The harmonic mean χ^2 test to substantiate scientific findings

Leonhard Held

Epidemiology, Biostatistics and Prevention Institute (EBPI)

and Center for Reproducible Science (CRS)

University of Zurich

Hirschengraben 84, 8001 Zurich, Switzerland

Email: leonhard.held@uzh.ch

21st February 2020

Abstract: Statistical methodology plays a crucial role in drug regulation. Decisions by the FDA or EMA are typically made based on multiple primary studies testing the same medical product, where the two-trials rule is the standard requirement, despite a number of shortcomings. A new approach is proposed for this task based on the (weighted) 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 one 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 significance level and the number of studies. A real example with 5 clinical trials investigating the effect of Carvedilol for the treatment of patients with moderate to severe heart failure patients is used to illustrate the methodology. As a by-product, the approach provides a calibration of the sceptical p-value recently proposed for the analysis of replication studies.

Key Words: Combining p-values; drug regulation; evidence synthesis; Lévy distribution; Type-I error control; two-trials rule; Sceptical p-value

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 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 result may occur in a meta-analysis even if some of the individual studies have not been convincing on its 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 5 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 (*cf*. Fisher, 1999a, Table 1). One-sided p-values (from log-rank tests) and hazard ratios (HR) are shown in Table 1, indicating a consistent reduction in mortality between 28 and 78% across the different studies.

study number	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

Table 1: Results from 5 clinical trials on the effect of Carvedilol for the treatment of patients with moderate to severe heart failure patients. Shown are one-sided p-values, estimated hazard ratios (HR), and the associated log hazard ratios (log HR) with standard errors (SE).

A meta-analysis could be applied to the data shown in Table 1, but the drug regulation industry (including the U.S. ”Food and Drug Administration, or FDA) typically relies instead on the "two-trials rule" (Kay, 2015, Section 9.4), 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" (FDA, 1998, p. 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 2 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. On the other hand, 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 different studies into one p-value (Fisher, 1999b; Darken and Ho, 2004; Shun et al., 2005). Ronald 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 one with $p = 0.0001$ would give Fisher's $p = 0.0005 < 0.000625 = 0.025^2$, so a claim of success with respect to the Type I error rate of the two-trials rule. On the other hand, two trials both with $p = 0.01$ would not be considered as successful with Fisher's $p = 0.001$. Both decisions seem undesirable from a regulator's perspective. Another problem is that Fisher's method treats large and small studies equally. It can be extended to incorporate weights (Good, 1955), but then the null distribution does no longer have a convenient form.

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 one 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 while the other just shows a trend. He combines 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 Stouffer's pooled 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 proposed test comes to opposite conclusions for the examples mentioned above: In contrary to Fisher's method, it leads to approval of two trial both with $p = 0.01$, but not to approval if one has $p = 0.0001$ and the other one $p = 0.5$. Contrary to the two-trials rule, it leads to approval of one trial with $p = 0.026$ and the other one 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. Furthermore, the new approach implies useful bounds on the individual study-specific p-values $p_1, \ldots, p_n$, 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 to the two-trials rule and illustrated on the Carvedilol data in Section 3. I close with some discussion in Section 4.

2. The harmonic mean χ^2 test

Suppose 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 (2007) compares some of the more prominent papers on this topic. Among them is Stouffer's method, which is based on the Z-scores $Z_i = \Phi^{-1}(1 - p_i)$, here $\Phi^{-1}(.)$ denotes the quantile function of the standard normal distribution. Under the assumption of no effect, the test statistic $Z = \sum_{i=1}^n Z_i / \sqrt{n}$ is standard normally distributed. The corresponding p-value forms the basis of the "pooled-trials rule" and is equivalent to investigate significance of the overall effect estimate from a fixed-effects meta-analysis (Senn, 2007, Section 12.2.8). It can also be extended to include weights. Fisher's method is also commonly used and compares $-2 \sum_{i=1}^n \log p_i$ with a χ^2-distribution with $2\,n$ degrees of freedom to compute a combined p-value. 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); 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 $Z_H^2 = n / \sum_{i=1}^n 1/Z_i^2$ of the squared Z-scores:

$$\chi^2 = n\, Z_H^2 = \frac{n^2}{\sum\limits_{i=1}^{n} 1/Z_i^2}. \tag{1}$$

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 χ^2 does not depend on n.

Weights $w_1, \ldots, w_n$ can also be introduced in (1), then the test statistic

$$\chi^2_w = \frac{w^2}{\sum\limits_{i=1}^{n} w_i / Z_i^2} \text{ where } w = \sum_{i=1}^{n} \sqrt{w_i} \tag{2}$$

should be used. Multiplication with w^2 ensures that the null distribution of χ^2_w does not depend on the weights $w_1, \ldots, w_n$ nor on n.

The specific form of (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 effective sample size of study i. If we use weights $w_i = 1/\sigma_i^2$ equal to the precision of the effect estimates, (2) can be written as the unweighted harmonic mean $\hat{\theta}^2_H$ of the squared effect estimates $\hat{\theta}^2_i$ times a scaling factor w^2/n:

$$\chi^2_w = w^2/n \cdot \hat{\theta}^2_H \text{ where } w = \sum_{i=1}^{n} \sqrt{m_i}. \tag{3}$$

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

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, (1) will therefore also tend to increase if there is a true non-zero effect. However, the test statistic (3) is based on the harmonic mean $\hat{\theta}^2_H$ of the squared study-specific effect estimates $\hat{\theta}^2_i$, 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 (1) and (2) is χ^2 with one degree of freedom, see Appendix A for details. We can thus compute an overall p-value p_H from (1) or (2) based on the $\chi^2(1)$ distribution function. However, we have to be careful since (1) does not take the direction of the effects into account. Usually we are interested in a pre-defined 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 (1) and (2) can be large for any of the 2^n possible combinations of the signs on $Z_1, \dots, Z_n$, 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 have to adjust the p-value accordingly.

To be specific, suppose all studies have a positive effect and the observed test statistic (1) or (2) is $\chi^2 = y$, respectively $\chi^2_w = y$. The overall p-value from the proposed significance test is then

$$p_H = \Pr(\chi^2(1) \geq y)/2^n = \left[1 - \Phi(\sqrt{y})\right]/2^{n-1}. \tag{4}$$

Likewise we can obtain the critical value

$$c_H = \left[\Phi^{-1}(1 - 2^{n-1}\alpha_H)\right]^2 \tag{5}$$

for the test statistic (1) or (2) to control the Type-I error rate at some overall significance level 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 to obtain n positive results is $1/2^n$. We are only interested in this case, so if at least one of the studies has a negative effect we suggest to report the inequality $p_H > 1/2^n$, for example $p_H > 0.25$ for $n = 2$ studies.

In what follows I restrict attention to the unweighted test statistic χ^2 given in (1), similar results can be obtained for χ^2_w given in (2). Let $Z_i = z_i$ denote the observed test statistic in the i-th study. I assume that $z_i > 0$ for all $i = 1, \dots, n$, *i. e.* all effects go in the right direction. First note that the smallest squared test statistic $z^2_{\text{min}} = \min\{z^2_1, \dots, z^2_n\}$ multiplied by the number of studies n is an upper bound on the harmonic mean $z^2_H = n / \sum_{i=1}^n 1/z^2_i$:

$$z^2_{\text{min}} \leq z^2_H \leq n\, z^2_{\text{min}} \leq n\, z^2_i,$$

where the last inequality holds for all $i = 1, \dots, n$. This implies $y \leq n^2\, z^2_i$ for the observed test statistic y and any study $i = 1, \dots, n$ and with equation (4) we obtain

$$\Pr\{\chi^2(1) \geq n^2\, z^2_i\}/2^n \leq p_H.$$

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

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

This is a necessary but not sufficient restriction on the study-specific p-values for a claim of success.

It is also possible to derive the corresponding sufficient bound. Assume all p-values are equal (*i. e.* $z^2_1 = \ldots = z^2_n$), then the condition $\chi^2 = n\, z^2_i \geq c_H$ implies $z_i \geq \sqrt{c_H}/\sqrt{n}$. Note that the sufficient bound on z_i differs from the corresponding necessary bound only by the multiplicative factor $\sqrt{n}$. The restriction on the corresponding p-values is

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

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

α_H	bound	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
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

Table 2: Necessary and sufficient bounds on the one-sided study-specific p-values for overall significance level α_H and different number of studies n

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 (6) and (7), respectively, are shown in Table 2 for $\alpha_H = 1/1600$ (the two-trials rule), 1/31574 (the one-sided four-sigma rule) and 1/3488556 (the one-sided five-sigma rule). The significance level of the one-sided k-sigma rule is based on a normally distributed test statistic $T \sim \mathrm{N}(0, \sigma^2)$ with zero mean and defined as $\Pr(T > k\,\sigma) = 1 - \Phi(k)$. The five-sigma 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 significance level of the four-sigma rule is between the two-trials rule and the five-sigma rule.

The first line of Table 2 reveals that for level 1/1600, the requirement $p_i \leq 0.065$, $i = 1, 2$, is necessary for claiming success based on $n = 2$ studies. If one of the two studies has a p-value 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, $i = 1, 2$, 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 0.37 while it is sufficient that each p-value is smaller 0.20. Decreasing the significance level 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 also 0.19 for $\alpha_H = 1/3488556$ and $n = 5$. For the five-sigma level $1/3488556$, the necessary bound is 0.0075 for $n = 2$ and 0.058 for $n = 3$ studies. The sufficient bound is 0.00029 for $n = 2$ and 0.0032 for $n = 3$.

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, so are one-sided. The two-sided p-value 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 suggest to report $p_S > 1/2^{n-1}$.

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 $\gamma \geq 1 - 1/2^{n-1}$ 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 3.2.

3. Comparison with the two-trials rule

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

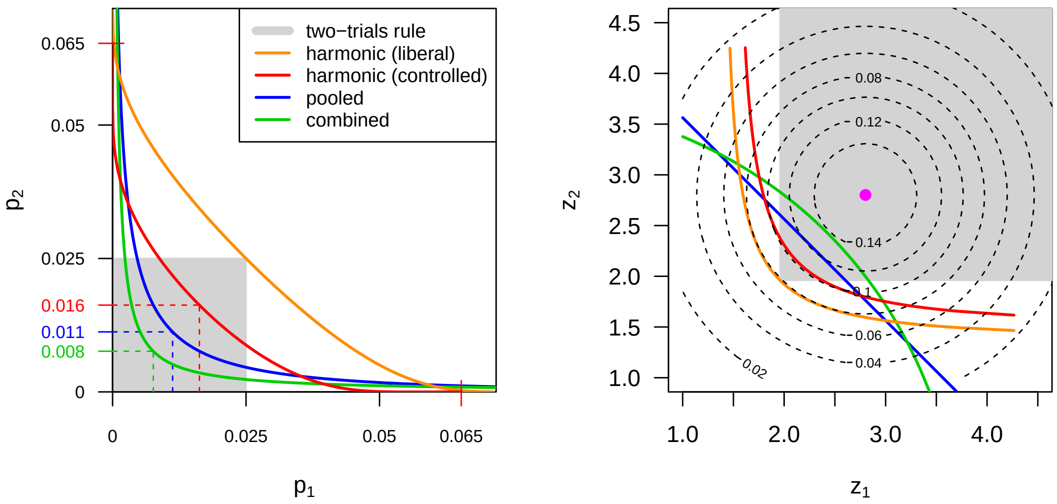

Figure 1: Comparison of different approaches for drug approval as function of the two p-values p_1 and p_2 (left) and the two Z-values Z_1 and Z_2 (right), respectively. The acceptance region of the two-trials rule is shown in grey. The acceptance regions of the other methods is below (left) or above (right) 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 the right plot 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% significance level.

Figure 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. This has been computed by equating the right-hand side of (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 to the $\alpha^2 = 0.025^2$ level.

Also shown in Figure 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, particularly good to see in the z-scale shown in the right plot of Figure 1. The latter two suffer from the possibility of approval if one of the p-values is very small while the other one is far away 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, shown in the left plot of Figure 1, is considerably smaller for the pooled (0.011) and combined (0.008) method than for the controlled harmonic mean χ^2 test (0.016). These features make both the pooled and the combined method less suitable for drug approval.

For comparison, the two-trials rule has the necessary and sufficient conditions $p_i \leq 0.025$, $i = 1, 2$. The harmonic mean χ^2 test can be significant only if both p-values are small (< 0.065). This has been discussed in Section 2 and can also be seen from Figure

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 tests (calculated as described in Held (2020, Section 4)) are smooth, quickly approaching zero at $p_1 = 0.065$ respectively $p_1 = 0.083$. 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 $c = \Pr(\chi^2(4) \geq \alpha_H)$. The conditional power of the pooled method turns out to be $1 - \Phi[2\,\Phi^{-1}(p_1) - \sqrt{2}\,\Phi^{-1}(\alpha_H)]$.

3.1. Project power

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 study, the distribution of Z_1 and Z_2 is $\mathrm{N}(\mu, 1)$ where $\mu = \Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)$. We can thus simulate independent Z_1 and Z_2 for $\alpha = 0.025$ and different values of the power $1 - \beta$ and compute the proportion of trial results with drug approval at level α^2. This is shown in Table 3 for the different 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 to 7 percentage points larger, depending on the power of the two studies. 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 trials was not sufficiently convincing on its own.

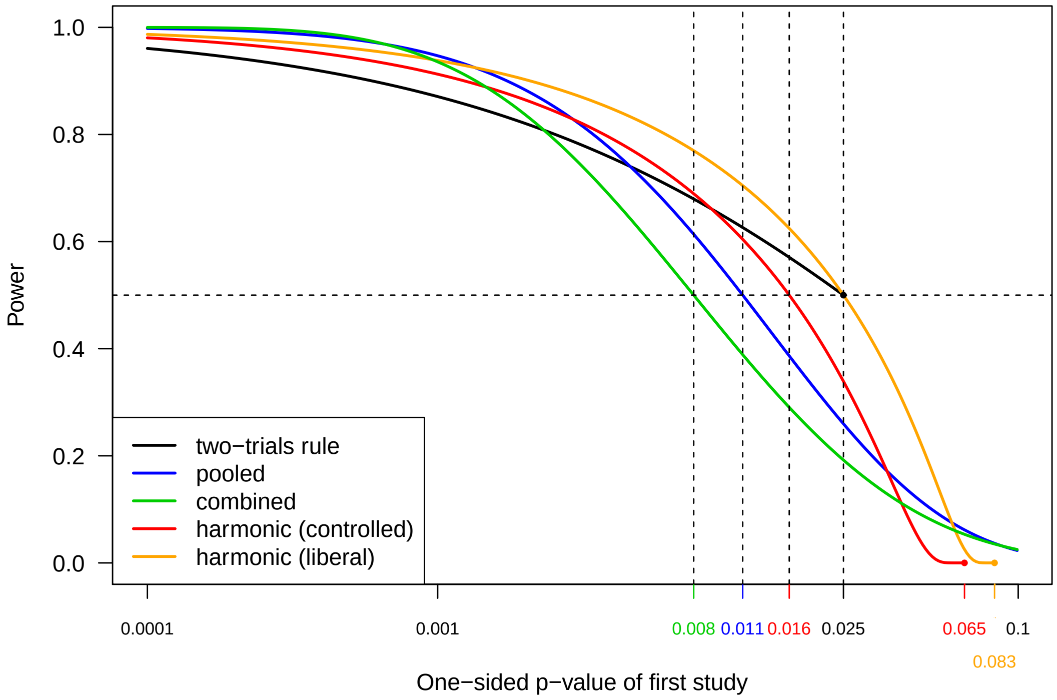

Figure 2: Power for drug approval conditional on the one-sided p-value of the first study. Power values of exactly zero are omitted.

3.2. Application

Two advantages of the proposed method are that it allows for weighting and is readily applicable to the case where results from more than 2 studies are available. For practical illustration, I revisit 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 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) reports Fisher's combined p-value, which is 0.00013. Stouffer's unweighted pooled test gives the p-value 0.00009, the weighted version gives $p = 0.00018$.

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

Table 3: The probability of drug approval (in %) as a function of the original power of the two studies

For the latter the weights have been chosen inversely proportional to the squared standard errors of the associated log hazard ratios 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 somewhat 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 runs 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 $1 - 2 \cdot 0.000625 = 0.99875$. The more standard 95% confidence interval for the hazard ratio runs from 0.21 to 0.74. For comparison, a random-effects meta-analysis gives the 95% confidence interval 0.25 to 0.77 (two-sided $p = 0.004$). A fixed-effects meta-analysis gives the 95% confidence interval 0.32 to 0.72. The corresponding two-sided p-value is 0.00035, twice as large as the p-value from Stouffer's weighted test.

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 away 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 large effect on the proposed method: 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 only by 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 the case 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.

4. 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 proposed method is different from the harmonic mean p-value (Good, 1958; Wilson, 2019), where the null distribution is more difficult to compute (Wilson, 2019, Section 1 of Supplementary Material).

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 a means to ensure sufficiently powered and properly conducted individual studies. Meta-analytic techniques may be more suitable if some of the studies 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 Medical Agency, 2016), where a temporary license is 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., 2019). This setting has much in common with replication studies that try to confirm original results in independent investigations, so it would be interesting to apply the sceptical p-value in this setting (Held, 2020; Roes, 2020).

Acknowledgments 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.

References

Berk, R. H. and Cohen, A. (1979). Asymptotically optimal methods of combining tests. *Journal of the American Statistical Association*, 74(368):812–814.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Chapman & Hall, 2nd edition.

Cousins, R. D. (2007). Annotated bibliography of some papers on combining significances or p-values. 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. *Pharmaceutical Statistics*, 3(1):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. *JAMA: The Journal of the American Medical Association*, 311(4):368–377. https://dx.doi.org/10.1001/jama.2013.282034.

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

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

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

Fisher, L. D. (1999a). Carvedilol and the Food and Drug Administration (FDA) approval process: The FDA paradigm and reflections on hypothesis testing. *Controlled Clinical Trials*, 20(1):16 – 39.

Fisher, L. D. (1999b). One large, well-designed, multicenter study as an alternative to the usual FDA paradigm. *Drug Information Journal*, 33(1):265–271.

Fisher, R. A. (1958). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, 13th ed.(rev.) edition.

Good, I. J. (1955). On the weighted combination of significance tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):264–265.

Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, 53(284):799–813. http://dx.doi.org/10.1080/01621459.1958.10501480.

Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine*, 11(7):875–879. https://doi.org/10.1002/sim.4780110705.

Heard, N. A. and Rubin-Delanchy, P. (2018). Choosing between methods of combining p-values. *Biometrika*, 105(1):239–246.

Held, L. (2020). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society, Series A*, 183:431–469.

Hlavin, G., Koenig, F., Male, C., Posch, M., and Bauer, P. (2016). Evidence, eminence and extrapolation. *Statistics in Medicine*, 35(13):2117–2132. https://doi.org/10.1002/sim.6865.

Infanger, D. and Schmidt-Trucksäss, A. (2019). P value functions: An underused method to present research results and to promote quantitative reasoning. *Statistics in Medicine*, 38(21):4189–4197.

Johnson, V. E. (2013). Uniformly most powerful Bayesian tests. *Ann. Stat.*, 41(4):1716–1741.

Kay, R. (2015). *Statistical Thinking for Non-Statisticians in Drug Regulation*. John Wiley & Sons, Chichester, U.K., second edition.

Littell, R. C. and Folks, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests II. *Journal of the American Statistical Association*, 68(341):193–194.

Maca, J., Gallo, P., Branson, M., and Maurer, W. (2002). Reconsidering some aspects of the two-trials paradigm. *Journal of Biopharmaceutical Statistics*, 12(2):107–119.

Matthews, J. N. (2006). *Introduction to Randomized Controlled Clinical Trials*. Chapman & Hall/CRC, second edition.

Nolan, J. P. (2018). *Stable Distributions - Models for Heavy Tailed Data*. Birkhauser, Boston. In progress, Chapter 1 online at http://fs2.american.edu/jpnolan/www/stable/chap1.pdf.

Pillai, N. S. and Meng, X.-L. (2016). An unexpected encounter with Cauchy and Lévy. *Annals of Statistics*, 44:2089–2097.

Roes, K. C. B. (2020). Discussion of "A new standard for the analysis and design of replication studies" by Leonhard Held. *Journal of the Royal Statistical Society, Series 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 Information Journal*, 36(1):875–879.

Senn, S. (2007). *Statistical Issues in Drug Development*. John Wiley & Sons, Chichester, U.K., second edition.

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. *Statistics in Medicine*, 24(11):1619–1637.

Uchaikin, V. V. and Zolotarev, V. M. (1999). *Chance and Stability: Stable Distributions and Their Applications.* Walter de Gruyter.

Westberg, M. (1985). Combining independent statistical tests. *The Statistician*, 34:287–296.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1814092116.

Zhang, A. D., Puthumana, J., Downing, N. S., Shah, N. D., Krumholz, H., and Ross, J. S. (2019). Clinical trial evidence supporting FDA approval of novel therapeutic agents over three decades, 1995-2017: Cross-sectional analysis. *medRxiv*. http://dx.doi.org/10.1101/19007047.

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 with 1 degree of freedom, *i. e.* a gamma $\mathrm{G}(1/2, 1/2)$ distribution. The random variable $X_i = 1/Z_i^2$ is therefore inverse gamma distributed, $X_i \sim \mathrm{IG}(1/2, 1/2)$, also known as the standard Lévy distribution: $X_i \sim \mathrm{Levy}(0, 1)$. More generally, the $\mathrm{Levy}(0, c)$ distribution corresponds to the $\mathrm{IG}(1/2, c/2)$ distribution and belongs to the class of stable distributions (Uchaikin and Zolotarev, 1999, Section 2.3).

Now $Z_1, \ldots, Z_n$ are assume to be independent, so $X_1, \ldots, X_n$ are also independent and we are interested in the distribution of the sum $X = X_1 + \ldots + X_n$, compare equation (1). The standard Lévy distribution is known to be stable, which means that the sum of independent standard Lévy random variables is again a Lévy random variable: $X \sim \mathrm{Levy}(0, n^2)$, which corresponds to a $\mathrm{IG}(1/2, n^2/2)$ distribution. Therefore $1/X = 1/\sum_{i=1}^n 1/Z_i^2$ follows a $\mathrm{G}(1/2, n^2/2)$ distribution and $\chi^2 = n^2/X$ follows a $\mathrm{G}(1/2, 1/2)$, *i. e.* a χ^2 distribution with one degree of freedom.

The weighted version $X = w_1 X_1 + \ldots + w_n X_n$ is also a Lévy random variable, $X \sim \mathrm{Levy}(0, w^2)$ where $w = \sum_{i=1}^n \sqrt{w_i}$, see Nolan (2018, Proposition 1.17). Therefore $\chi^2_w = w^2/X$ also follows a χ^2 distribution with one degree of freedom. It is noteworthy that the $\chi^2(1)$ distribution of χ^2 respectively χ^2_w holds even under dependence of $Z_1, \ldots, Z_n$, as described by Drton and Xiao (2016, Conjecture 6.2) and proven by Pillai and Meng (2016, Theorem 2.2).

B. Further details on the Carvedilol example

The data shown in Table 1 are taken from Fisher (1999a, Table 1) for the outcome mortality. The discussion on Fisher (1999a, page 17) suggests that the p-values reported in the table come from a log-rank test. The relative risks reported in the table 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 also reported in the table. 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 runs 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 log-rank and Wald 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 only used 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 endpoint of the different studies, but Fisher (1999a) argues that "it is the most important endpoint" and "almost always of primary importance to patients and their loved ones".