CME

A Review of Analysis and Sample Size Calculation Considerations for Wilcoxon Tests

George Divine, PhD,* H. James Norton, PhD,† Ronald Hunt, MD,‡ and Jacqueline Dienemann, PhD, RN§

When a study uses an ordinal outcome measure with unknown differences in the anchors and a small range such as 4 or 7, use of the Wilcoxon rank sum test or the Wilcoxon signed rank test may be most appropriate. However, because nonparametric methods are at best indirect functions of standard measures of location such as means or medians, the choice of the most appropriate summary measure can be difficult. The issues underlying use of these tests are discussed. The Wilcoxon-Mann-Whitney odds directly reflects the quantity that the rank sum procedure actually tests, and thus it can be a superior summary measure. Unlike the means and medians, its value will have a one-to-one correspondence with the Wilcoxon rank sum test result. The companion article appearing in this issue of *Anesthesia & Analgesia* ("Aromatherapy as Treatment for Postoperative Nausea: A Randomized Trial") illustrates these issues and provides an example of a situation for which the medians imply no difference between 2 groups, even though the groups are, in fact, quite different. The trial cited also provides an example of a single sample that has a median of zero, yet there is a substantial shift for much of the nonzero data, and the Wilcoxon signed rank test is quite significant. These examples highlight the potential discordance between medians and Wilcoxon test results. Along with the issues surrounding the choice of a summary measure, there are considerations for the computation of sample size and power, confidence intervals, and multiple comparison adjustment. In addition, despite the increased robustness of the Wilcoxon procedures relative to parametric tests, some circumstances in which the Wilcoxon tests may perform poorly are noted, along with alternative versions of the procedures that correct for such limitations. (Anesth Analg 2013;117:699–710)

A common practice in designing and analyzing anesthesia and other clinical research projects is to assume that the data are normally distributed. When the data are ordinal with unknown differences between anchors (the response choices) and a small range such as 4 or 7, this assumption is inappropriate. Under these conditions, the Wilcoxon rank tests can provide meaningful analysis results. The Wilcoxon rank sum test (also called the Mann-Whitney *U* test) is the nonparametric alternative to the 2-sample *t* test. The Wilcoxon signed rank test is the nonparametric alternative to the paired *t* test. Starting with the example of a study of aromatherapy for postoperative nausea (PON) reduction published in this issue,[1] use of the Wilcoxon rank sum test and the Wilcoxon signed rank test is illustrated. Along with the issues surrounding the choice of a summary measure, considerations for the computation of sample size and power, confidence intervals, and multiple comparison adjustment are discussed. In addition, despite the increased robustness of the Wilcoxon procedures relative to parametric tests, there are circumstances in which the Wilcoxon tests may perform poorly. We describe alternative versions of the procedures that address such limitations.

THE ANALYSIS OF ORDINAL DATA

Many clinical research studies use measures with narrow integer outcome ranges such as a Visual Descriptive Scale with 4 anchors or a Likert-type scale with 5 or 7 anchors. For example, the aromatherapy PON trial cited[1] used a verbal descriptive scale with a range of 4 anchors: 0 = none, 1 = some, 2 = a lot, and 3 = severe. With only 4 possible values for the nausea scale and 6 possible values for its change, it is unreasonable to assume the outcomes will be consistent with the assumption of a normal distribution that goes with use of a *t* test. In addition, the difference between anchors can be unknown and unequal (e.g., the difference between *none* and *some* need not be exactly half the difference between *some* and *severe*). The mean might not be the optimal summary statistic for the data collected.

When an ordinal outcome variable is analyzed, hypothesis tests are often performed using a nonparametric procedure such as the Wilcoxon rank sum test (for 2 groups) or the Wilcoxon signed rank test (for paired data or for single group). Such analyses present unique issues for the data analyst to address and clinicians to understand. Foremost among such concerns is the determination of what summary measure(s) should be used to reflect the difference (or lack of difference) implied by a Wilcoxon test result. A common choice for a summary measure is the median. However, the connection between a difference in medians and the Wilcoxon rank sum test is more tenuous than is

From *Public Health Sciences, Henry Ford Hospital, Detroit, Michigan; †Dickson Advanced Analytics Group, Carolinas Medical Center, Charlotte; ‡Department of Anesthesia, Carolinas Medical Center University, Charlotte; and §School of Nursing, University of North Carolina Charlotte, Charlotte, North Carolina.

Accepted for publication November 15, 2012

The authors declare no conflicts of interest.Reprints will not be available from the authors.

Address correspondence to George Divine, PhD, Public Health Sciences, Henry Ford Health System, 1 Ford Place, 3E, Detroit, MI 48202-3450. Address e-mail to gdivine1@hfhs.org.

DOI: 10.1213/ANE.0b013e31827f53d7

often perceived. Consequently, summarizing with medians can lead to erroneous conclusions. Fortunately, there is a summary measure for the Wilcoxon rank sum test: the Wilcoxon-Mann-Whitney odds (the "WMWodds"), which has both (1) a direct connection to the Wilcoxon rank sum test statistic, and (2) an interpretation that is clinically relevant.

The situation with the Wilcoxon signed rank test is similar to that for the rank sum test, but with some additional difficulties. Its connection to the sample median is also tenuous, but a good alternative for a summary statistic is not available.

We present examples whereby the median falls short as a summary measure when either respective Wilcoxon test is used and discuss alternative approaches such as reporting and testing the means with either *t* tests or permutation tests.

WILCOXON RANK SUM TEST EXAMPLES

Because ordinal data can be ranked, a statistical test based on ranks, such as the Wilcoxon rank sum test, is a natural choice for generating a *P* value to test for a difference between 2 sets of ordinal values. Its application can be illustrated by two examples using data from the aromatherapy PON trial. Patients were eligible for inclusion in the trial if they reported any level of nausea (a PON score of 1, 2, or 3) after surgery. A research question of secondary interest for the trial involved risk factors for PON. One such risk factor is receipt of an opioid medication for reduction of postoperative pain. Figure 1 shows the distributions of PON scores for patients with and without exposure to opioids. As can be seen in the figure, more nausea was reported by the opioid-treated patients. The Wilcoxon rank sum test reflects this difference, $P = 0.001$. Notably, this Wilcoxon test confirms that there is a difference in nausea scores between the 2 groups, despite the fact that the median scores are the same.

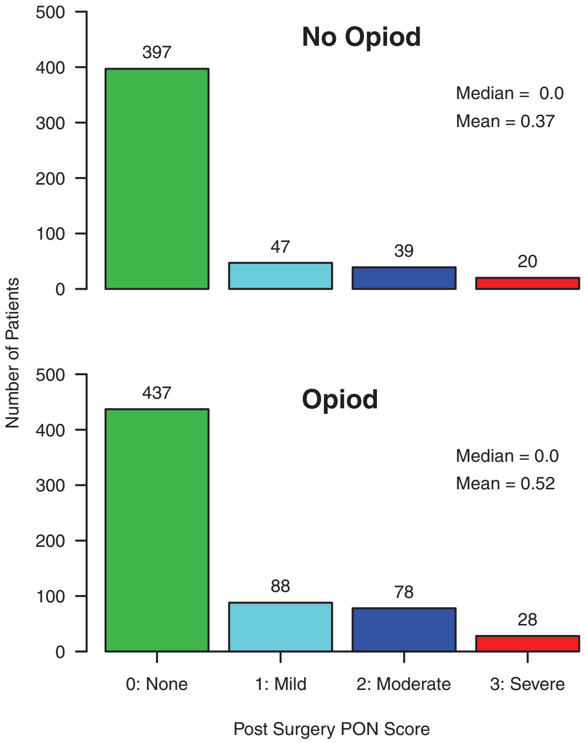

Figure 1. Distribution of postoperative nausea (PON) score by receipt of opioid medication.

A second example more directly relevant to the goals of the PON trial would be a comparison of the second and fourth panels of Figure 2, which show the posttreatment changes in nausea scores for the alcohol and blend aromatherapy groups. A greater shift toward improvement occurred with the blend treatment compared with alcohol (82% of blend-treated patients improved compared with only 51% of those treated with alcohol). The Wilcoxon rank sum test reflects this difference, giving $P < 0.001$. (Details of the test statistic calculations appear in Appendix 1, and a discussion of the special considerations raised by tied observations is found in Appendix 2.)

A summary statistic that reflects the difference implied by the Wilcoxon rank sum test result would be helpful. The mean might be used despite the limitations noted earlier, but the difference between means for the 2 groups may not correspond to the difference implied by the rank sum test result. The median is a common summary measure for an ordinal outcome, and the Wilcoxon rank sum test is frequently said to be a test of medians. Unfortunately, this may be just barely true in even a general sense and it can be just plain false in the most practical sense. In contrast to the median, the WMWodds can serve as a summary measure that is a direct function of what the Wilcoxon rank sum procedure tests. These points are illustrated by the examples above. In particular, the data show that the nausea scores for patients with and without opioid treatment, and the changes in the alcohol and blend aromatherapy groups, are different. However, the medians in each case are identical and equal to 0 or −1 for the respective pairs of groups being compared.

Hodges-Lehmann confidence interval estimates are frequently computed for differences in medians. However, for the differences between the median postsurgery nausea scores (Fig. 1) and their changes after treatment (Fig. 2), the Hodges-Lehmann confidence intervals are very inconsistent with the Wilcoxon rank sum test results. For the examples presented, the Hodges-Lehmann confidence intervals go from 0 to 0, and from −1 to 0, respectively. The first confidence interval is quite misleading and the second is indeterminant at best. In particular, the first interval suggests no effect for opioid, and the second suggests no difference between alcohol and blend. Given the shortcomings of point and interval estimates for the median, it is clear that a better summary measure is needed.

SUMMARY STATISTICS FOR THE WILCOXON RANK SUM TEST

p'' = Prob($X < Y$) and the "WMWodds"

Although means and medians may be unsuitable as summary measures for data tested using a Wilcoxon rank sum test, the test statistic does have a direct association with a quantity that can be useful. Appendix 1 reviews the direct correspondences among the rank sum statistic R, the Mann-Whitney *U* statistic, and p''. In particular, the test statistic can be formulated to be a function of the proportion of pairs of observations for which $X < Y$, where X and Y, respectively,

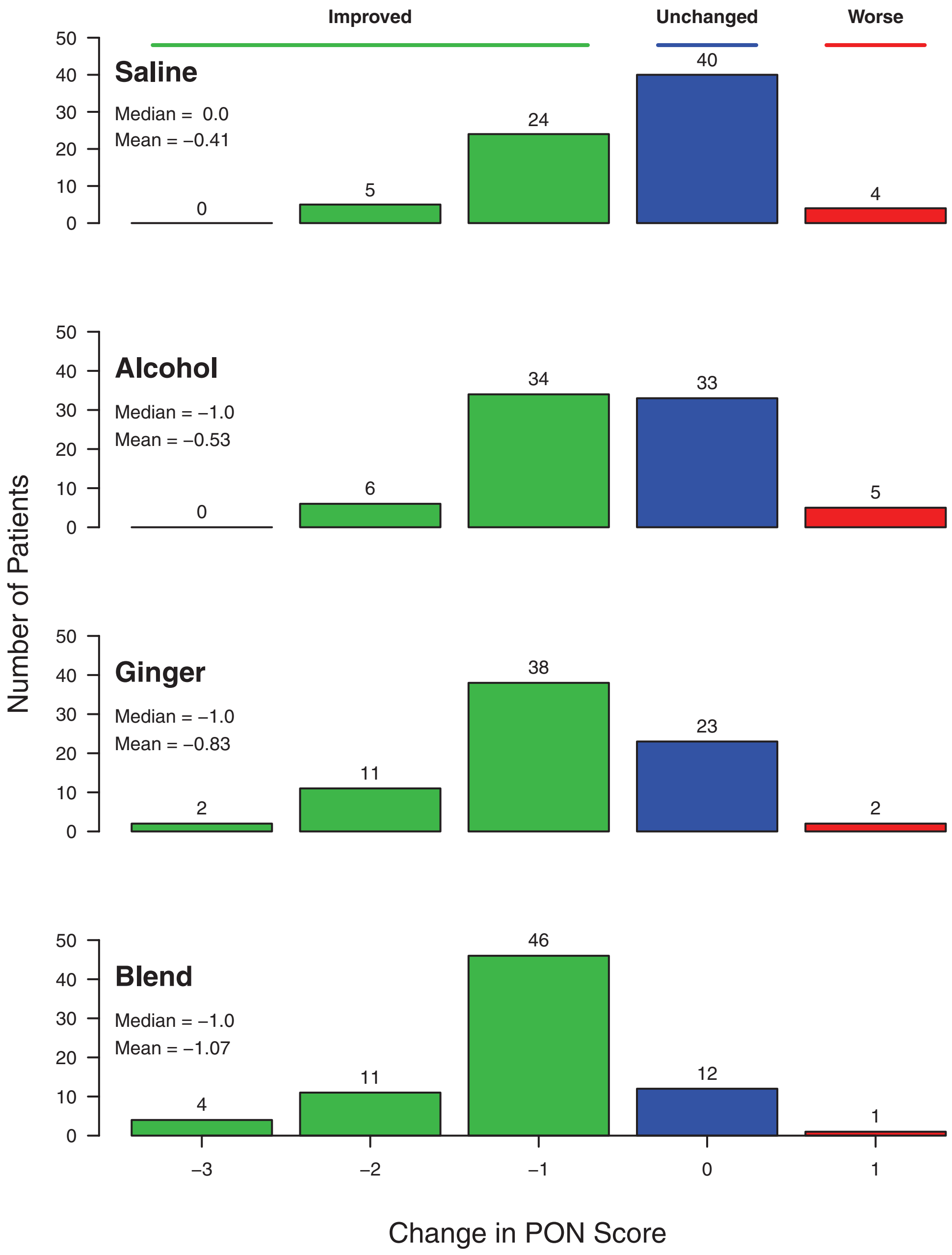

Figure 2. Changes in postoperative nausea (PON) score by aromatherapy group.

are random observations from the 2 groups being compared. This proportion is an estimate of the probability that a random observation from the second group will be larger than a random observation from the first. Under the null hypothesis, this probability will be 1/2. Following the notation used by Noether,[2] the symbol p″ will be used for the quantity $\Pr(X < Y)$. Noether only addressed the situation where the outcome variable is both continuous and has no ties. In practice, however, tied observations cannot be ignored, and they are counted as if they go half in one direction and half in the other. As an example of p″, for the comparison of the nausea score changes for alcohol versus blend, p″ is 0.68, which is substantially different from the null hypothesis value of 1/2.

Although the estimated probability that a patient who is given a blend aromatherapy will have a better nausea score than one treated with alcohol is clinically useful, its interpretation may be difficult to convey both succinctly and accurately. For instance, the statement that "the probability that a patient will report a more favorable nausea score change after aromatherapy with blend compared to after alcohol, is 0.68" may not suggest to the reader that this probability is relative to a null hypothesis value of 1/2, rather than relative to another potential null hypothesis probability that might also seem reasonable, such as 0. However, p″ can be transformed to an odds by use of the formula: $\text{odds} = p''/(1-p'')$. The null hypothesis odds of 1 (or 1:1) are more intuitive and clinically interpretable.

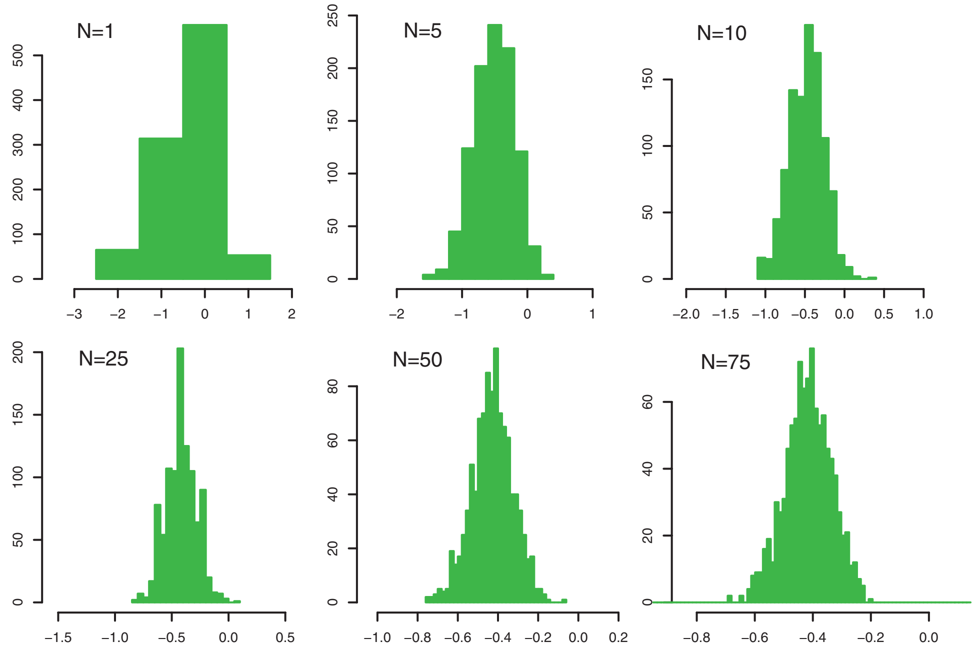

Figure 3. Distributions of mean postoperative nausea (PON) scores for 1000 random samples from the saline-treated group, with sample sizes of 1, 5, 10, 25, 50, and 75.

O'Brien and Castelloe[3] refer to this odds as the WMWodds[a] (pronounced "Whim Wads"), where "WMW" abbreviates "Wilcoxon-Mann-Whitney."

For the blend versus alcohol comparison in the aromatherapy nausea trial example, the 0.68 probability of less nausea with blend than with alcohol, translates to a WMWodds of 0.68/0.32 = 2.13. That is, the odds that a randomly selected subject from the blend-treated group has a larger reduction in nausea than a randomly selected subject from the alcohol-treated group are 2.13 to 1. The 95% confidence interval[b] for this WMWodds ranges from 1.50 to 3.17.

Clearly, the WMWodds of 2.13 is a better summary measure than the median difference of 0 in this scenario. Because the WMWodds is a one-to-one function of the Wilcoxon rank sum test's statistic, its magnitude and confidence interval directly reflect a clinically relevant quantity and the precision with which the quantity has been estimated. The strength of the WMWodds is such that it should always be considered as a summary measure for the Wilcoxon rank sum test. This is particularly the case for an ordinal outcome variable with a limited number of levels, when the median will likely fail, as in the example cited above. Despite the general utility of the WMWodds, there are other situations in which a mean or another summary measure might be preferred. The sections that follow discuss such alternatives.

ALTERNATIVE SUMMARY STATISTICS FOR THE WILCOXON RANK SUM TEST

Means

In the aromatherapy PON trial, the alcohol and blend aromatherapy groups had mean changes in the raw nausea scores (on the 0 to 3 scale) of −0.53 and −1.07, respectively. The difference (alcohol change minus blend change) between these means is 0.54 with a 95% confidence interval of 0.30, 0.78. This difference in means reflects the observed superiority of the blend group, but without a direct connection to the Wilcoxon rank sum test. (The next portion of this section further discusses the potential utility of using means for ordinal data.)

Two Options for Testing Means for Ordinal Data

Although a distributional form that is not normal may suggest that use of a *t* test is inappropriate, there are 2 paths to mitigating this constraint. The first is the central limit theorem, which essentially demonstrates that even though the distribution of individual observations is not normal, with an increasing sample size, the distribution of a mean (or the difference between 2 means) becomes more normal. The second path is through use of a permutation test based on the difference between means.

The Central Limit Theorem

To illustrate the behavior underlying the central limit theorem, a small simulation experiment is presented. Figure 3 shows the distributions of 1000 means for samples of size 1, 5, 10, 25, 50, and 75 drawn from the distribution of nausea score changes for the saline-treated group. As can be seen, with a small sample size, the distributions are lumpy, but as the sample size increases to 75 (i.e., the sample size for our main example), the distribution of the mean becomes very close to that expected for a normal distribution. Because the

[a] It should be noted that the WMWodds is a single odds and not an odds ratio. That is, although a single odds like the WMWodds is computed as the ratio of 2 *probabilities*, in biostatistics, an odds ratio is computed as the ratio of 2 distinct odds (and as such, an odds ratio is a function of 4 probabilities, not just 2).

[b] A confidence interval for the U statistic can be computed using ±2 times the standard error for U, and then computing the odds for U and the confidence interval limits. Alternatively, it can be noted that U is equal to the concordance index (the area under a receiver operating characteristic curve) and as was done here, PROC LOGISTIC in SAS can be used to compute a standard error and 95% confidence interval. For the standard error calculation, SAS uses a slightly better formula presented by DeLong et al.[4]

Table 1. Raw P Values by Asymptotic and Exact Wilcoxon Rank Sum Test, t Test, and Permutation Test on Means

	Asymptotic Wilcoxon rank sum test	Exact Wilcoxon rank sum test	*t* test	Permutation test on means
Saline versus alcohol	0.252	0.253	0.329	0.366
Saline versus ginger	<0.001	<0.001	<0.001	<0.001
Saline versus blend	<0.001	<0.001	<0.001	<0.001
Alcohol versus ginger	0.019	0.017	0.012	0.013
Alcohol versus blend	<0.001	<0.001	<0.001	<0.001
Ginger versus blend	0.076	0.073	0.080	0.097

t test is a function of a mean (or means), normality of the distribution of the mean is generally enough for a *t* test to perform well.

Permutation Tests and Permutation Tests on Means

A *P* value for a Wilcoxon rank sum test can be obtained either from a *z* (or χ^2) statistic as shown in Appendix 1, or by comparing the observed rank sum with tabled or computer-generated values from the exact distribution of that statistic under the null hypothesis. The exact distribution is derived from all possible permutations of the data for a given sample size. Such an exact permutation test is also possible for other statistics, including the difference between means.[c]

A permutation test on means does not require normally distributed data to be valid, so this option can be used when means are the most relevant summary statistic for the data being compared. For instance, cost data often are highly skewed, but if one condition decreases cost for most cases in a group, while increasing it drastically for a few others, basing an analysis on ranks (as the Wilcoxon does) might miss such an effect (or reach a contradictory conclusion). For testing for such an effect, a permutation test directly reflects the difference in means and allows a valid conclusion despite using data that are anything but normal.

The first 2 columns of Table 1 show *P* values from the standard asymptotic χ^2 statistic from the Wilcoxon rank sum test and its exact or permutation form for the 6 group-to-group comparisons in the aromatherapy PON trial. The third and fourth columns show *P* values for *t* tests and for permutation tests on the means, respectively.

Comparison of the *P* values in Table 1 shows that results for the Wilcoxon rank sum test and tests on means (*t* test and permutation test) are very similar, whether they are from the asymptotic tests or from the permutation versions of these procedures. In particular, all of the hypothesis test conclusions are the same.

Medians and the Wilcoxon Rank Sum Test

The perception that the Wilcoxon rank sum procedure is a test of medians is so broadly held that it is worthwhile to address exactly what connection (or lack of connection) exists between medians and the Wilcoxon. To do this, we review the issues involved.

Sample Medians or Population Medians?

A fundamental concept in data analysis is the difference between a *sample* and a *population*. In general, a data sample (or samples) is analyzed in order to try to reach conclusions about the population (or populations) from which the data are presumed to come. Often it is not essential to specify whether the results of an analysis refer to a sample or a population. However, when addressing the relationship between the Wilcoxon rank sum test and medians, the distinction is crucial. Specifically, the Wilcoxon rank sum procedure tests for a difference between population medians when it is assumed that a "shift alternative" holds. Under a shift alternative, everything about the 2 distributions being compared is the same: the shape, the variance, the skewness, etc., except for location. Mathematically, the shift alternative, the null hypothesis is stated as: H_o: Distribution F = Distribution G, and the alternative hypothesis is stated as: $H_a : G(x) = F(x + \Delta)$.

Under the shift alternative formulation, the 2 population medians differ by an amount that is equal to Δ. In this sense, the Wilcoxon rank sum procedure tests medians. Unfortunately, in many real-world situations such as those shown in Figures 1 and 2, the shift alternative assumption is unrealistic. This can be attributable to boundaries beyond which a shift cannot occur (e.g., an ordinal outcome such as a nausea score). This can also apply to an outcome like a count that cannot assume values less than zero, or intubation grade that cannot have a score less than 1 (i.e., the intubation cannot be easier than 1: easy). Thus, without a valid shift alternative possibility, the Wilcoxon rank sum test fails to be a test of population medians.

As noted earlier, in the aromatherapy trial, the sample median nausea scores observed for patients treated with and without opioids were both 0, and the median changes observed for the alcohol and blend groups both equaled −1. Yet, the Wilcoxon rank sum tests in both instances were highly significant. Either real-world example by itself is more than sufficient proof (a proof by counter-example) that the procedure is not a test based on the *sample* medians.

SAMPLE SIZE CALCULATION FOR THE WILCOXON RANK SUM TEST

The sample size calculation presented in the aromatherapy trial report reflects the computation performed in planning the study. That calculation used a formula derived by Noether[2] and is shown as Equation A3.1 in Appendix 3 of this report. For equal group sizes, α = 0.05 and power of 80% to detect a value of p″ equal to 0.63 requires a total sample size of 156, or 78 for each of 2 groups. Availability of the data now in hand could facilitate the planning for potential follow-up studies.

[c] In SAS, the procedure PROC NPAR1WAY will perform exact/permutation tests, and if the "exact scores=data" option is used, it will do a permutation test on the means of the variable of interest. This procedure was used to compute the *P* values shown in the fourth column of Table 1. The software package StatXact will do permutation tests for a wide range of analyses.

If the basic sample size calculation was repeated to consider multiple comparisons (i.e., for $\alpha = 0.05/3 = 0.017$), the sample size requirement would increase from 78 to 103. Alternatively, the value of p″ detectable with 80% power would increase from 0.63 to 0.65 for a sample size of 75 to 77 per group. However, because there are a large number of ties for each of the 6 possible changes (−3, −2, −1, 0, 1, or 2), and because such ties constrain how much variability can be seen in the sample, the variance will be smaller than expected for the formula used with a continuous outcome. An adjustment to the variance estimate used in the sample size calculation is appropriate. Given the observed proportions of ties for the different outcome (change) values, and using a formula by Zhao et al.[5] (simplified somewhat by Divine et al.[6]), the variance and the sample size requirement would be ~15% lower, so a sample size of 85 per group would yield 80% power to detect p″ = 0.63.

The data presented in the lowest 2 panels of Figure 2 could be used to plan a new clinical trial. For example, consider a 2-group trial comparing ginger versus blend. Let X represent the change in nausea score for a ginger-treated patient, and Y the change for a blend-treated patient. To 3 decimal places, the observed value for p″ = 0.576. Using formula A3.2, with α = 0.05, to achieve 80% power, the required sample size would be 226 per group if there were no ties. However, the proportions tied at the nausea scores of −3, −2, −1, 0, and 1 are 0.04, 0.146, 0.56, 0.233, and 0.02, respectively. The variance will be reduced by the sum of these terms cubed. The result of that calculation for these data is 0.0001 + 0.0031 + 0.1756 + 0.0127 + 0.000 = 0.1915. Therefore, after applying formula A3.3, the required sample size will be reduced by 19.15%, yielding a sample size requirement of 183 per group in order for the trial to have 80% power (details for these calculations are shown in Appendix 3). Finally, it should be noted that nQuery Advisor and the SAS procedure PROC POWER perform accurate sample size and power calculations for the Wilcoxon rank sum test, and both packages will appropriately account for ties.

ANALYSIS, SAMPLE SIZE, AND POWER CALCULATION FOR THE WILCOXON SIGNED RANK TEST

Analysis

Researchers should plan the design, outcomes, and analysis before the calculation of sample size and power. The importance of this sequence is shown in the aromatherapy PON trial. The study design required all randomized patients to report nausea, which constrained all the initial nausea scores to be larger than zero. The posttreatment scores were unconstrained and could include zero. Consequently, a "regression to the mean" bias would impact the changes observed for all 4 groups. Because of this bias, the null hypothesis of no change in nausea score in a particular group is of reduced interest. (Of course, this bias is one of the very strong reasons that use of a randomized control group was essential for the trial.)

The top panel of Figure 2 shows that the *saline*-treated controls had a large reduction in nausea scores. Whereas 40% of patients (29 of 73) improved (a negative change), only 5% (4 of 73) experienced more nausea (a positive change). This improvement was so predominant that a significant outcome for any formal statistical hypothesis test should be expected. For the Wilcoxon signed rank test, $P < 0.001$. For other applications, tests for within-group changes would be of value. Even for an ordinal outcome such as the nausea score, the Wilcoxon signed rank test is an obvious choice for such testing. Unfortunately, unlike the rank sum test, the signed rank test statistic is not a nice function of a single quantity that can be expressed in a form that is of direct clinical interest.

To review use of the signed rank procedure, its test statistic is computed by ranking the absolute values of the observations (changes in this case), summing the ranks of the positive observations and the negative observations, and using the smaller of those 2 quantities as the test statistic "T." (Note: A feature of the Wilcoxon signed rank test is that all the changes that are equal to zero are omitted from the calculation.) For small samples, T is compared with exact critical values either obtained from a published table or calculated by computer. For large samples, the expected value (under the null hypothesis) for T is subtracted from its observed value, and then divided by its standard error. The resulting *z* statistic is then used to find the *P* value.

A Summary Statistic for the Wilcoxon Signed Rank Test, p′ = Pr(X + X′ < 0)

If X and X′ are 2 random observations from the sample, using Noether's notation, 2 quantities, $\mathrm{p} = \Pr(\mathrm{X} < 0)$ and $\mathrm{p}' = \Pr(\mathrm{X} + \mathrm{X}' < 0)$, may be defined. The basic Wilcoxon signed rank test statistic T may then be expressed as:

$$\mathrm{T} = N\mathrm{p} + \tfrac{1}{2}N(N-1)\mathrm{p}' \qquad (1)$$

Where under the null hypothesis, both p and p′ are equal to 1/2.

Whereas p is easy to understand, the meaning of p′ is not as intuitive. Unfortunately, the term involving p′ contributes much more to the value of T than the one involving p (indeed, for very large *N*, the contribution of the p term is negligible). This is partly illustrated in Table 2, which shows the values of p, p′, T, and the signed rank test *z* statistic for the 4 groups in the companion article. As the table shows,

Table 2. Quantities Underlying the Wilcoxon Signed Rank Test

Group	*N*	p = Prob(X < 0)	p′ = P(X + X′ < 0)	T = *N*p + ½*N*(*N* − 1)p′	*z* statistic
Saline	33	0.878	0.898	503 = 28.97 + 474.1	6.40
Alcohol	45	0.889	0.904	935 = 40.0 + 895.0	7.89
Ginger	53	0.962	0.972	1390 = 51.0 + 1339.0	12.59
Blend	62	0.984	0.988	1929 = 61.0 + 1868.0	15.93

X and X′ are any 2 observations from the sample being tested.

the observed values of T track with p′ more closely than they track p, and this relationship becomes stronger with increasing sample size. Because p′ is not very satisfactory as a summary measure to go with the signed rank test, other choices should be considered.

The Mean as Summary Statistic, or as the Basis for an Alternative Test

Although it does not relate directly to the Wilcoxon signed rank test, the mean of the observations (usually the differences between paired values), may still be a reasonable summary statistic to report when the signed rank test is used. Alternatively, the central limit theorem may support the use of a 1-sample or paired *t* test, instead of the signed rank test. A permutation test for a difference between means is also an option. Although actual performance of a *t* test will depend on how much a distribution diverges from normal, it is common to find statistical textbook authors such as McClave and Sincich[7] suggesting that "for most sampled populations, sample sizes of $n \geq 30$ will suffice for the normal approximation to be reasonable."

A permutation test can also work well, even if normality cannot be assumed. However, if normality holds, a *t* test will yield more power than a permutation test.

The Median

If it is assumed that there is a symmetric distribution for the underlying population (usually of differences), the Wilcoxon signed rank procedure is a test of the population median. However, if we consider the changes for the saline group (Fig. 2), the sample median is equal to 0, and the 95% confidence interval is −1, 0. Yet, the signed rank test is quite significant $(P < 0.001)$. Once again, the data from the companion study show that the Wilcoxon signed rank procedure does not test the *sample* median.

Given the lack of an ideal general summary measure for the Wilcoxon signed rank test, one must choose among second-best alternatives. The choice depends on circumstances. For the aromatherapy trial, our first choice would be the percentage showing improvement. For instance, for the saline group, 29 of 73 patients (39.7%) improved, whereas only 4 of 74 (5.5%) had increased nausea after 5 minutes. A second option would be to report the mean changes. The saline group's mean change and its *t* statistic–based confidence intervals are −0.41 and −0.58, −0.25, respectively. Noting that a negative change implies a reduction in nausea score, the mean change tells the same story as the signed rank test.

For the saline group, reporting the median and its confidence interval contradicts the Wilcoxon signed rank test result (not to mention common sense). Despite the median's failure in this example, however, it can be suitable in other circumstances.

Sample Size Calculation

Relatively few software packages address sample size computation for the Wilcoxon signed rank test. The SAS procedure PROC POWER[8] does not, nor does *nQuery Advisor*. (Details for a sample size calculation for the Wilcoxon signed rank test are shown in Appendix 4.)

WILCOXON PROCEDURE LIMITATIONS AND VARIATIONS

A Limitation of the Standard Wilcoxon Rank Sum Test: Equality of Variances Required

Although the Wilcoxon rank sum test does not require normal distributions to be valid, there is a requirement that the variances for the 2 groups being compared be equal.[d] In 1964, Pratt[9] reported results of a simulation study that showed for normally distributed data, when there is a fourfold difference in variances and a sample size ratio of 4:1, the type I error rate under the null hypothesis can increase to >0.10, instead of the expected 0.05. This limitation cannot be overcome by use of the exact/permutation test version of the Wilcoxon rank sum. However, there are alternatives that can do a good job of preserving the type I error rate.

Fligner and Policello[10] and Brunner and Munzel,[11] respectively, proposed variations of the Wilcoxon rank sum test that adjust the Wilcoxon rank test statistic to consider the unequal variances. (The process is roughly analogous to the way a Welch's *t* test gives a more valid *t* test when variances are unequal.) In a simulation study with 10,000 simulated samples undertaken for another project, for the conditions studied by Pratt, instead of a type I error rate >0.10, the Fligner and Policello and Brunner and Munzel versions of the Wilcoxon rank sum test better approximated the required value of 0.05 (0.062 and 0.047, respectively). If the variances for the 2 groups to be compared are quite different, use of one of these alternative forms of the Wilcoxon rank sum test is recommended.[e]

A Limitation of the Wilcoxon Signed Rank Test: Observations at "Zero"

The Wilcoxon signed rank test excludes from the calculation any values (or paired differences) that are equal to zero. The rationale for this is that zeros provide no information about the direction in which a set of observations tend to go. However, removing the zeros may give rise to 2 concerns. One of the concerns is that removing the zeros can be unintuitive. The second, more serious concern is that removing the zeros can sometimes lead to results that are *illogical*.

To illustrate the unintuitive impact of removing zeros, consider the following pair of hypothetical research projects using the Intubation Difficulty Scale (IDS) to compare 2 different intubation devices. In the first project, there are 60 consecutive patients who are each intubated with 2 devices and the IDS for each device is obtained and the differences in IDS scores are compared. For 54 patients, the IDS scores are the same, but for the remaining 6, all have better (easier) scores for the second device and those 6 differences {IDS for Device 1 minus IDS for Device 2} are {1, 2, 3, 5, 8, and 9}. In the second project, just 6 consecutive patients are intubated with the 2 devices, and for all 6, the second device has a better IDS score and the differences in scores are again 1, 2,

[d] Although the term "distribution free" is sometimes used to describe nonparametric methods such as the Wilcoxon procedures, it is not the case that the distributions do not matter.

[e] There is a SAS macro written by Paul von Hippel that implements the Fligner-Policello test (http://www.sociology.ohio-state.edu/people/ptv/macros/fligner_policello.htm), and there is R code in the lawstat package that will do the Brunner-Munzel test (http://rss.acs.unt.edu/Rdoc/library/lawstat/html/brunner.munzel.test.html).

3, 5, 8, and 9. In both cases, the Wilcoxon signed rank test is significant, and the P values are identical ($P = 0.031$ in each case). Although this lack of a difference between the Wilcoxon signed rank test results is unintuitive, this reflects a feature of how the Wilcoxon signed rank test works more than a limitation. One perspective on the test behavior is to consider that dropping zeros is intended to optimize sensitivity to statistical significance rather than to clinical or practical significance. That is, the test focuses only on those observations that provide some information about the difference between the 2 devices. The statistical significance is in contrast to the practical difference in the 2 results, where using device 2 is of benefit only 10% of the time in the first study, whereas it is always beneficial in the second study.

The second difficulty with removing zeros in computing the Wilcoxon signed rank test was noted by Pratt.[12] He observed that for some sets of observations that include one or more zeros, shifting all observations slightly to one side can change the Wilcoxon signed rank test statistic to a new value that would imply that the data had shifted in the direction opposite to the one that actually occurred. As an example, Pratt used the set {−18, 0, 2, 3, 4, 6, 7, 8, 9, 11, 14, 15, 17}, which clearly has a predominance of positive numbers. The signed rank test result is $P = 0.034$, reflecting this predominance. Pratt noted, however, that if all the observations were reduced by 1, the resulting set of numbers {−19, −1, 1, 2, 3, 5, 6, 7, 8, 10, 13, 14, 16} gives a more significant $P = 0.028$, despite the decrease in every observation. To address this illogical result, Pratt proposed an alternative form of the Wilcoxon signed rank test (usually referred to as Pratt's test) that includes any zeros at the initial step when the observations are ranked, and omits them only when the test statistic is computed in the next step. For the 2 sets of data above, Pratt's test gives $P = 0.0239$ and $P = 0.0281$, respectively, which do align properly with the shift in the data. Although Pratt's test is not provided in standard commercial statistical software packages such as SAS and SPSS, SAS macros for it exist,[f] and it is available with the wilcoxsign_test program in the R "coin" package. (Finally, it is important to note that Pratt's test does not address the unintuitive issue with zeros, as Pratt's test gives that same P value for both hypothetical IDS studies $[P = 0.031]$.)

MULTIPLE COMPARISONS AND CONFIDENCE INTERVALS

Multiple Comparisons

Although the aromatherapy clinical trial can be regarded as a trial of 3 aromatherapy agents undertaken in parallel, it can also be interpreted as a trial of aromatherapy as a general treatment method, with 3 separate opportunities to succeed. To the extent that the latter is or might be true, there is a need to protect against the inflated potential type I error rate. Ideally, addressing this issue would take place in the planning stage of the trial and the sample size would be adequate to maintain power. However, in this instance, it was first addressed during the manuscript review process, when it was decided to protect against a type I error for the 3 primary tests involving each of the 3 active arms versus the control/saline group. It should be noted that given the strength of the effects observed, the question of whether or not additional power is needed is a moot issue.

Selecting the "Family" for the Multiple Comparisons Adjustment

A fundamental concept underlying multiple comparison adjustment involves the family of comparisons for which the adjustment is needed and applied. Most textbooks assume that knowing which comparisons belong in the family is easy or obvious, so they explain only how the α level is adjusted and maintained. However, selecting the family can be at least as important as these other, more mathematical details. For an observational study, identifying the number of comparisons belonging to a family can be particularly challenging because many of the comparisons made will depend on the data observed or on intermediate analysis results. Even for a formal randomized clinical trial, this issue may not be clear cut. For instance, there were 6 group-to-group comparisons made in the aromatherapy PON trial design.

Because the trial had 3 aromatherapy agents that were each expected to be potentially beneficial, plus a control (saline) group, the family of comparisons selected for the primary analysis consisted of the 3 individual active treatment group comparisons against the control group. This is precisely the family addressed by Dunnett's multiple comparisons procedure[13] that tests for a difference between a control group and each of N treatments. Dunnett's test makes use of the distribution and correlations expected among N such comparisons between means within a 1-way analysis of variance. The correlation arises because the difference between mean A and mean B must be correlated with the difference between mean A and mean C, as both differences are a function of mean A. However, this correlation is not appropriate for multiple comparisons among Wilcoxon rank sum tests where the rank sums have a somewhat different correlation structure.[g]

Selecting the Multiple Comparison Method

As noted above, Dunnett's test is not quite appropriate for comparisons when rank-based methods are used, because there is additional correlation among the rank sum due to the groups being ranked together. That is, ranking the groups together implies that if one or more groups have high ranks, another group or groups must have lower ranks. Steel[14] gives a method that incorporates such extra correlation. However, implementation of Steel's method is not easy, and in this situation, generic multiple comparisons methods such as Bonferroni[13] or one of its variations, such as those of Holm[15] or Hochberg,[16] are often used. The Bonferroni adjustment simply divides α by the number of comparisons. For instance, with Bonferroni, in order to maintain the type I error rate at 0.05 for 3 comparisons, a P value must be less than $0.05/3 = 0.017$ to be considered significant. Bonferroni is easiest to implement, but the simplicity comes at a cost of some power.

Some procedures such as Holm's (also called Holm-Bonferroni) can restore some power; however, under the Holm framework, for 3 comparisons if the smallest P value is <0.017, the next smallest need only be less than $0.05/2 = 0.025$

[f] See http://www.egms.de/static/de/journals/mibe/2010-6/mibe000104.shtml.

[g] It should be noted that for an analysis of 4 groups, the Kruskal-Wallis procedure might be used to test for an overall difference among the groups, and the multiple comparisons might be performed by comparing groups based on their ranks when all 4 are ranked together. However, because the aromatherapy for PON study involves a placebo arm and 3 potentially active treatments, comparing the groups 2 at a time was considered more appropriate than using the Kruskal-Wallis procedure.

in order to be considered significant. In that case, the largest *P* value would only need to be <0.05 to be significant. The Hochberg method can provide even more power, in that with Hochberg, all 3 *P* values will be considered significant as long as the largest is <0.05 (even if the smallest is not <0.017). Also, with Hochberg, even if the largest *P* value is not <0.05, as long as the larger of the remaining 2 is <0.025, both may be considered significant. Finally, even if the second largest number is <0.025, the last *P* value will still be significant if it is <0.017.

For the aromatherapy PON trial, Bonferroni was selected as the most straightforward method applicable. With 3 primary comparisons and an overall α of 0.05, the *P* value required for significance for any single comparison is 0.017 = 0.05/3. Accordingly, the confidence level was set at 98.3% instead of 95% for the 3 primary comparisons. If the Holm or Hochberg method had been used, corresponding confidence intervals would have been difficult[17] or impossible[18] to compute.

Confidence Intervals

Whether adjusted for multiple comparisons or not, the *P* value is used to determine whether a result is statistically significant. However, it can be just as important to determine practical significance, where an estimate of the effect of interest and the precision of that estimate are directly relevant. In the aromatherapy PON trial, the estimated odds of having a greater reduction in nausea with ginger versus saline were 1.86 with a 98.3% confidence interval of 1.22 to 3.00. The confidence level was set at 98.3% (= 1 – 0.05/3) to reflect the Bonferroni adjustment for the 3 primary comparisons. The lower confidence interval bound for these odds, 1.22, is large enough that even a pessimistic interpretation of these results would support at least a modest superiority of ginger over saline.

DISCUSSION

Using examples from the aromatherapy PON trial published in this month's issue of *Anesthesia & Analgesia*, we have demonstrated that Wilcoxon tests are not functions of sample medians. The sample medians can perform poorly as summary measures for an ordinal scale that has unknown differences and a short range (i.e., a small number of possible values). This shortcoming is attributable to the fact that the median and its confidence interval estimates are constrained to assume either one of the observable values or a point midway between 2 of them. Essentially, this does not leave room for the medians or the Hodges-Lehmann confidence interval estimates to have the resolution to show clear differences between the groups. Another circumstance in which the median performs poorly is for count data with a preponderance of zeros (zero inflation).

A suitable alternative summary statistic exists that is clinically relevant and directly related to what the Wilcoxon rank sum procedure actually tests: the odds that a patient receiving one treatment will do better than one treated with another. These odds are a function of $p'' = \text{Prob}(X < Y)$. They can be used as the basis for sample size calculation for Wilcoxon rank sum testing and to give a clinically relevant summary measure: the WMWodds.

We also illustrated that the Wilcoxon signed rank test does not have such a nice potential summary quantity underlying it. Instead, Wilcoxon signed rank test is a function of the quantity $p' = \Pr(X + X' < 0)$, whose interpretation is not intuitive. Selection of a good summary measure for the signed rank test depends on the particular data being compared. Graphical presentation, therefore, is especially helpful. In this instance, changes in each of the 4 groups are shown in the respective 4 panels of Figure 2.

Despite the robustness of the Wilcoxon rank sum test, it can perform poorly when the variances and sample sizes of the groups being compared are quite different. For such data, researchers should consider using either the Fligner-Policello or Brunner-Munzel methods if this is a characteristic of their data. Similarly, the Wilcoxon signed rank test may not behave well when the distribution of values (differences) being tested includes many zeros. In that case, Pratt's test should be considered.

APPENDIX 1: THE ONE-TO-ONE RELATIONSHIPS AMONG THE WILCOXON RANK SUM, THE MANN-WHITNEY U STATISTIC, AND p″ = Prob(X < Y) AND THE WMWodds

The Mann-Whitney U_1 statistic associated with group 1 may be computed from the rank sum R_1, for group 1, using the formula:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \qquad \text{(A1.1)}$$

$p'' = \text{Prob}(X < Y) = U_1 / n_1 n_2$ where n_1 and n_2 are the sample sizes of the 2 groups being compared. The $\text{WMWodds} = p''/(1-p'')$. Thus, p″ and the WMWodds can be expressed as direct functions of the standard Wilcoxon rank sum test statistics.

The calculations are illustrated below for the alcohol versus blend comparison. The Wilcoxon rank sum χ^2 test statistic is equal to:

$$X^2 = \left[\frac{R_1 - E(R_1)}{se(R_1)}\right]^2 \qquad \text{(A1.2)}$$

where R_1 is the rank sum in group 1, $E(R_1) = n_1(N+1)/2$ is its expected value under the null hypothesis, and the standard error of $R_1[se(R_1)]$ is equal to:

$$\sqrt{\frac{n_1 n_2(N+1)}{12}\left[1 - \frac{\sum_{i=1}^{k} t_i^3 - t}{N^3 - N}\right]} \qquad \text{(A1.3)}$$

where n_1 and n_2 are the sample sizes in groups 1 and 2 and $N = n_1 + n_2$, k is the number of categories with ties, and t_i is the number of observations (from both groups combined) tied at the *i*th level.

Filling in values for this example as shown in Table A1, we have n_1 = 78, n_2 = 74, N = 152, R_1 = 7008.5, and $E(R_1) = 78(152+1)/2 = 5967$. The standard error is equal to:

$$\sqrt{\frac{(78)74(152+1)}{12}\left[1 - \frac{608318 - 152}{3511808 - 152}\right]} = \sqrt{\frac{883116}{12}[1 - 0.173]} = \sqrt{60844.61} = 246.67 \qquad \text{(A1.4)}$$

Table A1. Example of Computational Details for the Wilcoxon Rank Sum and Mann-Whitney *U* Tests

			Rank sum calculations				Tie calculations		*U* statistic calculations		
A	B		C	D	E		F	G	H	I	J
Nausea score change	Patient count		Rank range	Average rank	Rank sums		Tie count	Tie count cubed	No. blend < alcohol	No. blend = alcohol	*U*
	Alcohol	Blend			Alcohol	Blend					
−3	0	4	1–4	2.5	0	10	4	64	4*78 = 312	0	312
−2	6	11	5–21	13	78	143	17	4913	11*72 = 792	11*6 = 66	792 + 66/2 = 825
−1	34	46	22–101	61.5	2091	2829	80	512,000	46*38 = 1748	46*34 = 1564	1748 + 1564/2 = 2530
0	33	12	102–146	124	4092	1488	45	91,425	12*5 = 60	12*33 = 396	60 + 396/2 = 258
+1	5	1	147–152	149.5	747.5	149.5	6	196	0	5	5/2 = 2.5
Sum	78	74			7008.5	4619.5	152	608,318			3867.5

The χ^2 statistic is equal to $[(7008.5 - 5967)/246.67]^2 = [(1041.5)/246.67]^2 = [4.22]^2 = 17.8$

The Mann-Whitney U χ^2 test statistic is equal to:

$$\chi^2 = \left[\frac{U_1 - E(U_1)}{se(U_1)}\right]^2 \qquad \text{(A1.5)}$$

where U_1 is the number of times an observation from group 1 (alcohol) is less than an observation in group 2 (blend) (column H in Table A1) plus one-half the number of ties (column I). (That is, ties are treated as if they are half above and half below each other.) $E(U_1) = (n_1)(n_2)/2$, and the standard error of U_1 is the same as the standard error of R_1.

For the example, we have $U_1 = 3867.5$, $E(U_1) = 2886$, and $se(R_1) = 246.67$. The χ^2 statistic is equal to $[(3867.5 - 2886)/246.67]^2 = [(1041.5)/246.67]^2 = 17.8$.

APPENDIX 2: TIES AND WILCOXON TESTS

There are 3 main considerations regarding ties and Wilcoxon tests: (1) the impact of ties on the numerator of the test statistic (i.e., reflecting the impact of ties on the estimate of the magnitude of the difference of interest), (2) the impact of the ties on the variance (denominator) of the test statistic, and (3) for the Wilcoxon signed rank test applied to differences between paired observations, pairs that are "tied" with each other result in a zero difference.

Ties and Wilcoxon Test Statistics (Numerators)

As illustrated in Appendix 1 for the Wilcoxon rank sum test, tied observations are given the average of the range of ranks they would have been assigned if they could have been ordered. For instance, the 4 blend-treated patients who had a 3-point decrease in their nausea scores (−3) each gets a rank of 2.5, so that their contribution to the rank sum for the blend group is 4 times 2.5 = 10, which is the same as if they had received the ranks 1 + 2 + 3 + 4 = 10. Because there is no reason to penalize the blend group for having these ties, this equivalence is reasonable and appropriate. The same principle holds when the tied observations involve both groups. For instance, for a 2-level decrease in nausea score (−2), 17 patients are tied with this score change (6 in the alcohol-treated group and 11 who were blend-treated). Each of them is assigned the average of the ranks from 5 to 21 (i.e., 13).

Analogously, for the Mann-Whitney U statistic calculations, the 4 blend group observations with a nausea score change of −3 will be lower than all 78 of the scores for the alcohol group whether they are tied or not. The 17 patients tied with a nausea score change of −2 will all be counted as being less than patients with scores of −1, 0, and 1, but they are also treated as being half above and half below each other [i.e., the 11 blend-treated patients with a change of −2 contribute 11(72) = 792, *plus* ½(11)6]. Thus, the total contribution from those 11 patients to the U statistic total is 792 + 33 = 825.

For the signed rank test, the absolute values are ranked, and tied observations are assigned the average rank for the range of ranks involved. Again, the rank total will be the same as it would have been if there had been no ties. (The signed rank test treatment of one or more observations equal to zero is addressed following the next section.)

All in all, other than the special case of zeros for the Wilcoxon signed rank test, the rank sum *total* levels are unchanged whether or not ties are present. However, this is not true for the variance of the test statistics.

Ties and Wilcoxon Statistic Variances

The presence of ties in the observed distribution results in less room for the values to spread out or to have variance. Therefore, the variance must be reduced to take this into account. This variance reduction for the Wilcoxon rank sum (or Mann-Whitney U) test is represented by the right-hand components of Equations A1.3 and A3.3. These terms:

$$1 - \frac{\sum_{i=1}^{k} t_i^3 - t}{N^3 - N}$$

and

$$(1 - \sum_{i=1}^{k} P_i^3) \qquad \text{(A2.1)}$$

are equivalent.

The variance reduction due to ties for the Wilcoxon signed rank test has the same form, but it is only one-fourth as large, i.e., it is:

$$(1 - \sum_{i=1}^{k} P_i^3 / 4) \qquad \text{(A2.2)}$$

Zero Differences for the Wilcoxon Signed Rank Test

Paired observations that are tied, resulting in a difference of zero, are given special treatment in the Wilcoxon signed rank test. Although it can be unintuitive, such zero differences are assumed to provide no information about whether or not the observations tend to be above or below zero, thus they are excluded from the calculation of the signed rank statistic. This treatment of differences that do not go one way or the other is analogous to how McNemar's test works for paired binary outcomes.

APPENDIX 3: SAMPLE SIZE CALCULATION FOR THE WILCOXON RANK SUM TEST

Noether's formula[2] for Wilcoxon rank sum test sample size is

$$N=\frac{(Z_{\alpha/2}+Z_{\beta})^2}{12c(1-c)(p''-0.5)^2} \qquad \text{(A3.1)}$$

where c is the ratio of the sample size for the X population (m) relative to that for the Y population (n) (i.e., $N = m + n$ and $c = m/N$). If m = n, c = 0.5 and the formula becomes:

$$N=\frac{(Z_{\alpha/2}+Z_{\beta})^2}{3(p''-0.5)^2} \qquad \text{(A3.2)}$$

For $\alpha = 0.05$, $Z_{\alpha/2} = 1.96$, and for 80% power $(\beta = 0.20)$, $Z_{\beta} = 0.84$, so $Z_{\alpha/2} + Z_{\beta} = 2.8$ and $(Z_{\alpha/2} + Z_{\beta})^2 = 7.84$.

As shown by Zhao et al.[5] and simplified somewhat by Divine et al.,[6] to take into account the ties that can occur with an ordinal outcome variable (or that must occur when there are only a limited number of possible outcome values), the formula becomes:

$$N=\frac{(Z_{\alpha/2}+Z_{\beta})^2}{3(p''-0.5)^2}\left(1-\sum_{i=1}^{k}P_i^3\right) \qquad \text{(A3.3)}$$

where k is the number of outcome categories, and P_i is the proportion of all of the observations tied in the i^{th} category. For instance, for a comparison of ginger versus blend, if it is assumed that a new trial would have tied proportions similar to those reported by Hunt et al.,[1] these proportions would be approximately 0.04, ~0.147, 0.56, ~0.233, and 0.02 (= 6/150, 22/150, 84/150, 35/150, 3/150), and the sum of the cubed terms becomes 0.000 + 0.003 + 0.176 + 0.013 + 0.000 = 0.192.

To illustrate the calculations of the sample size requirement for a potential trial with a goal of determining definitively whether blend aromatherapy is superior to ginger alone, if p″ = 0.576 as was observed in the current trial, the total sample size for both groups combined that would give 80% power with $\alpha = 0.05$ is:

$$\begin{aligned}N&=\frac{(Z_{\alpha/2}+Z_{\beta})^2}{3(p''-0.5)^2}\left(1-\sum_{i=1}^{k}P_i^3\right)=\frac{7.84}{3(0.576-0.5)^2}(1-0.192)\\&=\frac{7.84}{3(0.076)^2}(0.808)=365.6\approx 366 \qquad (5)\end{aligned}$$

or 183 per group.

APPENDIX 4: SAMPLE SIZE CALCULATION FOR THE WILCOXON SIGNED RANK TEST

The numerator (T) for the Wilcoxon signed rank test can be expressed as:

$$T = Np + ½N(N-1)p'$$

where $p = \Pr(X > 0)$ and $p' = \Pr(X + X' > 0)$, and $p = p' = ½$ under the null hypothesis. Asymptotically, the term $p' = \Pr(X + X' > 0)$ will dominate, so a reasonable approximation can be based on $p' = \Pr(X + X' > 0)$. Specifically, Noether's sample size formula[2] is:

$$N=\frac{(Z_{\alpha/2}+Z_{\beta})^2}{3(p'-0.5)^2} \qquad \text{(A4.1)}$$

As with the rank sum test, the variance is reduced by ties, but the adjustment is only one-fourth as large in this case.

$$N=\frac{(Z_{\alpha/2}+Z_{\beta})^2}{3(p'-0.5)^2}\left(1-\sum_{i=1}^{k}P_i^3/4\right) \qquad \text{(A4.2)}$$

If we needed to select a sample size to detect a change similar to that seen for the saline group, formula A4.2 would apply. There the observed value of p′ was 0.897, and the tied proportions were 5/33 at an absolute value of 2 (−2) and 28/33 at an absolute value of 1 (−1 or 1). When these tie proportions are cubed, the result is $(5/33)^3 + (28/33)^3 = 0.0035 + 0.6109 = 0.6144$. Inserting these into the above formula, to get 80% power, the required *usable* sample size is $7.84/[3(0.397)^2](1 - 0.6144/4) = [7.84/3(0.158)](1 - 0.154) = 16.54(0.846) = 14.0$. Because 40 of 73 of the saline changes were zero and therefore are omitted from the Wilcoxon signed rank test calculation, the usable *N* must be inflated to account for the proportion of omitted observations expected. In this case, the final target sample size will need to be 14/[40/73] = 25.6 ≅ 26 in order to get the 14 usable observations needed to give 80% power for the hypothesized reduction.

Twenty-six is a relatively small sample size, but it is consistent with the large reduction in nausea scores observed and assumed to hold for this calculation.

DISCLOSURES

Name: George Divine, PhD.
Contribution: This author helped analyze the data and write the manuscript.
Attestation: George Divine approved the final manuscript.
Name: H. James Norton, PhD.
Contribution: This author helped analyze the data and write the manuscript.
Attestation: H. James Norton approved the final manuscript.
Name: Ronald Hunt, MD.
Contribution: This author helped analyze the data and write the manuscript.
Attestation: Ronald Hunt approved the final manuscript.
Name: Jacqueline Dienemann, PhD, RN.
Contribution: This author helped analyze the data and write the manuscript.
Attestation: Jacqueline Dienemann approved the final manuscript.
This manuscript was handled by: Franklin Dexter, MD, PhD.

REFERENCES

1. Hunt R, Dienemann J, Norton HJ, Hartley W, Hudgens A, Stern T, Divine G. Aromatherapy as treatment for postoperative nausea: a randomized trial. Anesth Analg 2012 Mar 5. [Epub ahead of print]

2. Noether GE. Sample size determination for some common nonparametric tests. J Am Stat Assoc 1987;82:645–7
3. O'Brien RG, Castelloe JM. Exploiting the link between the Wilcoxon-Mann-Whitney test and a simple odds statistic. Proceedings of the Thirty-First Annual SAS Users Group International Conference 2006. Cary, NC: SAS Institute Inc., 2006:209–31
4. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45
5. Zhao YD, Rahardja D, Qu Y. Sample size calculation for the Wilcoxon-Mann-Whitney test adjusting for ties. Stat Med 2008;27:462–8
6. Divine G, Kapke A, Havstad S, Joseph CL. Exemplary data set sample size calculation for Wilcoxon-Mann-Whitney tests. Stat Med 2010;29:108–15
7. McClave JT, Sincich T. Statistics. 9th ed. New Jersey: Prentice, 2003
8. SAS Institute Inc. 2004 SAS/STAT 9.1 Users Guide. Cary, NC: SAS Institute Inc., 2004
9. Pratt JW. Robustness of some procedures for the two-sample location problem. J Am Stat Assoc 1959;59:665–80
10. Fligner MA, Policello GEII. Robust rank procedures for the Behrens–Fisher problem. J Am Stat Assoc 1981;76:162–8
11. Brunner E, Munzel U. The nonparametric Behrens–Fisher problem: asymptotic theory and a small-sample approximation. Biometrical J 2000;42:17–25
12. Pratt JW. Remarks on zeros and ties in the Wilcoxon signed rank procedures. Je Am Stat Assoc 1959;54:655–67
13. Miller RG Jr. Simultaneous Statistical Inference. New York: Springer-Verlag, 1981
14. Steel RGD. A multiple comparison rank sum test: treatments versus control. Biometrics 1959;15:560–72
15. Holm S. A simple sequentially rejective Bonferroni test procedure. Scand J Stat 1979;6:65–70
16. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800–2
17. Dmitrienko A, Tamhane AC, Bretz F, eds. Multiple Testing Problems in Pharmaceutical Statistics. New York: CRC Press, 2009
18. Strassburger K, Bretz F. Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. Stat Med 2008;27:4914–27