Conditional Drug Approval with the Harmonic Mean Chi-Squared Test

Manja Deforth\textsuperscript{1*}, Charlotte Micheloud\textsuperscript{1*}, Kit C Roes\textsuperscript{2}, Leonhard Held\textsuperscript{1}

\textsuperscript{1}Department of Biostatistics at the Epidemiology, Biostatistics and Prevention Institute (EBPI) and Center for Reproducible Science (CRS), University of Zurich, Zurich, Switzerland
\textsuperscript{2}Department of Health Evidence, Section Biostatistics, Radboud University Medical Center, Radboud University, The Netherlands
\textsuperscript{*}Manja Deforth and Charlotte Micheloud contributed equally to this work.

Correspondence
Manja Deforth, University of Zurich, Department of Biostatistics at the Epidemiology, Biostatistics and Prevention Institute (EBPI), Hirschengraben 84, 8001 Zurich, Switzerland
Email: manjaelisabeth.deforth@uzh.ch

Funding information
This work was partially funded by the Swiss National Science Foundation (project 189295).

Software and data availability
All analyses were performed in the R programming language version 4.1.2 (2021-11-01).
Data and code to reproduce all analyses are available at https://gitlab.uzh.ch/charlotte.micheloud/repro-material-conditional-drug-approval

December 23, 2021
Abstract
Approval of treatments in areas of high medical need may not follow the two-trials paradigm, but might be granted under conditional approval. Under conditional approval, the evidence for a treatment effect from a pre-market clinical trial has to be substantiated in an independent post-market clinical trial or a longer follow-up duration. Several ways exist to quantify the overall evidence provided by the two trials. We study the applicability of the recently developed harmonic mean \( \chi^2 \)-test to the conditional drug approval framework. The proposed approach can be used both for the design of the post-market trial and the analysis of the combined evidence provided by both trials. Other methods considered are the two-trials rule, Fisher’s criterion and Stouffer’s method. For illustration, we apply the harmonic mean \( \chi^2 \)-test to a drug which received conditional (and eventually final) market licensing by the EMA. A simulation study is conducted to study the operating characteristics in more detail. We finally investigate the applicability of the harmonic mean \( \chi^2 \)-test to compute the power at interim of an ongoing post-market trial. In contrast to some of the traditional methods, the harmonic mean \( \chi^2 \)-test always requires a post-market clinical trial. In addition, if the \( p \)-value \( p_1 \) from the pre-market clinical trial is \( \ll 0.025 \), a smaller sample size for the post-market clinical trial is needed than with the two-trials rule.

KEYWORDS:
Conditional drug approval, harmonic mean chi-squared test, interim power
1 INTRODUCTION

In case of an “unmet medical need”, the European Medicines Agency (EMA) may grant a conditional market approval (MA). This will give patients earlier access to promising new treatments. Certain requirements must be met for regular MA, e.g. the conduct of a post-market clinical trial or an extension of the follow-up duration. A recent example is the conditional marketing authorization of COVID-19 Vaccine AstraZeneca by the EMA. Another example is the drug Fampridine, a drug for patients with Multiple Sclerosis. The drug Fampridine received conditional approval in July 2011 and regular drug market approval in May 2017. The EMA requirements for regular drug market approval were “To conduct a double-blinded, placebo-controlled, long-term efficacy and safety study to investigate a broader primary endpoint clinically meaningful in terms of walking ability and to further evaluate the early identification of responders in order to guide further treatment [...].” An overview about the pre- and post-market clinical trial(s) is given in Table 1.

The primary outcome was improvement in walking speed (T25FW Test) in the pre-market clinical trial and improvement in walking ability (MSWS-12 score) in the post-market clinical trial. A change in the primary outcome between the pre- and post-market clinical trial is quite common in the conditional drug approval framework. Specifically, a surrogate outcome is often used in the pre-market clinical trial, whereas the follow-up time is usually longer and the evaluation of a clinical outcome is preferred in the post-market clinical trial.

From a regulatory point of view, the question may arise if a conditional MA can be granted and, during reevaluation, if the initial conditional licensing decision was correct. Additionally, if the interim results of the post-market clinical trial are available but the final results are not, the power at interim can be calculated. It might be used as a decision tool for stopping the post-market clinical trial prematurely for futility. Here we consider the common scenario that pre- and post-market clinical trials are independent. In this case, there are several ways to quantify the overall evidence provided by the two trials. Traditional methods include for example the two-trials rule, Fisher’s criterion, and Stouffer’s method. The objective of this paper is to study the applicability of the harmonic mean \( \chi^2 \)-test, developed to combine the result of multiple independent clinical trials, and to compare it with the more traditional methods. We evaluated the performance of each statistical method based on a simulation study and by using the drug Fampridine. The methods may not only be used for the analysis of the two clinical trials, but also at an earlier stage: the design of the post-market clinical trial.

In Section 2 the statistical methods are described. The description of the simulation study and the results thereof are given in Section 3. Section 4 contains the extension of the power at interim. The conclusion is given in Section 5.

2 STATISTICAL METHODS

A short introduction to some notation and assumptions is given in Section 2.1. The statistical methods, namely the harmonic mean \( \chi^2 \)-test (Section 2.3), two-trials rule (Section 2.2), Fisher’s criterion (Section 2.4) and Stouffer’s method (Section 2.5) are summarized and compared (Section 2.6).

2.1 Notation and assumptions

In many problems it is reasonable to assume a normal distribution with known variance for the estimate \( \hat{\theta} \) of the true treatment effect \( \theta \): \( \hat{\theta} \sim N(\theta, \text{se}^2) \). Here \( \text{se} = \sigma / \sqrt{n} \) denotes the standard error of \( \hat{\theta} \) based on a sample of size \( n \) and a residual standard deviation \( \sigma \). This formulation can be extended to the two-sample case of (balanced) controlled clinical trials where \( \text{se} = \sqrt{2\sigma / \sqrt{n}} \) and \( n \) is the sample...
size per group. It applies to binary outcomes (where \( \theta \) is the log odds ratio), survival outcomes (where \( \theta \) is the log hazard ratio) and continuous outcomes (where \( \theta \) is the mean difference). The exact form of \( \sigma \) and \( n \) for the different types of outcomes can be found in Spiegelhalter, Abrams and Myles in Section 2.4.

Here we consider a pre-market clinical trial (index 1) and an independent post-market clinical trial (index 2) with two groups each to estimate the corresponding true, unknown, treatment effects \( \theta_1 \) and \( \theta_2 \), respectively. The pre- and post-market clinical trial might have different primary outcomes (for example a surrogate and a clinical outcome) and therefore different treatment effects, i.e. \( \theta_1 \neq \theta_2 \). Another possibility is that the primary outcome is identical, but the treatment effect \( \theta_2 \) is lower in the post-market clinical trial due to a different study population.

The estimated treatment effects of the pre- and post-market clinical trials are denoted by \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \), respectively, and are independent and normally distributed with known variances: \( \hat{\theta}_1 \sim N(\theta_1, \sigma_1^2) \) and \( \hat{\theta}_2 \sim N(\theta_2, \sigma_2^2) \). Let \( \sigma_1 = \sqrt{2\sigma_1^2} / \sqrt{m_1} \) and \( \sigma_2 = \sqrt{2\sigma_2^2} / \sqrt{m_2} \) denote the corresponding standard errors and \( m_1, m_2 \) the corresponding sample sizes. The variance ratio \( c \) is equal to \( \sigma_1^2 / \sigma_2^2 \). The test statistics of the pre- and post-market clinical trials are \( z_1 = \hat{\theta}_1 / \sigma_1 \) and \( z_2 = \hat{\theta}_2 / \sigma_2 \). The one-sided \( p \)-values for the null hypothesis \( H_0:\ \theta = 0 \) vs. the alternative \( H_1:\ \theta > 0 \) are denoted as \( p_1 = 1 - \Phi(z_1) \) and \( p_2 = 1 - \Phi(z_2) \) where \( \Phi(\cdot) \) denotes the cumulative distribution function of the standard normal distribution.

### 2.2 Two-trials rule

The two-trials rule requires “at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness” \(^8(p. 3)\). In the conditional drug approval framework this might be achieved by independently replicating the result of a pre-market clinical trial in a post-market clinical trial, both significant at one-sided level \( \alpha = 0.025 \). The overall Type-I error rate of this simple decision rule is hence \( \gamma = \alpha^2 = 0.000625 \).

The sample size of the post-market clinical trial can be calculated such that the power to detect \( \hat{\theta}_1 \) at level \( \alpha \) reaches a pre-specified value \( 1 - \beta \). In practice, the estimate \( \hat{\theta}_1 \) is often inflated as compared to the true \( \theta_1 \) due to various phenomena such as the winner’s curse and publication bias. \(^9\) These phenomena will be explained in details in Section 3. As a result, calculating the sample size of the post-market trial to detect \( \hat{\theta}_1 \) might lead to an underpowered trial. A simple method to address this has been proposed in the context of replication studies. \(^10\) In short, a shrinkage factor \( s \) can be chosen and represents the scepticism about the estimate \( \hat{\theta}_1 \) from the pre-market trial, with \( s = 0 \) representing no expected inflation of \( \hat{\theta}_1 \). It follows that the required variance ratio \( c \) (pre-to post-market squared standard errors \( \sigma_1^2 / \sigma_2^2 \)) only depends on the \( z \)-value \( z_1 \) from the pre-market clinical trial \(^10\) and the shrinkage factor \( s \):

\[
\sigma_2^2 = c \sigma_1^2 \quad \text{and} \quad \sigma_2 = c \sigma_1 \]

\[
c = \Phi^{-1} \left( \frac{z_1 - \beta + z_{1-a}}{(1-s)^2 z_1^2} \right),
\]

where \( z_{1-a} = \Phi^{-1}(1-p) \).

The sample size \( n_2 \) of the post-market trial can then be calculated as \( n_2 = cn_1(\sigma_2^2 / \sigma_1^2) \). If the pre- and post-market standard deviations \( \sigma_1 \) and \( \sigma_2 \), respectively, are the same, the variance ratio \( c \) reduces to the relative sample size \( c = n_2 / n_1 \). Formula (1) can also be used when different endpoints are of interest in the pre- and post-market clinical trials, i.e. when \( \theta_1 \neq \theta_2 \). To do so, the researchers need to judge the expected shrinkage \( s \) in the post-market \( z \)-value \( z_2 \) (or standardized mean difference \( \hat{\theta}_2 / \sigma_2 \)) as compared to the pre-market \( z \)-value \( z_1 \) (or standardized mean difference \( \hat{\theta}_1 / \sigma_1 \)) and use it in (1). In the two-trials rule, the required significance level \( \alpha \) does not depend on the result from the pre-market clinical trial.
2.3 Harmonic mean $\chi^2$-test

The harmonic mean $\chi^2$-test was developed to combine the $p$-values from independent clinical trials with the same research hypothesis into one overall $p$-value. In the conditional approval framework, the primary outcome may differ between the pre-market and the post-market clinical trial. The harmonic mean $\chi^2$-test can still be applied as long as the hypotheses of the pre- and post-market trials are the same.

The test statistic $X^2$ of the harmonic mean $\chi^2$-test is twice the harmonic mean $I_H^2$ of the squared test-statistics:

$$X^2 = 2 I_H^2 = \frac{4}{1/z_1^2 + 1/z_2^2}.$$  \hfill (2)

Weights $w_1$ and $w_2$ can also be introduced, then

$$X^2 = \frac{w^2}{w_1/z_1^2 + w_2/z_2^2} \text{ with } w = \sqrt{w_1} + \sqrt{w_2}. \hfill (3)$$

Under the null hypothesis $H_0$ where $z_1 \sim N(0, 1)$ and $z_2 \sim N(0, 1)$, both (2) and (3) are $\chi^2$-distributed with one degree of freedom. This property can be used to derive the one-sided $p$-value $p_H$ of the harmonic mean $\chi^2$-test under the assumption that the alternative hypothesis is $H_1$: $\theta > 0$. If $z_1 > 0$ and $z_2 > 0$ both hold, then $p_H = \Pr(\chi_1^2 \geq X^2)/4 = [1 - \Phi(X)]/2$ where $X = +\sqrt{X^2}$. Likewise we can obtain the critical value

$$c_H = \chi^2_1(1 - 4\gamma)$$

for the test statistic (2) and (3), where $\gamma$ is the required significance level and $\chi^2_{v}(.)$ denotes the quantile function of the $\chi^2$-distribution with $v$ degrees of freedom.

The $p$-value from the harmonic $\chi^2$-test is usually compared against $\gamma = \alpha^2$, the overall Type-I error rate of the two-trials rule at one-sided significance level $\alpha$, with critical value $c_H = 9.14$ for $\gamma = 0.025^2$. In the unweighted case, an overall significant result at this level is only possible if both $p_1 \leq 0.065$ and $p_2 \leq 0.065$ hold.

The harmonic mean $\chi^2$-test can also be applied to determine the required $z$-value $z_2$ (respectively $p$-value $p_2$) in the post-market trial given the $z$-value $z_1$ (respectively $p$-value $p_1$) from the pre-market clinical trial, to achieve overall significance at level $\gamma$. For the more general weighted version, rearranging (3) yields the requirement

$$z_2 \geq \frac{z_2}{\sqrt{w_2/c_H} - w_1/z_1^2}. \hfill (4)$$

This implies an upper bound $p_2 = 1 - \Phi(z_2)$ for the $p$-value $p_2$ from the post-market trial to achieve overall significance. The necessary requirement $p_2 \leq 0.065$ for significance in the unweighted case at level $\gamma = 0.025^2$ can be obtained from (4) for $w_1 = w_2 = 1$ (so $w = 2$) and $z_1^2 \to \infty$, where $z_2 \downarrow \sqrt{0.14/2} = 1.51$ and so $p_2 \uparrow 0.065$. This means that - no matter how convincing the pre-market clinical trial is - the post-market clinical trial always needs to be conducted and has to provide sufficient evidence against the null hypothesis on its own. We will also consider the case where the weight of the post-market trial is 40% rather than 50%, then the necessary requirement on $p_2$ is $p_2 \leq 0.087$. The necessary condition on $p_1$ can also be derived from (4). The upper bound on $p_1$ is $p_1 = 0.065$ in the unweighted case and $p_1 = 0.048$ in the weighted case.

The upper bound $p_2$ can be used as an adaptive significance level $\alpha = \alpha(p_1)$ (depending on the result from the pre-market trial) in the sample size computation (1) for the post-market trial.
2.4 Fisher’s criterion

Another way to combine $p$-values is Fisher’s criterion.\textsuperscript{11,12} Overall significance is achieved if $p_1 p_2 \leq c_F = \exp\left(-\frac{1}{2} \chi^2_F(1-\alpha)\right)$, where $\chi^2 = \chi^2_F$. Given the $p$-value $p_1$ from the pre-market trial, the post-market trial $p$-value $p_2$ has to fulfill $p_2 \leq \tilde{p}_2 = c_F / p_1$ to achieve overall significance. Similar to the harmonic mean $\chi^2$-test, the sample size calculation (1) for the post-market trial is now based on the adaptive significance level $\tilde{p}_2$.

If the evidence from the pre-market clinical trial is large, $p_1$ is sufficiently small and Fisher’s bound $\tilde{p}_2$ will be larger than one. For example, for $\gamma = 0.025^2$ we have $c_F = 0.000058$. If the pre-market $p$-value $p_1$ is smaller than $c_F$, the post-market clinical trial does not need to be conducted at all. This is an undesired property for conditional drug approval, where regulators usually want additional evidence from a post-marketing trial, no matter how convincing the pre-market trial was. In some cases, drug approval may even be granted based on one study with overwhelming evidence.\textsuperscript{13,14} Using only one study might be viewed as an application of Fisher’s criterion; however in this case the replication requirement is dropped.

2.5 Stouffer’s method

Another $p$-value combination approach is Stouffer’s method, also known as inverse-normal method, which is based on the $z$-scores $z_i = \Phi^{-1}(1-p_i)$. Under the assumption of no effect, the test statistic $z = (z_1 + z_2)/\sqrt{2}$ follows a standard normal distribution. Overall significance at level $\gamma$ is achieved if $z \geq z_{1-\gamma} = \Phi^{-1}(1-\gamma)$. Weights can also be introduced.

Suppose the $p$-value $p_1$ with corresponding $z$-score $z_1$ from the pre-market clinical trial is known. The post-market $z$-score $z_2$ then needs to fulfill

$$z_2 \geq \sqrt{2} z_{1-\gamma} - z_1$$

to achieve overall significance at level $\gamma$. This can be translated to an upper bound for the $p$-value $p_2$ from the post-market trial:

$$p_2 \leq \tilde{p}_2 = 1 - \Phi(\sqrt{2} z_{1-\gamma} - z_1),$$

which serves as an adaptive significance level in the sample size calculation (1) for the post-market trial. For $\gamma = 0.025^2$ we have $\tilde{p}_2 = 1 - \Phi(4.56 - z_1)$, so the upper bound can be close to 1 if the pre-market $z$-value $z_1$ is relatively large.

2.6 Comparison of the different methods

A comparison of the methods described above is given in Figure 1 (upper panel). For each method, the upper bound $p$-value ($\tilde{p}_2$) is shown as a function of the $p$-value ($p_1$) from the pre-market clinical trial. This means that overall significance is achieved if and only if $p_2 \leq \tilde{p}_2$. We set the significance level to $\alpha = 0.025$. If $p_1 \leq 0.025$, the two-trials rule is fulfilled for any $p_2 \leq 0.025$, regardless of the value of $p_1$. For the other three methods (harmonic mean $\chi^2$-test, Fisher’s criterion and Stouffer’s method), the smaller $p_1$, the larger $\tilde{p}_2$. In other words, the more convincing the pre-market clinical trial, the less evidence is required in the post-market clinical trial. However, as mentioned earlier, in case of a very small $p_1$ there would be no need for a post-market clinical trial with Fisher’s criterion. This is an unwanted property in the conditional approval framework where a post-market trial is always required to confirm the result of the pre-market trial. Stouffer’s method will always require a post-market trial, but the adaptive significance level $\tilde{p}_2$ can become very large. Due to this, we decided to not investigate these two methods further. The harmonic mean $\chi^2$-test does not suffer from this drawback: if the necessary condition on $p_1$ is fulfilled ($p_1 \leq 0.048$ for $w_1 = 3$ and $w_2 = 2$, or $p_1 \leq 0.065$ in the unweighted case), the harmonic mean $\chi^2$-test always requires a post-market trial.
Figure 1 (lower panel) shows the variance ratio \( c \) calculated with (1) for the two-trials rule, and the unweighted and weighted harmonic mean \( \chi^2 \)-test (with the suitable significance level) as a function of the \( p \)-value \( p_1 \) of the pre-market clinical trial. The sample size \( n_2 \) of the post-market trial can be calculated from \( c \) via \( n_2 = cn_1(s^2 / \sigma^2_1) \). Two scenarios are considered. In the first one, no inflation of the effect estimate \( \hat{\theta}_1 \) is expected (\( s = 0 \)), while a 50\% shrinkage (\( s = 0.5 \)) is used in the calculation of \( c \) in the second scenario. The power is always fixed at 90\%. Compared to the two-trials rule, the harmonic mean \( \chi^2 \)-test requires a smaller sample size for the post-market clinical trial if the \( p_1 \leq \alpha = 0.025 \) and a larger sample size if \( p_1 \approx 0.025 \). Shrinking the pre-market trial effect estimate \( \hat{\theta}_1 \) results in a larger \( c \) for the same level of power.

### 2.7 Application

We now revisit the Fampridine clinical trial results given in Table 1 to illustrate the applicability of the different methods to the conditional drug approval setting. We apply the variance-stabilizing arc-sine square root transformation\(^{15}(\text{Sec. 3.4})\) to the different proportions in order to achieve approximate normality with a constant standard deviation of \( \sigma_1 = \sigma_2 = 0.5 \). The corresponding \( z \)-values (and \( p \)-values) of the pre- and post-market clinical trial pair turn out to be \( z_1 = 8.6 (p_1 < 0.00001) \) and \( z_2 = 2.5 (p_2 = 0.014) \). The pre-market \( p \)-value \( p_1 \) is so small that Fisher’s method does not require a post-market clinical trial at all at the standard two-trials rule significance level \( \gamma = 0.025^2 \). Stouffer’s method does in principle require a post-market, but the \( p \)-value threshold \( \hat{p}_2 = 0.999976 \) is very close to 1. This illustrates that both Fisher’s and Stouffer’s method are not suitable for the conditional drug approval setting. In sharp contrast, application of (4) with \( z_1 = 8.6 \) and \( \gamma = \alpha^2 = 0.025^2 \) shows that the harmonic mean \( \chi^2 \)-method has the requirement \( p_2 \leq 0.062 \) (respectively \( p_2 \leq 0.083 \) in the weighted case) for the post-market \( p \)-value \( p_2 \). Both requirements are met by the observed post-market \( p \)-value \( p_2 = 0.014 \).

It is also of interest to compare sample size calculation based on the harmonic mean \( \chi^2 \)-test with the traditional two-trials rule. The post-market trial has already been conducted, so we are in a position to reconstruct the post-market sample size calculation based on the almost universal \( \alpha = 0.025 \) standard and to compute the required sample size if the adjusted level \( \hat{p}_2 \) based on (4) would had been used instead. The post-market clinical trial aimed to reach a sample size of 590 patients in total, allowing for 15\% drop out.\(^5\) The corresponding sample size of 251 per group (before drop-out) has a power of 90\% to detect the relative effect size \( d = 0.29 \) at the one-sided significance level \( \alpha = 0.025 \). Sample size calculations based on the harmonic mean \( \chi^2 \)-test are summarized in Table 2. The evidence from the pre-market clinical trial is very strong (with \( z \)-value = 8.6). The adaptive level based on the unweighted harmonic mean \( \chi^2 \)-test is \( p_2 = 0.062 \). If this significance level would have been used in the post-market trial, only 444 patients would have to be included in the post-market clinical trial. This corresponds to a reduction in sample size of 25\%. The 3:2 weighted harmonic mean \( \chi^2 \)-test has the adaptive significance level \( \hat{p}_2 = 0.083 \) so the reduction in sample size would have been even more pronounced (32\% reduction down to 400 patients).

[Table 2 about here.]

### 3 SIMULATION STUDY

**Method** A simulation study was conducted in order to obtain a better understanding of the different methods and how they compare. The details are specified below, and the simulation has been carried out according to the recommendations of Burton et al.\(^{16}\) and Morris, White and Crowther.\(^{17}\) The simulation study was conducted in R Project for Statistical Computing.\(^{18}\) The performance of each method was measured based on the rejection percentage,\(^{17}\) i.e. the proportion of simulations where overall significance can be declared. Under the null, the rejection percentage is the Type-I error rate
Figure 1: The upper bound for the post-market clinical trial $p_2$ and the variance ratio $c$ calculated with the different methods. The upper bound $p_2$ is calculated based on the $p$-value $p_1$ of the pre-market clinical trial. In the weighted version of the harmonic mean $\chi^2$-test, the weights are $w_1 = 3$ and $w_2 = 2$. 
while it is the power in the other scenarios. The number of repetitions $n_{\text{sim}}$ for the simulation was calculated as follows:\textsuperscript{17}

$$n_{\text{sim}} = \frac{E(\text{Power}) \cdot (1 - E(\text{Power}))}{(\text{Monte Carlo se})^2}.$$  

With $n_{\text{sim}} = 10,000$ repetitions, the Monte Carlo standard error $\text{se}_{\text{req}}$ is smaller than 0.5\% for any value of the power.

One-sided, superiority, balanced two-armed (treatment vs. placebo) pre- and post-market clinical trials were simulated assuming different values for the true effect $\theta$: 0, 0.25 and 0.5. We chose four different combinations of $\theta_1$ and $\theta_2$ leading to four scenarios (Figure 2): 1) $\theta_1 = \theta_2 = 0$, 2) $\theta_1 = \theta_2 = 0.25$, 3) $\theta_1 = \theta_2 = 0.5$ and 4) $\theta_1 = 0.5, \theta_2 = 0.25$. The last scenario represents the case where the true effect in the pre-market and the post-market clinical trial are different, maybe due to a switch from a surrogate endpoint to a clinically meaningful outcome. For simplicity, the standard deviation $\sigma_1 = \sigma_2 = \sigma = 1$ is assumed to be the same in the pre- and post-market trial, and so $c$ can be directly interpreted as the relative sample size $n_2/n_1$.

Pre-market clinical trials were assumed to have evidence for a treatment effect ($p_1 \leq \alpha$). The test statistics $z_1$ were simulated from a normal distribution

$$z_1 \sim \text{TN}(\mu, \sigma^2, z_{1-\alpha}, \infty),$$  

with mean $\mu = \theta_1 \cdot \sqrt{n_1}/(\sqrt{2} \sigma) = z_{1-\alpha} + z_{1-\beta}$\textsuperscript{15}(Sec. 3.3) where $1 - \beta$ is the power from the pre-market trial, and truncated to the interval $[z_{1-\alpha}, \infty)$.

The sample size of $n_1 = 85$ (per group) was chosen in such a way that the pre-market clinical trial had a power of 90\% to detect $\theta_1 = 0.5$ and so the power is reduced to only 37\% for $\theta_1 = 0.25$.

The sample size of the post-market clinical trial was calculated to detect the effect $\hat{\theta}_1$ estimated in the pre-market clinical trial with a power of 90\% and with the four methods as discussed in Section 2, assuming no shrinkage ($s = 0$). As can be seen in Equation (1), it is equivalent to know the $z$-value $z_1$ instead of the value of the effect estimate $\hat{\theta}_1$ for the sample size calculation.

Figure 3 compares the sample sizes obtained with the (weighted and unweighted) harmonic mean
Conditional Drug Approval (Preprint) Deforth, Micheloud, Roes, Held

Figure 3: Distribution of the sample sizes $n_2$ of the post-market clinical trial based on the estimated treatment effect $\hat{\theta}_1$ of the pre-market clinical trial and 90\% conditional power. The different methods are unweighted ($H_u$) and weighted ($H_w$) harmonic mean $\chi^2$-test and the two-trials rule ($T$).

$\chi^2$-test and the two-trials rule method. For all three methods there is an upper bound for the sample size regardless of the value of true effect. It corresponds to draws where the one-sided $p$-value $p_1$ from the pre-market clinical trial was borderline significant, i. e. $p_1 \approx 0.025$. The bound based on the weighted harmonic mean $\chi^2$-test ($n_{2\text{max}}^w = 325$) is larger than with the unweighted harmonic mean $\chi^2$-test ($n_{2\text{max}}^u = 293$), which is larger compared to the two-trials rule ($n_{2\text{max}}^w = 230$). With all methods, the median sample size $n_2$ of the post-market clinical trial decreases with increasing true effect $\theta_1$. The larger the true effect $\theta_1$, the fewer patients are necessary to detect it. The test statistic $z_2$ of the post-market clinical trials was simulated from a normal distribution:

$$z_2 \sim N \left( \theta_2 \cdot \sqrt{\frac{n_2}{\sigma^2}}, 1 \right),$$

with $\theta_2$ and $n_2$ being the treatment effect, and the sample size (per group) of the post-market clinical trial, respectively. The choice of selecting only the significant clinical trials for the pre-market trials and all clinical trials for the post-market trials reflects what happens in practice: only the persuasive clinical trials are granted conditional approval, while the results of all the post-market clinical trials need to be reported.

**Type-I error and power of each method**

Table 3 displays the rejection rate (percentage of study pairs where overall significance is declared) for each method and under each scenario. The median sample sizes (absolute $n_2$ and relative $c$) are included for completeness; a more exhaustive picture can be found in Figure 3.

In the first scenario, the true effect is null in both the pre- and post-market clinical trials. The rejection rate hence represents the proportion of clinical trials where overall significance is, in this scenario, incorrectly declared. As the pre-market clinical trial has $p_1 \leq \alpha$ by construction, the rejection rate under this scenario is not the overall Type-I error rate, but the Type-I error rate of the post-market trial only. By definition, the Type-I error rate is $\alpha = 0.025$ with the two-trials rule. With the harmonic mean $\chi^2$-test, in contrast, the Type-I error rate is smaller than $\alpha$ as only $p_1 \leq \alpha$ were considered. In the unweighted (resp. weighted) case, $p_1 \leq 0.065$ (resp. 0.048) might have led to overall significance.

For the other three scenarios, the same logic prevails: the rejection rate can be interpreted as the power of the post-market trial (and not the overall power) as the power of the pre-market clinical trial is fixed (90\% for scenarios 3 and 4 and 37\% for scenario 2). Under every scenarios, the power is the smallest with the two-trials rule as compared to the harmonic mean $\chi^2$-test. One might wonder why...
the power is always below 90%, although every post-market trial was powered for this value. This is due to the truncation \( p_1 \leq 0.025 \), which causes a phenomenon called ‘the winner’s curse’,\(^{19}\) which “occurs when thresholds, such as statistical significance, are used to determine the presence of an effect and is most severe when thresholds are stringent and clinical trials are too small and thus have low power”. This phenomenon creates estimates \( \hat{\theta}_1 \) that are on average overestimated; \textit{i.e.} larger than the true effect \( \theta_1 \). A similar problem occurs when the aim is to conduct a replication of a promising subgroup finding in an overall non-significant trial.\(^{20}\) As the post-market trials were powered to detect \( \theta_1 \), the sample size \( n_2 \) is too small and the post-market trials are underpowered. The effect of the winner’s curse on the power is only moderate under the third scenario, it is in contrast substantial under the second scenario as the pre-clinical trial is underpowered (37%). In addition to this, the fourth scenario suffers from another problem: the true effects \( \theta \) suffers from another problem: the true effects \( \theta \) are on average overestimated; \textit{i.e.} larger than the true effect \( \theta \). A similar problem occurs when the aim is to conduct a replication of a promising subgroup finding in an overall non-significant trial.\(^{20}\) As the post-market trials were powered to detect \( \theta_1 \), the sample size \( n_2 \) is too small and the post-market trials are underpowered. The effect of the winner’s curse on the power is only moderate under the third scenario, it is in contrast substantial under the second scenario as the pre-clinical trial is underpowered (37%). In addition to this, the fourth scenario suffers from another problem: the true effects \( \theta \) are on average overestimated; \textit{i.e.} larger than the true effect \( \theta \).

In the following we compare the probability that the unweighted harmonic mean \( \chi^2 \)-test method leads to a smaller sample size \( n_2^{Hu} \) and simultaneously a larger power

\[
\text{power}_{Hu} = 1 - \Phi \left( z_2 - \frac{\theta \sqrt{n_2^{Hu}}}{\sigma \sqrt{2}} \right)
\]

(6)

to detect the true effect \( \theta \) than the two-trials rule. For simplicity, we assume that the outcomes and standard deviations are the same in pre- and post-market trials, \textit{i.e.} \( \theta_1 = \theta_2 = \theta \) and \( \sigma_1 = \sigma_2 = \sigma \), and apply no shrinkage in the sample size calculation (1). We denote this probability \( \Pr(H_u \text{ superior}) \), namely the probability that the harmonic mean \( \chi^2 \)-test is superior than the two-trials rule. Similarly, the harmonic mean \( \chi^2 \)-test is inferior if the sample size \( n_2^{2TR} \) is smaller and the power

\[
\text{power}_{2TR} = 1 - \Phi \left( z_{1-\alpha} - \frac{\theta \sqrt{n_2^{2TR}}}{\sigma \sqrt{2}} \right)
\]

(7)

to detect the true effect \( \theta \) is larger with the two-trials rule than with the harmonic mean \( \chi^2 \)-test method. We also calculate the probability of inconclusive results, \textit{i.e.} the harmonic mean \( \chi^2 \)-test method leads to a smaller sample size, but also to a smaller power, and the other way around.

In the calculation of the sample size \( n_2 \) (or equivalently the relative sample size \( c \)), the only difference between the two approaches is the threshold for the \( z \)-statistics of the post-market trial: \( z_{1-\alpha} \) for the two-trials rule and \( z_2 \) for the harmonic mean \( \chi^2 \)-test method (see (4)). Using (1), a smaller sample size \( n_2^{Hu} \) is required with the harmonic mean \( \chi^2 \)-test method as compared to the two-trials rule if \( z_2 < z_{1-\alpha} \) or equivalently \( p_2 > a \). From (4), we see that \( p_2 > a \) holds whenever \( z_1 > 1/\sqrt{4/c_H} - 1/z_{1-\alpha}^2 := b \). If \( a = 0.025 \), this happens when \( p_1 < 0.009 \), as can also be observed in Figure 1.

Furthermore, solving \( \text{power}_{Hu} > \text{power}_{2TR} \) with \( p_2 > a \) indicates that the power of the post-market trial to detect the true treatment effect is larger with the harmonic mean \( \chi^2 \)-test method if \( z_1 > \mu \), the mean of the normal distribution defined in (5). The condition for the harmonic mean \( \chi^2 \)-test to be superior (simultaneously smaller sample size and larger power) than the two-trials rule is therefore

\[
z_1 > \max(\mu, b).
\]

(8)
The harmonic mean $\chi^2$-test is inferior to the two-trials rule if

$$\mu < z_1 < b.$$ (9)

The probabilities of (8) and (9) only depend on the power $1 - \beta$ from the pre-market clinical trial (and the level $\alpha$) and can be computed with numerical integrations as follows:

$$\Pr(H_u \text{ superior}) = \int_{\max(\mu, b)}^{\infty} \text{TN}(z_1; \mu, 1, z_1 - \alpha, \infty) dz_1$$ (10)

$$\Pr(H_u \text{ inferior}) = \begin{cases} \int_{\mu}^{b} \text{TN}(z_1; \mu, 1, z_1 - \alpha, \infty) dz_1, & \text{if } \mu < b \\ 0, & \text{otherwise} \end{cases}$$ (11)

where TN is the truncated normal distribution defined in (5).

The results are depicted in Figure 4. For $\alpha = 0.025$, the condition $\mu < b$ in (11) is fulfilled if $1 - \beta < 66.1\%$, so if the power of the pre-market clinical trial is larger than 66.1%, the harmonic mean $\chi^2$-test is never inferior to the two-trials rule. Moreover, if the power of the pre-market trial is 50% or less, $z_1 > \mu$ always holds (as $z_1 > z_1 - \alpha$ by construction), and so the probability of an inconclusive result is null. The probability that the harmonic mean $\chi^2$-test is inferior to the two-trials rule is larger than 50% only if the pre-market clinical trial was severely underpowered to detect the true effect $\theta$ (power of pre-market trial < 16.6%). An extended Figure with a differentiation between the two types of inconclusive results can be found in the appendix (Figure 6). Moreover, similar results can be obtained for the weighted harmonic mean $\chi^2$-test.
4 EXTENSION: POWER AT INTERIM

The methods discussed above combine the final results of the pre- and post-market clinical trials. Another scenario is that the post-market clinical trial is still ongoing but the results of an interim analysis are already available. In this scenario, the power at interim can be calculated. The power at interim is “the power of a [post-market clinical trial] taking into account the data from an interim analysis”. 10(p. 2) If the power at interim is very low, it might be advisable to consider a premature termination of the clinical trial for futility. The interim power can be the power to detect the effect \( \hat{\theta}_1 \) from the pre-market clinical trial (conditional power, CPI), the power to detect \( \hat{\theta}_1 \) acknowledging its uncertainty (informed predictive power, IPPi) or can ignore the original result (predictive power, PPi). The description and formulas of the three types of interim power can be found Micheloud and Held. 10 We will focus on the IPPi in this work; we have adapted it to the drug market licensing framework and applied it to the simulation study for illustrative purposes.

**Method** Similar to the previous sections, two methods are used to calculate the interim power: the harmonic mean \( \chi^2 \)-test (we restrict on the unweighted case) and the two-trials rule. The interim power is thus either the probability that \( p_2 \leq \bar{p}_2 \) as calculated in (4) (harmonic mean \( \chi^2 \)-test) or the probability that \( p_2 \leq \alpha \) (two-trials rule) at the final analysis, given the samples collected so far at the time of the interim analysis. We assume that the interim analysis occurs after half of the data was available; and so the sample size at interim \( n_{2,i} = n_2/2 \), with \( n_2 \) as in Figure 3. The test statistics \( z_{2,i} \) at interim of the post-market clinical trial was simulated based on a normal distribution:

\[
z_{2,i} \sim N \left( \frac{\theta_2 \cdot \sqrt{n_{2,i}}}{\sqrt{2}\sigma}, 1 \right).
\]

![Figure 5: Informed predictive predictive power with the two-trials rule (T) and the unweighted harmonic mean \( \chi^2 \)-test (Hu) under the four scenarios.](#)

**Results and discussion** The interim power under each scenario and with \( n_{2,i} = n_2/2 \) is presented in Figure 5. Some authors 10,21,22 suggest to use a futility boundary of 20%, i.e. the trial should be stopped at interim is the interim power falls below 20%. With this threshold, the proportion of clinical trials
stopped for futility under the null (scenario 1) is 67.7% ($H_u$) and 65.3% ($T$), so more clinical trials would be correctly stopped using the harmonic mean $\chi^2$-test rather than the two-trials rule. In contrast, the proportion of clinical trials (incorrectly) stopped at interim under the alternative (scenarios 2, 3 and 4) is smaller with the harmonic mean $\chi^2$-test rather than with the two-trials rule (scenario 2: 12.5% ($H_u$) vs 13.8% ($T$), scenario 3: 2% ($H_u$) vs 2.5% ($T$) and scenario 4: 9.3% ($H_u$) vs 11.6% ($T$)). In addition, the median interim power is smaller with $H_u$ as compared to $T$ under the null, and larger under the alternative.

5 CONCLUSIONS

We evaluated the applicability of four statistical methods within the conditional or accelerated drug approval framework. In contrast to Fisher’s criterion or Stouffer’s method, the harmonic mean $\chi^2$-test always requires a post-market clinical trial to be conducted. The harmonic mean $\chi^2$-test takes into account the evidence from the pre-market clinical trial with an adaptive significance level for the post-market trial. In contrast to the two-trials rule, the harmonic mean $\chi^2$-test requires a smaller sample size for the post-market clinical trial if the $p$-value $p_1$ from the pre-market clinical trial is $\ll 0.025$ and a larger sample size if $p_1 \approx 0.025$ (for $\alpha = 0.025$). The harmonic mean $\chi^2$-test can also be used to calculate the power at interim of the post-market clinical trial and can easily be extended to more than two clinical trials. The harmonic mean $\chi^2$-test method used for the design and the analysis of the post-market study tends to be superior to the two-trials rule, i.e. the proposed method tends to require a smaller sample size while still achieving larger power. Finally, we conclude with some discussion points.

In the two-trials rule, the primary function of replication by conducting a second study is not precisely to reach a combined error rate of $\alpha^2$ across studies. The suggestion of assessing an overall error rate across studies also does not appear in guidance from regulators. The primary goal is to consider independent replication, preferably under different circumstances, as part of the scientific method as a whole - confirming the robustness of observed effects. However, we do argue that it is valuable to compare alternatives for clinical trials proposed to be conducted to promote conditional drug approval to full approval in a standardized way, i.e. making sure that they achieve the same level of evidence. This paper provides useful insights how to do so efficiently in terms of statistical tests and sample size, which is very relevant in those settings for which conditional approval is considered appropriate. For final regulatory assessment, achieving statistical significance under appropriate Type-I error control is in essence a ticket to entry of full evaluation, because there is a clear signal of confirmed efficacy. Full assessment of all aspects (efficacy and safety in the broad sense) ultimately leads to regulatory decisions for approval, which are rarely simple acceptance or not: in almost all cases refinements in indication and target population(s) are included, as well as conditions for manufacturing and use are part of the decision.

6 Appendix

6.1 Detailed version of Figure 4

Figure 6 is a more detailed version of Figure 4, where the inconclusive results are separated into two categories: the harmonic mean $\chi^2$-test method is superior to the two-trials rule either with respect to the sample size $n_2$ (smaller) or the power (larger).
Figure 6: Probability of the harmonic mean $\chi^2$-test to be superior/inferior than the two-trials rule as a function of the true power from the pre-market clinical trial.

References

1. European Medicines Agency (EMA). Guideline on the scientific application and the practical arrangements necessary to implement Commission Regulation (EC) No 507/2006 on the conditional marketing authorisation for medicinal products for human use falling within the scope of Regulation (EC) No 726/2004 https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-scientific-application-practical-arrangements-necessary-implement-commission-regulation-2006-conditional-marketing-authorisation-medicinal-products-human-use-falling_en.pdf 2016. Accessed 05 Oct 2021. Last update 25 Feb 2016.

2. European Medicines Agency (EMA). EMA receives application for conditional marketing authorisation of COVID-19 Vaccine AstraZeneca https://www.ema.europa.eu/en/news/ema-receives-application-conditional-marketing-authorisation-covid-19-vaccine-astrazeneca 2021. Accessed 20 Sep 2021. Last update 12 Jan 2021.

3. European Medicines Agency (EMA). Fampyra https://www.ema.europa.eu/en/medicines/human/EPAR/fampyra#overview-section 2020. Accessed 13 Sep 2021. Last update 24 Aug 2020.

4. European Medicines Agency (EMA). Assessment report Fampyra https://www.ema.europa.eu/en/documents/assessment-report/fampyra-epar-public-assessment-report_en.pdf 2011. First published 04 Aug 2011. Last update 23 Jun 2011.

5. Hobart J, Ziemssen T, Feys P, et al. Assessment of clinically meaningful improvements in self-reported walking ability in participants with Multiple Sclerosis: Results from the randomized, double-blind, Phase III ENHANCE trial of prolonged-release Fampridine CNS Drugs. 2019;33:61-79.

6. Held L. The harmonic mean $\chi^2$-test to substantiate scientific findings Journal of the Royal Statistical Society: Series C (Applied Statistics). 2020;69:697-708.

7. Spiegelhalter D J, Abrams K R, Myles J P. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester: John Wiley & Sons Ltd 2004.

8. U.S. Food and Drug Administration (FDA). Providing clinical evidence of effectiveness for human
9. Ioannidis J P A. Why most discovered true associations are inflated *Epidemiology*. 2008;19:640–648.
10. Micheloud C, Held L. Power calculations for replication studies *Statistical Science*. 2021. To appear, preprint available at https://arxiv.org/abs/2004.10814.
11. Fisher R A. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd 13th ed. (rev.) ed. 1958.
12. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses *Biometrics*. 1994;50:1029-1041.
13. Fisher L D. One large, well-designed, multicenter study as an alternative to the usual FDA paradigm *Drug Information Journal*. 1999;33:265–271. https://doi.org/10.1177/009286159903300130.
14. Shun Z, Chi E, Durrleman S, Fisher L. Statistical consideration of the strategy for demonstrating clinical evidence of effectiveness—one larger vs two smaller pivotal studies *Statistics in Medicine*. 2005;24:1619–1637.
15. Matthews J N S. *Introduction to Randomized Controlled Clinical Trials*. Chapman & Hall/CRC 2nd ed. 2006.
16. Burton A, Altman D G, Royston P, Holder R L. The design of simulation studies in medical statistics *Statistics in Medicine*. 2006;25:4279-4292.
17. Morris T P, White I R, Crowther M J. Using simulation studies to evaluate statistical methods *Statistics in Medicine*. 2019;38:2074-2102.
18. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria 2021.
19. Button K S, Ioannidis J P A, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience *Nature Reviews Neuroscience*. 2013;14:365.
20. Tanniou J, van der Tweel I, Teerenstra S, Roes K C B. Level of evidence for promising subgroup findings in an overall non-significant trial *Statistical Methods in Medical Research*. 2016;25:2193–2213.
21. DeMets D L. Futility approaches to interim monitoring by data monitoring committees *Clinical Trials*. 2006;3:522–529.
22. Bauer P, König F. The reassessment of trial perspectives from interim data—a critical view *Statistics in Medicine*. 2006;25:23–36. https://doi.org/10.1002/sim.2180.
23. Altman D G, Machin D, Bryant T N, Gardner M J. *Statistics with Confidence*. BMJ Books 2nd ed. 2000.
List of Tables

1 Results from pre- and post-market clinical trials on the efficacy of Fampridine for the treatment of patients suffering from Multiple Sclerosis. ........................................... 18
2 Sample size calculation for the post-market clinical trial based on the two-trials rule (left) and the harmonic mean $\chi^2$-test (right). ................................................................. 19
3 Median sample size $n_2$ of the post-market clinical trial, median relative sample size $c$, and rejection rate under the four scenarios with the three methods: two-trials rule ($T$), weighted ($H_w$) and unweighted ($H_u$) harmonic mean $\chi^2$-test. ......................... 20
Table 1: Results from pre- and post-market clinical trials on the efficacy of Fampridine for the treatment of patients suffering from Multiple Sclerosis.

| clinical trial   | Time   | No. of Patients | Primary Outcome                                         | Estimates (in %)                      |
|------------------|--------|-----------------|----------------------------------------------------------|--------------------------------------|
| pooled analysis  | pre    | Fampridine: 394 | difference in responder rates (improved walking speed)   | Fampridine: 37.3                      |
| (MS-F202/3/4)    |        | Placebo: 237    |                                                          | Placebo: 8.9                         |
|                  |        | Total: 631      |                                                          | Difference: 28.4                      |
|                  |        |                 |                                                          | 95% CI: 22.1 to 34.2                 |
|                  | post   | Fampridine: 315 | difference in responder rates (improved walking ability) | Fampridine: 43.2                      |
| 218MS305         |        | Placebo: 318    |                                                          | Placebo: 33.6                        |
| phase III        |        | Total: 633      |                                                          | Difference: 9.5                       |
|                  |        |                 |                                                          | 95% CI: 1.9 to 17                    |

Source: 45. Shown are unadjusted risk differences with 95% CIs 23.
Table 2: Sample size calculation for the post-market clinical trial based on the two-trials rule (left) and the harmonic mean $\chi^2$-test (right).

|                          | Two-trials rule | Harmonic Mean $\chi^2$-Test |
|--------------------------|-----------------|-------------------------------|
|                          | unweighted      | weighted                      |
| Required significance level | 0.025           | 0.062                         | 0.083                        |
| Required sample size      | 590             | 444                           | 400                          |
| Sample size reduction     | 25%             | 32%                           |                              |
Table 3: Median sample size $n_2$ of the post-market clinical trial, median relative sample size $c$, and rejection rate under the four scenarios with the three methods: two-trials rule ($T$), weighted ($H_w$) and unweighted ($H_u$) harmonic mean $\chi^2$-test.

| Method | $n_2$ | $c$ | Rejection rate [%] |
|--------|-------|-----|--------------------|
| scenario 1 | $H_u$ | 184 | 2.2 | 2.0 |
| | $H_w$ | 182 | 2.2 | 2.2 |
| | $T$   | 175 | 2.1 | 2.5 |
| scenario 2 | $H_u$ | 131 | 1.6 | 55.0 |
| | $H_w$ | 125 | 1.5 | 56.0 |
| | $T$   | 138 | 1.6 | 53.6 |
| scenario 3 | $H_u$ | 66  | 0.8 | 85.6 |
| | $H_w$ | 60  | 0.7 | 86.2 |
| | $T$   | 78  | 0.9 | 84.6 |
| scenario 4 | $H_u$ | 66  | 0.8 | 42.7 |
| | $H_w$ | 60  | 0.7 | 45.0 |
| | $T$   | 78  | 0.9 | 37.9 |