Published in final edited form as:

Stat Med. 2018 September 10; 37(20): 2923–2937. doi:10.1002/sim.7799.

Causal Estimands and Confidence Intervals Associated with Wilcoxon-Mann-Whitney Tests in Randomized Experiments

Michael P. Fay¹, Erica H. Brittain¹, Joanna H. Shih², Dean A. Follmann¹, and Erin E. Gabriel³

¹Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, MD

²Biometric Research Branch, DCTD, National Cancer Institute, Rockville, MD

³Unit of Biostatistics at Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

Abstract

Although the p-value from a Wilcoxon-Mann-Whitney (WMW) test is used often with randomized experiments, it is rarely accompanied with a causal effect estimate and its confidence interval. The natural parameter for the WMW test is the Mann-Whitney parameter, ϕ , which measures the probability that a randomly selected individual in the treatment arm will have a larger response than a randomly selected individual in the control arm (plus an adjustment for ties). We show that the Mann-Whitney parameter may be framed as a causal parameter and show that it is not equal to a closely related and non-identifiable causal effect, ψ , the probability that a randomly selected individual will have a larger response under treatment than under control (plus an adjustment for ties). We review the paradox, first expressed by Hand, that the ψ parameter may imply that the treatment is worse (or better) than control, while the Mann-Whitney parameter shows the opposite. Unlike the Mann-Whitney parameter, ψ is non-identifiable from a randomized experiment. We review some nonparametric assumptions which rule out Hand's paradox through bounds on ψ , and use bootstrap methods to make inferences on those bounds. We explore the relationship of the proportional odds parameter to Hand's paradox, showing that the paradox may occur for proportional odds parameters between 1/9 and 9. Thus, large effects are needed to ensure that if treatment appears better by the Mann-Whitney parameter, then treatment improves responses in most individuals. We demonstrate these issues using a vaccine trial.

Keywords

Causal estimand; Clinical trial; Mann-Whitney U test; Probabilistic index; Wilcoxon rank sum test

1 Introduction

The Wilcoxon-Mann-Whitney (WMW) test ([1, 2]), also known as the Wilcoxon rank sum test or the Mann-Whitney U test, is an important test used for the two-sample problem. The WMW test can be derived as a permutation test on the difference in mean ranks between the

two samples. The WMW test remains popular because it is robust to outliers, it can handle responses with ties, it can be implemented for small or large sample sizes, it is palindromically invariant (i.e., the two-sided inferences are the same if you reverse the ranking), and it can often be more efficient than alternative tests such as the t-test [3, 4, 5]. Finally, the WMW test is valid with minimal assumptions, since we only need to assume equality of the distributions of the two groups under the null. (Sometimes the WMW test is interpreted under more restrictive assumptions, such as a location shift for continuous responses, but those more restrictive assumptions are not necessary for validity of the p-value from the test. See [5], Section 5.1.1.)

Although the use of the WMW test is common, presenting associated effect estimates and confidence intervals with the WMW test is not. Presenting p-values without effect estimates and confidence intervals can lead to common misinterpretations of p-values [6]. Ideally, we want an estimand and an accompanying confidence interval that is compatible with the WMW test; in other words, the $100(1 - \alpha)\%$ confidence interval excludes the null hypothesis parameter value(s) (in this case, the parameter value implied when the distributions of the two groups are equal) if and only if the WMW test rejects the null at level α .

An important estimand related to the WMW test is the Mann-Whitney parameter, say ϕ , which we define as the probability that a random response from one group is larger than an independent random response from the other group plus one half the probability that the two responses are equal. We use a tie correction so that when the distributions of the two groups are equal, then $\phi = 1/2$ for either ordinal data or continuous data. The ϕ parameter is also known as the probabalistic index [7]. When $\phi \neq 1/2$ then the WMW test is consistent (i.e., power goes to one as the sample size goes to infinity) [8]. Other estimands do not have this consistency without extra assumptions. The difference in medians does not give consistency of the WMW test for non-null values (i.e., when the difference is not equal to zero), unless we assume continuous responses with a location shift. The proportional odds parameter requires the proportional odds assumption to give consistency for non-null values for the WMW test. Thus, in this paper we focus on confidence intervals for ϕ .

Fay and Malinovsky [9] studied confidence intervals on ϕ that are compatible with the WMW test. They show that extra assumptions on the distributions are needed (besides the null that the two distributions are equal), in order to get confidence intervals with proper coverage. In [9] the focus is on confidence intervals on ϕ under the proportional odds assumption, since that assumption may be applied to either continuous or discrete ordinal responses, unlike the Hodges-Lehmann intervals under the location shift assumption which require the continuity assumption. Other papers focus on confidence intervals for ϕ that require fewer assumptions under the null but are not compatible with the WMW test [10, 11]. Although these papers [9, 10, 11] give details about assumptions for statistical inferences (tests of association and confidence intervals), they give no guidance on causal interpretations, even in the relatively simple randomized experiment case. That is the focus of this paper.

In the context of causally defined treatment effects, Imbens and Rubin noted that “[u]nlike the simple difference in means, or the difference in logarithms, the rank-based statistics do not have a direct interpretation as a meaningful treatment effect.” ([12], p. 68). Although there has been some closely related recent work studying causal estimands for ordinal outcomes [13] and work studying Neyman versus Fisher null hypotheses with causal inference [14], this paper more directly explores causal estimands related to the WMW test (continuous and discrete versions) in randomized experiments.

For ease of exposition, throughout the paper we refer to the two groups as treatment and control, and assume higher responses are a better outcome. In Section 2 we give a causal interpretation of the Mann-Whitney parameter, ϕ , and explore its relationship with a closely related parameter, ψ , which is the probability that a randomly selected subject will respond better on treatment than on control plus one half the probability of responding the same on both. As noted by Hand [15], treatment effects for ψ can be in the opposite direction from those of ϕ . For example, it is possible that $\psi > 1/2$ (treatment appears better than control) and $\phi < 1/2$ (control appears better than treatment). When those two parameters imply opposite directional effects, we call this Hand’s paradox. Without making strong assumptions ψ is not identifiable in a randomized experiment; therefore, without making some additional assumptions, we cannot be certain that Hand’s paradox is not present. Wu, *et al.* [16] study causal inference for the WMW test in nonrandomized studies; however, their statements about ϕ and ψ in randomized trials seem to imply that Hand’s paradox cannot occur in randomized experiments (see [16], p. 1264).

To explore the difference between ϕ and ψ , we give two simple examples of Hand’s paradox in Section 3. For the bulk of the paper, we review known assumptions and explore new ones that, if true, rule out Hand’s paradox. We review some not very restrictive assumptions in Section 4, using results from Lu *et al.* [17], and Fan and Park [18] to obtain bounds on ψ . Using bootstrap methods to estimate those bounds, we can make some inferences on whether Hand’s paradox is possible in any specific data set. In Section 5 we review stronger assumptions that have been applied that preclude the paradox [19, 20]. The stronger assumptions (e.g., the change in potential responses is the same for all individuals) may not fit many data applications and some of those assumptions are not testable. Section 6 gives new work, avoiding Hand’s paradox using semi-parametric assumptions on the marginal distributions (e.g., proportional odds or proportional hazards), which are a compromise between the weaker assumptions of Section 4 and the stronger assumptions of Section 5. We show that even with these semi-parametric assumptions, it can be hard to rule out Hand’s paradox. For example, under continuous responses and the proportional odds assumption, we require that the proportional odds parameter, θ , has either $\theta > 9$ or $\theta < 1/9$ to ensure no Hand’s paradox exists. Translated into ϕ , this means that Hand’s paradox is possible even when we can assume proportional odds, provided that $\phi \in (0.184, 0.816)$. We end with an application in Section 7, demonstrating how to causally interpret results from a two-sample randomized experiment using the WMW test, ϕ , and its confidence interval, as well as giving estimated bounds on ψ .

2 Notation and Definition of Two Causal Parameters

Let z_i be the indicator for the randomized intervention performed on the i th individual (we assume perfect compliance with the random allocation), with $z_i = 0$ denoting control and $z_i = 1$ denoting treatment. Let $y_i(z)$ be the potential outcome that individual i would have had if she had received $z_i = z$, so each individual has two potential outcomes. Let $[Y_i(0), Y_i(1)]$ be the random variables representing the pair of potential outcomes for the i th individual. We assume that all pairs of potential responses are independent and are distributed with the bivariate distribution, F . For the i th individual we only observe $y_i(z_i)$. Let the marginal distributions be F_0 and F_1 , where $Y_i(0) \sim F_0$ and $Y_i(1) \sim F_1$. Since we only observe one element of each potential outcome pair, we cannot estimate F , but can only estimate F_0 and F_1 .

One can define the Mann-Whitney parameter as a functional of F_0 and F_1 ,

$$\phi = h_{MW}(F_0, F_1) = \Pr[Y_i(0) < Y_i(1)] + \frac{1}{2} \Pr[Y_i(0) = Y_i(1)] \quad (1)$$

where $Y_i(0) \sim F_0$ and independently $Y_i(1) \sim F_1$. A natural but non-identifiable causal parameter is the proportion of the population that would do better on treatment than control (corrected for ties),

$$\psi = h_{PBT}(F) = \Pr[Y_i(0) < Y_i(1)] + \frac{1}{2} \Pr[Y_i(0) = Y_i(1)] \quad (2)$$

where jointly $[Y_i(0), Y_i(1)] \sim F$. The value ψ can be reexpressed using the proportion who do better on treatment minus the proportion that do better on control,

$$\psi = \frac{1}{2} + \frac{1}{2} \{ \Pr[Y_i(0) < Y_i(1)] - \Pr[Y_i(0) > Y_i(1)] \}. \quad (3)$$

Although ψ is not identifiable, we can estimate bounds on it. We explore these bounds in later sections.

By examining the definitions of ϕ and ψ one might mistakenly assume that ψ is the causal parameter associated with the effect represented by ϕ ; however, ϕ may be expressed as a distinctly different causal parameter than ψ .

$$\phi = E\{\dot{G}(Y_i(1)) - \dot{G}(Y_i(0))\} + \frac{1}{2} \quad (4)$$

where

$$\dot{G}(y) = \frac{\dot{F}_0(y) + \dot{F}_1(y)}{2}$$

and

$$\dot{F}_0(y) = \frac{F_0(y) + F_0(y-)}{2} = Pr[Y(0) < y] + \frac{1}{2} Pr[Y(0) = y]$$

and $\dot{F}_1(y)$ is defined analogously. We show the proof of equation 4 separately for continuous and discrete data in Appendix A.

Using equation 4, and the fact that

$$E\{\dot{G}(Y_i(1)) - \dot{G}(Y_i(0))\} = E\{\dot{G}(Y_i(1))\} - E\{\dot{G}(Y_i(0))\}$$

we can show that this causal estimand does not depend on the specific potential outcome pairings (analogous to the causal effect on the mean). Thus, this causal estimand is identifiable from a randomized experiment, because F_0 and F_1 are identifiable.

Interpreting ϕ as a causal estimand is not straightforward. For interpretation, let the random variable $D_i = \dot{G}(Y_i(1)) - \dot{G}(Y_i(0))$, represent the “quantile difference” causal effect for the i th subject, so that $\phi = E(D_i) + \frac{1}{2}$. The value $\dot{G}(Y_i(1))$ represents where the i th subject’s treatment response falls, in terms of quantiles, among all potential responses, under either arm, in the population. For example, $\dot{G}(Y_i(1)) = 0.90$ and $\dot{G}(Y_i(0)) = 0.70$ means that the responses for the i th subject when on treatment are about as good or better than 90% of all responses (including both treatment and control potential responses), but her responses on control are only about as good or better than 70% of all responses. Because the potential outcome pairs are independent, the individual quantile difference causal effects, $D_1, \dots, D_n$, are also independent when $\dot{G}$ is known. Nevertheless, because of the nature of $\dot{G}$, although $-1 \leq D_i \leq 1$, we have $-1/2 \leq E(D_i) \leq 1/2$, which gives the needed result that $\phi \in (0, 1)$.

Although, we have shown ψ and ϕ are different estimands, an even more surprising result is that the two estimands can have opposite qualitative implications for the treatment effect: one can show that treatment is better than control while the other shows the opposite. Loosely speaking, this can occur when a majority of subjects would do slightly worse on treatment than control, and the remaining subjects would do substantially better on treatment than control, or vice versa. Because this phenomenon (to the best of our knowledge) was first noted by Hand [15], we define Hand’s paradox as when either $(\phi < 1/2 \text{ and } \psi > 1/2)$, or $(\phi > 1/2 \text{ and } \psi < 1/2)$.

3 Two Examples of Hand's Paradox

3.1 Simple Example with Three Classes

Consider a toy example to make the point. Suppose there are three types of people with equal proportions in the population, types A,B and C. Suppose the responses under treatment, $Y(1)$, and control, $Y(0)$, are:

Type	$Y(0)$	$Y(1)$
A	1	6
B	3	2
C	5	4

Thus, in the true population no individual responds the same on treatment and control, and only $\psi = 1/3$ of the population (Type A) does better on treatment, while $1 - \psi = 2/3$ of the population is better on control (Types B and C). But if we do an infinite sample size randomized trial, the Kaplan-Meier curves would look like Figure 1, and the treatment looks like it leads to larger responses. If we did a Wilcoxon-Mann-Whitney test on the infinite trial, we would estimate $\phi = P[Y_f(0) < Y_f(1)] + 0.5P[Y_f(0) = Y_f(1)] = 2/3$, seeming to imply that 2/3 of the population would do better on treatment. In fact, the proper interpretation is that the expected change in quantile going from control to treatment is an increase of $2/3 - 1/2 = 1/6$, i.e., an average percentile point change of +16.7%. Let us explore this interpretation if we could observe the above table. There are three types of people, Type A have the best responses on treatment, $\hat{G}(Y(1)) = 11/12$, and the worse responses on control, $\hat{G}(Y(0)) = 1/12$, while Types B and C do slightly worse on treatment than on control (Type B: $\hat{G}(Y(1)) = 3/12$ and $\hat{G}(Y(0)) = 5/12$, Type C: $\hat{G}(Y(1)) = 7/12$ and $\hat{G}(Y(0)) = 9/12$). So the expected change in quantile for going on treatment instead of control is positive,

$$E(D_i) = \frac{1}{3}\left(\frac{10}{12}\right) + \frac{2}{3}\left(\frac{-2}{12}\right) = \frac{1}{6}. \quad (5)$$

and $\phi = 1/2 + 1/6 = 2/3$.

3.2 A Continuous Example of Mixture of Normals

Now we consider a more elaborate toy example with continuous responses. Suppose in a randomized trial patients are randomized to get either treatment ($z = 1$) or control ($z = 0$). Suppose larger values are better, so that for the continuous case ψ represents the proportion of patients that are better off on the treatment than on the control. In Figure 2 we plot distributions from a hypothetical example where the treatment helps about 35% of individuals with substantial improvement, but the other 65% are slightly worse off on the treatment.

The bivariate potential outcome distribution is a mixture of two normal distributions, and the parameters were chosen because they are a particularly striking example of Hand's paradox (see Appendix B). The value of ψ for this example is 0.358, so that in truth only 35.8% of individuals would respond better on the treatment arm. The value of $\phi = 0.654$, so the average quantile difference favors the treatment arm. A naive interpretation of ϕ as measuring ψ would incorrectly conclude that more patients would be helped on the new treatment. The proper interpretation is that the expected quantile difference is $\phi - 1/2 = E(\dot{G}(Y_A(1))) - E(\dot{G}(Y_A(0))) = 0.154$, and since it is positive the average response quantile is larger on treatment than on control.

4 Sharp Bounds for ψ with Minimal Assumptions

4.1 Sharp Bounds when Marginal Distributions are Known

Sharp bounds for ψ have been developed in the literature. When F_0 and F_1 are continuous distributions (see [18] Lemma 2.1; [21], p. 312), then the sharp bounds on ψ are

$$\psi_L \equiv \sup_y \{F_0(y) - F_1(y)\} \leq \psi \leq 1 - \sup_y \{F_1(y) - F_0(y)\} \equiv \psi_U. \quad (6)$$

For example, by numerically finding the largest difference between the survival curves in Figure 2 (right panel), we find that $\psi_L = 0.293$ and $\psi_U = 0.999$. The true value $\psi = 0.358$ is within those bounds, and is fairly close to the lower bound, partly because the parameters were chosen to highlight Hand's paradox. When F_0 and F_1 are discrete distributions, we combine the sharp bounds of Lu *et al.* [17] for $P[Y_A(1) \leq Y_A(0)]$ and $P[Y_A(1) > Y_A(0)]$ to get bounds for ψ . We write sharp bounds that apply to either continuous or discrete responses as,

$$\psi_L \equiv \sup_y \{\dot{F}_0(y) - F_1(y-)\} \leq \psi \leq 1 - \sup_y \{\dot{F}_1(y) - F_0(y-)\} \equiv \psi_U. \quad (7)$$

For the example of Section 3.1, we get $\psi_L = 1/3$ and $\psi_U = 1$, and the true value $\psi = 1/3$ equals the lower bound.

4.2 Estimating Bounds

A simple way to estimate the bounds is to use empirical distribution functions of F_0 and F_1 in the bounds of expression 7. As the sample sizes get larger, the estimates will get closer to the true bounds. A simple way to examine the variability is to use a nonparametric bootstrap.

Let us return to the example of Section 3.1. Suppose we perform a randomized trial under that distribution with $n_0 = 100$ and $n_1 = 100$. Here is a contingency table of some simulated responses:

	1	2	3	4	5	6
$y_i(0)$	33	0	31	0	36	0
$y_i(1)$	0	36	0	30	0	34

Using the empirical distributions, the bounds are $\hat{\psi}_L = 0.34$ and $\hat{\psi}_U = 0.97$. Using a percentile bootstrap confidence interval for both bounds using 10^4 resamples, we get 95% confidence intervals for ψ_L of (0.29, 0.45) and for ψ_U of (0.83, 1.00). As expected, both confidence intervals cover the true values of $\psi_L = 1/3$ and $\psi_U = 1$.

Now return to the continuous example of Section 3.2 where $\psi_L = 0.293$ and $\psi_U = 0.999$. We simulate data with $n_0 = n_1 = 100$. Then using a percentile bootstrap method with 10^4 replications, we get 95% confidence intervals for ψ_L of (0.22, 0.41) and for ψ_U of (0.93, 1.00).

5 Strong Assumptions That Exclude Hand's Paradox

For completeness we list some strong assumptions that guarantee that Hand's paradox does not exist. Let $\tau_i = Y_i(1) - Y_i(0)$. The additive treatment effect assumption is that $\tau_1 = \tau_2 = \dots = \tau_n = \tau$, which implies that $\text{cor}(Y_i(0), Y_i(1)) = 1$. The paradox is excluded if we assume an additive treatment effect, where the treatment will change every individual's response by the same amount (see e.g., [20]). When the responses are continuous, under the additive treatment effect we can get Hodges-Lehmann confidence intervals on τ using a series of permutation test p-values. Let $p(\tau_0) \equiv p(\mathbf{y}, \mathbf{z}, \tau_0)$ be the two-sided p-values for testing $H_0: \tau = \tau_0$ using a permutation test on the shifted the potential outcomes $y_i^*(z_i) = y_i(z_i) + z_i\Delta_0$. The $100(1 - \alpha)\%$ Hodges-Lehmann confidence interval is $\{\tau_0: p(\tau_0) > \alpha\}$. Heckman *et al.* [22] (p. 498, bottom, and p. 500 top) use the Hoeffding-Frechet bounds (p. 497, bottom) to estimate a lower bound on $\text{Var}(\tau)$ using the most highly correlated pairing, and uses a bootstrap to make inference about that statistic. If the bootstrap inferences imply that $\text{Var}(\tau)$ is bounded away from 0, then we must reject the additive assumption. For a randomization test of $H_0: \text{Var}(\tau) = 0$ and other issues relating to estimating $\text{Var}(\tau)$, see Ding *et al.* [23].

A less stringent assumption that rules out Hand's paradox is that $\tau_i \geq 0$ for all i . We call this the "no harm" assumption, since it implies that the treatment will never cause any individual to do worse than when on control. Rosenbaum [19] uses this assumption in Section 4 of that paper to make confidence intervals from randomization inference of the WMW test on attributable treatment effects which are functions of a large dimensional study specific parameter. This parameter is rarely used because it is difficult to interpret.

6 Semiparametric Assumptions and Hand's Paradox

6.1 Hand's Paradox and the Proportional Odds Model

6.1.1 Continuous Data—Now consider the case where the marginals are related by a proportional odds model, i.e.,

$$\frac{F_0(y)}{1 - F_0(y)} = \frac{\theta F_1(y)}{1 - F_1(y)}$$

for all y , where $\theta \in (0, \infty)$ is the proportional odds parameter. Letting $z_i = 0$ for control and $z_i = 1$ for treatment, we can rewrite this model as

$$\text{logit}(P[Y_i(z_i) \leq y]) = \text{logit}(F_0(y)) - z_i\beta \quad (8)$$

where $\beta = \log(\theta)$ and $\text{logit}(q) = \log(q/(1 - q))$ (see [9]). Since our test and parameter estimates are based on ranks, we get the same inferences if we use any strictly increasing transformation of the responses. We can show that for the continuous proportional odds model there is a strictly increasing transformation of the responses, $h(y) = G_0(F_0^{-1}(y))$, where G_0 is the standard logistic distribution. So without loss of generality, under a proportional odds model we can assume that F_0 is a standard logistic and $F_1(y) = F_0(y - \log(\theta))$. Using that expression and solving through integration $P[Y_i(0) < Y_i(1)]$ we get

$$\phi = \begin{cases} 0 & \text{if } \theta = 0 \\ \frac{1}{2} & \text{if } \theta = 1 \\ 1 & \text{if } \theta = \infty \\ \frac{\theta\{\theta - 1 - \log(\theta)\}}{(\theta - 1)^2} & \text{otherwise} \end{cases} \quad (9)$$

To rule out Hand's paradox, first assume that $\theta > 1$. In this case, $\phi > 0.5$. We want to find the value of θ where the lower bound on ψ is equal to 0.5 such that the paradox does not exist. We can get sharp lower bounds on ψ using $\psi_L = \sup_y \{F_0(y) - F_1(y)\}$ as in equation 6. Because f_0 (the associated probability density function of F_0) is unimodal and f_1 is just a location shifted version of f_0 , there is only one local maximum and it is the global maximum. So to find the supremum we take the derivative of $F_0(y) - F_1(y)$ with respect to y , set that to 0, and solve for y . This is the same as solving for the y where $f_0(y) = f_1(y)$. Thus, the supremum occurs at $y = 0.5 \log(\theta)$. Solving for $F_0(0.5 \log(\theta)) - F_1(0.5 \log(\theta)) = 0.5$ gives $\theta = 9$. In other words, if $\theta \geq 9$ then $\phi > 0.5$ and $\psi \geq 0.5$. Using equation 9 we can express the bound in terms of ϕ , under the proportional odds assumption on the marginals, as long as $\phi \geq 0.816$ then the paradox cannot occur. The analogous bound assuming $\theta < 1$ is $\theta = 1/9$ or in terms of the Mann-Whitney parameter $\phi = 1 - 0.816 = 0.184$.

6.1.2 Discrete Proportional Odds Model—Now suppose we observe ordinal responses with K possible outcomes that follow a proportional odds model. Let $Y_i(z_i)$ be the random variable associated with the ordinal response for subject i randomized to group z_i . For convenience we define the sample space for the K possible ordinal outcomes as $\{1, 2, \dots, K\}$. One way to create such data is to suppose there is a latent continuous response associated with $Y_i(z_i)$, say $Y_i^*(z_i)$, and suppose that those latent responses follow a proportional odds model. As described in the previous section, we assume without loss of generality that $Y_i^*(0) \sim F_0^*$ and $Y_i^*(1) \sim F_1^*$, where here F_0^* is the standard logistic and $F_1^*(y) = F_0^*(y - \log(\theta))$. We relate the two types of random variables by $Y_i(z_i) = k$ if $Y_i^*(z_i) \in (\gamma_{k-1}, \gamma_k]$, where $-\infty \equiv \gamma_0 < \gamma_1 < \dots < \gamma_{K-1} < \gamma_K \equiv \infty$. The γ parameters are unknown parameters that determine how the continuous variable is grouped into ordered categories. The proportional odds model of equation 8 becomes

$$\begin{aligned} \text{logit}(P[Y_i^*(z_i) \leq \gamma_k]) &= \text{logit}(F_0^*(\gamma_k)) - z_i \log(\theta) \quad \Rightarrow \quad \text{logit}(P[Y_i(z_i) \leq k]) = \gamma_k \\ &- z_i \log(\theta), \end{aligned} \quad (10)$$

because the logit function is F_0^{*-1} . Because of the grouping, ties are introduced, and ϕ no longer follows equation 9. Nevertheless, we can use equation 9 to get the latent continuous Mann-Whitney parameter, say ϕ^* from θ .

Here is an example of the proportional odds model on the marginals with an ordinal responses with $K = 3$ categories where the $\phi > 0.5$, but $\psi < 0.5$. Suppose the unobservable joint distribution of $[Y(0), Y(1)]$ is F with a 3×3 probability mass function, f ,

$$f(y_0, y_1) = \begin{bmatrix} f(1,3) & f(2,3) & f(3,3) \\ f(1,2) & f(2,2) & f(3,2) \\ f(1,1) & f(2,1) & f(3,1) \end{bmatrix} = \begin{bmatrix} 0.2231 & 0 & 0.0769 \\ 0 & 0.0965 & 0.2035 \\ 0.2 & 0.2 & 0 \end{bmatrix}$$

This is a proportional odds model with $\theta = 1.1$ and $\gamma = [-0.3101, 0.9425]$, so that $F_0 = [F_0(1), F_0(2), F_0(3)] = [0.4231, 0.7196, 1]$ and $F_1 = [F_1(1), F_1(2), F_1(3)] = [0.4, 0.7, 1]$. In this model, $\phi = 0.5140$, and the latent continuous Mann-Whitney parameter from equation 9 equals $\phi^* = 0.5159$. The ties have moved ϕ^* to ϕ , closer to 0.5. Since $0.184 < \phi^* < 0.816$ (from the previous section) it is possible that $\phi > 0.5$ and $\psi < 0.5$, and in fact, this occurs for this example: $\psi = 0.4098$.

6.2 Hand's Paradox and the Proportional Hazards Model

The proportional hazards model has $S_1(t) = S_0(t)^\lambda$, where $S_a(t) = 1 - F_a(t)$ for $a = 0, 1$, and is often used to analyze time to event data. Just as the proportional odds model can be treated as a shift in logistic after some unknown monotonic transformation, without loss of generality the two-sample proportional hazards model can be expressed as a standard

exponential (with mean 1), and an exponential with mean $1/\lambda$. Thus, under the proportional hazards model on the marginal random variables $Y(0)$ (for control) and $Y(1)$ (for treatment), we can calculate λ using those exponential distributions,

$$\phi = \Pr[Y(0) < Y(1)] = \int_0^{\infty} S_1(t)f_0(t)dt = \frac{1}{\lambda + 1} \quad (11)$$

See [24] for a more extensive discussion of estimators of hazard ratios, that hold even without the proportional hazard assumption holding. So when $\lambda > 1$ then $\phi < 0.5$ and the treatment has shorter times to event, and when $\lambda < 1$ then $\phi > 0.5$ and the treatment has longer times to event. Analogously to equation 10, we can also write the model using the complementary log-log link as a generalized linear model [25].

Using equation 6 analogously to in the proportional odds model, we can find the upper and lower bound of the continuous proportional hazards which exclude Hand's paradox. Using the standard exponential for F_0 , taking derivatives and setting to zero, we find that the maximum of $F_0(t) - F_1(t)$ occurs at $\log(\lambda)/(\lambda - 1)$. We numerically find that $\lambda = 0.2271$ solves

$$F_0(\log(\lambda)/(\lambda - 1)) - F_1(\log(\lambda)/(\lambda - 1)) = \frac{1}{2}.$$

So if the proportional hazards model holds on the marginals, and $\lambda < 0.2271$ or $\lambda > 1/0.2271 = 4.403$ then Hand's paradox cannot occur. In terms of ϕ from the continuous proportional hazard model, if $\phi > 1/(1 + 0.2271) = 0.8149$ or $\phi < 1/(1 + 4.403) = 0.1851$ then the paradox is avoided. For discrete data there are analogous results to Section 6.1.2, and the same proportional hazards parameter can apply to discrete data and an unknown latent continuous model. Using the log-log link and the estimates of the cutpoints, we can convert the proportional hazards parameter from discrete data into ϕ^* , the Mann-Whitney parameter on the latent continuous proportional hazards variates.

6.3 Hand's Paradox and the Probit Model

Now we consider the probit model. Consider first the continuous response case. For this model suppose there exists some monotonic transformation of the responses such that the distribution of the control group on the transformed responses is standard normal, and the distribution of the treatment group is $\mathcal{N}(\mu, 1)$. Hand [15] showed that as long as $\mu \in (-1.349, 1.349)$ then the paradox could occur. Translating this into ϕ values and using equation 6, under this model as long as $\phi \in (0.170, 0.830)$ then the paradox can occur.

Analogous to Section 6.1.2 we can write this as a discrete model, except now we use a probit link. We can again convert the probit parameter from discrete data using the estimates of the cutpoints to get an estimate of the latent ϕ^* value.

6.4 Comparing Semiparametric Models

Notice the similarity in the bounds on ϕ^* for allowing Hand's paradox: for the proportional odds model the bounds are $0.184 \leq \phi^* \leq 0.816$, which are very close to the bound for the proportional hazards model $0.185 \leq \phi^* \leq 0.815$. The bounds for the probit model are a bit larger $0.170 \leq \phi^* \leq 0.830$. Unlike the proportional hazards model, both the proportional odds model and the probit model are palindromically invariant because the latent continuous distribution for F_0^* is symmetric.

7 Application

Mullen *et al.* [26] describes a phase I randomized trial of an experimental malaria vaccine. The vaccine used AMA1-C1 (the AMA1 antigen combined from two different strains of the *Plasmodium falciparum* parasite) together with Alhydrogel (a standard adjuvant). Some of the arms of the study also added a novel adjuvant, CPG 7909. The study had 3 vaccine arms:

v80: 80 μg of AMA-C1/Alhydrogel,

v20+CPG: 20 μg of AMA-C1/Alhydrogel+ 564 μg CPG 7909, and

v80+CPG: 80 μg of AMA-C1/Alhydrogel+ 564 μg CPG 7909.

Subjects were randomized to the three arms, and given the allocated vaccine for that arm at days 0, 28 and 56. Although one of the primary purposes of the study was safety, we focus on the efficacy as measured by the growth inhibition assay (GIA) for the 3d7 strain on day 70. The GIA is an *in vitro* assay that aims to measure how much the antibodies from the subjects are inhibiting the growth of the parasite (100% GIA means that no parasites grow). Subjects were randomized using a 2 : 1 : 2 ratio ($n = 30$ for v80, $n = 15$ for v20+CPG, and $n = 30$ for v80+CPG), but for various reasons, there is only GIA data on 58 subjects ($n = 23$ for v80, $n = 10$ for v20+CPG, and $n = 25$ for v80+CPG). To avoid long discussions on missingness, we act as if those were the only subjects randomized to the three arms. Additionally, we use only the asymptotic methods and the resulting confidence intervals to avoid issues of small sample sizes (see [9] for discussions of that issue). For this illustration, we focus on testing for an effect of CPG by comparing v80 to v80+CPG and on testing for a dose effect (when CPG is present) by comparing v20+CPG to v80+CPG.

We plot the empirical distributions of the responses for the three arms in Figure 3. The raw data are given in Appendix C. Graphically, it appears that both arms with CPG have better growth inhibition. The inferences using the WMW test and its compatible confidence interval and the methods of this paper are listed in Table 1. The WMW test shows that the v80 and v80+CPG are significantly different ($p < 0.0001$) but the v20+CPG and v80+CPG are not significantly different ($p = 0.956$). The hypotheses that are being tested are $H_0 : F_0 = F_1$ versus $H_1 : F_0 \neq F_1$. For the proper interpretation of the p-values we do not need to make any assumptions on the F_0 and F_1 , however, those p-values and tests give us information on the direction of the effects. For confidence intervals on ϕ , if we make the additional proportional odds assumption, we can use the `wmwTest` function in the `asht` R package which estimates $\hat{\phi}$ using the h_{MW} functional on the empirical distributions: $\hat{\phi} = h_{MW}(\hat{F}_0, \hat{F}_1)$ [9]. Under that assumption, the causal effect of the CPG appears strong, $\hat{\phi} = 0.960$

(measuring $P[Y(v80+CPG) > Y(v80)] + 0.5P[Y(v80+CPG) = Y(v80)]$) with 95% confidence interval (0.847, 0.990). Thus, the average quantile difference ($\phi - 1/2$) is 0.460 (95% CI: 0.347, 0.490), which is close to the theoretical maximum of 0.5. Additionally, since the lower confidence limit for ϕ is > 0.816 , and since the confidence limits are central (the error is bounded at 2.5% on either side), as long as the proportional odds assumption holds we can conclude with 97.5% confidence that $\psi > 0.5$ (i.e., most subjects will improve their grow inhibition response by getting CPG added to their vaccine of 80 μg of AMA1-C1/Alhydrogel). We can examine that last statement nonparametrically by estimating the lower bound on ψ and then bootstrapping that estimate; we get $\hat{\psi}_L = 0.873$ (two-sided central 95% CI on ψ_L by percentile bootstrap with 10^4 replications: 0.743, 1.000). By this nonparametric method, we conclude with 97.5% confidence that $\psi > 0.743$, that is, at least 74.3% of subjects will improve their GIA response with CPG added to the 80 μg dose of the vaccine.

Now consider the ϕ effect of the higher dose by comparing v20+CPG to v80+CPG (measuring $P[Y(v80+CPG) > Y(v20+CPG)] + 0.5P[Y(v80+CPG) = Y(v20+CPG)]$). The p-value tells us that there is no significant difference between the groups, but high p-values could be due to insufficient evidence to reject the null hypothesis. We can supplement that p-value by adding a proportional odds assumption to get confidence intervals on ϕ , giving $\hat{\phi} = 0.508$ (95% CI: 0.310, 0.703). This gives a range of plausible values of ϕ , and we have ruled out very large effects in either direction. The sharp lower bound estimate, $\hat{\psi}_L = 0.200$ (95% CI: 0.020, 0.520) does not rule out Hand's paradox because it may be less than 0.5.

To illustrate the discrete inferences, suppose that the GIA values are grouped into a 4 point ordinal scale (low= $y < 25$], moderate= $[25 \leq y < 50]$, substantial= $[50 \leq y < 75]$, and strong= $[75 \leq y]$). The grouped data are given by Table 2. In Table 3 we show the results, which are similar to the results of the non-grouped data. There are significant differences between v80 and v80+CPG ($p < 0.0001$), but the Mann-Whitney effect is slightly different than in the continuous analysis, $\hat{\phi} = 0.934$, and assuming proportional odds, (95% CI: 0.815, 0.977), where we again use the `wmwTest` function [9]. The difference could be due to the ϕ parameter getting pulled closer to 1/2 because of ties, or they could be due to randomness and the small sample size. Similar to Agresti and Kateri [28], we can estimate the latent Mann-Whitney effect using a proportional odds model with its associated confidence intervals. We use `wmwTest` with `latentContinuous=TRUE` to estimate the latent Mann-Whitney values, ϕ^* , which uses a method described in [9] that is more closely tied to $\hat{\phi}$ than the methods of [28]; this gives $\hat{\phi}^* = 0.957$ (95% CI: 0.836, 1.000). This $\hat{\phi}^*$ is closer to $\hat{\phi} = 0.960$ from the continuous data. Finally, using the nonparametric estimate of ψ_L from equation 7, we get $\hat{\psi}_L = 0.826$ and a 95% bootstrap percentile interval on ψ_L with 10^4 replications of (0.696, 0.935), so we have 97.5% confidence that at least 69.6% of subjects will improve their GIA response on by adding CPG to the 80 μg dose vaccine.

Now consider the dose effect. There is no significant difference by WMW test ($p = 0.588$). The Mann-Whitney estimate $\hat{\phi} = 0.558$, and the 95% confidence interval (by [9] requiring the additional proportional odds assumption) is (0.363, 0.735). Using `wmwTest` with `latentContinuous=TRUE`, we estimate the latent Mann-Whitney parameter as $\hat{\phi}^* = 0.565$

(95% CI : 0.344, 0.758). As expected $\hat{\phi}^*$ is further from 1/2 than $\hat{\phi}$. Finally, the sharp lower bound estimate, $\hat{\psi}_L = 0.200$ (95% CI : 0.050, 0.490) does not rule out Hand's paradox because it may be less than 0.5.

8 Discussion

We have reframed the Mann-Whitney parameter as one-half plus the expected quantile difference between two groups. To the best of our knowledge, this causal interpretation has not previously been shown. This allows us to interpret the Mann-Whitney parameter for randomized experiments causally. We have explored the closely related parameter, ψ . In the treatment versus control experiment, with high responses more favorable, the ψ parameter represents the probability that a random individual in the population will have better responses on treatment than on control plus half the probability of responding the same. We have studied Hand's paradox [15]: when $\phi > 0.5$ (treatment is better by Mann-Whitney) but $\psi < 0.5$ (treatment helps fewer people), or vice versa. By showing that the two parameters are different under the potential outcome framework, we have shed light on when Hand's paradox can and cannot occur.

In order to exclude the possibility of Hand's paradox, we traditionally require some very strong assumptions such as: treatment affects all subjects by the same constant change in response, or treatment never hurts any subjects. However, we can explore bounds on ψ non-parametrically or semiparametrically. We estimate a nonparametric lower bound on ψ (previously developed in [17], see also [21]) using empirical distributions, and make inferences on that bound using bootstrap methods. Further, we study two important semi-parametric models (proportional odds and proportional hazards), and show the bounds needed to exclude Hand's paradox. For example, Hand's paradox can occur whenever proportional odds parameter, θ , is in (1/9, 9).

While we have referred to disagreement in direction of these two causal parameters as a paradox, in truth, these parameters reflect different aspects of the treatment effect. Often the parameters will be concordant. However, generally speaking, if a majority of subjects would do slightly worse with treatment than control, and the remaining subjects would experience substantial benefit, then ϕ may imply a beneficial treatment while ψ implies a harmful treatment, and vice versa. While the interpretation of ϕ is less intuitive than that of ψ , ϕ is arguably a more sensitive and relevant metric of treatment effect. The relationship of ϕ to ψ is roughly analogous to the relationship of a signed rank test to a sign test. That is, the former is affected by the magnitude of the benefit within an individual, whereas the latter is affected only by the direction of benefit. Since only ϕ can be directly estimated in a randomized study, and since its direction does not imply the direction of ψ , it is important that any causal interpretation is made carefully.

In randomized trials with imperfect compliance, there are several ways to use the treatment assignment as an instrumental variable to make inferences on causal effects related to treatment taken (see e.g., [12], Chapters 23–25). Further, in this situation, partially identified treatment effects have been developed for binary responses and continuous outcomes (see e.g., references cited in [29]). Future work could expand those methods and other causal

inference methods for dealing with the Mann-Whitney parameter, ϕ (which we have shown may be interpreted causally) and the parameter ψ (for which we can only estimate bounds from randomized trials).

A: Equivalence of Two Forms of ϕ

A.1 Continuous Case

In the continuous case,

$$\phi = \Pr[Y_i(0) < Y_j(1)] = \int (1 - F_1(y)) dF_0(y)$$

For continuous data, $\dot{G}(y) = G(y)$, where $G(y) = \frac{F_1(y) + F_0(y)}{2}$ is the combined distribution function of the two types of potential outcomes. Let ϕ_G be the expected value of the combined distribution function evaluated under treatment minus the combined distribution function evaluated under control plus 1/2,

$$\phi_G = E\{G(Y_i(1))\} - E\{G(Y_i(0))\} + \frac{1}{2}.$$

Rewriting, using $f_a(y) = dF_a(y)$ we get

$$\begin{aligned} \phi_G &= E\{G(Y_i(1))\} - E\{G(Y_i(0))\} + \frac{1}{2} \\ &= \int G(y)f_1(y)dy - \int G(y)f_0(y)dy + \frac{1}{2} \\ &= \int \frac{F_1(y) + F_0(y)}{2} f_1(y)dy - \int \frac{F_1(y) + F_0(y)}{2} f_0(y)dy + \frac{1}{2} \\ &= \frac{1}{2} \left[\int F_1(y)f_1(y)dy + \int F_0(y)f_1(y)dy - \left[\int F_1(y)f_0(y)dy + \int F_0(y)f_0(y)dy \right] \right] + \frac{1}{2} \\ &= \frac{1}{2} \left[\frac{1}{2} + \int F_0(y)f_1(y)dy - \left[\int F_1(y)f_0(y)dy + \frac{1}{2} \right] \right] + \frac{1}{2} \\ &= \frac{1}{2} \left[\int F_0(y)f_1(y)dy - \int F_1(y)f_0(y)dy \right] + \frac{1}{2} \\ &= \frac{1}{2} \left[\Pr[Y_i(0) < Y_j(1)] - \Pr[Y_j(1) < Y_i(0)] \right] + \frac{1}{2} \\ &= \frac{1}{2} \left[2\Pr[Y_j(1) > Y_i(0)] - 1 \right] + \frac{1}{2} \\ &= \Pr[Y_i(0) < Y_j(1)] = \phi, \end{aligned}$$

where going from the 4th to the 5th step we use a change of variables in the integration,

$$\int_{-\infty}^{\infty} F(y)dF(y) = \int_0^1 u \, du = \frac{1}{2}.$$

A.2 Discrete Case

For the discrete case

$$\begin{aligned}\phi &= Pr[Y_i(1) > Y_j(0)] + \frac{1}{2}Pr[Y_i(1) = Y_j(0)] \\ &= \sum F_0(y)f_1(y) - \frac{1}{2}\sum f_0(y)f_1(y)\end{aligned}$$

where the summations are over the entire sample space.

Let $\dot{F}(y) = F(y) - \frac{1}{2}f(y) = \frac{1}{2}\{F(y) + F(y-)\}$ for any distribution function. Let

$\dot{G}(y) = \frac{1}{2}\dot{F}_0(y) + \frac{1}{2}\dot{F}_1(y)$ Let

$$\begin{aligned}\phi_G &= E\{\dot{G}(Y_i(1))\} - E\{\dot{G}(Y_i(0))\} + \frac{1}{2} \\ &= \sum \dot{G}(y)f_1(y)dy - \sum \dot{G}(y)f_0(y)dy + \frac{1}{2} \\ &= \sum \frac{\dot{F}_1(y) + \dot{F}_0(y)}{2}f_1(y)dy - \sum \frac{\dot{F}_1(y) + \dot{F}_0(y)}{2}f_0(y)dy + \frac{1}{2} \\ &= \frac{1}{2}\left\{\sum \dot{F}_1(y)f_1(y)dy + \sum \dot{F}_0(y)f_1(y)dy - \left[\sum \dot{F}_1(y)f_0(y)dy + \sum \dot{F}_0(y)f_0(y)dy\right]\right\} + \frac{1}{2} \\ &= \frac{1}{2}\left\{\frac{1}{2} + \sum \dot{F}_0(y)f_1(y)dy - \sum \dot{F}_1(y)f_0(y)dy - \frac{1}{2}\right\} + \frac{1}{2} \\ &= \frac{1}{2}\left\{\sum \dot{F}_0(y)f_1(y)dy - \sum \dot{F}_1(y)f_0(y)dy\right\} + \frac{1}{2} \\ &= \frac{1}{2}\left\{2\sum \dot{F}_0(y)f_1(y)dy - 1\right\} + \frac{1}{2} \\ &= \sum \dot{F}_0(y)f_1(y)dy = \phi,\end{aligned}$$

where the 4th step uses $\sum \dot{F}_1(y)f_1(y)dy = \frac{1}{2}$ (see Appendix of Shih and Fay, 2017, Biometrics), and the 6th step uses,

$$\begin{aligned}\sum \dot{F}_0(y)f_1(y)dy &= Pr[Y_i(0) < Y_j(1)] + \frac{1}{2}Pr[Y_i(0) = Y_j(1)] \\ &= 1 - \left\{Pr[Y_j(1) < Y_i(0)] + \frac{1}{2}Pr[Y_i(0) = Y_j(1)]\right\} \\ &= 1 - \sum \dot{F}_1(y)f_0(y)dy.\end{aligned}$$

B: Mixture of Two Normal Distributions

First consider a bivariate normal distribution with

$$\begin{bmatrix} Y(0) \\ Y(1) \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix} \right)$$

Then $Y(1) - Y(0) \sim N(\mu_1 - \mu_0, \sigma_0^2 + \sigma_1^2 - 2\rho\sigma_0\sigma_1)$. So

$$\begin{aligned}
\psi &= Pr[Y(1) - Y(0) > 0] \\
&= Pr\left[\frac{(Y(1) - Y(0)) - (\mu_1 - \mu_0)}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\rho\sigma_0\sigma_1}} > \frac{-(\mu_1 - \mu_0)}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\rho\sigma_0\sigma_1}}\right] \\
&= 1 - \Phi\left[\frac{-(\mu_1 - \mu_0)}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\rho\sigma_0\sigma_1}}\right] \\
&= \Phi\left[\frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\rho\sigma_0\sigma_1}}\right]
\end{aligned}$$

The value ϕ is equal to the same expression except with ρ set to zero,

$$\phi = \Phi\left[\frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2 + \sigma_1^2}}\right].$$

Because $\Phi(x) < 0.5$ when $x < 0$, whenever $\mu_1 - \mu_0 < 0$ then $\psi < 0.5$ and $\phi < 0.5$, and similarly whenever $\mu_1 - \mu_0 > 0$ then $\psi > 0.5$ and $\phi > 0.5$. Thus, Hand's Paradox cannot occur if F is bivariate normal.

Now suppose $\mathbf{Y} = [Y(0), Y(1)]$ is distributed F which is the mixture of two bivariate normals, and the two marginals are F_0 and F_1 . In other words, $\mathbf{Y} = \lambda\mathbf{Y}_G + (1 - \lambda)\mathbf{Y}_H$ and $\mathbf{Y}_G \sim G$ and $\mathbf{Y}_H \sim H$, where $0 < \lambda < 1$ and G and H are both bivariate normal. Let the marginal distributions be G_0 , G_1 , H_0 and H_1 . Then $F_0 = \lambda G_0 + (1 - \lambda)H_0$, and $F_1 = \lambda G_1 + (1 - \lambda)H_1$, and $\psi = h_{PTB}(F) = \lambda h_{PTB}(G) + (1 - \lambda)h_{PTB}(H)$, and

$\phi = h_{mw}(F_0, F_1) = \lambda^2 h_{mw}(G_0, G_1) + \lambda(1 - \lambda)h_{mw}(G_0, H_1) + \lambda(1 - \lambda)h_{mw}(H_0, G_1) + (1 - \lambda)^2 h_{mw}(H_0, H_1)$. The model of

Figure 2 has $\lambda = 0.35$ with

$$G \equiv N\left(\begin{bmatrix} 0 \\ 5 \end{bmatrix}, \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

and

$$H \equiv N\left(\begin{bmatrix} 3 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$$

C: Mullen, et al data

Here are the raw *in vitro* percent growth inhibition data from [26]:

v80 arm: -4, -3, -2, -1, 2, 2, 3, 3, 6, 7, 7, 10, 14, 14, 17, 18, 24, 26, 29, 30, 31, 55, 72.

v20+CPG arm: 16, 45, 49, 67, 69, 72, 73, 75, 87, 94.

v80+CPG arm:

30,33,36,47,47,48,48,49,58,60,64,64,67,72,75,76,77,79,84,86,86,88,88,89,96.

References

- [1]. Wilcoxon F Individual comparisons by ranking methods. *Biometrics bulletin* 1945; 1(6):80–83.
- [2]. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* 1947; 18(1):50–60.
- [3]. Lehmann E *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, Inc.: Oakland, Ca, 1975.
- [4]. McCullagh P Regression models for ordinal data (with discussion). *Journal of the royal statistical society. Series B (Methodological)* 1980; :109–142.
- [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. URL <http://www.i-journals.org/ss/viewarticle.php?id=51>. [PubMed: 20414472]
- [6]. Wasserstein RL, Lazar NA. The asa's statement on p-values: context, process, and purpose. *Am Stat* 2016; 70(2):129–133.
- [7]. Thas O, Neve JD, Clement L, Ottoy JP. Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2012; 74(4):623–671.
- [8]. Lehmann EL. Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics* 1951; :165–179.
- [9]. Fay MP, Malinovsky Y. Confidence intervals of the Mann-Whitney parameter that are compatible with the Wilcoxon-Mann-Whitney test. unpublished manuscript 2018; .
- [10]. Chung E, Romano JP. Asymptotically valid and exact permutation tests based on two-sample u-statistics. *Journal of Statistical Planning and Inference* 2016; 168:97–105.
- [11]. Pauly M, Asendorf T, Konietzschke F. Permutation-based inference for the auc: A unified approach for continuous and discontinuous data. *Biometrical Journal* 2016; 58(6):1319–1337. [PubMed: 27502845]
- [12]. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences, An Introduction*. Cambridge University Press: New York, 2015.
- [13]. Volfovsky A, Airoldi EM, Rubin DB. Causal inference for ordinal outcomes. *arXiv preprint arXiv:1501.01234* 2015; .
- [14]. Ding P, et al. A paradox from randomization-based causal inference (with discussion). *Statistical science* 2017; 32(3):331–366.
- [15]. Hand DJ. On comparing two treatments. *The American Statistician* 1992; 46(3):190–192.
- [16]. Wu P, Han Y, Chen T, Tu X. Causal inference for Mann-Whitney-Wilcoxon rank sum and other nonparametric statistics. *Statistics in medicine* 2014; 33(8):1261–1271. [PubMed: 24132928]
- [17]. Lu J, Ding P, Dasgupta T. Sharp bounds of causal effects on ordinal outcomes. *arXiv preprint arXiv:1507.01542* 2015; .
- [18]. Fan Y, Park SS. Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory* 2010; 26(03):931–951.
- [19]. Rosenbaum PR. Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika* 2001; :219–231.
- [20]. Rosenbaum PR, et al. Covariance adjustment in randomized experiments and observational studies. *Statistical Science* 2002; 17(3):286–327.
- [21]. Huang EJ, Fang EX, Hanley DF, Rosenblum M. Inequality in treatment benefits: Can we determine if a new treatment benefits the many or the few? *Biostatistics* 2017; 18(2):308–324. [PubMed: 28025183]
- [22]. Heckman JJ, Smith J, Clements N. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies* 1997; 64(4):487–535.

- [23]. Ding P, Feller A, Miratrix L. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2016; 78(3):655–671.
- [24]. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted cox regression. *Statistics in medicine* 2009; 28(19):2473–2489. [PubMed: 19472308]
- [25]. McCullagh P, Nelder JA. *Generalized Linear Models*, second edition Chapman & Hall: London, 1989.
- [26]. Mullen GE, Ellis RD, Miura K, Malkin E, Nolan C, Hay M, Fay MP, Saul A, Zhu D, Rausch K, et al. Phase 1 trial of ama1-c1/alhydrogel plus cpg 7909: an asexual blood-stage vaccine for plasmodium falciparum malaria. *PLoS One* 2008; 3(8):e2940. [PubMed: 18698359]
- [27]. Shao J, Tu D. *The Jackknife and Bootstrap*. Springer: New York, 1995.
- [28]. Agresti A, Kateri M. Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics* 2017; 73(1):214–219. [PubMed: 27438478]
- [29]. Siddique Z Partially identified treatment effects under imperfect compliance: the case of domestic violence. *Journal of the American Statistical Association* 2013; 108(502):504–513.

Figure 1:
Toy example, treatment responses 2, 4, 6 and control responses 1, 3, 5.

Figure 2:

Bivariate distribution (left panel) and marginal distributions (right panel) for mixture of two normal distributions example. The marginal distributions are plotted as survival curves, for example $1 - F_0(y) = \Pr[Y(0) > y]$. The dotted gray line on the left panel is the line of equality for reference.

Figure 3:
Empirical distributions for the growth inhibition for the 3d7 strain for the three arms ($n = 23$ for v80, $n = 10$ for v20+CPG, and $n = 25$ for v80+CPG) [26]

Table 1:

Inferences on raw growth inhibition data from [26] (see Appendix C).

	CPG Effect	Dose Effect	Assumptions
$z_i = 0$	v80	v20+CPG	
$z_i = 1$	v80+CPG	v80+CPG	
p-value			$H_0 : F_0 = F_1$
WMW Test	$p < 0.0001$	$p = 0.956$	$H_1 : F_0 \neq F_1$
$\hat{\phi}$	0.960	0.508	
(95% CI)	(0.847, 0.990)	(0.310, 0.703)	proportional odds
$\widehat{E(D_i)}$	0.460	0.008	
(95% CI)	(0.347, 0.490)	(-0.190, 0.203)	proportional odds
sharp lower bound: $\hat{\psi}_L$	0.873	0.200	
(95% CI)	(0.743, 1.000)	(0.020, 0.520)	usual regularity *

* Assumptions needed for bootstrap to work, large sample sizes (see references in [27] p. 85).

Table 2:

Grow inhibition responses of Figure 3 grouped into 4 ordinal categories.

	low	moderate	substantial	strong
c80	17	4	2	0
c20+CPG	1	2	4	3
c80+CPG	0	8	6	11

Table 3:

Inferences on grouped growth inhibition data (see Table 2).

	CPG Effect	Dose Effect	Assumptions
$z_i = 0$	v80	v20+CPG	
$z_i = 1$	v80+CPG	v80+CPG	
p-value			$H_0 : F_0 = F_1$
WMW Test	$p < 0.0001$	$p = 0.588$	$H_1 : F_0 \neq F_1$
$\hat{\phi}$	0.934	0.558	
(95% CI)	(0.815, 0.977)	(0.363, 0.735)	proportional odds
$\hat{\phi}^*$	0.957	0.565	(latent continuous)
(95% CI)	(0.836, 1.000)	(0.344, 0.758)	proportional odds
$\widehat{E(D_i)}$	0.434	0.058	
(95% CI)	(0.315, 0.477)	(-0.137, 0.235)	proportional odds
sharp lower bound: $\hat{\psi}_L$	0.826	0.200	
(95% CI)	(0.696, 0.935)	(0.050, 0.490)	usual regularity*

* Assumptions needed for bootstrap to work, large sample sizes (see references in [27] p. 85).