Published in final edited form as:

Contemp Clin Trials. 2014 November ; 39(2): 294–302. doi:10.1016/j.cct.2014.09.005.

Formulating Appropriate Statistical Hypotheses for Treatment Comparison in Clinical Trial Design and Analysis

Peng Huang¹, Ou Ai-hua², Steven Piantadosi³, and Ming Tan⁴

¹Johns Hopkins University

²Guangdong Provincial Hospital of Traditional Chinese Medicine

³Cedars Sinai Medical Center

⁴Georgetown University

Summary

Ideally designed null and alternative hypotheses should correspond to a partition of all possible true probability models such that the alternative hypothesis can be inferred upon the rejection of null hypothesis. However, tests are often carried out and recommendations are made without a precise specification of alternative hypothesis. This can lead to different superiority claims depending on which test is used instead of scientific plausibility. Such differences are often overlooked. We provide relationship of different superiority specifications in typical hypothesis testing. This can help investigators to select proper hypotheses for treatment comparison in clinical trial design.

Keywords

efficacy; clinical translational research; sample size and power; p-values; global treatment effect; treatment superiority

1. INTRODUCTION

Clinical trials are often designed specifically to evaluate whether a new treatment is superior to a standard. For a chosen endpoint, the establishment of treatment superiority in a clinical trial depends on how the hypotheses are selected. If we use $\mathcal{P} = \{P_\omega; \omega \in \Omega\}$ to denote all possible claims of underlying true probability models P_ω to quantify the difference between two treatments on the chosen endpoint, the null hypothesis $H_0 = \{P_\omega; \omega \in \Omega_0\}$ typically asserts that new treatment is no better than or equivalent to the standard treatment. The sample size is chosen so that the statistical test has adequate power to show a clinically meaningful improvement as claimed by the alternative hypothesis $H_a = \{P_\omega; \omega \in \Omega_a\}$. An

© 2014 Elsevier Inc. All rights reserved.

Corresponding author: Peng Huang, phuang12@jhmi.edu, Tel: 410-502-0944, Fax: 410-955-0859.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ideally designed hypothesis should constitute a partition of $\mathcal{P}$: $\Omega_0 \cap \Omega_a$ is an empty set, and $H_0 \cup H_a = \mathcal{P}$ (or equivalently $\Omega_0 \cup \Omega_a = \Omega$). This condition ensures that the alternative hypothesis H_a can be inferred upon the rejection of null hypothesis H_0 . However, many applications have $H_0 \cup H_a$ strictly smaller than $\mathcal{P}$ or even do not tell what H_a is. A typical example is when a null hypothesis of no treatment difference is used. Tests are carried out or recommendations are made without specifying the intended alternative hypothesis ([6], [7], [8]). Chi-square tests on contingency tables for ordinal endpoints are frequently used to detect positive associations, although what conclusion could be made regarding treatment preference upon rejection of the null hypothesis may not be clear. Furthermore, what investigators can assert in the one-sided alternative hypothesis is not precisely defined ([4]), although the need to do so has been well recognized by many authors (e.g., [1], [2], [3], [4]).

Different hypothesis pairs (H_0, H_a) can be used to test treatment superiority. Each pair corresponds to a partition of $\mathcal{P}$. The consequence of such multiplicity is often overlooked. Statistical tests are generally chosen empirically to compare treatments without examining whether the probability models claimed by H_a are appropriate measure of superiority for the given endpoint. Several mathematical methods to define superiority have been proposed in the literature. For ordinal outcomes, Grove ([1]) gave a variety of such definitions relevant to two-way contingency tables. Agresti and Coull ([3]) surveyed order-restricted methods and proposed 4 different definitions of “superiority” using odds ratios in contingency tables. Douglas et al. ([2]) proposed a number of definitions of positive dependence that include notions proposed by Grove. Cohen and Sackrowitz ([4]) proposed various definitions of superiority and complete class theorems for each superiority definition. Fay and Proschan ([5]) proposed 15 pairs of (H_0, H_a) for treatment comparison and gave an extensive evaluation on how these pairs are related. It is noted that only 3 out of 15 pairs (A_4 , A_{10} , and A_{15}) proposed by Fay and Proschan satisfy the condition $H_0 \cup H_a = \mathcal{P}$, and these 3 pairs are not designed for testing superiority.

On the other hand, scientists may not be clear which probability models are most appropriate to reflect the biological properties they discovered in the preclinical studies. They are likely to be less knowledgeable about the implications of various pairs of hypotheses (H_0, H_a) and how to select among multiple candidate hypotheses. The purpose of this article is to provide a theoretical and practical framework for evaluating the statistical properties of different superiority claims in typical hypothesis test setting. We consider only the case of $H_0 \cup H_a = \mathcal{P}$ where $\mathcal{P}$ is the collection of all possible true models as defined before. Since $\mathcal{P}$ is fixed, comparing pairs (H_0, H_a) is equivalent to comparing alternative hypothesis H_a . This is what we will discuss in the rest of the paper. We show theoretically the interrelationship of several commonly used superiority claims for alternative hypotheses and their relative strength in representing superiority. This will help investigators to understand how these hypothesis pairs are related, which one is the most appropriate statistical formulation to evaluate treatment improvement in clinical trial design.

2. STATISTICAL FRAMEWORK

In preclinical studies, a biological model may suggest that a new treatment could yield a “better” outcome than the other one. However, when translating this type of finding into

clinical studies, little attention is given to the most appropriate and robust way to quantify superiority for a new discovery. Superiority is often defined as a shift of population means. When the endpoint is an event time, the improvement claim is often defined by a hazard ratio or difference in median survival time primarily for statistical convenience instead of scientific plausibility. Some claims of superiority are made through stochastic ordering such as proportional odds order, hazard rate order, or likelihood function order, etc. Although various models can be used to claim superiority, not all yield the same strength of superiority. That is to say, they correspond to different alternative hypotheses $H_a = \{P_\omega: \omega \in \Omega_a\}$. If one claim is more stringent than another, then the two claims are nested subsets in $\mathcal{P}$. For example, let $H_{a1} = \{P_\omega: \omega \in \Omega_{a1}\}$ be a collection of all probability models that specify proportional hazards rate $h_1(t)/h_2(t) < 1$ for a new treatment T_1 as compared to a control treatment T_2 at any time t , and let $H_{a2} = \{P_\omega: \omega \in \Omega_{a2}\}$ be a collection of all probability models for which T_1 has a longer median survival time than T_2 . Then, we have $\Omega_{a1} \subseteq \Omega_{a2}$ (or equivalently, $H_{a1} \subseteq H_{a2}$). This relation shows that, a claim of treatment superiority defined by a hazard ratio implies the claim of treatment superiority defined by median survival time. The converse is not necessarily true. In other words, a superiority claim in terms of hazard ratios is stronger than a superiority claim in terms of median survival time.

As a real data example, we have collected survival times from 207 patients who all died of prostate cancer ([9], see Figure 1). The Kaplan-Meier estimates of overall survival suggest that, for most part of time interval, group 1 has longer survival than group 2. The two curves cross simply because 3 patients in group 2 had much longer survival time than anyone else. However, the logrank test yields p -value = 0.13. The $R^2 = 0.028$ from the Cox model shows a poor fit under the proportional hazards assumption. On the other hand, Wilcoxon test and t -test of log survival time give p -values of 0.007 and 0.01 respectively, suggesting group 1 has significantly longer overall survival time than group 2. If the null hypothesis is rejected by the logrank test, the alternative hypothesis $H_{a1} = \{P: h_1(t)/h_2(t) = HR < 1, \forall t > 0\}$ is implicit. If the Wilcoxon test is used, the alternative hypothesis $H_{a2} = \{P: P\{X < Y\} - P\{X > Y\} > 0\}$ is implicit. Here X is the survival time of group 1, and Y is the survival time of group 2. Interpretation of $\theta = P\{X < Y\} - P\{X > Y\}$ is given in the next section. The logrank test assumes the hazard ratio does not change over time. This is a much stronger claim than $\theta > 0$ that leads to $H_{a1} \subseteq H_{a2}$. This example shows that a survival benefit from group 1 can be claimed through a weaker claim “ $\theta > 0$ ” but not through a stronger claim of hazard ratio “ $HR < 1$ ”.

If claim A is stronger than claim B (i.e., claim A implies claim B), their corresponding alternative hypotheses H_{aA} and H_{aB} in $\mathcal{P}$ are nested: $H_{aA} \subseteq H_{aB}$. In practice, the inherent biological property of a new treatment only allows one to demonstrate improvement through certain claims in H_{aB} but not all its nested subsets H_{aA} corresponding to stronger claims. How to choose a subset in $\mathcal{P}$ for the claim of superiority is encountered when formulating hypotheses in study design. The null hypothesis often specifies that a new treatment is no better than the control or they are equivalent. A study is typically powered to reject the null hypothesis and make a claim described under the alternative hypothesis. If the claim in an alternative hypothesis is too strong, the study (as well as future ones) is more likely to be negative. If the claim in the alternative hypothesis is too weak, it fails to provide the most

informative conclusion about the treatment effect. This failure is fundamentally different than the lack of power – it is not the same as statistical efficiency. The former refers to the selection of a subset H_a in $\mathcal{P}$ to define superiority while the latter refers to the sample size needed to claim the pre-defined subset H_a in $\mathcal{P}$.

Not all claims of superiority are nested. In such cases, careful selection of the alternative hypothesis is necessary to support a robust study conclusion. For example, if both random variables X and Y are survival times having Weibull density $f(x) = \frac{1}{2\sqrt{x}}e^{-\sqrt{x}}$ and $g(y) = e^{-y}$ respectively, then Y has longer median survival time than X but with shorter mean survival time. If the alternative hypothesis states that Y has longer mean survival time than X , it is less likely to be supported by the data. Increasing sample size does not help to increase power to launch such a claim from the alternative hypothesis. In fact, a larger sample size is more likely to show no increase in population mean. Even if the null hypothesis is rejected by the current data with $p < 0.05$ and a one-sided alternative is claimed (i.e., a type I error is encountered), this finding is unlikely to reproduce in future or in larger studies.

Although different claims can be used to define superiority, it would be helpful for investigators to know how they are related when formulating hypotheses for their studies. This paper presents relationships among 5 commonly used claims to define superiority through population means, medians, stochastic ordering, and relative treatment effect. We provide the relative strength of each claim, conditions needed for one claim to imply another one, and conditions for two claims to be equivalent. Examples to illustrate these relationships are presented. We provide suggestions how to select a claim for the alternative hypothesis to reflect the underlying biological properties and reach a robust conclusion.

The rest of the paper is arranged as follows. Section 3 gives precise statements of 5 commonly used claims to define superiority in alternative hypotheses and how they arise in practice. Section 4 gives theoretical relationships and examples of the 5 claims. Proofs of all theorems are provided in the Appendix. We conclude the paper with some recommendations.

3. FIVE CLAIMS TO DEFINE SUPERIORITY IN ALTERNATIVE HYPOTHESES

For any chosen endpoint, we first need to define a complete collection of probability models $\mathcal{P} = \{P_\omega; \omega \in \Omega\}$ that includes all possible scenarios of true difference between the two treatments. For example, if we are comparing mean tumor size reduction between treatments T_X and T_Y , $\mathcal{P}$ should consist of all possible situations about the true mean tumor size difference between T_X and T_Y . Any claim of such difference is a subset of $\mathcal{P}$. If claim A says that “mean tumor size reduction μ_Y in treatment T_Y is 20% larger than the mean tumor size reduction μ_X in treatment T_X ”, then claim A is a set in $\mathcal{P}$ that includes all subsets for which $\mu_Y - \mu_X > 20\%$ is true. Obviously, this claim A is stronger than the claim B which says “ $\mu_Y - \mu_X > 10\%$ ”. In this case, we say that claim A is nested within claim B , denoted as $A \subset B$. For any two nested claims A and B , it is harder to prove a stronger claim A than a weaker claim B . This is because a stronger claim automatically implies a weaker claim, but not vice versa.

Without a loss of generality, we assume that larger observations correspond to a better clinical outcomes. If we use X and Y to denote outcomes from treatments T_X and T_Y respectively, we say that Y is stochastically larger than X (denoted as $Y \succ X$) if $P\{Y > t\} \geq P\{X > t\}$ for all t , and there exists at least one t_0 such that $P\{Y > t_0\} > P\{X > t_0\}$. A simple way to interpret stochastic ordering is that, for any cutoff point t , treatment T_Y is more likely to yield observations greater than t than treatment T_X is. The 5 claims to be compared in this paper are:

- (c1) Y is stochastically larger than X , denoted as $Y \succ X$;
- (c2) the mean of treatment T_Y is greater than the mean of treatment T_X , denoted as $\mu_Y > \mu_X$;
- (C3) the median of treatment difference ($Y - X$) is greater than zero, denoted as $med_{(Y - X)} > 0$;
- (C4) $\theta = P\{X < Y\} - P\{X > Y\} > 0$;
- (C5) the median of treatment T_Y is greater than the median of treatment T_X , denoted as $med_Y > med_X$.

The θ defined in (c4) is a value between -1 and $+1$ that measures how likely that Y gives a greater value than X . When $\theta = -1$, Y is 0% likely to be greater than X ; When $\theta = +1$, Y is 100% likely to be greater than X . When $\theta > 0$, Y has larger probability to be greater than X , and when $\theta < 0$, Y has smaller probability to be greater than X .

Parameter θ was called the “relative treatment effect” by Brunner, Munzel, and Puri ([10]), and was further extended by Huang et al. ([11], [12], [13]) to define “global treatment effect” when comparing multiple outcomes. The Wilcoxon-Mann-Whitney rank test is sometimes referred to as a test of medians. It is better viewed as a test for relative treatment effect θ . Examples of using claim (c1) to define better include: the hazards ratio of treatment T_Y versus treatment T_X is less than 1; the odds ratio of T_Y versus T_X is greater than 1; and positive treatment effect in a number of regression models where treatment benefit is defined through a location shift: $P\{Y \geq t\} = P\{X \geq t - \delta\}$ for some $\delta > 0$. When outcome measures are non-negative, such as tumor size, a scale shift model $P\{Y \geq t\} = P\{X \geq t/\beta\}$ for some $\beta > 1$ is another example of stochastic ordering in (c1). When random variables X and Y are assumed to have distributions in the same monotone likelihood ratio family, such as when a likelihood ratio test is used, the stochastic ordering assumption is implicitly made. Claim (c2) is most frequently used in clinical trials to define improvement. Claim (c3) is sometimes used when differences between matched pairs are of interest. Examples include studies on how tumor size changes before and after neoadjuvant therapy from the same patient, difference between treated and untreated eyes from the same subject, and similarly in matched case-control studies. Claim (c4) is often used when performing nonparametric tests such as the Wilcoxon rank test. If data have a continuous distribution, claim $\theta > 0$ is equivalent to the claim of $p = P\{Y > X\} > 1/2$ because $\theta = 2p - 1$. Claim (c4) is often chosen when investigators have little knowledge about outcome distributions. Claim (c5) is often used when data are highly skewed or when quantiles are of interest. For example,

comparison of median survival between two treatments is performed in cancer studies. (c5) is also used in sign test and quantile regression.

4. RELATIONSHIP AMONG CLAIMS (c1) – (c5)

We investigate claims (c1) – (c5) by showing whether one is nested within another, and whether and when two claims are equivalent. Let X and Y be two random variables denoting the outcomes from treatments T_X and T_Y respectively. Their corresponding cumulative distribution functions are $F(x)$ and $G(y)$. We use Serfling's ([14]) method to define the u th quantile $q(u)$ of a cumulative distribution F :

$$Q_F(u) = \inf\{x: F(x) \geq u\}. \quad (1)$$

The median is defined as the 0.5th quantile. Notations med_X , med_Y and med_{Y-X} are used to denote the medians of X , Y and $(Y-X)$ in claims (c3) and (c5) respectively. Theorem 1 below shows that claims (c1) and (c2) are always nested, claims (c1) and (c4) are nested if two treatments do not give the same result, and claims (c1) is stronger than claims (c2), (c3) and (c4) for data with continuous distributions.

Theorem 1

1. Claim (c1) is always stronger than claim (c2);
2. Claim (c1) is stronger than claims (c3) and (c4) if the probability that the two treatments give the same outcome is zero;
3. Claim (c1) is stronger than all other 4 claims (c2 through c5) if data from both treatment groups have continuous distributions and $P\{Y > t\} > P\{X > t\}$ for all t .

When both X and Y are continuous random variables, the condition $P(X = Y) = 0$ is naturally satisfied. Thus, claim (c1) is stronger than claim (c4) for all continuous endpoints. The reverse is generally not true: claim (c2) or (c4) do not necessarily imply claim (c1). The prostate cancer data described in the introduction demonstrates this. Under the same $\alpha = 0.05$ significance level (type I error), the stochastic ordering defined by a hazard ratio cannot be claimed by the logrank test even though both (c2) and (c4) can be claimed by the t -test and Wilcoxon test respectively. Clearly, claim $(c1^*) = P\{Y > t\} > P\{X > t\}$ for all t is stronger than claim (c1) [i.e., $(c1^*)$ is a subset of (c1)]. Claim $(c1^*)$ is frequently seen in location shift linear regression models for continuous response variables where the binary indicator of treatment assignment is included as a linear covariate. In fact, most examples of (c1) given in previous sections actually belong to $(c1^*)$. To see this, examine the location shift model $P\{Y > t\} = P\{X > t - \delta\}$ for two random variables X and Y with normal distributions. Apparently, $(c1^*)$ is true when $\delta > 0$. An example of (c1) that does not belong to $(c1^*)$ is when the two distribution function curves meet. For example, when event times from all patients are observed in a finite time interval $[0, T]$. Then the survival functions (or distribution functions) between the two treatment groups are the same in time interval (T, ∞) .

Claim (c4) is typically concluded following a nonparametric test while claim (c2) is concluded more often after a parametric test or a semiparametric test. It is natural to guess that they are not related. In general, neither is stronger than the other one. There are many examples when (c2) is true but (c4) is not, and vice versa. For example, if outcome measures from both treatment groups are ordinal with the following distributions then $\mu_Y - \mu_X = 0.1 > 0$ but $\theta = -0.02 < 0$. If we switch the distributions of X and Y , we will get $\mu_Y - \mu_X < 0$ but $\theta > 0$. However, following theorem shows that (c2) and (c4) are equivalent when the distributions of endpoints in both groups are from Koopman-Darmois family.

Theorem 2

Suppose observations from two treatments are independent to each other. Claims (c2) and (c4) are equivalent if data from both treatment groups have distributions from the Koopman-Darmois family, i.e., their probability distribution functions have the form $p(t) = \exp[\eta_i t - b(\eta_i) - a(t)]$ for treatment group $i = 1, 2$ respectively for some convex function $b(\cdot)$ and continuous function $a(\cdot)$.

Distributions in the Koopman-Darmois family include many outcome measures in practice. It includes: the normal distribution with common variance, chi-square distribution, exponential distribution, gamma distribution with common shape parameter, and Weibull distribution with common shape parameter. For discrete outcomes, it includes: the binomial distribution, geometric distribution, and Poisson distribution. Although (c2) and (c4) are equivalent in all these outcomes, claim (c4) imposes weaker assumptions than (c2) because the mean value for some distributions may not be finite or even not exist. Also, (c4) is invariant to any monotone transformation of the data while (c2) is not. This is because θ depends only on the relative rank order of the two random variables, not the actual values. For the example in Table 1, we have $\mu_Y - \mu_X = E[Y - X] = 0.1 > 0$. However, if we use a square root transformation, we have $E[\sqrt{Y} - \sqrt{X}] = -0.01712 < 0$. Before and after the square root transformation, $\theta = -0.02$. Note that, claim (c1) is never true in both situations. In fact, there is no stochastic ordering between the two treatment groups no matter how we transform the data. This shows that the sign of θ in claim (c4) does not tell whether claim (c1) is true. That is to say, (c1) may not be true even when (c4) is true.

Like claim (c4), claim (c3) is also invariant to any monotone transformation of the data. Following Theorem 3 shows that claim (c3) is also stronger than (c4), and they are the same if there are no ties in outcome measures between the two treatment groups.

Theorem 3

If claim (c3) is true, then claim (c4) must be true. On the other hand, if the probability that the two treatments give the same outcome is zero, then both (c3) and (c4) are equivalent. In general, (c4) implies only $\text{med}_{(Y-X)} = 0$ but not $\text{med}_{(Y-X)} > 0$ in (c3).

Based on Theorem 3, claims (c3) and (c4) are equivalent when the endpoint from both treatment groups follows continuous distributions. However, they are not equivalent when the endpoint is not continuous. As a simple counter example, let $T = Y - X$ have mass function $P(T = -1) = 1/10$, $P(T = 0) = 3/5$, $P(T = 1) = 3/10$, then its median $m = 0$. However,

$\theta = P(T = 1) - P(T = -1) = 1/5 > 0$. This means that (c4) is strictly weaker than (c3), i.e., (c3) is nested within (c4). Since both (c3) and (c5) are claims about medians, one may guess that claims (c3) and (c5) are the same. This is not true in general. For example, if $P\{X = 1\} = P\{Y = 2\} = 0.6$ and $P\{X = 3\} = P\{Y = 0\} = 0.4$, then (c5) is true but (c3) is not. If we switch X and Y 's distributions, then (c3) is true but (c5) is not. However, the following theorem shows (c3) and (c5) are equivalent when outcomes from both treatment groups follow some symmetric distributions.

Theorem 4

If outcomes from both treatment groups follow symmetric distributions, then

1. claims (c2), (c3), and (c5) are equivalent,
2. claim (c4) is at least as strong as claims (c2), (c3), and (c5),
3. If (c2) is true and condition $P\{0 \leq Y - X < 2(\text{med}_Y - \text{med}_X)\} > 0$ is satisfied, then claims (c3), (c4), and (c5) are all true.

Based on theorem 4(2), claim (c4) is generally stronger than claims (c3), (c3) and (c5) when the endpoints have symmetric distributions. This does not exclude the case when (c4) is equivalent to them. For example, when paired endpoints follow normal distributions, then (c3) and (c4) are equivalent based on Theorem 3 and Theorem 4(2).

5. HOW TO SELECT A CLAIM WHEN CONSTRUCTING HYPOTHESES?

An appropriate claim for the alternative hypothesis is essential before data analysis. It can also help determine the best way to collect data. When formulating hypotheses for comparing treatments, it is tempting to think first of convenient statistical software rather than the best way to quantify biological properties. Different statistical models or tests are developed under different assumptions, and their claims correspond to different subsets in $\mathcal{P}$. Failure to examine assumptions can prevent some conclusions or sacrifice reproducibility. In translational research when little information is available about treatment effects, making overly strong assumptions or testing strong claims is not optimal, because it can contribute to abandoning a promising treatment. The stochastic ordering claim (c1) is the strongest and most likely to be violated. If the biological principle of the treatment suggests that treatment T_Y could give better clinical outcomes than T_X no matter how we dichotomize the data, then it is relatively safe to formulate the hypothesis using claim (c1). If the treatment is likely to be effective for patients with certain positive biomarkers, the use of claim (c1) could be too strong if the study is going to enroll subjects with both positive and negative biomarkers. When study subjects are heterogeneous, the stochastic ordering assumes treatment effect is the same across different subgroups. If we are not confident in such an assumption, it is better to begin with more homogeneous study subjects. For example, we can start with a simple trial that enrolls only subjects with positive biomarkers. If treatment efficacy in this subtype is established, further studies can test if the treatment effect extends to those with negative biomarkers. If it is difficult to start with homogeneous study subjects due to cost, a weaker claim should be used when constructing the hypothesis. This strategy gives investigators a greater chance to succeed and avoids abandoning some promising treatments prematurely.

When two claims are equivalent, we suggest choosing the test with higher statistical power. For example, when claim (c2) is equivalent to claim (c4), then we can choose a test that has higher power for either (c2) or (c4) depending on the shape of the distribution. We can choose a t -test when the shape is not too skewed and the Wilcoxon test when the shape is skewed.

When $H_0 \cup H_a = \mathcal{P}$ is not satisfied for the chosen hypotheses, one should check whether or not the alternative hypothesis can be automatically inferred upon the rejection of the null. For example, the rejection of equal hazard functions does not necessarily imply that the ratio is constant, because the two distribution functions may not be stochastically ordered. It is possible that the null hypothesis of equal hazard functions is too restrictive and easily rejected. Furthermore, failure to reject a strong null hypothesis does not necessarily imply that a meaningful conclusion cannot be made should the null hypothesis be formed properly. This point is demonstrated from the cognitive behavioral therapy example in figure 4 of Acion et al. ([15]). The study was to compare the efficacy of a cognitive behavioral therapy (CBT) for medication resistant symptoms with a standard treatment (STD). One of the outcome variables was the change in global functioning score after 6 months of treatment. Higher score indicates better outcome. All 8 patients in the CBT group had positive changes while there was only one positive change among 9 patients in the STD group. The mean difference between the two groups is 0.83 times its standard deviation. The p -value from the t -test was 0.12, failing to reject the null hypothesis of equal means. However, the p -value from a Wilcoxon test was 0.01, indicating strong evidence that relative treatment effect $\theta = P(Y > X) - P(X > Y)$ is not zero.

6. CONCLUSION

We have discussed the problem of properly defining treatment superiority in clinical trials and the associated difficulties in statistical inference, namely, the need to precisely define the notion of superiority in a one-sided hypothesis test. We provided a theoretical framework for evaluating the statistical properties of different specifications of superiority in typical hypothesis testing. We proved the order of strength of five typical claims for alternative hypotheses. A summary of the relations among these five claims is given in Figure 2. Some suggestions of how to select statistical parameters for treatment comparison were presented. This method can be applied to compare superiority for trials of different phases from early stage to comparative effectiveness. Because a two-sided hypothesis that two treatments are unequal can be viewed as a combination of two one-sided hypotheses, results presented in this paper would be useful for trials aimed to show inequality. Furthermore, it is worth noting that a Bayesian approach to compute the predictive probabilities $P(X < Y)$ in both the parametric and the nonparametric model would be a natural alternative to what is presented. We plan to report findings using that approach in a future paper.

Acknowledgments

We want to thank the 4 reviewers for their insightful comments and suggestions that greatly improved the presentation of this paper. This work is partially supported by research grants 1R21NS043569-01, P50CA103175, MCRF-FHA05CRF, 1P01AG023630-049002, and P30CA006973.

References

1. Grove DM. Positive association in a two way contingency table: likelihood ratio tests. *Communications in Statistics; Part A; Theory and Methods*. 1984; 13:931–945.
2. Douglas, R.; Fienberg, SE.; Lee, M-LT.; Sampson, AR.; Whitaker, LR. Positive dependence concepts for ordinal contingency tables. In: Block, HW.; Sampson, AR.; Savits, TH., editors. *Topics in Statistical Dependence*. 1991. p. 189-202. *IMS Lecture Notes – Monograph Series*
3. Agresti A, Coull BA. Order-restricted inference for monotone trend alternatives in contingency tables. *Computational Statistics & data analysis*. 1998; 28:139–155.
4. Cohen A, Sackrowitz HB. Testing whether treatment is “better” than control with ordered categorical data: Definitions and complete class theorems. *Statistics and Decisions*. 2000; 18(1):1–25.
5. Fay MP, Proschan MA. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*. 2010; 4:1–39. [PubMed: 20414472]
6. Moses LE, Emerson JD, Hosseini H. Analyzing data from ordered categories. *New England J of Medicine*. 1984; 311:442–448.
7. Emerson JD, Moses LE. A note on the Wilcoxon-Mann-Whitney test for 2k ordered tables. *Biometrics*. 1985; 41:303–309. [PubMed: 2408689]
8. Rahlfs VW, Zimmerman H. Scores: Ordinal data with few categories - how they should be analyzed. *Drug Information Journal*. 1993; 27:1227–1240.
9. Goyal J, Nuhn P, Huang P, Tyagi P, Oh D, Carducci MA, Eisenberger MA, Antonarakis E. The effect of clinical trial participation versus non-participation on overall survival in men receiving first-line docetaxel-containing chemotherapy for metastatic castration-resistant prostate cancer. *British Journal of Urology International*. 2012; 110:E575–582.
10. Brunner E, Munzel U, Puri ML. The multivariate nonparametric Behrens-Fisher problem. *Journal of Statistical Planning and Inference*. 2002; 108 (1–2):37–53.
11. Huang P, Tilley B, Woolson R, Lipsitz S. Adjusting O’Brien’s test to control type I error for the generalized nonparametric Behrens-Fisher problem. *Biometrics*. 2005; 61:532–539. [PubMed: 16011701]
12. Huang P, Woolson RF, O’Brien PC. A rank-based sample size method for multiple outcomes in clinical trials. *Statistics in Medicine*. 2008; 27(16):3084–3104. [PubMed: 18189338]
13. Huang P, Goetz CG, Woolson RF, Tilley B, Kerr D, Palesch Y, Elm J, Ravina B, Bergmann KJ, Kieburtz K. Parkinson Study Group. Using global statistical tests in long-term Parkinson’s disease clinical trials. *Movement Disorders*. 2009; 24(12):1732–1739. [PubMed: 19514076]
14. Serfling, RJ. *Approximation theorems of mathematical statistics*. John Wiley & Sons; New York: 1980.
15. Acion L, Peterson J, Temple S, Arndt S. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*. 2006; 25:591–602. [PubMed: 16143965]

Appendix I. Some Lemmas

Lemma 1

Let F be any cumulative distribution function, $Q_F(u)$ is defined by (1). For any real constant t and $0 \leq u \leq 1$, we have

$$F(t) < u \text{ if and only if } t < Q_F(u) \quad (2)$$

Proof

Since CDF F must be a right continuous function, we have $\{x : F(x) \leq u\}$ is a closed convex set. Also, $Q_F(u) = \inf\{x : F(x) \leq u\}$ is a limit point of this closed set, we have $\{x : F(x) \leq u\} = [Q_F(u), +\infty)$.

To prove (2), suppose $t < Q_F(u)$ holds. Then $t \notin [Q_F(u), +\infty) = \{x : F(x) \leq u\}$. This implies $F(t) < u$.

On the other hand, suppose $F(t) < u$ holds. Then $t \notin \{x : F(x) \leq u\} = [Q_F(u), +\infty)$. This implies $t < Q_F(u)$.

Lemma 2

If random variable X has cumulative distribution function F and quantile function defined by (1), U is a random variable with uniform distribution on closed interval $[0, 1]$, then $X_1 = Q_F(U)$ has the same distribution as X , denoted as $X_1 =_D X$.

Proof

Since

$$\begin{aligned} P\{X_1 > t\} &= P\{Q_F(U) > t\} \\ &= P\{U > F(t)\} \text{ (from (2))} \\ &= 1 - F(t) \end{aligned}$$

X_1 has the same cumulative distribution function as X . Thus $X_1 =_D X$.

Lemma 3

If two cumulative distribution functions F and G satisfy $F(t_0) > G(t_0)$ for some t_0 . Then there exists a closed interval (t_2, t_1) with $t_2 < t_1$ and a positive constant $\varepsilon > 0$ such that $F(t) - G(t) \geq \varepsilon$ for all t with $t_2 < t < t_1$.

Proof

Let $\delta = [F(t_0) - G(t_0)]/4$. Then $\delta > 0$. Two open intervals $A = (G(t_0) - \delta, G(t_0) + \delta)$ and $B = (F(t_0) - \delta, F(t_0) + \delta)$ do not overlap with $G(t_0) + \delta < F(t_0) - \delta$. Since both F and G are right continuous and monotone non-decreasing, there exists sequences $s_1 > s_2 > \dots > s_n > \dots$ with $G(s_n) \in A$, $F(s_n) \in B$ such that $\lim_{s_n \downarrow t_0} G(s_n) = G(t_0) < F(t_0) = \lim_{s_n \downarrow t_0} F(s_n)$. As open intervals A and B do not overlap and both F and G are monotone non-decreasing, we have

$$F(s_m) - G(s_n) \geq F(t_0) - [G(t_0) + \delta] = 3\delta > 0, \quad \forall m, n = 1, 2, \dots \quad (3)$$

Let $\varepsilon = \delta > 0$, t_1, t_2 be any two values in $s_1 > s_2 > \dots > s_n > \dots$, say, $t_1 = s_1$ and $t_2 = s_2$. Now, for any t with $t_2 < t < t_1$, we have $F(t) - G(t) \geq F(t_2) - G(t_1) = F(s_2) - G(s_1) \geq 3\delta > \varepsilon > 0$ according to (3).

Lemma 4

If $Y \succ X$, i.e., $P\{Y > t\} > P\{X > t\}$ for all t , and there exists at least one t_0 such that $P\{Y > t_0\} > P\{X > t_0\}$. Then there exist random variables X_1 and Y_1 such that

1. $X_1 =_D X$ and $Y_1 =_D Y$;
2. $X_1 < Y_1$, and $P\{X_1 < Y_1\} > 0$.

Proof

Let U be a random variable with uniform distribution on closed interval $[0, 1]$, $F(t) = P\{X \leq t\}$, $G(t) = P\{Y \leq t\}$, $X_1 = Q_F(U)$, and $Y_1 = Q_G(U)$. Results in (1) can be obtained from Lemma 2.

According to $Y \succ X$ assumption, $F(t) < G(t)$ for all t . This implies that $\{t : F(t) \leq u\} \supseteq \{t : G(t) \leq u\}$. Hence $X_1 = \inf\{t : F(t) \leq u\} \leq \inf\{t : G(t) \leq u\} = Y_1$. Now it remains to show that $P\{X_1 < Y_1\} > 0$. If this is not true, i.e., $P\{X_1 < Y_1\} = 0$, then for any t ,

$$\begin{aligned} 1 &= P\{X_1 \leq Y_1\} = P\{X_1 < Y_1\} + P\{X_1 = Y_1\} \\ &= P\{X_1 = Y_1\} \\ &= P\{X_1 \leq t, X_1 = Y_1\} + P\{X_1 > t, X_1 = Y_1\} \\ &= P\{X_1 \leq t, X_1 = Y_1\} \\ &= P\{Y_1 \leq t, X_1 = Y_1\} \\ &\leq P\{Y_1 \leq t\} \\ &= G(t) \quad \text{for any } t. \end{aligned}$$

This contradicts $F(t_0) < G(t_0)$ assumption. Thus, we must have $P\{X_1 < Y_1\} > 0$.

Lemma 5

Let G be the cumulative distribution function of some random variable Y . Then (i) $P\{G(Y) \leq u\} = u$ if u is in the range of G ; $P\{G(Y) \leq u\} < u$ if $u \in (0, 1)$ is not in the range of G ; (ii) $E[G(Y)] \leq 1/2$; (iii) Suppose $t_1, \dots, t_n$ are discontinuous points of G . Then $E[G(Y)] \geq 1 + \sum_{i=1}^n [G(t_i) - G(t_i-)]^2$.

Proof

- (i) We first prove that $P\{G(Y) \leq u\} = u$ for $\forall u \in (0, 1)$. For $\forall u \in (0, 1)$, define

$$G^{-1}(u) = \inf\{y : G(y) \geq u\}, \quad q(u) = \inf\{y : G(y) \geq u\}, \quad p(u) = G(G^{-1}(u)).$$

Then $p(u) \geq u$ (Serfling, 1980, page 3), $q(u) \leq G^{-1}(u)$, and $G(q(u)) \leq G(G^{-1}(u)) = p(u)$. Let's prove (i) under the following three cases: (ia). $p(u) > u$; (ib). $p(u) = u$ and $G(q(u)) > u$; (ic). $p(u) = u$ and $G(q(u)) = u$.

- (ia) If $G(G^{-1}(u)) = p(u) > u$, the definition of $G^{-1}(u)$ implies $G(G^{-1}(u) - \varepsilon) < u$ for $\forall \varepsilon > 0$; so $\{G(Y) = u\} = \varnothing$ (an empty set) and $G(G^{-1}(u) - 0) = u$. Now we have $P\{G(Y) = u\} = P\{G(Y) < u\} = P\{Y < G^{-1}(u)\}$ (Serfling, 1980, page 3) = $G(G^{-1}(u) - 0) = u$.
- (ib) If $p(u) = u$ and $G(q(u)) > u$. Definition of $q(u)$ implies $G(q(u) - 0) = u$. Since $0 = P\{G^{-1}(u) < Y < q(u)\} = G(q(u) - 0) - G(G^{-1}(u)) = u - u = 0$, i.e., $G(q(u) - 0) = G(G^{-1}(u)) = u$, we have $P\{G(Y) = u\} = P\{Y < q(u)\} = G(q(u) - 0) = u$.
- (ic) If $p(u) = u$ and $G(q(u)) = u$. We have $P\{G(Y) = u\} = P\{Y = q(u)\} = u$.

In summary, we have showed that $P\{G(Y) = u\} = u$ for $\forall u \in (0, 1)$. Suppose u is in the range of G : there exists y_0 such that $G(y_0) = u$. Then $P\{G(Y) = u\} = P\{G(Y) = G(y_0)\} = P\{Y = y_0\} = G(y_0) = u$. On the other hand, if u is not in the range of G , then there exists $\delta > 0$ such that $P\{G(Y) = u\} = P\{G(Y) = u - \delta\} = u - \delta < u$.

- (ii) From (i), $E[G(Y)] = 1 - \int_0^1 P\{G(Y) \leq u\} du \geq 1 - \int_0^1 u du = 1/2$.
- (iii) Denote $a_i = G(t_i^-)$; $b_i = G(t_0) > a_i$; $a_{n+1} = 1$; $b_0 = 0$ ($i = 1, \dots, n$). Then $P\{G(Y) = u\} = a_i$ for $\forall u \in (a_i, b_i]$, and $P\{G(Y) = u\} = u$ for $\forall u \in (b_i, a_{i+1}]$. Now

$$\begin{aligned} \int_0^1 P\{G(Y) \leq u\} du &= \sum_{i=0}^n \int_{b_i}^{a_{i+1}} P\{G(Y) \leq u\} du + \sum_{i=1}^n \int_{a_i}^{b_i} P\{G(Y) \leq u\} du \\ &\leq \sum_{i=0}^n \int_{b_i}^{a_{i+1}} u du + \sum_{i=1}^n \int_{a_i}^{b_i} a_i du \\ &= \sum_{i=0}^n \frac{1}{2} (a_{i+1}^2 - b_i^2) + \sum_{i=1}^n a_i (b_i - a_i) \\ &= \frac{1}{2} - \frac{1}{2} \sum_{i=1}^n (b_i - a_i)^2; \\ E[G(Y)] &= 1 - \int_0^1 P\{G(Y) \leq u\} du \\ &\geq \frac{1}{2} + \frac{1}{2} \sum_{i=1}^n (b_i - a_i)^2. \end{aligned}$$

Appendix II. Proof of Theorems

1. Proof of Theorem 1

- Let F and G be the cumulative distribution functions of X and Y respectively. Then $F(t) \geq G(t)$ and $G^{-1}[F(t)] \leq t$ for all t since $Y \succ X$. From Lemma 3, there exists some closed interval (t_2, t_1) with $t_2 < t_1$ and a positive constant $\varepsilon > 0$ such that $F(t) - G(t) \geq \varepsilon$ for all t with $t_2 < t < t_1$. As both F and G are monotone right continuous distributions, this implies that $G^{-1}[F(t)] > t$ for $t_2 < t < t_1$. Now we have $E[X] = \int t dF(t) < \int G^{-1}[F(t)] dF(t) = \int y dG(y) = E[Y]$ where transformation $y = G^{-1}[F(t)]$ is used in the integral. This shows that claim (c1) is stronger than claim (c2).
- To prove $\theta > 0$, first, we note that, for any cumulative distribution function G , its range is defined by a set $\{p: p = G(y)\}$. Since G is a monotone nondecreasing function, it has at most countable discontinuous points.

Let F and G be the cumulative distribution functions of random variables X and Y respectively. By definition of $Y \succ X$, there exists a finite point t_0 such that $F(t_0) > G(t_0)$.

Define $t_1 = \inf\{t > t_0: G(t_0) < G(t)\}$. Then $t_0 < t_1 < +\infty$, $G(t_1 - \delta) = G(t_0)$, and $G(t_1 + \delta) > G(t_0)$ for $\forall \delta > 0$. We first show $E[F(Y)] > 1/2$. If G has jump at t_1 . Lemma 5 (iii) implies $E[F(Y)] - E[G(Y)] > 1/2$.

If G is continuous at t_1 , then $G(t_0) - G(t_1) = \lim_{n \rightarrow \infty} G(t_1 - 1/n) - G(t_0)$, i.e., $G(t_1) = G(t_0) < F(t_0)$. Definition of t_1 and the right continuous of G imply that there exist constants $\delta_1 > 0$ and $\varepsilon_1 > 0$ such that $G(t_1 + \delta_1) < F(t_0) - \varepsilon_1$. For any $t \in (t_0, t_1 + \delta_1]$: $G(t) - G(t_1 + \delta_1) < F(t_0) - \varepsilon_1 - F(t) - \varepsilon_1$. Thus

$$\begin{aligned} E[F(Y)] &= E[F(Y)1_{\{Y \in (t_0, t_1 + \delta_1)\}}] + E[F(Y)1_{\{Y \notin (t_0, t_1 + \delta_1)\}}] \\ &\geq E[G(Y)1_{\{Y \in (t_0, t_1 + \delta_1)\}}] + E[G(Y)1_{\{Y \notin (t_0, t_1 + \delta_1)\}}] + \varepsilon_1 P\{t_0 < Y \leq t_1 + \delta_1\} \\ &= E[G(Y)] + \varepsilon_1 [G(t_1 + \delta_1) - G(t_0)] \\ &\geq 1/2 + \varepsilon_1 [G(t_1 + \delta_1) - G(t_0)] \quad (\text{from Lemma 5}) \\ &> 1/2. \end{aligned}$$

Now we have $\theta = 2P\{Y < X\} - 1 = 2E[F(Y)] - 1 > 0$ (since $P\{X = Y\} = 0$). Thus claim (c1) is stronger than claim (c4).

To prove that claim (c1) is stronger than claim (c3), we refer to Theorem 3 where the equivalence between claim (c3) and (c4) under the assumption of $P\{X = Y\} = 0$ is proved.

3. When data from both treatment groups have continuous distributions, we must have $P\{X = Y\} = 0$. From (1) and (2) above, we have showed that claim (c1) is stronger than claims (c2), (c3), and (c4). We now prove that claim (c1) is also stronger than claim (c5). Assumption $P\{Y > t\} > P\{X > t\}$ for all t means that $F(t) > G(t)$ for all t . Since $F(\text{med}_X) = 1/2 = G(\text{med}_Y)$ by definition of median for continuous random variable, we have $F(\text{med}_X) = G(\text{med}_Y) < F(\text{med}_Y)$, i.e., $F(\text{med}_X) < F(\text{med}_Y)$. Since F is a monotone increasing function, we must have $\text{med}_X < \text{med}_Y$, i.e., claim (c5) is true.

2. Proof of Theorem 2

Let X_i be the random variable for treatment group i , $i = 1, 2$. Since $b(\cdot)$ is a convex function, $E[X_1] - E[X_2] = b(\eta_1) - b(\eta_2) > 0$ if and only if $\eta_1 > \eta_2$. Then

$$\begin{aligned} P(X_1 > X_2) - P(X_1 < X_2) &= \iint_{x_1 > x_2} \exp[-b(\eta_1) - b(\eta_2) - a(x_1) - a(x_2)] \cdot \{\exp[\eta_1 x_1 + \eta_2 x_2] - \exp[\eta_1 x_2 + \eta_2 x_1]\} dx_1 dx_2 \\ &= \iint_{x_1 > x_2} \exp[-b(\eta_1) - b(\eta_2) - a(x_1) - a(x_2) + \eta_1 x_2 + \eta_2 x_1] \cdot \{\exp[(\eta_1 - \eta_2)(x_1 - x_2)] - 1\} dx_1 dx_2 \end{aligned}$$

This implies $P(X_1 > X_2) - P(X_1 < X_2) > 0$ if and only if $\eta_1 > \eta_2$, completing the proof.

3. Proof of Theorem 3

Define $m = \text{med}_{(Y-X)}$, $T = Y - X$ and its distribution function H . Then

$$H(m) \geq 1/2; H(m-\delta) < 1/2 \text{ for } \forall \delta > 0. \quad (4)$$

If (c3) is true, i.e., $m > 0$:

$$\begin{aligned} \theta &= P(T > 0) - P(T < 0) \\ &= 1 - 2H(0) + P(T = 0) \\ &\geq 1 - 2H(m/2) + P(T = 0) \\ &> 1 - 2 \cdot \frac{1}{2} + P(T = 0) \quad (\text{use (4)}) \\ &\geq 0. \end{aligned}$$

Thus, (c4) is true.

On the other hand, if $P(T = 0) = 0$ and $\theta > 0$, we use (4) to obtain $H(0) = [1 - \theta + P(T = 0)]/2 < 1/2 = H(m)$. Hence $m > 0$.

In general, if $\theta > 0$, then, for any $\delta > 0$, $H(-\delta) - H(0) - P(T = 0) = 1/2 - \theta/2 - P(T = 0)/2 < 1/2$, we have $-\delta \notin \{t: H(t) \geq 1/2\}$. Hence $m \geq 0$.

4. Proof of Theorem 4

1. Under the symmetric distribution assumption, the median and the mean coincide. Thus (c2) and (c5) are equivalent.

Let m_X and m_Y be the medians of X and Y respectively. The symmetry assumption implies that

$$\forall t: \begin{cases} P\{X < m_X - t\} = P\{X > m_X + t\} \\ P\{Y < m_Y - t\} = P\{Y > m_Y + t\}. \end{cases} \quad (5)$$

Then

$$\begin{aligned} P\{Y - X > m_Y - m_X + t\} &= P\{Y - m_Y > t + X - m_X\} \\ &= E[P\{Y - m_Y > t + X - m_X | X\}] \\ &= E[P\{Y - m_Y < -t - X + m_X | X\}] \quad \text{from (5)} \\ &= P\{Y - m_Y < -t - X + m_X\} \\ &= E[P\{X - m_X < -t - Y + m_Y | Y\}] \\ &= E[P\{X - m_X > t + Y - m_Y | Y\}] \quad \text{from (5)} \\ &= P\{Y - X < m_Y - m_X - t\}, \end{aligned}$$

i.e.,

$$P\{Y - X > m_Y - m_X + t\} = P\{Y - X < m_Y - m_X - t\}, \quad \forall t. \quad (6)$$

This means that $Y - X$ is symmetric about $m_Y - m_X$. By symmetry, we have $med_{Y-X} = m_Y - m_X = med_Y - med_X$. Thus (c3) and (c5) are equivalent. In summary, we have (c2), (c3), and (c5) are equivalent.

2. Substituting $t = -m_Y + m_X$ into (6) to obtain $P\{Y - X > 0\} = P\{Y - X < 2(m_Y - m_X)\}$. Thus $\theta = P\{Y - X > 0\} - P\{Y - X < 0\} = P\{Y - X < 2(m_Y - m_X)\} - P\{Y - X > 0\}$. $\theta > 0$ implies $P\{Y - X < 2(m_Y - m_X)\} > P\{Y - X > 0\}$. We must have $2(m_Y - m_X) > 0$, i.e., $m_Y > m_X$ which is (c5). Based on claims in (1), we conclude that claim (c4) is at least as strong as claims (c2), (c3) and (c5).
3. We only need to show that (c4) is true. From algebra in the proof of (2), $\theta = P\{Y - X < 2(m_Y - m_X)\} - P\{Y - X > 0\} = P\{0 < Y - X < 2(m_Y - m_X)\} > 0$. Thus $\theta > 0$.

Figure 1.
Overall survival

Figure 2.
Relation among 5 claims

Table 1

	0	1	2
X	0.6	0.4	0
Y	0.7	0.1	0.2