-based tests can be constructed using $T^{w}_{G(\alpha)} = \sum_{i=1}^{n} F^{-1}_{G(w_i\alpha)}(1-P_i)$, where w_i is the weight assigned to study i $(i = 1, \cdots, n)$. Based on the properties of gamma distributions, it is not difficult to show that $\lim_{\alpha\to\infty} T^{w}_{G(\alpha)} \equiv \sum_{i=1}^{n} w_i \Phi^{-1}(1-P_i) / \sqrt{\sum_{i=1}^{n} {w_i}^2}$, the weighted z test. Therefore, the class of weighted gamma distribution-based tests $T^{w}_{G(\alpha)}$ are generalizations of the weighted z test. Likewise, the weighted log-likelihood function with given weights becomes $l^{w}(\alpha, c) = \alpha ln(1-c) \sum_{i=1}^{n} w_i + c \sum_{i=1}^{n} F^{-1}_{G(w_i\alpha)}(1-p_i)$, from which the corresponding weighted CLRT-based optimal test T^{W}_{CLRT} can be constructed accordingly.

Our proposed tests have much broader applications than in meta-analysis. In fact, they can be applied to almost all statistical testing problems when (asymptotically) independent p-values from individual components are available. For instance, in model selection, a typical step is to test whether a set of variables (or a single categorical variable with multiple levels) should be included in the final model. Often the time, the parameters, and the covariances of their estimates are estimated simultaneously through maximum likelihood estimation. Then the LRT via comparing the log-likelihood values from two models with and without the candidate variables, or the Wald chi-square test of the weighted sum of the squared estimated effect sizes, can be applied. For both LRT and the Wald test, a set of asymptotically independent p-values can be obtained through their asymptotically independent components (see, e.g., Chapter 16 of [39]). Hence, our proposed p-value combination methods, such as T_{CLRT}, can be applied and may result in a better final model.

Another example is the association test for two categorical variables in a two-way contingency table to which the Pearson chi-square test is usually applied. It is known

that the Pearson's chi-square test statistic with k df can be partitioned into k asymptotically independent components whose null distributions are asymptotically *iid* chi-square distribution with 1 df [40]. For instance, the partition can be done through the Lancaster approach [41]. Hence, we can calculate a set of asymptotically independent p-values to which our proposed CLRT-based test is applicable.

The performance of the proposed approaches can be improved if the p-values to be combined are obtained from an individual study using more powerful tests. For instance, if we already know the direction of the effect (positive or negative) when we compare two group means, we can use a one-sided rather than a two-sided test to obtain the individual p-value. However, it should be pointed out that sometimes one-sided tests may not be always applicable to individual studies. Nevertheless, our proposed approaches can still be used.

In this paper, we focus on using gamma distribution to combine independent p-values. Our future direction will be developing gamma distribution-based methods to combine dependent p-values. The difficulty in this direction is how to choose the "optimal"-shape parameter so that the resulting test has good detection power and can control type I error rate for arbitrary dependency structure of the p-values to be combined. Our preliminary results indicate that this direction is promising. A follow-up paper will be published.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/app12010322/s1, Figure S1: some densities of $f_{a,c}(t)$; Figure S2: Histogram and the estimated density from simulated data when $\mu_i = 0$; Figures S3–S21: Histograms and the estimated densities from simulated data; Table S1: Empirical power from simulation under scenario 1 using n = 10 and α = 0.05; Table S2: Empirical power from simulation under scenario 2 using n = 10 and α = 0.05; Table S3: Empirical power from simulation under scenario 3 using n = 10 and α = 0.05; Table S4: Empirical type I error rates from simulation study with 10,000 replicates using different significant levels.

Funding: This work was partially supported by the National Institutes of Health grants 1R03DE030259, UL1TR002529, and the Indiana University Open Access Article Publishing Fund.

Data Availability Statement: Data is contained within the article or supplementary material.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A Proof of Theorems

To prove Theorem 1, we need the following results:

Lemma A1 (Theorem 1 of Liu, Martin, and Syring 2017). *If* $Y_\alpha = Gamma(\alpha, 1)$, *then* $\lim_{\alpha \to 0+} -\alpha ln(Y_\alpha) \sim Exp(1)$ *in distribution [42].*

Corollary A1. $Pr\left(\lim_{\alpha \to 0+} Y_\alpha > 1\right) = 0.$

Proof of Corollary A1. From Lemma A1, $\Pr\left(\lim_{\alpha \to 0+} Y_\alpha > 1\right) = \Pr(\lim_{\alpha \to 0+} \ln(Y_\alpha) > 0) = \Pr(\lim_{\alpha \to 0+} -\alpha \ln(Y_\alpha) < 0) = \Pr(Exp(1) < 0) = 0.$ □

Corollary A2. *Let* $Y = \lim_{\alpha \to 0+} Y_\alpha^{-\alpha}$, *where* $Y_\alpha = Gamma(\alpha, 1)$, *then the PDF of* Y *is* $f_Y(y) = 1/y^2$ for $y \in (1, \infty)$.

Proof of Corollary A2. Notice that $Y = \lim_{\alpha \to 0+} Y_\alpha^{-\alpha} = \exp\left[\ln(\lim_{\alpha \to 0+} Y_\alpha^{-\alpha})\right] = \exp[\lim_{\alpha \to 0+} -\alpha \ln(Y_\alpha)]$. But from Lemma A1, $\lim_{\alpha \to 0+} -\alpha \ln(Y_\alpha) \sim Exp(1)$. Hence, the PDF of Y is $f_Y(y) = \exp(-\ln(y))/y = 1/y^2$ for $y \in [1, \infty)$. □

Lemma A2. *Let* $0 < p_1 < p_2 < 1$ *and* $q_i^{(\alpha)} = F_{G(\alpha)}^{-1}(p_i)$ $(i = 1, 2)$. *Denote* $r_\alpha = q_2^{(\alpha)}/q_1^{(\alpha)}$, *then we have* $\lim_{\alpha \to 0+} r_\alpha = \infty$. □

Proof of Lemma A2. Suppose $\lim_{\alpha \to 0+} r_\alpha = \infty$ does not hold; then, there exists a constant R such that $r_\alpha < R$ for any $\alpha > 0$. However, $0 < p_2 - p_1 = \Pr\left(q_1^{(\alpha)} < Y_\alpha < q_2^{(\alpha)}\right) = \Pr\left[-\alpha \ln\left(q_2^{(\alpha)}\right) < -\alpha \ln(Y_\alpha) < -\alpha \ln\left(q_1^{(\alpha)}\right)\right] = \Pr[-\alpha \ln\left(q_1^{(\alpha)}\right) - \alpha \ln(r_\alpha) < -\alpha \ln(Y_\alpha) < -\alpha \ln\left(q_1^{(\alpha)}\right)] < \Pr\left[-\alpha \ln\left(q_1^{(\alpha)}\right) - \alpha \ln(R) < -\alpha \ln(Y_\alpha) < -\alpha \ln\left(q_1^{(\alpha)}\right)\right] \to \Pr[-\alpha \ln\left(q_1^{(\alpha)}\right) < -\alpha \ln(Y_\alpha) < -\alpha \ln\left(q_1^{(\alpha)}\right)] = 0$ $(\alpha \to 0+)$, a contradiction. □

Corollary A3. $\lim_{\alpha \to 0+} T_{G(\alpha)}(P_1, P_2, \cdots, P_n) = \lim_{\alpha \to 0+} F_{G(\alpha)}^{-1}\left(1 - P_{(1)}\right)$, *where* $P_{(1)}$ *is the smallest value of* $P_1, P_2, \cdots, P_n$.

Proof of Corollary A3. This is a direct consequence of Lemma A2. □

Proof of Theorem 1. Now, we prove Theorem 1:

(i) Denote $Q_1 = \lim_{\alpha \to 0+} F_{G(\alpha)}^{-1}\left(1 - P_{(1)}\right)$. From Lemma A1 and Corollary A3, $P_{T_{G(0)}} = \lim_{\alpha \to 0+} \Pr(\text{Gamma}(\text{n}\alpha, 1) > Q_1) = \lim_{\alpha \to 0+} \Pr(Y_{n\alpha} > Q_1) = \lim_{\alpha \to 0+} \Pr(-n\alpha \ln(Y_{n\alpha}) < -n\alpha \ln(Q_1)) = 1 - \exp[n\alpha \ln(Q_1)]$. But, $\exp[n\alpha \ln(Q_1)] = \exp\left[\ln\left(Q_1^{n\alpha}\right)\right] = Q_1^{n\alpha} = \left(Q_1^{\alpha}\right)^n = \{\exp[\alpha \ln(Q_1)]\}^n = \left[\lim_{\alpha \to 0+} \Pr(-\alpha \ln(Y_\alpha) > \alpha \ln(Q_1))\right]^n = [\lim_{\alpha \to 0+} \Pr(Y_\alpha < Q_1)]^n = \left(1 - P_{(1)}\right)^n$. Hence, $P_{T_{G(0)}} = 1 - \left(1 - P_{(1)}\right)^n = P_{T_p}$, and $T_{G(0)} \equiv T_p$.

(ii) From the property of gamma distribution, we know that $Gamma(v/2, 2) = \chi_v^2$, a chi-square distribution with ν df. However, the sum of n *iid* χ_v^2 is χ_{nv}^2. Hence, let $\nu = 1$ or $\alpha = 0.5$, $T_{G(0.5,2)} \equiv \chi_n^2$. However, from Proposition 2, $T_{G(0.5,2)} \equiv T_{G(0.5)}$; therefore, $T_{G(0.5)} \equiv \chi_n^2$.

(iii) As in (ii), when $\nu = 2$ and $\alpha = 1$, $Gamma(1, 2) = \chi_2^2$; therefore, $T_{G(1)} \equiv T_{G(1,2)} \equiv F_p$.

(iv) From the property of gamma distribution, we know that $Gamma(\alpha, \beta) \to N\left(\alpha/\beta, \alpha/\beta^2\right)$ $(\alpha \to \infty)$. Hence, for $\beta = 1$, $Gamma(\alpha, 1) \to N(\alpha, \alpha)$, and $Gamma(na, 1) \to N(n\alpha, n\alpha)$. If we define $T'_{G(\alpha)} = \left(T_{G(\alpha)} - na\right)/\sqrt{na} = \sum_{i=1}^n \left[F_{G(\alpha)}^{-1}(1 - P_i) - a\right]/\sqrt{na}$, then $T'_{G(\alpha)} \to N(0, 1)$ $(\alpha \to \infty)$. On the other hand, since $T'_{G(\alpha)}$ is a linear transformation of $T_{G(\alpha)}$, it is easy to show that $T_{G(\alpha)} \equiv T'_{G(\alpha)}$ for any $\alpha > 0$. Hence, $T_{G(\infty)} \equiv \lim_{\alpha \to \infty} T'_{G(\alpha)} = Z_p$. □

Proof of Theorem 2 .

First, we show that $f_{\alpha,c}(p) = (1 - c)^\alpha \exp\left[cF_{G(\alpha)}^{-1}(1 - p)\right]$ is a PDF. Let $y = F_{G(\alpha)}^{-1}(1 - x)$, then $x = 1 - F_{G(\alpha)}(y) = \int_y^\infty t^{\alpha-1} \exp(-t)/\Gamma(\alpha) dt$, and $dx = -y^{\alpha-1} \exp(-y)/\Gamma(\alpha) dy$. Hence, $\int_0^1 f_X(x) dx = \int_0^1 (1 - c)^\alpha \exp\left[cF_{G(\alpha)}^{-1}(1 - x)\right] dx = \int_0^\infty (1 - c)^\alpha \exp(cy) y^{\alpha-1} \exp(-y)/\Gamma(\alpha) dy = 1$ as $(1 - c)^\alpha \exp(cy) y^{\alpha-1} \exp(-y)/\Gamma(\alpha) = (1 - c)^\alpha y^{\alpha-1} \exp[-(1 - c)y]/\Gamma(\alpha)$, the PDF of $Gamma(\alpha, 1 - c)$. However, under the global null hypothesis, $P_i \sim \text{U}(0, 1)$, the log-likelihood ratio under the global null and alternative hypotheses is $\sum_{i=1}^n \ln(f_{\alpha,c}(P_i)) = a(\alpha) + c \sum_{i=1}^n F_{G(\alpha)}^{-1}(1 - P_i)$, where $a(\alpha) = -n\alpha \ln(1 - c)$, a constant. Therefore, by the Neyman–Pearson lemma [43], $T_{G(\alpha)}$ is UMP under the specified condition. □

Proof of Theorem 3. Since the unconstrained MLE for c is $\hat{c}_{\alpha_0} = 1 - n\alpha_0/\sum_{i=1}^n F_{G(\alpha_0)}^{-1}(1 - p_i)$, when $\hat{c}_{\alpha_0} \le 0$, $T_{CLRT,\alpha_0} = 0$. On the other hand, when $\hat{c}_{\alpha_0} > 0$, i.e., $T_{G(\alpha_0)} = \sum_{i=1}^n F_{G(\alpha_0)}^{-1}(1 - p_i) > n\alpha_0$, $\hat{c}_{CLRT,\alpha_0} = \hat{c}_{\alpha_0}$, and $T_{CLRT,\alpha_0} = 2n\alpha_0 ln\left(n\alpha_0/\sum_{i=1}^n F_{G(\alpha_0)}^{-1}(1 - p_i)\right) + 2\sum_{i=1}^n F_{G(\alpha_0)}^{-1}(1-$

$p_i) - 2n\alpha_0 = 2n\alpha_0 ln\left(n\alpha_0/T_{G(\alpha_0)}\right) + 2T_{G(\alpha_0)} - 2n\alpha_0 = 2n\alpha_0 ln(n\alpha_0) - 2n\alpha_0 ln\left(T_{G(\alpha_0)}\right) + 2T_{G(\alpha_0)} - 2n\alpha_0$. For any $t > 0$, let $A = \{t|t < T_{CLRT,\alpha_0}\}$; it is easy to show that $A = \{t|t < T_{CLRT,\alpha_0}\} = \left\{t\middle|2n\alpha_0 ln(n\alpha_0) - 2n\alpha_0 ln\left(T_{G(\alpha_0)}\right) + 2T_{G(\alpha_0)} - 2n\alpha_0 > t\right\} = \left\{t\middle|2\left(T_{G(\alpha_0)} - n\alpha_0\right) > t + 2n\alpha_0 ln\left(T_{G(\alpha_0)}/n\alpha_0\right)\right\}$. Let $f(x) = x - n\alpha_0 ln(x) - n\alpha_0 - t/2 + n\alpha_0 \ln(n\alpha_0)$, then for $x > n\alpha_0$, $f'(x) = 1 - n\alpha_0/x > 0$, and $f(x)$ is an increasing function of x, but $\lim_{x\to(n\alpha_0)+} f(x) = -t/2 < 0$, and $f(kn\alpha_0) = (k - 1 - \ln(k))n\alpha_0 - t/2 > 0$ for large k. Hence, there must exist a unique $x_0 \in (n\alpha_0, \infty)$ such that $f(x_0) = 0$. Accordingly, $Pr\left[T_{CLRT,\alpha_0} > t\right] = Pr\left[T_{CLRT,\alpha_0} > t, \hat{c}_{CLRT,\alpha_0} > 0\right] = Pr[2n\alpha_0 ln(n\alpha_0) - 2n\alpha_0 ln\left(T_{G(\alpha_0)}\right) + 2T_{G(\alpha_0)} - 2n\alpha_0 > t \text{ and } T_{G(\alpha_0)} > n\alpha_0] = Pr[T_{G(\alpha_0)} - n\alpha_0 ln\left(T_{G(\alpha_0)}\right) - n\alpha_0 - t/2 + n\alpha_0 ln(n\alpha_0) > 0,\ T_{G(\alpha_0)} > n\alpha_0] = Pr[T_{G(\alpha_0)} > t_0]$, where t_0 is the root of $f(x)$, i.e., $f(t_0) = 0$. This shows that when $T_{G(\alpha_0)} > n\alpha_0$, T_{CLRT,α_0} and $T_{G(\alpha_0)}$ have the same p-value. On the other hand, when $T_{G(\alpha_0)} \le n\alpha_0$, $T_{CLRT,\alpha_0} = 0$; hence, $Pr\left[T_{CLRT,\alpha_0} = 0\right] = Pr\left[T_{G(\alpha_0)} < n\alpha_0\right]$. □

References

1. Fisher, R.A. *Statistical Methods for Research Workers*, 4th ed.; Oliver and Boyd: Edinburgh, UK, 1932.
2. Pearson, K. On a New Method of Determining "Goodness of Fit". *Biometrika* **1934**, *26*, 425.
3. Stouffer, S.A.; Suchman, E.A.; DeVinney, L.C.; Star, S.A.; Williams, R.M., Jr. *The American Soldier: Adjustment during Army Life. (Studies in Social Psychology in World War II)*; Princeton University Press: Princeton, NJ, USA, 1949; Volume 1.
4. Tippett, L.H.C. *Methods of Statistics*; Williams Norgate: London, UK, 1931.
5. Chen, Z. Is the weighted z-test the best method for combining probabilities from independent tests? *J. Evol. Biol.* **2011**, *24*, 926–930. [CrossRef]
6. Loughin, T.M. A systematic comparison of methods for combining *p*-values from independent tests. *Comput. Stat. Data Anal.* **2004**, *47*, 467–485. [CrossRef]
7. Whitlock, M.C. Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **2005**, *18*, 1368–1373. [CrossRef]
8. Liu, Y.; Xie, J. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* **2020**, *115*, 393–402. [CrossRef] [PubMed]
9. Chen, Z. Robust tests for combining p-values under arbitrary dependency structures. 2021; unpublished.
10. Owen, A.B. Karl Pearson's meta-analysis revisited. *Ann. Stat.* **2009**, *37*, 3867–3892. [CrossRef]
11. Hedges, L.; Olkin, I. *Statistical Methods for Meta-Analysis*; Academic: San Diego, CA, USA, 1985.
12. Chen, Z.; Wang, K. Gene-based sequential burden association test. *Stat. Med.* **2019**, *38*, 2353–2363. [CrossRef] [PubMed]
13. Chen, Z.; Liu, Q.; Wang, K. A novel gene-set association test based on variance-gamma distribution. *Stat. Methods Med. Res.* **2018**, *28*, 2868–2875. [CrossRef]
14. Chen, Z.; Liu, Q.; Wang, K. A genetic association test through combining two independent tests. *Genomics* **2019**, *111*, 1152–1159. [CrossRef] [PubMed]
15. Chen, Z.; Lu, Y.; Lin, T.; Liu, Q.; Wang, K. Gene-based genetic association test with adaptive optimal weights. *Genet. Epidemiol.* **2018**, *42*, 95–103. [CrossRef]
16. Chen, Z.; Wang, K. A gene-based test of association through an orthogonal decomposition of genotype scores. *Hum. Genet.* **2017**, *136*, 1385–1394. [CrossRef]
17. Chen, Z.; Ng, H.K.T.; Li, J.; Liu, Q.; Huang, H. Detecting associated single-nucleotide polymorphisms on the X chromosome in case control genome-wide association studies. *Stat. Methods Med. Res.* **2017**, *26*, 567–582. [CrossRef]
18. Chen, Z.; Lin, T.; Wang, K. A powerful variant-set association test based on chi-square distribution. *Genetics* **2017**, *207*, 903–910. [CrossRef] [PubMed]
19. Chen, Z.; Han, S.; Wang, K. Genetic association test based on principal component analysis. *Stat. Appl. Genet. Mol. Biol.* **2017**, *16*, 189–198. [CrossRef]
20. Chen, Z. Testing for gene-gene interaction in case-control GWAS. *Stat. Its Interface* **2017**, *10*, 267–277. [CrossRef]
21. Choquet, H.; Melles, R.B.; Anand, D.; Yin, J.; Cuellar-Partida, G.; Wang, W.; Hoffmann, T.J.; Nair, K.S.; Hysi, P.G.; Lachke, S.A.; et al. A large multiethnic GWAS meta-analysis of cataract identifies new risk loci and sex-specific effects. *Nat. Commun.* **2021**, *12*, 3595. [CrossRef]
22. Schwantes-An, T.H.; Darlay, R.; Mathurin, P.; Masson, S.; Liangpunsakul, S.; Mueller, S.; Aithal, G.P.; Eyer, F.; Gleeson, D.; Thompson, A.; et al. Genome-wide Association Study and Meta-analysis on Alcohol-Associated Liver Cirrhosis Identifies Genetic Risk Factors. *Hepatology* **2021**, *73*, 1920–1931. [CrossRef]

23. Birnbaum, A. Combining Independent Tests of Significance. *J. Am. Stat. Assoc.* **1954**, *49*, 559–574.
24. Bonferroni, C. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*; Bardi: Rome, Italy, 1935; pp. 13–60.
25. Lancaster, H. The combination of probabilities: An application of orthonormal functions. *Aust. J. Stat.* **1961**, *3*, 20–33. [CrossRef]
26. Chen, Z.; Nadarajah, S. On the optimally weighted z-test for combining probabilities from independent studies. *Comput. Stat. Data Anal.* **2013**, *70*, 387–394. [CrossRef]
27. Berk, R.H.; Cohen, A. Asymptotically optimal methods of combining tests. *J. Am. Stat. Assoc.* **1979**, *74*, 812–814. [CrossRef]
28. Birnbaum, A. Characterizations of complete classes of tests of some multiparametric hypotheses, with applications to likelihood ratio tests. *Ann. Math. Stat.* **1955**, *26*, 21–36. [CrossRef]
29. Bahadur, R.R. Rates of Convergence of Estimates and Test Statistics. *Ann. Math. Stat.* **1967**, *38*, 303–324. [CrossRef]
30. Self, S.G.; Liang, K.Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **1987**, *82*, 605–610. [CrossRef]
31. Bachmann, S.; Finger, C.; Huss, A.; Egger, M.; Stuck, A.E.; Clough-Gorr, K.M. Inpatient rehabilitation specifically designed for geriatric patients: Systematic review and meta-analysis of randomised controlled trials. *BMJ* **2010**, *340*, c1718. [CrossRef]
32. Riley, R.D.; Higgins, J.; Deeks, J. Interpretation of random effects meta-analyses. *BMJ* **2011**, *342*, d549. [CrossRef]
33. Chen, Z.; Zhang, G.; Li, J. Goodness-of-fit test for meta-analysis. *Sci. Rep.* **2015**, *5*, 16983. [CrossRef]
34. The Digitalis Investigation Group. The effect of digoxin on mortality and morbidity in patients with heart failure. *N. Engl. J. Med.* **1997**, *336*, 525–533. [CrossRef]
35. Qiu, P.; Sheng, J. A two-stage procedure for comparing hazard rate functions. *J. R. Stat. Soc. Ser. B* **2007**, *70*, 191–208. [CrossRef]
36. Chen, Z.; Huang, H.; Qiu, P. Comparison of multiple hazard rate functions. *Biometrics* **2015**, *72*, 39–45. [CrossRef]
37. Mosteller, F.; Bush, R.; Lindzey, G. *Handbook of Social Psychology*; Addison-Wesley: Cambridge, MA, USA, 1954; pp. 289–334.
38. Good, I. On the weighted combination of significance tests. *J. R. Stat. Soc. Ser. B* **1955**, *17*, 264–265. [CrossRef]
39. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 2000; Volume 3.
40. Agresti, A. *Categorical Data Analysis*; Wiley-Interscience: Hoboken, NJ, USA, 2002.
41. Lancaster, H. The derivation and partition of $\chi 2$ in certain discrete distributions. *Biometrika* **1949**, *36*, 117. [CrossRef] [PubMed]
42. Liu, C.; Martin, R.; Syring, N. Efficient simulation from a gamma distribution with small shape parameter. *Comput. Stat.* **2017**, *32*, 1767–1775. [CrossRef]
43. Casella, G.; Berger, R.L. *Statistical Inference*; Duxbury: Pacific Grove, CA, USA, 2002; Volume 2.