American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Review Article

Statistical Test for Composite Hypothesis in Clinical Research

Fuyu Song[1], Xinyi Ma[2] and Shein-Chung Chow[3]*

[1]*Center for Food and Drug Inspection, National Medical Products Administration, China*

[2]*Amherst College, Amherst, Massachusetts, USA*

[3]*Professor of Biostatistics and Bioinformatics, Duke University School of Medicine, USA*

***Corresponding author:** Shein-Chung Chow, Professor of Biostatistics and Bioinformatics, Duke University School of Medicine, 2424 Erwin Road, Room 11037, Durham, NC 27705, USA.

To Cite This Article: *Shein-Chung Chow, Statistical Test for Composite Hypothesis in Clinical Research. 2020 - 10(2). AJBSR.MS.ID.001485.*

DOI: 10.34297/AJBSR.2020.10.001485.

Received: August 24, 2020; **Published:** September 03, 2020

Abstract

In clinical evaluation of the safety and efficacy of a test treatment under investigation, a typical approach is to test for the null hypothesis of no treatment difference in efficacy in randomized clinical trials (RCT). The investigator would reject the null hypothesis of no treatment difference and then conclude the alternative hypothesis that the treatment is efficacious. In practice, however, this typical approach based on test for efficacy alone may not be appropriate for a full assessment of both efficacy and safety of the test treatment under study. Alternatively, [1] suggested testing composite hypothesis by taking both safety and efficacy into consideration. In this article, appropriate statistical test for a composite hypothesis of non-inferiority in efficacy and superiority in safety is derived. The impact on power calculation for sample size requirement when switching from a single hypothesis (e.g., for efficacy) to a composite hypothesis (e.g., for both safety and efficacy) is also examined.

Keywords: Randomized Clinical Trial (RCT); Composite Hypothesis; Non-inferiority; Superiority; Power Calculation

Introduction

In clinical evaluation of a test treatment under investigation, a traditional approach is to first test for a null hypothesis that there is no treatment difference in efficacy alone in randomized clinical trials (RCTs) [2]. The investigator would reject the null hypothesis of no treatment difference and then conclude the alternative hypothesis that there is a difference and in favor of the test treatment under investigation. As a result, if there is a sufficient power for correctly detecting a clinically meaningful difference if such a different truly exists, the test treatment is then claimed to be efficacious. The test treatment will be reviewed and approved by the regulatory agency if the test treatment is well tolerated and there appears no safety concerns. In practice, the intended clinical trial is often powered for achieving the study objective with a desired power (say 80%) at a pre-specified level of significance (say 5%).

This traditional approach, however, may not be appropriate because one single primary efficacy endpoint cannot fully assess the performance of the treatment with respect to both efficacy and safety under study. Statistically, the traditional approach based on single primary efficacy endpoint for clinical evaluation of both safety and efficacy is a conditional approach (i.e., conditional on safety performance). It should be noted that under the traditional (conditional) approach, the observed safety profile may not be of any statistical meaning (i.e., the observed safety profile could be by chance alone and may not be reproducible). In addition, the traditional approach for clinical evaluation of both efficacy and safety may have inflated the false positive rate of the test treatment in treating the disease under investigation.

In the past several decades, the traditional approach is found to be inefficient because many drug products have been withdrawn

from the marketplace because of the unreasonable risks to patients. For illustration purpose, Table 1 provides a list of significant withdrawn drugs between 2000-2010 [3]. As it can be seen from Table 1, most drugs withdrawn from the marketplace are due to safety concern (i.e., unreasonable risks to the patients). These unreasonable risks to the patients include unexpected adverse effects that were not detected during late phase clinical trials and were only apparent from post-marketing surveillance data from the wider patient population (Table 1).

To take both safety and efficacy into consideration in clinical trials, [1] suggested testing composite hypothesis by testing non-inferiority/superiority or equivalence of safety and efficacy as compared to a control (e.g., placebo control or active control). As an example, a commonly considered composite hypothesis is that H_0 : not NS versus H_a : NS where N represents testing for non-inferiority in efficacy and S is for testing superiority in safety. Under the composite hypothesis, it is of interest to examine the impact of power calculation for sample size requirement when switching from testing a single hypothesis (i.e., for efficacy alone) to testing a composite hypothesis (i.e., for both safety and efficacy).

In the next section, several composite hypotheses which will take both safety and efficacy into consideration are proposed. In Section 3, for illustration purpose, statistical methods for testing the composite hypothesis that H_0 : not NS versus H_a : NS are derived. Section 4 studies the impact on power calculation for sample size requirement when switching from testing for a single hypothesis (for efficacy alone) to testing for a composite hypothesis (for both safety and efficacy). In Section 5, some concluding remarks are provided.

Hypotheses for Clinical Evaluation

In clinical trials, for clinical evaluation of efficacy, commonly considered approaches include tests for hypotheses of superiority (S), non-inferiority (N), or (therapeutic) equivalence (E). For safety assessment, the investigator usually examines the safety profile in terms of adverse events and other safety parameters such as laboratory testing to determine whether the test treatment is either better (superiority), non-inferior (non-inferiority) or similar (equivalence) as compared to the control. As an alternative to the traditional approach, [1] suggested testing composite hypothesis that will take into consideration both safety and efficacy. For illustration purpose, Table 2 provides a summary all possible scenarios of composite hypotheses for clinical evaluation of safety and efficacy of a test treatment under investigation.

Statistically, we would reject the null hypothesis at a pre-specified level of significance and conclude the alternative hypothesis with a desired power. For example, the investigator may be interested in testing non-inferiority in efficacy and superiority in safety of a test treatment as compared to a control. In this case, we can consider testing the null hypothesis that H_0 : not NS, where N denotes the non-inferiority in efficacy and S represents superiority of safety. We would reject the null hypothesis and conclude the alternative hypothesis that H_a : NS , i.e., the test treatment is non-inferior to the active control agent and its safety profile is superior to the active control agent. To test the null hypothesis that H_0 : not NS, appropriate statistical tests should be derived under the null hypothesis. The derived test statistics can then be evaluated for achieving the study objectives with a desired power under the alternative hypothesis. The selected sample size will ensure that the intended trial will achieve the study objectives of (i) establishing non-inferiority of the test treatment in efficacy and (ii) showing superiority of the safety profile of the test treatment at a pre-specified level of significance.

Note that for testing H_0 : not NS versus H_a : NS, the alternative hypothesis is that the test treatment is non-inferior (N) in efficacy and superior (S) in safety. Thus, the null hypothesis is not NS, i.e., the test treatment is inferior in efficacy or the test treatment is not superior in safety. Thus, the null hypothesis actually consists of three subsets: (i) the test treatment is inferior in efficacy and superior in safety; (ii) the test treatment is non-inferior in efficacy and not superior in safety; (iii) the test treatment is inferior in efficacy and not superior in safety. It would be complicated to consider all these three subsets when derive appropriate statistical test under the null hypothesis.

It also should be noted that in the interest of controlling the overall type I error rate at the α level, appropriate α levels (say α_1 for efficacy and α_2 for safety) should be chosen. Switching from a single hypothesis testing to a composite hypothesis testing, sample size increase is expected.

Testing Composite Hypothesis of Safety and Efficacy

For illustration purpose, consider the following composite hypothesis that

$$H_0 : \text{not } NS \text{ versus } H_a : NS, \quad (1)$$

where represents testing for non-inferiority in efficacy and S is for testing superiority in safety

Derivation of Statistical Test Under the Composite Null Hypothesis

To test the null hypothesis that , appropriate statistical tests should be derived under the null hypothesis. Let X and Y be the efficacy and safety endpoint, respectively. Assume that (X, Y) follows a bi-variate normal distribution with mean (μ_x, μ_y) and variance-covariance matrix $\sum$ i.e., where

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

Suppose that the investigator is interested in testing non-inferiority in efficacy and superiority in safety of a test treatment as compared to a control. The following composite hypotheses may be considered:

$$H_0 : \mu_{X1} - \mu_{X2} \le -\delta_X \text{ or } \mu_{Y1} - \mu_{Y2} \le \delta_Y \quad \text{v.s.}$$

$$H_a : \mu_{X1} - \mu_{X2} > -\delta_X \text{ and } \mu_{Y1} - \mu_{Y2} > \delta_Y \qquad (2)$$

where (μ_{X1}, μ_{Y1}) and (μ_{X2}, μ_{Y2}) are the means of (X, Y) for the test treatment and the control, respectively, and δ_X and δ_Y are the corresponding non-inferiority margin and superiority margin. Note that δ_X and δ_Y are positive constants. If the null hypothesis is rejected based on a statistical test, we conclude that the test treatment is non-inferior to the control in efficacy endpoint X, and is superior over the control in safety endpoint Y.

To test the above composite hypotheses, suppose that a random sample of (X, Y) is collected from each treatment arm. In particular, $(X_1, Y_1), \ldots, (X_{1n_1}, Y_{1n_1})$ are $i.i.d.N((\mu_{X1}, \mu_{Y1}), \Sigma)$, which is the random sample from the test treatment, and $(X_{21}, Y_{21}), \ldots, (X_{2n_2}, Y_{2n_2})$ are $i.i.d.N((\mu_{X2}, \mu_{Y2}), \Sigma)$, which is the random sample from the control treatment, where i.i.d. stands for independent and identically distributed. Let $\bar{X}_1$ and $\bar{X}_2$ be the sample means of X in the test treatment and the control, respectively. Similarly, $\bar{Y}_1$ and $\bar{Y}_2$ are the sample means of Y in the test treatment and the control, respectively. It can be verified that the sample mean vector $(\bar{X}_i, \bar{Y}_i)$ follows a bi-variate normal distribution. In particular, $(\bar{X}_i, \bar{Y}_i)$ follows $N((\mu_{Xi}, \mu_{Yi}), n_i^{-1}\Sigma)$. Since $(\bar{X}_1, \bar{Y}_1)$ and $(\bar{X}_2, \bar{Y}_2)$ are independent bi-variate normal vectors, it follows that $(\bar{X}_1 - \bar{X}_2, \bar{Y}_1 - \bar{Y}_2)$ is also normally distributed as $N((\mu_{X1} - \mu_{X2}, \mu_{Y1} - \mu_{Y2}), (n_1^{-1} + n_2^{-1})\Sigma)$ For simplicity, we assume Σ is known, i.e., the values of parameters σ_X^2, σ_Y^2 and are known. To test the composite hypothesis H_0 for both efficacy and safety, we may consider the following test statistics

$$T_X = \frac{\bar{X}_1 - \bar{X}_2 + \delta_X}{\sqrt{(n_1^{-1} + n_2^{-1})\sigma_X^2}}$$

$$T_y = \frac{\bar{Y}_1 - \bar{Y}_2 + \delta_Y}{\sqrt{(n_1^{-1} + n_2^{-1})\sigma_Y^2}}$$

Thus, we would reject the null hypothesis H_0 for large values of T_X and T_Y. Let C_1 and C_2 be the critical values for T_X and T_Y, respectively. Then, we have

$$P(T_X > C_1, T_Y > C_2) = P\left[U_X > C_1 - \frac{\mu_{X1} - \mu_{X2} + \delta_X}{\sqrt{(n_1^{-1} + n_2^{-1})\sigma_X^2}}, U_Y > C_2 - \frac{\mu_{Y1} - \mu_{Y2} + \delta_Y}{\sqrt{(n_1^{-1} + n_2^{-1})\sigma_Y^2}}\right], \quad (3)$$

where (U_X, U_Y) is the standard bi-variate normal random vector, i.e., a bi-variate normal random vector with zero means, unit variances and a correlation coefficient of ρ.

Under the null hypothesis H_0 that $\mu_{X1} - \mu_{X2} \le -\delta_X \text{ or } \mu_{Y1} - \mu_{Y2} \le \delta_Y$, it can be shown that the upper limit of $P(T_X > C_1, T_Y > C_2)$ is the maximum of the two probabilities, i.e., $\max\{1 - \Phi(C_1), 1 - \Phi(C_2)\}$, where Φ is the cumulative distribution function of the standard normal distribution. A brief proof is given below. For given constants a_1 and a_2 and a standard bi-variate normal vector

$(U_X, U_Y) \sim N\left[(0,0), \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]$, we have

$$P(U_X > a_1, U_Y > a_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{a_1}^{+\infty}\int_{a_2}^{+\infty} \exp\left\{-\frac{x^2 + y^2 - 2\rho xy}{2(1-\rho^2)}\right\} dydx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{a_1}^{+\infty} \exp\left\{-\frac{x^2}{2}\right\} \int_{a_2}^{+\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right\} dydx$$

$$1 - \Phi(a_1) - \frac{1}{\sqrt{2\pi}} \int_{a_1}^{+\infty} \Phi\left(\frac{a^2 - \rho x}{\sqrt{1-\rho^2}}\right) \exp\left\{-\frac{x^2}{2}\right\} dx \quad (4)$$

Since the joint distribution of (U_X, U_Y) is symmetric, (4) is also equal to

$$1 - \Phi(a_2) - \frac{1}{\sqrt{2\pi}} \int_{a_2}^{+\infty} \Phi\left(\frac{a^2 - \rho y}{\sqrt{1-\rho^2}}\right) \exp\left\{-\frac{y^2}{2}\right\} dy \quad (5)$$

Based on (5), $P(T_X > C_1, T_Y > C_2)$ can be expressed by (4) and (5) with a_1 and a_2 replaced by

$$D_1 = C_1 - \frac{\mu_{X1} - \mu_{X2} + \delta_X}{\sqrt{(n_1^{-1} + n_2^{-1})\sigma_X^2}}$$

and

$$D_2 = C_2 - \frac{\mu_{Y1} - \mu_{Y2} + \delta_Y}{\sqrt{(n_1^{-1} + n_2^{-1})\sigma_y^2}}$$

respectively. Under the null hypothesis H_0 that $\mu_{X1} - \mu_{X2} \le -\delta_X \text{ or } \mu_{Y1} - \mu_{Y2} \le \delta_Y$, it's true that either $D_1 \ge C_1$ or $D_2 \ge C_2$. Since integrals in (4) and (5) are positive, it follows that

$$P(T_X > C_1, T_Y > C_2 \mid H_0) < \max(1 - \Phi(C_1), 1 - \Phi(C_2))$$

To complete the proof, we need to show for any $\varepsilon > 0$, δ_X and δ_Y (>0), and given values of other parameters, there exists values of $\mu_{X1} - \mu_{X2}$ and $\mu_{Y1} - \mu_{Y2}$ such that (6) is larger than $1 - \Phi(C_1) - \varepsilon$ and $1 - \Phi(C_2) - \varepsilon$. Let . Then $\mu_{X1} - \mu_{X2} = -\delta_X$ (5) becomes

$$1-\Phi(C_1)-\frac{1}{\sqrt{2\pi}}\int_{c_1}^{+\infty}\Phi\left(\frac{D_2-\rho x}{\sqrt{1-\rho^2}}\right)\exp\left\{-\frac{x^2}{2}\right\}dx \quad (6)$$

For $\rho > 0$, there exists a negative value K such that when $D_2 < K$, for any x in $[C_1+00], \Phi\left(\frac{D_2-\rho x}{\sqrt{1-\rho^2}}\right) < \varepsilon$

For sufficient large $\mu_{Y1}-\mu_{Y2}$, it can happen that $D_2 < K$. Therefore, for sufficient large $\mu_{Y1}-\mu_{Y2}$, (9) > $1-\Phi(C_1)-\varepsilon$. For $\rho \le 0$, express the integral in (7) as $I_1 + I_2$, where

$$I_1 = \int_{c_1}^{E}\Phi\left(\frac{D_2-\rho x}{\sqrt{1-\rho^2}}\right)\exp\left\{-\frac{x^2}{2}\right\}dx$$

and $I_2 = \int_{E}^{+\infty}\Phi\left(\frac{D_2-\rho x}{\sqrt{1-\rho^2}}\right)\exp\left\{-\frac{x^2}{2}\right\}dx$

ϵ is chosen such that $I_2 \le \int_{E}^{+\infty}\exp\left\{-\frac{x^2}{2}\right\}dx$ < 0.5 ϵ.The first inequality holds as the cumulative distribution is always ≤ 1. For a chosen value of ϵ, the argument for $\rho > 0$ can be applied to prove I_1 <0.5 ϵ for sufficient large $\mu_{Y1}-\mu_{Y2}$. Hence, $P(T_X > C_1, T_Y > C_2 \mid H_0)$ is greater than $1-\Phi(C_1)-\varepsilon$ for $\mu_{X1}-\mu_{X2} = -\delta_X$ and sufficient large $\mu_{Y1}-\mu_{Y2}$. Similarly, it can be proved that $P(T_X > C_1, T_Y > C_2 \mid H_0)$ is greater than $1-\Phi(C_2)-\varepsilon$ for $\mu_{Y1}-\mu_{Y2} \le \delta_Y$ and sufficient large $\mu_{X1}-\mu_{X2}$. This completes the proof.

Therefore, the type I error of the test based on T_X and T_Y can be controlled at the level of α by appropriately choosing corresponding critical values of C_1 and C_2. Denote by Z_α the upper α - percentile of the standard normal distribution. Then, the power function of the above test is $P(T_X > Z_{\alpha_1}, T_Y > Z_{\alpha_2})$, which can be calculated from (5) and the cumulative distribution function of the standard bi-variate distribution.

The Impact on Power Calculation for Sample Size

Fixed Power Approach

In practice, when switching from testing a single hypothesis (i.e., based on a single study endpoint such as the efficacy endpoint in clinical trials) to testing a composite hypothesis (i.e., based on two study endpoints such as both safety and efficacy endpoints in clinical trials), increase in sample size is expected. Let X be the efficacy endpoint in clinical trials. Consider testing the following single non-inferiority hypothesis with a non-inferiority margin of δ_X:

$$H_{01}: \mu_{X1}-\mu_{X2} \le -\delta_X \; v.s. \; H_{a_1}: \mu_{X1}-\mu_{X2} > -\delta_X$$

Then, a commonly used test is to reject the null hypothesis H_{01} at the α level of significance if $T_X > Z_\alpha$. The total sample size for concluding the test treatment is non-inferior to the control with $1-\beta$ power if the difference of mean $\mu_{X1}-\mu_{X2} > -\delta_X$ is

$$N_X = \frac{(1+r)^2(Z_\alpha+Z_\beta)^2\sigma_X^2}{r(\mu_{X1}-\mu_{X2}+\delta_X)^2}$$

Where $r = n_2/n_1$ is the sample size allocation ratio between the control and test treatment [4]? (Table 3) gives total sample size (N_X) for test of non-inferiority based on efficacy endpoint X and total sample size (N) for testing composite hypothesis based on both efficacy endpoint X and safety endpoint Y, for various scenarios. In particular, we calculated sample sizes for α =0.05, β =0.20, $\mu_{Y1}-\mu_{Y2}-\delta_Y = 0.3$, = 1, and several values of $\Delta = \mu_{X1}-\mu_{X2}+\delta_X$ and other parameters. For a hypothesis of superiority of the test treatment in safety, i.e., the component with respect to safety in the composite hypothesis, the preceding specified values of type I error rate, power, and $\mu_{Y1}-\mu_{Y2}-\delta_Y$ and σ_Y requires a total sample size N_Y =275.

For many scenarios in Table 3, the total sample size N for test of the composite hypothesis is much larger than the sample size for test of non-inferiority in efficacy (N_X). However, it happens in some cases that they are the same or their difference is quite small. Actually, N is associated with the sample sizes for individual test of non-inferiority in efficacy (N_X) and of superiority in safety (N_Y), and the correlation coefficient (ρ) between X and Y. When large difference exists between N_X and N_Y, N is quite close to the larger of N_X and N_Y and has little change along with change in ρ. In this numerical study, for N_X = 69 and 39 (<< 275), N is mostly equal to 275; for N_X = 1392 and 619 (>>275), the difference between N and N_X is 0 or negligible compared with the size of N. At the preceding four scenarios, change in correlation coefficient between X and Y has little impact on N. On the hand, the larger of N_X and N_Y is not always close to N, especially when N_X and N_Y close to each other. For example in Table 3, when both N_X is equal to 275 (=N_Y), N is 352 for ρ =0.5, and 373 for ρ =0. In addition, the results in Table 3 suggest that the correlation coefficient between X and Y is unlikely to have great influence on N, especially when the difference between N_X and N_Y is quite substantial. The above findings consistent with the underlying 'rule': when the two sample sizes are substantially different, taking N as the larger of N_X and N_Y will ensure the powers of two individual tests for efficacy and safety is essentially 1 and $1-\beta$, 'resulting' in a power of $1-\beta$ for test of the composite hypotheses; when N_X and N_Y is close to each other, taking N as the larger of N_X and N_Y will power the test of composite hypotheses at about $(1-\beta)^2$. Therefore, a significant increment in N is required

for achieving a power of $1-\beta$.

Fixed Sample Size Approach

Based on the sample size in Table 3, power of the test of composite hypothesis H_0 were calculated with results presented in Table 3, where P is the power of test of composite hypothesis with N_X in Table 3. P_M is the power of the same test with max (N_X, 275). With sample size N_X, the power of test of composite hypothesis is always not greater than the target value 80% as N_X is always not larger than N in Table 4. In some cases that σ_X =1.5 > σ_Y =1.0, N_X =N. Hence the corresponding P=80%. However, P is less than 60% for many cases in our numerical study. The worst scenario is P = 4.3% when N_X=39 for σ_X =0.5, ρ =-1 ad Δ =0.4.Therefore, test of composite hypothesis of both efficacy and safety using sample size N_X which is for achieving a certain power in testing hypothesis of efficacy only, may not have enough power to reject the null hypothesis. Interestingly, testing the composite hypothesis with max(N_X, 275), the power P_M is close to the target value 80% in most scenarios. Some exceptions happen when N_X is close to 275 (corresponding to (Δ =0.3, σ_X =1.0), and (Δ =0.4, σ_X =1.5)) such that a significant increment in sample size from max(N_X, 275) to N is required. This suggest taking N as the larger of the two sample sizes N_X and N_Y for testing hypothesis of individual endpoint when one of the two is much larger, say, one-fold larger than the other (Table 4).

Table 1: Significant Withdrawals of Drug Products between 2000-2010.

Drug name	Withdrawn	Remarks
Troglitazone (Rezulin)	2000	Withdrawn because of risk of hepatotoxicity; superseded by pioglitazone and rosiglitazone
Alosetron (Lotronex)	2000	Withdrawn because of risk of fatal complications of constipation; reintroduced 2002 on a restricted basis
Cisapride (Propulsid)	2000s	Withdrawn in many countries because of risk of cardiac arrhythmias
Amineptine (Survector)	2000	Withdrawn because of hepatotoxicity, dermatological side effects, and abuse potential.
Phenylpropanolamine (Propagest, Dexatrim)	2000	Withdrawn because of risk of stroke in women under 50 years of age when taken at high doses (75mg twice daily) for weight loss.
Trovafloxacin (Trovan)	2001	Withdrawn because of risk of liver failure
cerivastatin (Baycol, Lipobay)	2001	Withdrawn because of risk of rhabdomyolysis
Rapacuronium (Raplon)	2001	Withdrawn in many countries because of risk of fatal bronchospasm
Rofecoxib (Vioxx)	2004	Withdrawn because of risk of myocardial infarction
mixed amphetamine salts (Adderall XR)	2005	Withdrawn in Canada because of risk of stroke. See Health Canada press release. The ban was later lifted because the death rate among those taking Adderall XR was determined to be no greater than those not taking Adderall.
hydromorphone extended-release (Palladone)	2005	Withdrawn because of a high risk of accidental overdose when administered with alcohol
Pemoline (Cylert)	2005	Withdrawn from U.S. market because of hepatotoxicity
Natalizumab (Tysabri)	2005-2006	Voluntarily withdrawn from U.S. market because of risk of Progressive multifocal leukoencephalopathy (PML). Returned to market July, 2006.
Ximelagatran (Exanta)	2006	Withdrawn because of risk of hepatotoxicity (liver damage).
Pergolide (Permax)	2007	Voluntarily withdrawn in the U.S. because of the risk of heart valve damage. Still available elsewhere.
Tegaserod (Zelnorm)	2007	Withdrawn because of imbalance of cardiovascular ischemic events, including heart attack and stroke. Was available through a restricted access program until April 2008.
Aprotinin (Trasylol)	2007	Withdrawn because of increased risk of complications or death; permanently withdrawn in 2008 except for research use
Lumiracoxib	2007-2008	Progressively withdrawn around the world because of serious side effects, mainly liver damage
Rimonabant (Accomplia)	2008	Withdrawn around the world because of risk of severe depression and suicide
Efalizumab (Raptiva)	2009	Withdrawn because of increased risk of progressive multifocal leukoencephalopathy; to be completely withdrawn from market by June 2009
Sibutramine (Reductil)	2010	Withdrawn in Europe because of increased cardiovascular risk

Table 2: Composite Hypotheses for Clinical Evaluation.

	Safety		
Efficacy	**N**	**S**	**E**
N	NN	NS	NE
S	SN	SS	SE
E	EN	ES	EE

Note: N=Non-inferiority; S=Superiority; E=Equivalence

Table 3: Comparison of Sample Size between Tests for Composite Hypothesis and Single Hypothesis.

	Δ = 0.2			Δ = 0.3			Δ = 0.4		
σ_x ρ	N_X	N	N/N_X	N_X	N	N/N_X	N_X	N	N/N_X
0.5 -1.0	155	304	1.96	69	276	4	39	275	7.05
-0.5	155	303	1.95	69	276	4	39	275	7.05
0	155	300	1.94	69	276	4	39	275	7.05
0.5	155	289	1.86	69	275	3.99	39	275	7.05
1	155	275	1.77	69	275	3.99	39	275	7.05
1.0 -1.0	619	647	1.05	275	381	1.39	155	304	1.96
-0.5	619	646	1.04	275	381	1.39	155	303	1.95
0	619	642	1.04	275	373	1.36	155	300	1.94
0.5	619	629	1.02	275	352	1.28	155	289	1.86
1	619	619	1	275	275	1	155	275	1.77
1.5 -1.0	1392	1392	1	619	647	1.05	348	433	1.24
-0.5	1392	1392	1	619	646	1.04	348	432	1.24
0	1392	1392	1	619	642	1.04	348	424	1.22
0.5	1392	1392	1	619	629	1.02	348	402	1.16
1	1392	1392	1	619	619	1	348	348	1

Table 4: Power (%) of Test of Composite Hypothesis.

	Δ = 0.2		Δ = 0.2		Δ = 0.4	
σ_x ρ	**P**	$\mathbf{P_m}$	**P**	$\mathbf{P_m}$	**P**	$\mathbf{P_m}$
0.5 -1.0	38.9	75.3	14.7	80	4.3	80
-0.5	41.9	75.4	22	80	14.2	80
0	47.1	76.2	27.7	80	19.2	80
0.5	52.9	78.1	32.3	80	22.8	80
1	58.8	80	34.5	80	23.9	80
1.0 -1.0	78.2	78.2	60.1	60.1	38.9	75.3
-0.5	78.2	78.2	60.9	60.9	41.9	75.4
0	78.6	78.6	64	64	47.1	76.2
0.5	79.4	79.4	68.8	68.8	52.9	78.1
1	80	80	80	80	58.8	80
1.5 -1.0	80	80	78.2	78.2	67.6	67.6
-0.5	80	80	78.2	78.2	68	68
0	80	80	78.6	78.6	70.1	70.1
0.5	80	80	79.4	79.4	73.7	73.7
1	80	80	80	80	80	80

Concluding Remarks

In clinical evaluation of a test treatment in randomized clinical trials, the traditional (conditional) approach of testing single hypothesis for efficacy alone is not efficient because the observed safety profile could be by chance alone and may not be reproducible. Thus, testing a composite hypothesis which takes both safety and efficacy into consideration (e.g., testing non-inferiority in efficacy and testing superiority of safety as compared to a control) is recommended. In practice, sample size is expected to increase when switching from a single hypothesis testing (the traditional approach) to a composite hypothesis testing for clinical evaluation of a test treatment under investigation.

For illustration purpose, in this article, we assume that both efficacy and safety data (X, Y) are continuous variables which follow a bi-variate normal distribution. Statistical tests were derived under the framework of bi-variate normal distribution. In practice, efficacy and safety data (X, Y) could be either a continuous variable, a binary response, or time-to-event data. Similar idea can be applied (i) to derive appropriate statistical test under the null hypothesis and (ii) to determine the impact on power calculation for sample size requirement when switching from testing a single hypothesis (for efficacy) to testing a composite hypothesis for both safety and efficacy [5,6]. It, however, should be noted that closed forms and/or formulas for sample size calculation and the relationships between the single hypothesis and the composite hypothesis may not exist. In this case, clinical trial simulation may be useful.

References

1. Chow SC, Shao J (2002) Statistics in Drug Research – Methodologies and Recent Development. Marcel Dekker Inc, New York, USA.
2. FDA (1988). Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications. U.S. Food and Drug Administration, Rockville MD.
3. Wikipedia (2010). List of withdrawn drugs.
4. Chow SC, Shao J, Wang H, Lokhnygina Y (2017) Sample Size Calculation in Clinical Research. 3rd Edition, Taylor & Francis, New York, USA.
5. Chow SC, Huang Z (2019) Innovative thinking on endpoint selection in clinical trials. Journal of Biopharmaceutical Statistics 29(5): 941-951.
6. Chow SC, Liu, JP (2013) Design and Analysis of Clinical Trials – Revised and Expanded, 3rd Edition, John Wiley & Sons, New York, USA.