The harmonic mean χ2 test to substantiate scientific findings

Held, Leonhard

DOI: https://doi.org/10.1111/rssc.12410

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2020)
69, *Part* 3, *pp.* 697–708

The harmonic mean χ^2-test to substantiate scientific findings

Leonhard Held

University of Zurich, Switzerland

[Received November 2019. Revised March 2020]

Summary. Statistical methodology plays a crucial role in drug regulation. Decisions by the US Food and Drug Administration or European Medicines Agency are typically made based on multiple primary studies testing the same medical product, where the two-trials rule is the standard requirement, despite shortcomings. A new approach is proposed for this task based on the harmonic mean of the squared study-specific test statistics. Appropriate scaling ensures that, for any number of independent studies, the null distribution is a χ^2-distribution with 1 degree of freedom. This gives rise to a new method for combining one-sided p-values and calculating confidence intervals for the overall treatment effect. Further properties are discussed and a comparison with the two-trials rule is made, as well as with alternative research synthesis methods. An attractive feature of the new approach is that a claim of success requires each study to be convincing on its own to a certain degree depending on the overall level of significance and the number of studies. The new approach is motivated by and applied to data from five clinical trials investigating the effect of carvedilol for the treatment of patients with moderate to severe heart failure.

Keywords: Combining p-values; Drug regulation; Evidence synthesis; Two-trials rule; Type I error

1. Introduction

Research synthesis has been characterized as the process of combining the results of multiple primary studies aimed at testing the same conceptual hypothesis. Meta-analysis is the preferred technique of quantitative research synthesis, as it provides overall effect estimates with confidence intervals and p-values through pooling of study results and allows for the incorporation of heterogeneity between studies. However, meta-analysis can be criticized as a too weak technique if the goal is to substantiate an original claim through one or more additional independent studies. Specifically, a significant overall effect estimate may occur even if some of the individual studies have not been convincing on their own, perhaps even with effect estimates in the wrong direction. This may be acceptable if the unconvincing studies have been small, but seems less tolerable if each study was well powered and well conducted.

For example, consider the results from five clinical trials on the effect of carvedilol, a beta- and alpha-blocker and an antioxidant drug for the treatment of patients with moderate to severe heart failure, on mortality (see Fisher (1999a), Table 1). One-sided p-values (from log-rank tests) and estimated hazard ratios HR are shown in Table 1, indicating a reduction in instantaneous risk of death between 28% and 78% across the various studies.

A meta-analysis could be applied to the data that are shown in Table 1, but the drug regu-

Address for correspondence: Leonhard Held, Epidemiology, Biostatistics and Prevention Institute and Center for Reproducible Science, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland.
E-mail: leonhard.held@uzh.ch

 0035–9254/20/69697

Table 1. Results from five clinical trials on the effect of carvedilol for the treatment of patients with moderate to severe heart failure†

Study	*p-value*	*HR*	*log(HR)*	*SE*
220	0.00025	0.27	−1.31	0.41
240	0.0245	0.22	−1.51	0.85
223	0.128	0.72	−0.33	0.29
221	0.1305	0.57	−0.56	0.51
239	0.2575	0.53	−0.63	1.02

†Shown are one-sided p-values, estimated hazard ratios HR and the associated log-hazard-ratios log(HR) with standard errors SE.

lation industry (including the US Food and Drug Administration) typically relies instead on the 'two-trials rule' (Senn, 2007; Kay, 2015), which is also known as the 'two-pivotal-study paradigm' (Hlavin *et al.*, 2016), for approval. This simple decision rule requires 'at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness' (Food and Drug Administration (1998), page 3). This is usually achieved by independently replicating the result of a first study in a second study, both significant at one-sided level $\alpha = 0.025$. However, in modern drug development often more than two trials are conducted and it is unclear how to extend the two-trials rule to this setting. Requiring at least two out of $n > 2$ studies to be significant is too lax a criterion if the results from the non-significant studies are not taken into account at all. However, requiring all n studies to be significant is too stringent. This problem applies to the carvedilol example, where two trials are significant at the 2.5% level (one just with $p = 0.0245$) but where it is unclear whether the remaining three studies (with p-values 0.128, 0.1305 and 0.2575) can be considered as sufficiently 'convincing on its own'.

This has led statistical researchers to discuss the possibility of pooling the results from the various studies into one p-value (Fisher, 1999b; Darken and Ho, 2004; Shun *et al.*, 2005). Fisher's method to combine p-values (Fisher, 1958) is often used for this task, e.g. in Fisher (1999a) for the carvedilol example. However, Fisher's method shares the problems of a meta-analysis as it can produce a significant overall result even if one of the trials was negative. For example, one completely unconvincing trial with (one-sided) $p = 0.5$ combined with a convincing second trial with $p = 0.0001$ would give Fisher's $p = 0.0005 < 0.000625 = 0.025^2$, and so a claim of success with respect to the type I error rate of the two-trials rule. In contrast, two trials both with $p = 0.01$ would give Fisher's $p = 0.001$ and so would not be considered as successful. Both decisions seem undesirable from a regulator's perspective.

The two-trials rule therefore remains the standard in drug regulation, but has additional deficiencies even for $n = 2$ studies, where independent p-value thresholding at 0.025 may lead to decisions that are the opposite to what the evidence warrants. For example, two trials both with $p = 0.024$ will lead to drug approval but carry less evidence for a treatment effect than one trial with $p = 0.026$ and the other trial with $p = 0.001$, which would, however, not pass the two-trials rule. Rosenkrantz (2002) has therefore proposed a method to claim efficacy if one of two trials is significant whereas the other just shows a trend. He combined the two-trials rule with Fisher's method and a relaxed criterion for significance of the two individual trials, say 2α. A similar approach has been proposed by Maca *et al.* (2002) using meta-analytic pooling rather than Fisher's combined method. The arbitrariness in the choice of the relaxed significance criterion

is less attractive, though, and it is not obvious how to extend the methods to results from more than two studies.

In this paper I develop a new method that addresses these issues and leads to more appropriate inferences: the harmonic mean χ^2-test described in Section 2. At the type I error rate 0.025^2 of the two-trials rule, the test proposed comes to opposite conclusions for the examples that were mentioned above: contrary to Fisher's method, it leads to approval of two trials both with $p = 0.01$, but not to approval if one has $p = 0.0001$ and the other has $p = 0.5$. Contrary to the two-trials rule, it leads to approval of one trial with $p = 0.026$ and the other with $p = 0.001$, but not to approval if both trials have $p = 0.024$. The work is motivated from a recent proposal how to evaluate the success of replication studies (Held, 2020) and is based on the harmonic mean of the squared Z-scores. It can include weights for the individual studies and can be calibrated to ensure exact type I error control and to compute an overall p-value; see Section 2.1. Furthermore, the new approach implies useful bounds on the study-specific p-values, thus formalizing the meaning of 'at least two adequate and well-controlled studies, *each convincing on its own*'. It can also be used to calculate a confidence interval for the overall treatment effect; see Section 2.2. The approach will be compared with the two-trials rule in Section 3 and applied to the carvedilol data in Section 4. I close with some discussion in Section 5.

2. The harmonic mean χ^2-test

Suppose that one-sided p-values $p_1, \ldots, p_n$ are available from n independent studies. How can we combine the p-values into one p-value? Cousins (2008) compared some of the more prominent references on this topic. Among them is Stouffer's method, which is based on the Z-scores $Z_i = \Phi^{-1}(1 - p_i)$; here $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution. Under the assumption of no effect, the test statistic $Z = \Sigma_{i=1}^n Z_i/\sqrt{n}$ is standard normally distributed. The corresponding p-value forms the basis of the 'pooled trials rule' and is equivalent to investigating significance of the overall effect estimate from a fixed effects meta-analysis (Senn (2007), section 12.2.8). Fisher's method is also commonly used and compares $-2\Sigma_{i=1}^n \log(p_i)$ with a χ^2-distribution with $2n$ degrees of freedom to compute a combined p-value. Both Stouffer's and Fisher's method can be extended to incorporate weights, where the null distribution of Fisher's method no longer has a convenient form (Good, 1955). There is a large literature on the comparison of these and other methods for the combination of p-values, such as Littell and Folks (1973), Berk and Cohen (1979), Westberg (1985) and Heard and Rubin-Delanchy (2018).

Here I propose a new approach to assess the overall evidence for a treatment effect based on the harmonic mean $Z_{\mathrm{H}}^2 = n/\Sigma_{i=1}^n 1/Z_i^2$ of the squared Z-scores:

$$X^2 = nZ_{\mathrm{H}}^2 = \frac{n^2}{\sum_{i=1}^n 1/Z_i^2}. \tag{1}$$

This form is motivated from the special case of $n = 2$ successive studies, one original and one replication, where a reverse Bayes approach for the assessment of replication success has recently been described (Held, 2020). If the two studies have equal precision (i.e. sample size), the assessment of replication success does not depend on the order of the two studies and is based on the test statistic $1/(1/Z_1^2 + 1/Z_2^2)$; compare Held (2020), equation (9). Equation (1) extends this to n studies with an additional multiplicative factor n^2, which ensures that the null distribution of equation (1) does not depend on n. Weights $w_1, \ldots, w_n$ can also be introduced in equation (1); then the test statistic

$$X_w^2 = \frac{w^2}{\sum_{i=1}^n w_i/Z_i^2}, \qquad w = \sum_{i=1}^n \sqrt{w_i}, \tag{2}$$

should be used. The factor w^2 ensures that the null distribution of equation (2) does not depend on the weights $w_1, \ldots, w_n$; nor on n.

The specific form of equation (2) deserves some additional comments. In practice we often have $Z_i = \hat{\theta}_i/\sigma_i$ where $\sigma_i = \kappa/\sqrt{m_i}$ is the standard error of the effect estimate $\hat{\theta}_i$, κ^2 is the one-unit variance and m_i the sample size of study i. If we use weights $w_i = 1/\sigma_i^2$ equal to the precision of the effect estimates, equation (2) can be written as the unweighted harmonic mean $\hat{\theta}_{\mathrm{H}}^2$ of the squared effect estimates $\hat{\theta}_i^2$ times a scaling factor w^2/n:

$$X_w^2 = (w^2/n)\hat{\theta}_{\mathrm{H}}^2, \qquad w = \sum_{i=1}^n \sqrt{m_i}. \tag{3}$$

In the special case of equal sample sizes $m_1 = \ldots = m_n = m$, the scaling factor reduces to nm.

There is a subtle difference between the two formulations (1) and (3). The unweighted test statistic (1) is based on the harmonic mean of the squared study-specific test statistics Z_i^2, $i = 1, \ldots, n$. If we increase the sample size of the different studies, statistics (1) will therefore also tend to increase if there is a true non-zero effect. However, test statistic (3) is based on the harmonic mean $\hat{\theta}_{\mathrm{H}}^2$ of the squared study-specific effect estimates $\hat{\theta}_i^2$, which should not be much affected by any increase of study-specific sample sizes because the study-specific estimates $\hat{\theta}_i$ should then stabilize around their true values. It is the scaling factor w^2/n that will react to an increase in study-specific sample sizes. The test statistic (3) can thus be factorized into a component depending on sample sizes and a component depending on effect sizes.

2.1. *p*-values

Using properties of Lévy distributions it can be shown that, under the null hypothesis of no effect, the distribution of both statistic (1) and statistic (2) is χ^2 with 1 degree of freedom; see Appendix A for details. We can thus compute an overall p-value p_{H} from equation (1) or (2) based on the $\chi^2(1)$ distribution function. However, we must be careful since equation (1) does not take the direction of the effects into account. Usually we are interested in a predefined direction of the underlying effect, say $H_1 : \theta > 0$ against $H_0 : \theta = 0$, and we will have to adjust for the fact that statistics (1) and (2) can be large for any of the 2^n possible combinations of the signs of $Z_1, \ldots, Z_n$, with all these combinations being equally likely under the null hypothesis. Since we are interested only in the case where all signs are positive, we must adjust the p-value accordingly.

To be specific, suppose that all studies have a positive effect and the observed test statistic (1) or (2) is respectively $X^2 = x^2$ or $X_w^2 = x^2$ and let $x = \sqrt{x^2}$. The overall p-value from the significance test proposed is then

$$p_{\mathrm{H}} = \frac{\Pr\{\chi^2(1) \geqslant x^2\}}{2^n} = \frac{1 - \Phi(x)}{2^{n-1}}. \tag{4}$$

Likewise we can obtain the critical value

$$c_{\mathrm{H}} = \{\Phi^{-1}(1 - 2^{n-1}\alpha_{\mathrm{H}})\}^2 \tag{5}$$

for the test statistic (1) or (2) to control the type I error rate at some overall level of significance α_{H}. Note that the overall p-value (4) cannot be larger than $1/2^n$ as it should, since under the null hypothesis the probability of obtaining n positive results is $1/2^n$. We are interested only in

this case, so if at least one of the studies has a negative effect I suggest reporting the inequality $p_H > 1/2^n$, e.g. $p_H > 0.25$ for $n = 2$ studies.

In what follows I restrict attention to the unweighted test statistic X^2 given in equation (1). Let $Z_i = z_i$ denote the observed test statistic in the ith study. I assume that $z_i > 0$ for all $i = 1, \ldots, n$, i.e. all effects go in the right direction. First note that the smallest squared test statistic $z^2_{min} = \min\{z_1^2, \ldots, z_n^2\}$ multiplied by the number of studies n is an upper bound on the harmonic mean $z^2_H = n/\Sigma_{i=1}^n 1/z_i^2$:

$$z^2_H \leqslant n z^2_{min} \leqslant n z_i^2,$$

where the second inequality holds for all $i = 1, \ldots, n$. This implies that $x^2 \leqslant n^2 z_i^2$ for the observed test statistic x^2 and any study $i = 1, \ldots, n$ and with equation (4) we obtain

$$\Pr\{\chi^2(1) \geqslant n^2 z_i^2\}/2^n \leqslant p_H.$$

If $p_H \leqslant \alpha_H$ is required for a claim of success at level α_H, then obviously $\Pr\{\chi^2(1) \geqslant n^2 z_i^2\}/2^n \leqslant \alpha_H$ must hold, which can be rewritten as $z_i \geqslant \sqrt{c_H}/n$ with c_H given in equation (5). The restriction on the corresponding p-values is

$$p_i \leqslant 1 - \Phi(\sqrt{c_H}/n). \tag{6}$$

The right-hand side of inequality (6) is thus a necessary but not sufficient bound on the study-specific p-values for a claim of success.

It is also possible to derive the corresponding sufficient bound. Assume that all p-values are equal (i.e. $z_1^2 = \ldots = z_n^2$); then the condition $X^2 = n z_i^2 \geqslant c_H$ implies that $z_i \geqslant \sqrt{c_H}/\sqrt{n}$. Note that the sufficient bound on z_i differs from the necessary bound by a factor of $\sqrt{n}$. The restriction on the corresponding p-values is now

$$p_i \leqslant 1 - \Phi(\sqrt{c_H}/\sqrt{n}). \tag{7}$$

For $n = 1$ the necessary and sufficient bounds in inequalities (6) and (7) both reduce to α_H, as they should.

The two-trials rule for drug approval is usually implemented by requiring that each study is significant at the one-sided level $\alpha = 1/40 = 0.025$, so the probability of $n = 2$ significant positive trials when there is no treatment effect is $\alpha^2 = 1/1600 = 0.000625$. The necessary and sufficient bounds in inequalities (6) and (7) respectively are shown in Table 2 for $\alpha_H = 1/1600$ (the two-trials rule), 1/31574 (the 4σ-rule) and 1/3488556 (the 5σ-rule). The level of significance of the $k\sigma$-rule is based on a normally distributed test statistic $T \sim N(0, \sigma^2)$ with zero mean and defined as $\Pr(T > k\sigma) = 1 - \Phi(k)$. The 5σ-rule ($k = 5$) was used to declare the discovery of the Higgs boson (Johnson (2013), section 3.2.1). The two-trials rule corresponds to $k = 3.23$, so the level of significance of the 4σ-rule is between the two-trials rule and the 5σ-rule.

The first row of Table 2 reveals that, for level 1/1600, the requirement $p_i \leqslant 0.065$, $i = 1, 2$, is necessary for claiming success based on $n = 2$ studies. If one of the two studies has a p-value that is larger than 0.065, a claim of success at level $\alpha_H = 1/1600$ is thus impossible, no matter how small the other p-value is. Both p-values being smaller than 0.016 is sufficient for a claim of success at that level. With increasing n both bounds increase; for example for $n = 6$ studies it is necessary that each p-value is smaller than 0.37 whereas it is sufficient that each p-value is smaller than 0.20. Decreasing the level of significance from 1/1600 to 1/31574 gives similar bounds for $n + 1$ rather than n studies, and likewise for another decrease from 1/31574 to 1/3488556. For example, the necessary bound is 0.17 for $\alpha_H = 1/1600$ and $n = 3$, 0.19 for $\alpha_H = 1/31574$ and $n = 4$, and again 0.19 for $\alpha_H = 1/3488556$ and $n = 5$.

Table 2. Necessary and sufficient bounds on the one-sided study-specific p-values for overall significance level α_H and various numbers of studies n

α_H	*Bound*	*Results for the following values of n:*				
		$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
1/1600	Necessary	0.065	0.17	0.26	0.32	0.37
	Sufficient	0.016	0.053	0.099	0.15	0.20
1/31574	Necessary	0.028	0.11	0.19	0.26	0.30
	Sufficient	0.0034	0.017	0.041	0.071	0.10
1/3488556	Necessary	0.0075	0.058	0.13	0.19	0.24
	Sufficient	0.00029	0.0032	0.011	0.024	0.04

2.2. Confidence intervals

The harmonic mean χ^2-test is not directly linked to an overall effect estimate and a confidence interval. However, the test can be inverted to obtain a confidence interval. Two extensions of the method are required to do so. First, we need to consider test statistics $Z_i = (\hat{\theta}_i - \mu)/\sigma_i$ for the more general point null hypothesis $H_0 : \theta = \mu$. Second, to compute a two-sided confidence interval we need to calculate a two-sided rather than one-sided p-value. A two-sided p-value defined as twice the one-sided p-value (4) represents the common scenario that an initial study is two sided and all following studies aim to substantiate the effect of the first study including its direction and so are one sided. The two-sided p-value $2p_H$ can hence be evaluated not only if all effect estimates are positive, but also if all effect estimates are negative. If the effect estimates are not all in the same direction I now suggest reporting $2p_H > 1/2^{n-1}$.

We can now calculate a p-value function (see Infanger and Schmidt-Trucksäss (2019) for a recent review), displaying the two-sided harmonic mean p-value as a function of μ. A two-sided confidence interval at any level $\gamma > 1 - 1/2^{n-1}$ can then be defined as the set of μ-values where the two-sided p-value is larger than $1 - \gamma$. An example is given in Section 4.

3. Comparison with the two-trials rule

Suppose that both studies have a positive effect in the right direction and the observed test statistic (1) is $X^2 = x^2$. The harmonic mean χ^2 p-value (4) now reduces to $p_H = \{1 - \Phi(x)\}/2$. A critical value for the test statistic (1) can also be calculated by using inequality (5). For $\alpha_H = 0.025^2$ and $n = 2$ we obtain the critical value $c_H = 9.14$.

Fig. 1 compares the region for drug approval based on the two-trials rule with the proposed harmonic mean χ^2-test. Shown are two versions of the latter, the 'controlled' version based on $\alpha_H = 0.025^2$, i.e. critical value $c_H = 9.14$, and a 'liberal' version with critical value 7.68. The liberal version has been computed by equating the right-hand side of inequality (7) with 0.025 and solving for c_H. The liberal version thus ensures that approval by the two-trials rule always leads to approval by the harmonic mean χ^2-test. The type I error rate of the liberal version is 0.00139, inflated by a factor of 2.23 compared with the $\alpha^2 = 0.025^2$-level.

Also shown in Fig. 1 is the corresponding region for drug approval of the pooled and combined method, both controlled at type I error 0.025^2. Both methods compensate smaller intersections with the two-trials rejection region with additional regions of rejection where one of the trials shows only weak or even no evidence for an effect. It is interesting to see that the harmonic mean χ^2-test is closer to the two-trials rule than Stouffer's pooled or Fisher's combined method, which

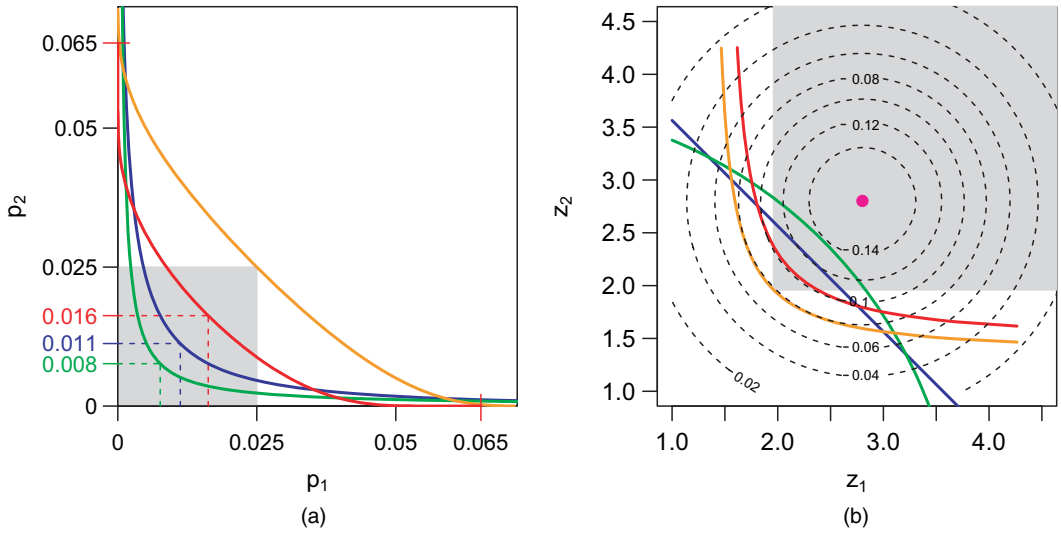

Fig. 1. Comparison of various approaches for drug approval depending on (a) the p-values p_1 and p_2 and (b) the Z-values Z_1 and Z_2 (■, rejection region of the two-trials rule; ——, harmonic (liberal); ——, harmonic (controlled); ——, pooled; ——, combined): the rejection region of the other methods is (a) below or (b) above the corresponding curves; all methods control the type I error rate at 0.000625 except for the liberal version of the harmonic mean χ^2-test, which has type I error rate 0.00139; the contour lines in (b) represent the distribution of Z_1 and Z_2 under the alternative if the two studies have 80% power at the one-sided 2.5% level of significance

is particularly good to see in the z-scale shown in Fig. 1(b). Stouffer's and Fisher's methods suffer from the possibility of approval if one of the p-values is very small whereas the other is far from traditional significance. A highly significant p-value may actually guarantee approval through Fisher's method, no matter how large the p-value from the other study is. This is not possible for Stouffer's method, but it may still happen that the effects from the two studies go in different directions with the combined effect being significant. As a consequence, the sufficient p-value bound, which is shown in Fig. 1(a), is considerably smaller for the pooled (0.011) and combined (0.008) method than for the harmonic mean χ^2-test (0.016) with the same type I error rate. These features make both the pooled and the combined method less suitable for drug approval.

The harmonic mean χ^2-test can be significant only if both p-values are small (less than 0.065). This has been discussed in Section 2 and can also be seen from Fig. 2, which shows the conditional power for drug approval given the p-value p_1 from the first study. The values represent the power to detect the observed effect from the first study with a second study of equal design and sample size. The two-trials rule has conditional power as described by Goodman (1992), but with a discontinuity at 0.025. The power curves of the two harmonic mean χ^2-tests (calculated with the results given in Held (2020), section 4) are smooth, quickly approaching zero at $p_1 = 0.065$ and $p_1 = 0.083$. Both the combined and the pooled method have longer tails with non-zero conditional power even for a larger p-value of the first study. Here the conditional power of the combined method can be derived as $1 - \Phi[\Phi^{-1}(p_1) - \Phi^{-1}\{\min(1, c/p_1)\}]$ where $c = \Pr\{\chi^2(4) \geqslant \alpha_H\}$. The conditional power of the pooled method turns out to be $1 - \Phi\{2\Phi^{-1}(p_1) - \sqrt{2}\Phi^{-1}(\alpha_H)\}$.

Of central interest in drug development is often the 'project power' for a claim of success before the two trials are conducted (Maca *et al.*, 2002). It is well known (Matthews, 2006) that, under the alternative that was used to power the two trials, the distribution of Z_1 and Z_2 is

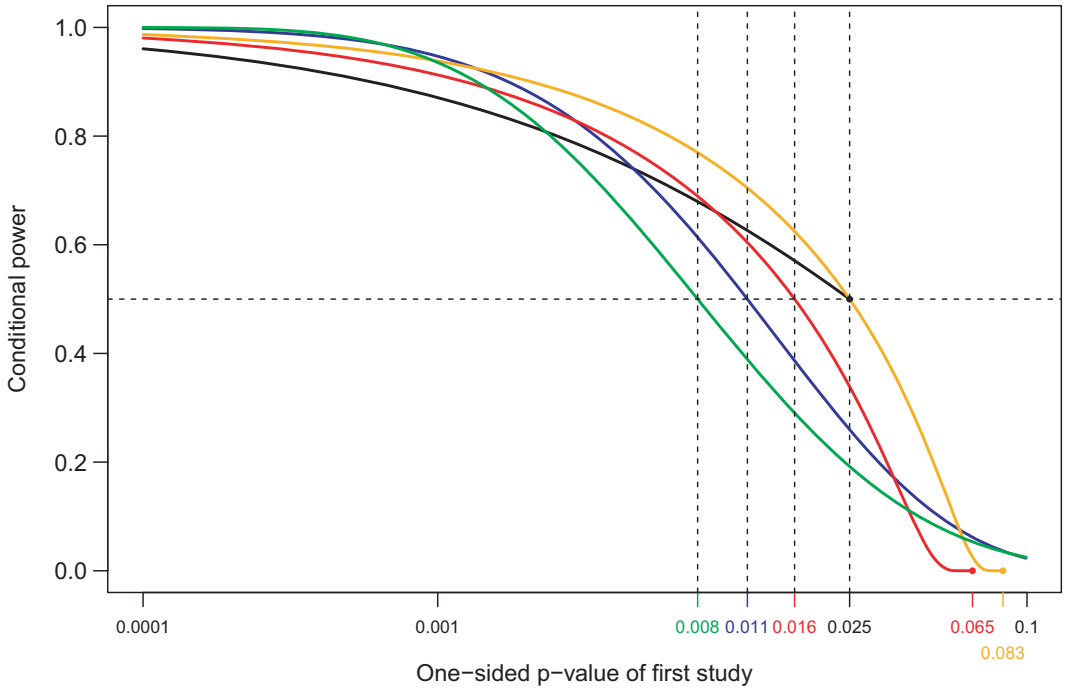

Fig. 2. Power for drug approval conditional on the one-sided p-value of the first study (power values of exactly 0 have been omitted: ——, two-trials rule; ——, pooled; ——, combined; ——, harmonic (controlled); ——, harmonic (liberal))

Table 3. Individual trial power and project power of the various methods for drug approval

Trial power (%)	*Project power (%)*			
	Two-trials rule	*Harmonic*	*Combined*	*Pooled*
70	49	56	58	61
80	64	71	74	77
90	81	87	90	91
95	90	94	96	97

$N(\mu, 1)$ where $\mu = \Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)$, where $1-\beta$ is the power of each trial. We can thus simulate independent Z_1 and Z_2 for $\alpha = 0.025$ and different values of the individual trial power $1-\beta$ and compute the proportion of results with drug approval at level α^2. This is shown in Table 3 for the various methods.

As expected, the two-trials rule gives project power equal to $(1-\beta)^2$, since the two trials are assumed to be independent, each significant with probability $1-\beta$. The project power of the type I error controlled harmonic mean χ^2-test is 4–7 percentage points larger, depending on the power of the two trials. The project power of the combined and pooled methods are even larger but this comes at the price that approval may be granted even if one of the two trials was not sufficiently convincing on its own.

4. Application

Two advantages of the method proposed are that it allows for weighting and is readily applicable to the case where results from more than two studies are available. Consider again the data shown in Table 1 on the effect of carvedilol on mortality. Note that all p-values are below the necessary success bound 0.32 at the level of the two-trials rule; compare with Table 2. Only the p-value of study 239 is above the sufficient bound 0.15; otherwise we could already claim success with the unweighted harmonic mean χ^2-test.

Fisher (1999a) reported Fisher's combined p-value, which is 0.00013. Stouffer's unweighted pooled test gives the p-value 0.00009; the weighted version gives $p = 0.00018$. For the latter the weights have been chosen inversely proportional to the squared standard errors of the associated log-hazard-ratios that are also shown in Table 1; see Appendix B for further details. The harmonic mean χ^2-test gives 0.00048 (unweighted) and 0.00034 (weighted); so slightly larger values. Note that all these p-values are smaller than the threshold 0.000625 of the two-trials rule.

I have also calculated two confidence intervals based on the inversion of the weighted harmonic mean χ^2-test as described in Section 2.2. The 99.875% confidence interval for the hazard ratio θ goes from 0.17 to 0.97. The confidence level is selected to be compatible with the one-sided type I error rate $\alpha_H = 0.000625$ of the two-trials rule, as $1 - 2 \times 0.000625 = 0.99875$. The more standard 95% confidence interval for the hazard ratio goes from 0.21 to 0.74. For comparison, a random-effects meta-analysis gives the 95% confidence interval 0.25–0.77 (two-sided $p = 0.004$). A fixed effects meta-analysis gives the 95% confidence interval 0.32–0.72. The corresponding two-sided p-value is 0.00035.

Suppose now that the p-value in study 223 (the largest study with the smallest standard error) is twice as large, i.e. 0.256 rather than 0.128. This would be considered as unimportant by many scientists, as both p-values are non-significant anyway and far from the standard 0.025 significance threshold. Keeping the standard error of the log-relative-risk fixed, the estimated hazard ratio in this study is now 0.83 rather than 0.72.

This change has a noticeable effect on the method proposed: the unweighted and weighted harmonic mean χ^2-test p-values increase by a factor of 2.5 and 7.9 to 0.0012 and 0.0027 respectively, so both would now fail the $0.025^2 = 0.000625$-threshold for drug approval. The p-values of the unweighted and weighted Stouffer's test increase by only a factor of 2.3 and 3.5 to 0.00021 and 0.00061 respectively. Both p-values are still below the 0.000625-threshold, and this is also so for Fisher's combined p-value, which increases by a factor of 1.7 to 0.00022. This illustrates that the harmonic mean χ^2-test is more sensitive to studies with unconvincing results, i.e. relatively small effect sizes with large p-values.

5. Discussion

There is considerable variation of clinical trial evidence for newly approved therapies (Downing *et al.*, 2014). New methods are required to provide better inferences for the assessment of pivotal trials supporting novel therapeutic approval. The harmonic mean χ^2-test is an attractive alternative to the two-trials rule as it has more power at the same type I error rate and avoids the evidence paradoxes that may occur close to the 0.025-threshold. It provides a principled extension to substantiate research findings from more than two trials, requesting each trial to be convincing on its own, and allows for weights. It is worth noting that the method proposed is different from the harmonic mean p-value (Good, 1958; Wilson, 2019), where the null distribution is more difficult to compute.

The method implicitly assumes that each of the individual trials is well powered for realistic treatment effects. The risk that the harmonic mean test fails increases substantially, if some of the trials have low power. Implementation of this new method may therefore be seen as an incentive to use sufficiently powered and properly conducted individual studies. Meta-analytic techniques may be more suitable if some of the studies that are considered are underpowered or if there is substantial heterogeneity between studies.

The two-trials rule is the standard for many indications, including many neurogenerative and cardiovascular diseases. However, approval of treatments in areas of high medical need may not follow the two-trials rule. An alternative approach is conditional approval based on 'adaptive pathways' (European Medicines Agency, 2016), where a temporary licence is granted based on an initial positive trial. A second post-marketing clinical trial is then often required to confirm or revoke the initial decision (Zhang *et al.*, 2020). This setting has much in common with replication studies that try to confirm original results in independent investigations (Held, 2020; Roes, 2020).

6. Availability of software

Software to perform the methodology that is described in this paper is available in the R package `ReplicationSuccess` which is available from R-Forge; use the R command `install.packages (''ReplicationSuccess'',repos= ''http://R-Forge.R-project.org'')`.

Acknowledgements

I am grateful to Mathias Drton, Karen Kafadar, Meinhard Kieser and Martin Posch for helpful discussions and suggestions. I also appreciate comments by two referees on an earlier version of this paper. Support by the Swiss National Science Foundation (project 189295) is gratefully acknowledged.

Appendix A: The null distribution of the harmonic mean χ^2-test statistic

Under the null hypothesis, Z_i, $i = 1, \ldots, n$, is standard normal distributed, so Z_i^2 is χ^2 distributed with 1 degree of freedom, i.e. a gamma $G(\frac{1}{2}, \frac{1}{2})$ distribution. The random variable $Y_i = 1/Z_i^2$ is therefore inverse gamma distributed, $Y_i \sim \text{IG}(\frac{1}{2}, \frac{1}{2})$, also known as the standard Lévy distribution: $Y_i \sim \text{Levy}(0, 1)$. More generally, the Levy$(0, c)$ distribution corresponds to the IG$(\frac{1}{2}, c/2)$ distribution and belongs to the class of stable distributions (Uchaikin and Zolotarev (1999), section 2.3).

Now $Z_1, \ldots, Z_n$ are assumed to be independent, so $Y_1, \ldots, Y_n$ are also independent and we are interested in the distribution of the sum $Y = Y_1 + \ldots + Y_n$; compare with equation (1). The standard Lévy distribution is stable, which means that the sum of independent standard Lévy random variables is again a Lévy random variable: $Y \sim \text{Levy}(0, n^2)$, which corresponds to an IG$(\frac{1}{2}, n^2/2)$ distribution. Therefore $1/Y = 1/\Sigma_{i=1}^n 1/Z_i^2$ follows a $G(\frac{1}{2}, n^2/2)$ distribution and $X^2 = n^2/Y$ in equation (1) follows a $G(\frac{1}{2}, \frac{1}{2})$ distribution, i.e. a χ^2-distribution with 1 degree of freedom.

The weighted version $Y = w_1 Y_1 + \ldots + w_n Y_n$ is also a Lévy random variable, $Y \sim \text{Levy}(0, w^2)$ where $w = \Sigma_{i=1}^n \sqrt{w_i}$; see Nolan (2018), proposition 1.17. Therefore $X_w^2 = w^2/Y$ in equation (2) also follows a χ^2-distribution with 1 degree of freedom. It is noteworthy that the $\chi^2(1)$ distribution of X^2 and X_w^2 holds even under dependence of $Z_1, \ldots, Z_n$, as described by Drton and Xiao (2016), conjecture 6.2, and proved by Pillai and Meng (2016), theorem 2.2.

Appendix B: Further details on the carvedilol example

The data that were shown in Table 1 are taken from Fisher (1999a), Table 1, for the outcome mortality. The discussion in Fisher (1999a), page 17, suggests that the p-values that are reported in Table 1 come from

a log-rank test. The relative risks that are reported in Table 1 appear to be 'instantaneous relative risks', i.e. hazard ratios. I have calculated the standard error of the log-hazard-ratios from the limits of the 95% confidence intervals that are also reported in Table 1. Note that there is an apparent discrepancy between the p-value and the confidence interval reported for study 240, with the one-sided log-rank p-value being just significant ($p = 0.0245$) whereas the 95% confidence interval for the hazard ratio is from 0.04 to 1.14 and includes the reference value 1. Leaving rounding errors aside, the corresponding one-sided p-value from a Wald-test is $p = 0.038$. This does not much affect the harmonic mean χ^2-test but the two-trials rule would obviously no longer be fulfilled. The difference between the log-rank and Wald test is still surprising, but a similar example has been reported in Collett (2003), example 3.3. I have decided to use the log-rank p-values as reported, whereas the standard errors of log-hazard-ratios are used only to weight the harmonic mean χ^2- and Stouffer's test. Likewise, the fixed and random-effects meta-analytic estimates are based on effect estimates calculated from the p-values and the log-hazard-ratio standard errors reported in Table 1, but the hazard ratios themselves are not used. Finally note that mortality was not the primary end point of the various studies, but Fisher (1999a) argued that 'it is the most important endpoint' and 'almost always of primary importance to patients and their loved ones'.

References

Berk, R. H. and Cohen, A. (1979) Asymptotically optimal methods of combining tests. *J. Am. Statist. Ass.*, **74**, 812–814.

Collett, D. (2003) *Modelling Survival Data in Medical Research*, 2nd edn. London: Chapman and Hall.

Cousins, R. D. (2008) Annotated bibliography of some papers on combining significances or p-values. *Preprint*. University of California at Los Angeles, Los Angeles. (Available from `https://arxiv.org/abs/0705.2209`.)

Darken, P. F. and Ho, S.-Y. (2004) A note on sample size savings with the use of a single well-controlled clinical trial to support the efficacy of a new drug. *Pharm. Statist.*, **3**, 61–63.

Downing, N. S., Aminawung, J. A., Shah, N. D., Krumholz, H. M. and Ross, J. S. (2014) Clinical trial evidence supporting FDA approval of novel therapeutic agents, 2005-2012. *J. Am. Med. Ass.*, **311**, 368–377.

Drton, M. and Xiao, H. (2016) Wald tests of singular hypotheses. *Bernoulli*, **22**, 38–59.

European Medicines Agency (2016) Adaptive pathways workshop—report on a meeting with stakeholders held at EMA on Thursday 8 December 2016. *Report*. European Medicines Agency. (Available from `https://www.ema.europa.eu/en/documents/report/adaptive-pathways-workshop-report-meeting-stakeholders-8-december-2016.en.pdf`.)

Fisher, L. D. (1999a) Carvedilol and the Food and Drug Administration (FDA) approval process: the FDA paradigm and reflections on hypothesis testing. *Contr. Clin. Trials*, **20**, 16–39.

Fisher, L. D. (1999b) One large, well-designed, multicenter study as an alternative to the usual FDA paradigm. *Drug Inform. J.*, **33**, 265–271.

Fisher, R. A. (1958) *Statistical Methods for Research Workers*, 13th edn (revised). Edinburgh: Oliver and Boyd.

Food and Drug Administration (1998) Providing clinical evidence of effectiveness for human drug and biological products. *Technical Report*. US Food and Drug Administration, Rockville. (Available from `www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-clinical-evidence-effectiveness-human-drug-and-biological-products`.)

Good, I. J. (1955) On the weighted combination of significance tests. *J. R. Statist. Soc.* B, **17**, 264–265.

Good, I. J. (1958) Significance tests in parallel and in series. *J. Am. Statist. Ass.*, **53**, 799–813.

Goodman, S. N. (1992) A comment on replication, p-values and evidence. *Statist. Med.*, **11**, 875–879.

Heard, N. A. and Rubin-Delanchy, P. (2018) Choosing between methods of combining p-values. *Biometrika*, **105**, 239–246.

Held, L. (2020) A new standard for the analysis and design of replication studies. *J. R. Statist. Soc.* A, **183**, 431–448; discussion, 449–469.

Hlavin, G., Koenig, F., Male, C., Posch, M. and Bauer, P. (2016) Evidence, eminence and extrapolation. *Statist. Med.*, **35**, 2117–2132.

Infanger, D. and Schmidt-Trucksäss, A. (2019) *P* value functions: an underused method to present research results and to promote quantitative reasoning. *Statist. Med.*, **38**, 4189–4197.

Johnson, V. E. (2013) Uniformly most powerful Bayesian tests. *Ann. Statist.*, **41**, 1716–1741.

Kay, R. (2015) *Statistical Thinking for Non-statisticians in Drug Regulation*, 2nd edn. Chichester: Wiley.

Littell, R. C. and Folks, J. L. (1973) Asymptotic optimality of Fisher's method of combining independent tests II. *J. Am. Statist. Ass.*, **68**, 193–194.

Maca, J., Gallo, P., Branson, M. and Maurer, W. (2002) Reconsidering some aspects of the two-trials paradigm. *J. Biopharm. Statist.*, **12**, 107–119.

Matthews, J. N. (2006) *Introduction to Randomized Controlled Clinical Trials*, 2nd edn. Boca Raton: Chapman and Hall–CRC.

Nolan, J. P. (2018) *Stable Distributions—Models for Heavy Tailed Data*, ch. 1. Boston: Birkhäuser. To be published. (Available from `http://fs2.american.edu/jpnolan/www/stable/chap1.pdf`.)
Pillai, N. S. and Meng, X.-L. (2016) An unexpected encounter with Cauchy and Lévy. *Ann. Statist.*, **44**, 2089–2097.
Roes, K. C. B. (2020) Discussion on 'A new standard for the analysis and design of replication studies', by L. Held. *J. R. Statist. Soc.* A, **183**, 459.
Rosenkrantz, G. (2002) Is it possible to claim efficacy if one of two trials is significant while the other just shows a trend? *Drug Inform. J.*, **36**, 875–879.
Senn, S. (2007) *Statistical Issues in Drug Development*, 2nd edn. Chichester: Wiley.
Shun, Z., Chi, E., Durrleman, S. and Fisher, L. (2005) Statistical consideration of the strategy for demonstrating clinical evidence of effectiveness–one larger vs two smaller pivotal studies. *Statist. Med.*, **24**, 1619–1637.
Uchaikin, V. V. and Zolotarev, V. M. (1999) *Chance and Stability: Stable Distributions and Their Applications*. Berlin: de Gruyter.
Westberg, M. (1985) Combining independent statistical tests. *Statistician*, **34**, 287–296.
Wilson, D. J. (2019) The harmonic mean *p*-value for combining dependent tests. *Proc. Natn. Acad. Sci. USA*, **116**, 1195–1200.
Zhang, A. D., Puthumana, J., Downing, N. S., Shah, N. D., Krumholz, H. and Ross, J. S. (2020) Assessment of clinical trials supporting US Food and Drug Administration approval of novel therapeutic agents, 1995-2017. *JAMA Network Open*, **3**, no. 4, article e203284.