Joint Hypothesis Testing and Gatekeeping Procedures for Studies with Multiple Endpoints

Edward J. Mascha, PhD,* and Alparslan Turan, MD†

A claim of superiority of one intervention over another often depends naturally on results from several outcomes of interest. For such studies the common practice of making conclusions about individual outcomes in isolation can be problematic. For example, an intervention might be shown to improve one outcome (e.g., pain score) but worsen another (e.g., opioid consumption), making interpretation difficult. We thus advocate joint hypothesis testing, in which the decision rule used to claim success of an intervention over its comparator with regard to the multiple outcomes are specified a priori, and the overall type I error is protected. Success might be claimed only if there is a significant improvement detected in all primary outcomes, or alternatively, in at least one of them. We focus more specifically on demonstrating superiority on at least one outcome and noninferiority (i.e., not worse) on the rest. We also advocate the more general “gatekeeping” procedures (both serial and parallel), in which primary and secondary hypotheses of interest are a priori organized into ordered sets, and testing does not proceed to the next set, i.e., through the “gate,” unless the significance criteria for the previous sets are satisfied, thus protecting the overall type I error. We demonstrate methods using data from a randomized controlled trial assessing the effects of transdermal nicotine on pain and opioids after pelvic gynecological surgery. Joint hypothesis testing and gatekeeping procedures are shown to substantially improve the efficiency and interpretation of randomized and nonrandomized studies having multiple outcomes of interest. (Anesth Analg 2012;114:1304–17)

Comparative efficacy or effectiveness studies frequently have more than one primary outcome, and often several secondary. In designing such studies it is crucial to a priori specify the decision rule(s) that will be used to claim success of one intervention over another. Not doing so makes interpretation difficult and invites post hoc decision making. Just as important is the ability to make inference on the multiple endpoints without either increasing the chance for type I error (i.e., false positives) or overadjusting for multiple comparisons, which reduces power. Each of these goals can be achieved using joint hypothesis testing and gatekeeping methods.

When several outcomes are selected as primary, for either a nonrandomized or randomized study, a customary approach is to test hypotheses about the outcomes in isolation and make separate conclusions about each. For instance, in pain management studies, researchers might hypothesize that a new treatment reduces both pain score and opioid consumption.1,2 However, reaching a clear conclusion is difficult if results are discordant. For example, practitioners would not typically deem a pain management intervention better than its comparator if it reduced average pain at the cost of substantially increasing average opioid consumption, even though some patients might be content with that tradeoff on an individual basis. A joint assessment of the outcomes with a priori rules for claiming success is preferred to avoid confusion in interpretation and the increased type I error due to multiple comparisons.

Alternatively, researchers sometimes claim a single outcome as primary and relegate other important outcomes as secondary, perhaps driven by the engrained goal of choosing a single primary outcome. Not infrequently, though, 2 or more outcomes are equally important. For example, pain score and opioid consumption are so interrelated that reporting the treatment effect for one while either demoting the other to secondary status or ignoring it altogether is often insufficient. In assessing gabapentin’s effect on pain management in arthroscopic rotator cuff repair, for example, researchers chose a primary outcome of visual analog scale pain score and secondary outcomes of fentanyl consumption and side effects.3 They conclude that pain score (primary outcome) was reduced and fentanyl consumption (secondary outcome) was no different between groups. However, because conclusions should be based on primary outcome(s), a better approach would be to a priori claim both pain and fentanyl consumption as joint primary outcomes, and assess them jointly.

We focus on the joint testing of pain and opioid consumption as our main illustrative example throughout (see “Illustrative Data Example”) because these outcomes are strongly clinically related and yet are often not analyzed together. However, the discussed methods apply to any study for which efficacy or effectiveness of one treatment versus another depends on results for multiple outcomes.
For studies with multiple primary outcomes we demonstrate in this paper how joint hypothesis testing can facilitate the desired inference while protecting type I error (see section JOINT HYPOTHESIS TESTING). For example, success of an intervention might be claimed only if there is significant improvement detected in all primary outcomes (see “Option 1: Superiority on All Outcomes”). However, requiring all to be significant may be unnecessarily stringent in some studies; more practical may be to restrict claims of success to interventions that are at least no worse on any outcome. Therefore, we focus on assessing whether a treatment is superior over its comparator on at least one of the primary outcomes and not worse (i.e., noninferior) on the rest (see “Option 2: Noninferiority on All Outcomes, Superiority on at Least One”).4–6

Finally, we discuss the more general gatekeeping procedures, in which primary and secondary hypotheses of interest are a priori organized into ordered sets, and the significance level for a particular set depends on results from previous sets in the sequence (see “Gatekeeping Procedures”).7,8

### ILLUSTRATIVE DATA EXAMPLE: THE NICOTINE STUDY

In a randomized controlled trial, Turan et al.9 tested the hypothesis that transdermal nicotine would decrease postoperative pain and opioid analgesic usage, thereby improving the early recovery process after pelvic gynecological surgery.

#### Table 1. Two-Tailed Superiority Testing for Nicotine Study Primary and Secondary Outcome (Turan et al.2)

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Nicotine (N = 43)</th>
<th>Placebo (N = 42)</th>
<th>Nicotine − placebo (95% CI)</th>
<th>2-tailed superiority P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average pain</td>
<td>1.5 ± 0.9</td>
<td>1.1 ± 0.8</td>
<td>0.33 (−0.05, 0.70)</td>
<td>0.086</td>
</tr>
<tr>
<td>Total morphine</td>
<td>29 (22, 37)</td>
<td>38 (28, 50)</td>
<td>−7.6 (−13.6, −2.4)†</td>
<td>0.013</td>
</tr>
<tr>
<td>Log2 (total morph)</td>
<td>4.9 ± 0.6</td>
<td>5.3 ± 0.7</td>
<td>−0.35 (−0.60, −0.10)</td>
<td>0.023</td>
</tr>
<tr>
<td>Ratio of geometric means</td>
<td>N/A</td>
<td>N/A</td>
<td>0.78 (0.64, 0.97)</td>
<td>0.023</td>
</tr>
<tr>
<td>Secondary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality of recovery 72 hours</td>
<td>17.3 ± 1.2</td>
<td>15.6 ± 1.4</td>
<td>1.7 (1.1, 2.3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Satisfaction score 72 hours</td>
<td>96.5 ± 6.1</td>
<td>95.8 ± 5.8</td>
<td>0.68 (−1.48, 2.84)</td>
<td>0.60</td>
</tr>
<tr>
<td>Complete satisfaction 72 hours, no. (%)</td>
<td>31 (72)</td>
<td>26 (62)</td>
<td>10.2 (−9.7, 30)*</td>
<td>0.36</td>
</tr>
<tr>
<td>Oral intake (hours)</td>
<td>33.1 ± 9.5</td>
<td>28.6 ± 7.6</td>
<td>4.5 (1.35, 7.55)</td>
<td>0.019</td>
</tr>
<tr>
<td>Ambulation (hours)</td>
<td>20.7 ± 4.6</td>
<td>18.1 ± 2.9</td>
<td>2.53 (1.14, 3.92)</td>
<td>0.003</td>
</tr>
<tr>
<td>D/C score 72 hours†</td>
<td>3.9 ± 0.3</td>
<td>3.7 ± 0.4</td>
<td>0.19 (0.06, 0.32)</td>
<td>0.018</td>
</tr>
<tr>
<td>Bowel sounds (hours)</td>
<td>20.9 ± 6.2</td>
<td>20.2 ± 6.0</td>
<td>0.65 (−1.56, 2.85)</td>
<td>0.63</td>
</tr>
<tr>
<td>Flatus (hours)</td>
<td>29 [21, 35]</td>
<td>24 [22, 29]</td>
<td>2 (−1, 6)*</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Data are presented as mean ± so, 50th (25th, 75th) percentiles or N (%), and the P values are from the t test, Wilcoxon rank sum test, or chi-square test, as appropriate. CI = confidence interval; N/A = not applicable.

†D/C score: higher is better.

Clearly, the amount of correlation among the outcomes matters clinically as well as statistically. For example, the same treatment effect—say, a 20% reduction in mean opioid consumption versus control—might be more clinically important if the correlation between the outcomes was negative for the intervention group (i.e., higher opioid consumption associated with lower pain score) and positive for the control group (higher opioid consumption associated with higher pain score) than if the correlations were the same between groups. Using data from several of our studies we found the correlation between pain score and opioid consumption to range from almost no correlation to moderate (−0.50) (Appendix 1). We use methods that take advantage of the correlation among the outcomes, but focus on the marginal treatment effects, i.e., the difference in means or proportions for each outcome of interest.

Secondary outcomes were overall quality of recovery, patient satisfaction with pain management, resumption of normal activities of daily living, and recovery of bowel function. We use this study to demonstrate our methods throughout, and refer to it as the nicotine study.

In the original study, no significant nicotine effect was found for either pain or opioids in tests for superiority. For illustrative purposes we have modified the results so that nicotine reduces opioid consumption and also improves quality of recovery score (as might be expected with fewer opioids) at 72 hours postsurgery. All else was left as in Turan et al. Table 1 thus shows that nicotine reduces opioid consumption but not pain score versus placebo. Nicotine was also found worse on ambulation and oral intake and better on discharge score at 72 hours. Instead of analyzing pain and opioid consumption with 2-tailed superiority tests as in Turan et al. (and Table 1), and with no adjustment for analyzing multiple outcomes, we will use these data to demonstrate joint hypothesis testing methods.

### JOINT HYPOTHESIS TESTING

In joint hypothesis testing, individual hypotheses are combined into a larger testing framework, such that an intervention is only deemed preferred over its comparator(s) if it fulfills an a priori set of criteria involving all hypotheses in the framework. For example, criteria for success might be (1) superiority on all outcomes; (2) noninferiority, i.e., “not worse,” on all and superiority on at least one; or, less commonly, (3) superiority on some with no restrictions on the rest (not discussed further). We briefly discuss the first option, “Superiority on all outcomes,” and then devote most of this section (and the paper) to a more in-depth discussion of the second, “Noninferiority on all outcomes, superiority on at least one.”

#### Option 1: Superiority on All Outcomes

Superiority of intervention A over B might be claimed only if A is deemed superior to B on all outcomes, for example, on both pain score and opioid consumption. Any such joint hypothesis test requiring significance on all outcomes in the set being tested is called an intersection–union test (IUT). Here and throughout, K is the number of hypotheses being tested, and particular null and alternative hypotheses are H0i and H1i, respectively.
HAi, respectively, for i = 1 to K. In an IUT the null hypothesis is the union (read “or”) of several hypotheses, such that at least 1 null hypothesis is true (i.e., at least 1 of H01 or H02 or H03 or ... H0k is true). The alternative is the intersection (read “and”) of several hypotheses, such that all null hypotheses are false (i.e., each of H01 and H02 and H03 and ... H0k is false).

For example, in equation (1) below, the joint null hypothesis is that the Experimental treatment E is the same or worse than Standard care S on either mean pain or mean opioid consumption, and the alternative is that Experimental is superior to Standard (i.e., lower) on both.

H0: PainE ≥ PainS OR H0: OpioidE ≥ OpioidS

HA: PainE < PainS AND H02: OpioidE < OpioidS (1)

So for this joint hypothesis, superiority of Experimental to Standard will be claimed only if Experimental is found superior on both outcomes. The null and alternative hypotheses for an intersection–union test are thus

H0: At least one of nulls is true (Union) (2)

HA: All of nulls are false (Intersection)

For any IUT, because all null hypotheses must be rejected to claim success, no adjustment to the significance criterion for multiple comparisons is needed. In the above scenario, for example, if each of the 2 hypotheses are tested using α = 0.025, the type I error of the joint hypothesis test is still controlled at 0.025. Intuitively, only 1 decision is being made, superiority or not on all outcomes, so there is only 1 chance to make a false positive conclusion.

In our modified version of the nicotine study data, only opioid consumption was significantly improved for the nicotine group at the 0.025 significance level (P = 0.023), so that a joint hypothesis requiring both to be significant would not have concluded nicotine to be more effective. Turan et al. did not specify whether significance on both pain and opioids was required to claim nicotine more effective than placebo. In retrospect, they would have said “no,” that superiority on either outcome would have sufficed, as long as not worse on either as well. Such a design is the focus of “Option 2: Noninferiority on All Outcomes, Superiority on at Least One” below.

**Direction of Testing and Significance Level (α)**

In equation (1) we need to test all outcomes in the same direction, i.e., with a 1-tailed test, because we will only claim superiority of one intervention over the other if it is superior on all. For ease of explanation, we focus on assessing whether Experimental is superior to Standard. Often, however, testing would be done in the other direction as well, i.e., whether Standard is superior to Experimental, because the true effect is not known, and an effect in either direction would typically be of interest. To test the other direction, we simply repeat the 1-tailed hypothesis testing in Equation (1) after switching the Experimental and Standard therapies. The overall α for testing in both directions is double the α for a single direction. Throughout, we use α of 0.025 for a single direction, implying a combined α of 0.05 if both directions are tested.

**Sample Size and Power**

Requiring all tests to be significant can reduce power over requiring only a single test to be significant. For example, a study with 90% power to detect superiority in each of 2 outcomes in a joint hypothesis test has only 81% power to reject both nulls (i.e., 0.9 × 0.9 = 0.81). Using 0.949 for each test (since 0.025 would give approximately 90% power for the joint test with noncorrelated outcomes, and somewhat greater (or lesser) than 90% to the extent positively (or negatively) correlated. These same properties hold for requiring noninferiority on all outcomes in “Joint Hypothesis Testing, Option 2.”

A counter intuitively, though, is that no adjustment to the significance criterion for multiple testing is needed in an IUT, as noted above. For example, in our nicotine study, n = 86 patients are needed in a traditional design to have 90% power at the 0.025 significance level to detect differences of 5.7 mg (SD = 8) on opioids and 1.1 (SD = 1.5) on pain in separate 1-tailed tests. However, a needed Bonferroni correction for multiple testing (significance criterion = α/2 = 0.0125) boosts N to 102. For the joint hypothesis test, each outcome is appropriately tested at level α (0.025); 90% power to find superiority on both requires n = 106 if the outcomes are uncorrelated, and less for positive correlations.

Finally, in many studies, superiority is not required on all primary outcomes to claim success. When superiority on any one of the outcomes is sufficient, an accounting for multiple comparisons needs to be made, either by a completely closed testing procedure or through a multiple-comparison procedure. Both of these methods are addressed in the section “Step 2: Superiority on at Least One Outcome” below.

**Option 2: Noninferiority on All Outcomes, Superiority on at Least One**

Overview

We now focus on joint hypothesis testing methods to handle the setting where Experimental is preferred to Standard (or vice versa) only if it is found superior on at least one of the primary outcomes and not worse (i.e., noninferior) on the rest. Such a design is attractive because it ensures that “no harm” is done by the intervention concluded to be better; it must be shown to be at least as good as its competitor on each primary outcome, and better on one or more outcomes. These methods give substantially better interpretation than traditional designs in which multiple related outcomes are analyzed independently.

When assessing noninferiority on all outcomes and superiority on at least one, the whole procedure is an IUT (intersection-union test), because both noninferiority and superiority are required. Specifically, the joint null hypothesis is that either noninferiority on all outcomes or superiority on at least one does not hold (i.e., one of the null hypotheses H01 or H02 is true in equation (3) below), and the alternative is that both noninferiority on all outcomes (HA1) and superiority on at least one (HA2) are true, as follows:

H01: E inferior to S on n ≥ 1 OR H02: E superior on none

HA1: E noninferior to S on all AND HA2: E superior S on ≥ 1 (3)
In our pain management example, the alternative hypothesis (which we hope to conclude) is that Experimental is noninferior to Standard on both pain score and opioid consumption (HA1) and superior on at least one of them (HA2). Because both noninferiority and superiority are required, no adjustment for testing 2 sets of hypotheses is required. Each of noninferiority and superiority are tested at level \( \alpha \). The benefit of not needing to adjust the significance criterion is somewhat counterbalanced by the increased power needed to find both noninferiority on all and superiority on some outcomes.

Figure 1 summarizes in an intuitive way the main methods presented in this section. Displayed are the observed confidence intervals for the treatment effect in our nicotine study for both opioids and pain score. Noninferiority delta (NI delta) here is 1 for visual analog scale pain and a ratio of geometric means of 1.2 for opioid consumption. Direction of arrows would be reversed when low values of outcome desirable. CI = confidence interval.

**Step 1: Noninferiority on All Outcomes**

In this first of 2 steps in our joint hypothesis test in equation (3), noninferiority of one treatment over another—i.e., the state of being “at least as good” or “not worse”—is claimed only if that treatment can be shown to be noninferior on every individual outcome. Because noninferiority must be found on all outcomes, the noninferiority testing itself is also an IUT. Therefore, noninferiority testing on each outcome is conducted at the overall level \( \alpha \) (type I error) chosen a priori for noninferiority testing (hypothesis 1 in equation (3) above).

For each outcome of interest, say the \( \text{th} \) outcome is conducted at the overall level \( \alpha \) (type I error) chosen a priori for noninferiority testing (hypothesis 1 in equation (3) above).

\[
H0i: \text{Experimental is at least } \delta \text{ worse, or } \mu_{Ei} - \mu_{Si} \geq \delta_i
\]

versus

\[
HA1_i: \text{Experimental } < \delta \text{ worse, or } \mu_{Ei} - \mu_{Si} < \delta_i
\]

where \( \mu_{Ei} \) and \( \mu_{Si} \) are the population means on experimental and standard interventions, respectively, for the \( \text{th} \) outcome. The IUT for noninferiority in equation (3) including all outcomes is thus simply

\[
H0: \text{Lack of Noninferiority on at least one outcome (Union)}
\]

in which the intersection alternative hypothesis (HA1) indicates that noninferiority is required on all outcomes to reject the null hypothesis.

Assuming that smaller outcome values are desirable (e.g., lower pain better than higher), noninferiority testing for the \( \text{th} \) outcome is most intuitively assessed by simply observing whether the upper limit of the confidence interval (CI) (e.g., 95% CI for \( \alpha = 0.025 \)) for the difference in means (Experimental minus Standard) lies below the noninferiority \( \delta \).

**Step 2: Superiority on at least One Outcome**

In this second step, superiority of Experimental to Standard is assessed either using individual tests alone (Step 2-A) or with a combination of an overall “global” test and individual tests (Step 2-B). In Step 2-B we discuss a powerful “closed testing” procedure that allows testing of individual outcomes for superiority while making no adjustment for multiple comparisons, given that all global tests including the particular outcome are also significant. At each step and overall, the type I error is preserved at level \( \alpha \). If noninferiority of Experimental to Standard is found on all outcomes in Step 1, and superiority on at least one outcome is found in Step 2, we reject the joint null hypothesis in equation (3) and claim that Experimental is preferred over Standard.

**Noninferiority on All Outcomes**

**Superiority on at least One Outcome**

**Joint Hypothesis Testing and Gatekeeping Procedures**

**Volume 114 • Number 6 www.anesthesia-analgesia.org 1307**

Copyright © 2012 International Anesthesia Research Society. Unauthorized reproduction of this article is prohibited.
In the nicotine study we conclude noninferiority of nicotine to placebo for both pain and opioids because the upper limit of each confidence interval is below the corresponding noninferiority deltas (Fig. 1). For pain score, for example, we observe mean (SD) of 1.4 (0.9) and 1.1 (0.8) for the nicotine and placebo patients, respectively, for a mean difference of 0.33 with 95% CI of −0.05 to 0.71. We conclude noninferiority of nicotine to placebo on pain score because the upper limit of the CI, 0.71 is below the noninferiority δ of 1.0 point on the pain scale (Fig. 1, Table 2). In other

---

**Figure 2.** Algorithm for testing noninferiority on all outcomes and superiority on at least one. If noninferiority is detected on all outcomes in step 1, superiority testing begins. If global superiority is not required, individual outcomes are tested for superiority in step 2 using a multiple comparison procedure (path A). Otherwise, if global superiority is required and detected, individual superiority is assessed either through a completely closed testing procedure at level α (path C, detailed in Fig. 3) or using a multiple-comparison procedure (path B). The joint null hypothesis (i.e., equation (3)) is rejected, concluding noninferiority on all outcomes and superiority on at least one, if at least one individual superiority test is significant. Analysis path must be chosen a priori. *In practice, “any” can be any number from 1 to the number of outcomes (K), but must be set a priori.
words, we conclude that nicotine is < 1 point worse than placebo on pain score.

We test noninferiority of nicotine to placebo on opioid consumption using a noninferiority δ of 1.2 for the ratio of means (technically geometric means [or medians] because the data are log-normal). A ratio δ of 1.2 implies an alternative hypothesis for which mean opioid consumption for nicotine is no more than 20% higher than placebo. The null and alternative hypotheses are thus

$$\text{H}_01 ({\text{null}}): \frac{\mu_\text{E}}{\mu_\text{S}} \geq 1.2$$

versus

$$\text{H}_1 (\text{alt}): \frac{\mu_\text{E}}{\mu_\text{S}} < 1.2$$

where \(\mu_\text{E}\) and \(\mu_\text{S}\) are the geometric means of opioid consumption for the nicotine and placebo groups, respectively. As is noted in Table 2, after back-transforming from the log scale, we observe a ratio of geometric means (95% CI) of 0.79 (0.64, 0.97). Because the upper limit of 0.97 is less than the noninferiority δ of 1.2, we reject the null and claim noninferiority of nicotine to placebo on opioid consumption (Fig. 1, Table 2).

In addition to a confidence interval, a test giving a \(P\) value for noninferiority is usually desired. This is done, for example, by rearranging terms and expressing the alternative hypothesis \(\text{H}_1\) in equation (4) as \(\text{H}_1\): \(\frac{\mu_\text{E}}{\mu_\text{S}} - \delta < 0\), and constructing a 1-tailed \(t\) test to assess whether the difference in means is below the given \(\delta\). A significant \(P\) value from the 1-tailed \(t\) test will always correspond to the upper limit of the confidence interval falling below the noninferiority \(\delta\), as is seen in the noninferiority \(P\) values in Table 2 for both pain and opioids, both significant at \(P < 0.025\). In Appendix 2 we give details on conducting these \(t\) tests for noninferiority using the nicotine study data. See also Mascha and Sessler (2011).14

Because noninferiority is concluded for both pain and opioid consumption in the nicotine study, the intersection-union null hypothesis \(\text{H}_01\) for noninferiority in equation (3) is rejected, and superiority can then be assessed on each outcome in Step 2, which follows below.

**Step 2: Superiority on at Least One Outcome**

If and only if noninferiority is detected on all outcomes in Step 1, superiority testing is conducted to assess whether Experimental is superior to Standard on any of the outcomes. We present 2 acceptable but quite different methods for superiority testing: (1) individual testing using a stepwise multiple-comparison procedure (Step 2-A) and (2) global testing followed by tests of individual hypotheses (Step 2-B).15 Global testing involves testing for overall superiority across the outcomes in a single 1-tailed multivariate test (although it can be repeated in the other direction). As shown by Troendle and Legler (their Tables 6 and 7)16 in simulation studies, the choice of method makes a noticeable difference in type I error and power. Individual testing procedures tend to be conservative in type I error (i.e., fewer false positives than the planned \(\alpha\)). Power, the proportion of null hypotheses correctly rejected, generally increases for global methods in relation to stepwise as the proportion of true effects increases. With only 2 outcomes, as in our pain and opioids example, the global procedure is expected to be more powerful if superiority truly exists for both outcomes, whereas the individual testing procedure may be more powerful if superiority exists for only 1 of the 2.

Both methods are explained in detail below, and the algorithms depicted in the flow chart in Figure 2. Although we discuss the case in which “any” (≥ 1) outcome being superior is sufficient, these procedures generalize seamlessly to requiring a prespecified number of significant outcomes, anywhere from 1 to K. As always, it is important to decide on a method a priori to avoid choosing on the basis of the observed results.

**Step 2-A: individual superiority testing with no global test.** In many studies a global assessment of superior treatment effect across the outcomes is a priori deemed not important or not interesting, and superiority testing of individual outcomes suffices. In such studies, Figure 2 (path “A” under “No” to “Global superiority required?”) is followed. Rejection of the null hypothesis for any outcome would indicate superiority of Experimental to Standard, and thus rejection of the joint NI-superiority hypothesis in equation (3), because noninferiority on all outcomes has already been shown.

Assessing individual outcome superiority without requiring a significant global test may be the most powerful and appropriate method when outcomes are either only moderately correlated or substantively dissimilar, or when a homogeneous treatment effect across the outcomes is not

---

**Table 2. Nicotine Study: 1-Tailed Noninferiority and Superiority Tests**

<table>
<thead>
<tr>
<th>Outcome (72 hours)</th>
<th>Nicotine ((n = 43))</th>
<th>Placebo ((n = 42))</th>
<th>Difference ((N - P) 95% CI))</th>
<th>Delta</th>
<th>(P) value</th>
<th>Difference ((N - P) 97.5% CI))</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain score</td>
<td>1.5 (0.9)</td>
<td>1.1 (0.8)</td>
<td>0.33 (−0.05, 0.71)</td>
<td>1</td>
<td>0.002</td>
<td>0.33 (−0.10, 0.76)</td>
<td>0.96</td>
</tr>
<tr>
<td>Opioids (mg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw</td>
<td>33 (16)</td>
<td>44 (26)</td>
<td>−10.8 (−20.2, −1.3)</td>
<td>N/A</td>
<td>N/A</td>
<td>−10.8 (−22, 0.09)</td>
<td>N/A</td>
</tr>
<tr>
<td>Log (_{10})</td>
<td>4.9 (0.63)</td>
<td>5.3 (0.75)</td>
<td>−0.35 (−0.65, −0.05)</td>
<td>0.26</td>
<td>&lt;0.001</td>
<td>−0.35 (−0.69, −0.006)</td>
<td>0.011</td>
</tr>
<tr>
<td>Ratio of medians</td>
<td>N/A</td>
<td>N/A</td>
<td>0.79 (0.64, 0.97)</td>
<td>1.2</td>
<td>&lt;0.001</td>
<td>0.79 (0.62, 0.99)</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Data are from a modified version of the nicotine study (Turan et al., see Table 1). Overall alpha is 0.025 for both noninferiority and superiority testing. Noninferiority was found on both outcomes with the given noninferiority deltas (both \(P < 0.025\), and superiority on opioids \((P < 0.025/2 = 0.0125)\). \(N = \) nicotine; \(P = \) placebo. Results presented as mean (so), median [1st, 3rd quartiles], or mean (confidence interval). Ratio of medians is for nicotine/placebo.

\(a\) Noninferiority: significant if upper confidence limit (in boldface type) is less than delta.

\(b\) Superiority: significant if upper confidence limit (in boldface type) is less than 0 for pain and log-transformed opioids, and less than 1 for ratio of geometric means.

June 2012 • Volume 114 • Number 6  www.anesthesia-analgesia.org 1309

Copyright © 2012 International Anesthesia Research Society. Unauthorized reproduction of this article is prohibited.
expected. Researchers may simply want to know for which distinct outcome(s) an intervention is superior over standard. This would certainly be the case for our pain management example.

Superiority of Experimental to Standard is assessed for each individual outcome using appropriate 1-tailed tests, conducted in the same direction as noninferiority testing in Step 1. Distinct statistical methods can be used for different outcome types. For example, pain score might be assessed with a t test if normally distributed, whereas opioid consumption (say, in morphine equivalents) might be assessed using the Wilcoxon rank sum test because the data would likely be nonnormal. A Pearson χ² test might be used for binary outcomes.

The null and alternative hypotheses for testing superiority of Experimental to Standard on the ith outcome in the superiority portion of equation (3) can be expressed as

\[ H_0^i: \mu_{EI} - \mu_{SI} \leq 0 \]

versus

\[ H_A^i: \mu_{EI} - \mu_{SI} > 0. \]  

(7)

Because superiority will be claimed if Experimental is found superior to Standard on at least one outcome, the individual superiority testing, i.e., H2 in equation (3), is a union–intersection test. In a union–intersection test the null hypothesis is the intersection of all nulls (i.e., nonsuperiority on all outcomes), and the alternative is the union (read “or”), such that at least one null is false (superiority on at least one). It is thus the opposite of the IUT discussed earlier.

Because superiority on any of the outcomes will suffice, a multiple-comparison procedure is needed to maintain the overall type I error at level α. The traditional Bonferroni method in which all P values are compared to α/K is quite conservative (e.g., for K = 5 and α = 0.025, P values <0.005 are considered significant), especially if the treatment effects are fairly homogenous. Instead, we recommend the less conservative Holm–Bonferroni stepdown procedure,17 especially as the number of outcomes increases. Observed P values from the K univariable tests are first ordered from smallest to largest, as \( p_1 \leq p_2 \ldots \leq p_K \). If \( p_1 \) is significant at \( \alpha/K \), then \( p_2 \) is compared to \( \alpha/(K - 1) \), and then \( p_3 \) to \( \alpha/(K - 2) \), etc., until the largest \( P \) value, given that all others were significant, is compared to \( \alpha \). The sequential testing stops as soon as any \( P \) value is nonsignificant.

Using the Holm–Bonferroni procedure, the smallest of the 2 \( P \) values for pain score and opioid consumption would be tested at \( \alpha/2 \) (e.g., at 0.0125 for \( \alpha = 0.025 \)), and if significant, the other at 0.025. For example, with 2 \( P \) values of 0.012 and 0.024, superiority would be claimed on both using the Holm–Bonferroni method. Using the traditional Bonferroni method, only the first would be declared significant. In a more extreme example, suppose 5 \( P \) values ordered from smallest to largest of 0.004, 0.006, 0.007, 0.01, and 0.024 were observed. All would be significant using the Holm–Bonferroni method, but only the first (\( P < 0.005 \)) with the traditional Bonferroni method. The Holm–Bonferroni procedure is thus preferred.

In our nicotine study, the smallest 1-tailed superiority \( P \) value for the 2 primary outcomes (\( P = 0.011 \) for opioids) is smaller than \( \alpha/2 = 0.0125 \), so nicotine is declared superior to placebo (Table 2). No difference was found for pain score in a 1-tailed test (\( P = 0.96 \)). We use 97.5% CIs for these 1-tailed superiority tests because we are only interested in the upper limit and its \( \alpha = 0.0125 \), which sums to an overall \( \alpha = 0.025 \) for 2 tests (pain and opioids). Because noninferiority was claimed for both nicotine study outcomes in Step 1, and superiority was claimed here for at least 1 of the 2, nicotine is claimed more effective than placebo.

In general, we would reject the joint null hypothesis in equation (3) and claim Experimental better than Standard if superiority is claimed for any of the individual outcomes, because noninferiority was already claimed on all outcomes in Step 1.

Step 2-B: global superiority followed by individual tests. A second method to assess superiority is to first require a significant overall or “global” test for superiority of Experimental to Standard across the K outcomes. Individual testing follows only if the global test is significant (see “Individual superiority testing following significant global test,” below). Requiring a significant global association is intuitive when it is important to demonstrate superiority of Experimental to Standard across the set of outcomes, even if superiority on individual components cannot be shown.

For example, in the NINDS t-PA (tissue plasminogen activator) trial for ischemic stroke, the trial sponsor and the U.S. Food and Drug Administration required the drug to be beneficial across a vector of 4 neurological outcomes with a global test.18 For a similar reason, requiring global significance would make sense in some composite endpoint situations or when the endpoint consists of the components of a disease activity index (a score calculated from several variables that together summarize the degree of the targeted disease).15 However, in studies similar to our pain management example, researchers may typically not require global significance for superiority, because superiority on either pain or opioids would suffice.

Step 2-B1: global superiority testing. Global superiority across the outcomes typically consists of a 1-sided multivariate test (i.e., analyzing multiple outcomes per patient) at level α to assess the overall treatment effect across components, conducted in the same direction as noninferiority testing in Step 1. The 1-sided multivariate test for superiority against a global null hypothesis of no treatment effects can be expressed as

\[ H_0: \Delta = 0 \text{ versus } H_A: \Delta > 0, \]  

(8)

where \( \Delta \) is the set of population treatment effects between Experimental and Standard for the K outcomes (e.g., \( \mu_{EI} - \mu_{SI}, \mu_{E2} - \mu_{S2}, \ldots, \mu_{EK} - \mu_{SK} \)).

Care must be taken to choose a global test able to differentiate the direction of the treatment effects across the set of outcomes. For example, for continuous outcomes the traditional Hotelling T-squared multivariate test does not discriminate direction; 2 outcomes with effects in opposite
directions would just as easily reject the global null hypothesis as would 2 outcomes in the same direction. Instead, we recommend O’Brien’s rank sum test, a global test that is sensitive to direction. This simple yet powerful nonparametric global test is useful for either continuous or ordinal data (or mixed continuous/ordinal), and has no assumption of normality. Data for each outcome is first ranked from smallest to largest, ignoring treatment group. Ranks are then summed across the K outcomes within subject, and a 1-tailed t test or Wilcoxon rank sum test for superiority between groups is conducted on the sums. Global test options for continuous, binary and survival outcomes are described in Appendix 3.

Using the nicotine study data, we conducted O’Brien’s rank sum test by first assigning each patient their rank (across all patients) for each of pain score and opioid consumption. Treatment groups were then compared on the sum of the 2 ranks using a 1-tailed $t$ test (the sum of ranks passed tests for normality). Mean (SD) sums of ranks were 84 (37) and 88 (41) for nicotine and placebo, respectively, with 1-tailed $P = 0.29$. Thus, if we had made global superiority a criterion for proceeding to individual testing, we would stop here and conclude insufficient evidence to claim superiority of nicotine to placebo (i.e., cannot reject $H_0$ in equation (3)), and thus also insufficient evidence to reject the joint null hypothesis in equation (3).

**Step 2-B2: individual superiority testing following significant global test.** If global superiority is detected, individual outcomes are tested using either the multiple-comparison procedure described above (path B in Fig. 2), or else a completely closed testing procedure in which all tests are conducted at level $\alpha$ (path C on Fig. 2, and detailed in Fig. 3 and below). This choice also needs to be prespecified. Using either method, showing superiority of Experimental to Standard on at least one individual component would lead to rejection of the joint null hypothesis in equation (3), as in the final box in Figure 2.

In a completely closed testing procedure (CTP), an individual hypothesis is only claimed significant if all multiple-outcome hypotheses that include the individual hypothesis, beginning with the overall global test, are also significant at level $\alpha$. For our pain management example, superiority would only be claimed on either pain or opioid consumption individually if the global test across both outcomes were first found significant at level $\alpha$ (say, $P < 0.025$), and the individual test was also significant at level $\alpha$. With only 2 outcomes and global superiority required, this CTP for the individual outcomes (path C) should always be chosen (instead of a multiple-comparison procedure, path B) because the individual outcomes can directly be tested at level $\alpha$ after passing the global test.

Expanding to a scenario with 3 primary (say, ordinal scaled) outcomes, the closed testing procedure at level $\alpha$ proceeds as follows. First, overall superiority across the 3 outcomes is assessed using O’Brien’s global rank sum test at level $\alpha$. Each pair of the 3 outcomes is then tested using the same global test as for the overall test, at level $\alpha$, but is only claimed significant if the overall test is also significant. Finally, each individual outcome is assessed at level $\alpha$ with an appropriate univariate test, but is only deemed significant if the pairwise tests that include the particular outcome, as well as the overall global test, are all significant.

For example, in our nicotine study, suppose we had a priori required global superiority of nicotine to placebo on pain score, opioid consumption, and quality of recovery score (QORS). We could then test for superiority at the 0.025 significance level using the completely closed testing procedure displayed in Figure 3. Because the 1-tailed O’Brien rank sum global test across the 3 outcomes is significant in favor of the nicotine intervention at level $\alpha$ ($P = 0.001$), overall global superiority is claimed. Pairwise global superiority tests for pain and QORS ($P = 0.007$) and opioids and QORS ($P = 0.016$) are also significant (row 2) at level $\alpha$. Because both of the 2-outcome global tests that include QORS are significant, the individual test for QORS ($H_0^{(3)}$) can be considered, also at level $\alpha$, and is found significant ($P < 0.001$). However, because the remaining 2-outcome global test (pain and opioids) is nonsignificant at level $\alpha$ ($P = 0.29$), individual superiority for neither pain nor opioids can be claimed, even if $P$ values are less than level $\alpha$ (e.g., opioids $P = 0.011$).

Utilizing the completely CTP described above, the overall type I error for superiority testing is protected at level $\alpha$, and all tests can be conducted at level $\alpha$. A disadvantage is that as the number of outcomes increases, the CTP loses power. Because all higher-level tests that include a particular outcome must be significant, it becomes more difficult to reach down to the individual outcome tests. Therefore, the CTP is particularly powerful when the treatment effect

![Hypothesis testing diagram](image-url)
on the outcomes is consistent, and less powerful for detecting individual effects when heterogeneity is great.

Another advantage of the CTP is that conclusions about subsets of variables can be made, even though some of the individual outcomes in the subset cannot be tested or are not significant. Conclusions can be made on treatment effects for groups of outcomes that might naturally aggregate into subsets. In our modified nicotine study we can conclude that nicotine reduces the set of opioid consumption and QORS (global \( P = 0.007 \)), even though no significance can be claimed for opioids alone.

In Appendix 4 we give an example expanding the above joint hypothesis testing methods to the case of \( >2 \) treatments and nonspecified direction of hypothesis testing, such that both directions for each pair of interventions are of interest.

### GATEKEEPING PROCEDURES: TESTING SETS OF HYPOTHESES

In addition to the primary hypothesis, it is common in perioperative medicine and clinical research in general to include at least several secondary hypotheses. For example, the nicotine study had 8 secondary outcomes of interest for which the investigators desired valid inference about the treatment effect (Table 1). Each secondary outcome is typically evaluated at level \( \alpha \) (as in the original nicotine study), inviting a substantially increased type I error for the entire study. Secondary outcomes are also usually evaluated regardless of the primary outcome results, creating problems of interpretation and logic because study conclusions should be based on the primary outcomes.21 "Gatekeeping" closed testing procedures neatly address both issues, maintaining type I error at the nominal level across all primary and secondary outcomes (much less conservatively than does a simple Bonferroni correction) and assuring that secondary outcome assessment depends on primary outcome results.

Gatekeeping procedures require primary and secondary hypotheses of interest to be a priori organized into \( M \) ordered sets, and testing of each next ordered set follows a predefined ordered sequence. For the nicotine study we construct the following ordered sets of primary (set 1) and secondary (sets 2 to 4) outcomes for demonstration purposes. Two-tailed superiority \( P \) values from Table 1 are in parentheses:

- **Set 1**: Pain score and opioid consumption (noninferiority on both, superiority on at least one, i.e., opioids).
- **Set 2**: Quality of recovery (\( P < 0.001 \)) and completely satisfied with pain management at 72 hours (\( P = 0.36 \)).
- **Set 3**: Time to oral intake (\( P = 0.019 \)), ambulation (\( P = 0.003 \)), and 72-hour discharge criteria score (\( P = 0.018 \), favoring placebo).
- **Set 4**: Time to return of bowel function (\( P = 0.63 \)) and flatus (\( P = 0.22 \)).

We describe 2 general approaches of gatekeeping, serial and parallel.

### Serial Gatekeeping

In serial gatekeeping, testing proceeds to the next ordered set, i.e., through the "gate" (for example, from set 1 to set 2, and then from set 2 to set 3, etc.), only if all tests in the current set are shown to be significant at level \( \alpha \). Because significance is required on all tests before proceeding, the overall type I error across the sets is maintained at level \( \alpha \) (similar to an IUT, as described above; see equations (1) through (3) and related text).7,22 For example, in a study with 4 primary outcomes in the first set, all 4 would need to be significant at level \( \alpha \) before proceeding to the most important secondary outcome, also evaluated at level \( \alpha \), and so on.

In set 1 of the nicotine study (primary outcomes) we reject the joint null hypothesis at level \( \alpha \) and claimed nicotine more effective than placebo for pain management (see “Option 2: Noninferiority on All Outcomes, Superiority on at Least One”). Even though superiority was not found on both primary outcomes (only on opioids), we can proceed to set 2 because we used a joint hypothesis test (specificially an IUT) requiring noninferiority on both primary outcomes and superiority on at least one. For secondary outcomes in sets 2 to 4 we use 2-tailed superiority tests, each set tested at overall \( \alpha \) of 0.05. For set 2 we reject the null hypothesis for QORS (\( P < 0.001 \), in favor of nicotine) but not for satisfaction with pain management (\( P = 0.36 \)). Because only 1 of the 2 outcomes is significant in set 2, the serial gatekeeping procedure stops here, and none of the outcomes in sets 3 or 4 can be deemed significant or nonsignificant.

### Parallel Gatekeeping

In parallel gatekeeping, testing proceeds to the next ordered set of hypotheses if at least one outcome in the previous set is significant. To account for this added flexibility, all tests are not conducted at level \( \alpha \).7 The overall type I error is protected by reducing the significance level for each sequential set of tests (i.e., making it more difficult to reject), according to a rejection gain factor (\( \rho_m \)) that reflects the cumulative proportion of hypotheses rejected in previous sets. Each set of hypotheses is thus tested at the \( \rho_m \times \alpha \) significance level. The rejection gain factor for a current set is simply the product of the rejection proportions for the previously tested sets. If all previous hypotheses have been rejected, the current set is tested at \( \alpha \), because the rejection gain factor would be 1. More detail is given in Appendix 5.

We now apply parallel gatekeeping to the nicotine study at the overall \( \alpha = 0.05 \) level. First, on the basis of our results from "Option 2: Noninferiority on All Outcomes, Superiority on at Least One," the joint null hypothesis for set 1 is rejected (proportion rejected is \( 1/1 = 1 \)) so that set 2 is assessed using an \( \alpha \) of \( 0.05 \times (1) = 0.05 \). For set 2 we evaluate each of the 2 outcomes at 0.05/2, making a Bonferroni correction because both are not required to be significant. Quality of recovery is significant (\( P < 0.001 \)), but satisfaction with pain management is not (\( P = 0.36 \)), for a rejection proportion of 0.50. Because at least 1 outcome in set 2 is significant, we proceed to set 3 using an overall \( \alpha \) of \( 0.05 \times (1) \times (0.50) = 0.025 \). For set 3 we have 3 outcomes. Our significance criteria for the 3 ordered \( P \) values (smallest to largest) are \( P < 0.025/3 = 0.008 \), \( P < 0.025/2 = 0.0125 \), and \( P < 0.025/1 = 0.025 \), using a Holm–Bonferroni correction for multiple testing. With these criteria, only the most significant \( P \) value in the third set is significant (\( P = 0.003 <
Joint Hypothesis Testing and Gatekeeping Procedures

DISCUSSION

Multiple primary and secondary outcomes are a natural part of many comparative effectiveness studies. For such studies it is crucial to prespecify the decision rule(s) for claiming success of one intervention over another on the basis of the multiple outcomes, and to make sure the type I error is protected. We advocate use of joint hypothesis testing and gatekeeping methods to accomplish these goals.

Joint hypothesis tests requiring noninferiority on several outcomes and superiority on at least one would often give a stronger conclusion and more coherent interpretation than the traditional design, which considers all outcomes separately. For example, in many studies, authors have found a difference in only one of pain or opioid consumption, or else opposite effects. Then, for example, finding superiority on pain score and no difference on opioid consumption in isolated superiority tests makes interpretation difficult,1–3,23,24 because “no difference” in a nonsignificant superiority test cannot be interpreted as “the same” or “not worse”; claims of “not worse” require prespecifying a noninferiority δ and using an appropriate noninferiority test.12,13,25 When effects are in opposite directions, interpretation of the individuals test for pain and opioids are even more difficult. Because results are not known in advance, though, joint testing should be planned whenever the main conclusions will involve more than one outcome variable.

In our joint hypothesis testing we compare intervention groups on the means of each outcome, but what about outcomes for an individual? Analyzing means alone does not consider the correlation between the outcomes, i.e., whether patients who did well on opioids, for example, did well or poorly on pain. It might therefore seem useful to dichotomize the data and define a successful patient outcome as pain of no more than X and total opioid consumption of no more than Y (mg). However, such cutpoints are arbitrary and throw away many data, so are not desirable. We do, on the other hand, recommend reporting the correlation coefficient and a scatterplot of pain and opioids by treatment group. A strong negative correlation within a group would mean that patients receiving higher amounts of opioids tended to have lower pain, and vice versa. A positive correlation would mean that patients with higher amounts of opioids tend to have higher pain scores as well. In Appendix 1 we report correlations ranging from 0 to 0.50 from our clinical trials and database studies. However, the direction and strength of the correlation would not typically change our conclusions as to which intervention is preferred. We would still be most interested in differences in mean pain and opioids.

We describe 2 main methods for superiority testing after noninferiority is shown on all outcomes, individual superiority tests with no global test, or else a significant global test followed by individual tests (Fig. 2). After a significant global test, individual comparisons could follow either a completely closed testing procedure or the more traditional multiple-comparison procedure.15,16 We assessed the relative powers of these 2 joint hypotheses testing methods through simulations (Fig. 4). We found that requiring the global test for superiority (followed by individual superiority tests) tends to be the more powerful test when outcomes are negatively correlated, especially when the intervention has a similar effect on all outcomes (Fig. 4B), whereas the individual testing method tends to be more powerful with positive correlations and when fewer outcomes are affected (Fig. 4A).

A key benefit of the joint hypothesis testing framework discussed is that no adjustment to the significance criterion is needed for testing both noninferiority and superiority, because both are required to be significant to reject the joint null (i.e., an IUT15,16). Similarly, numerous individual

Figure 4. Power as function of pain–opioids correlation. Simulations assumed multivariate normal distribution for pain and log(opioids). Pain–opioid correlation was varied between −0.90 and 0.90, with 3000 simulation runs per correlation. NI = noninferiority. Joint hypothesis tests: NI on both outcomes (black), NI on both + superiority on at least one (at α/2) (red), and NI on both + global superiority + superiority on at least one (α) (green). Overall α was 0.05 for each of noninferiority and superiority. When no global test was used, superiority for individual outcomes assessed at 0.05/2 = 0.025. Bonferroni correction was not used for individual NI tests because both are required to be significant. NI δ is 1 point for pain and 10 mg for opioid consumption; n = 100 per group. A, Superiority on pain, no effect on opioids. Mean difference (SD) is 1 (2.5) for pain, 0 (25) for opioids. B, Superiority on both pain and opioids. Mean difference (SD) is 1 (2.5) for pain and 5 (25) for opioids.

0.008 for time to ambulation). Because at least 1 null hypothesis in set 3 was rejected, we proceed to set 4 using an α level of 0.05 × (1) × (0.50) × (0.33) = 0.0083. Neither P value in set 4 is rejected (smallest is P = 0.22).

Volume 114 • Number 6 www.anesthesia-analgesia.org

Copyright © 2012 International Anesthesia Research Society. Unauthorized reproduction of this article is prohibited.
noninferiority tests can be conducted at the same overall \( \alpha \) level, because all are required to be significant. This can greatly reduce the required sample size in comparison with making a Bonferroni correction for all comparisons. Only for superiority testing (our step 2) is an adjustment needed for testing multiple outcomes, because not all are required to be significant.

However, an equally important feature of joint hypothesis testing and gatekeeping procedures is that power is generally decreased over traditional designs to the extent that multiple hypotheses need to be rejected before the overall hypothesis can be rejected. We gave an example of requiring superiority on 2 uncorrelated outcomes in the nicotine study in which the sample size would need to be 24\% higher to have 90\% power to reject both outcomes in comparison with only one of them. Importantly, though, the increase was far less (only 4\%) after adjusting for multiple comparisons in the traditional approach. In any case, boosting sample size for such studies is prudent. In contrast, many studies are currently underpowered because researchers hope to gain significance on multiple primary outcomes, but the study is only powered for a single outcome. Software to assist in planning joint hypothesis testing designs is available from E. J. Mascha (first author of this paper).

Turk et al. give an excellent overview of design and analysis of pain management studies, but do not address our main topic of interest, combined noninferiority–superiority designs.\textsuperscript{26} Several authors do propose methodologies on simultaneously testing noninferiority on 2 outcomes and superiority on at least one; we largely follow the methods of Rohmelm.\textsuperscript{4} There are more complex options, some of which may be more powerful in certain situations, but also more controversial.\textsuperscript{5,6,27–29} For example, results on noninferiority testing can be directly used in the superiority testing in a simultaneous CI approach,\textsuperscript{5,6} but this has been criticized as making superiority results dependent on the chosen noninferiority \( \delta.\textsuperscript{4} \) Bloch et al. use bootstrapping to simultaneously assess noninferiority and superiority, allowing a more complex null hypothesis and directly incorporating the correlation among the outcomes, thus increasing power.\textsuperscript{29} Traditional bootstrap resampling is also expected to increase power,\textsuperscript{30–32} especially for binary outcomes because it takes advantage of the discreteness of the data.\textsuperscript{33} As well, superiority, equivalence, and noninferiority can all be tested together in a 3-way test.\textsuperscript{34}

We also advocate the more general “gatekeeping” procedures, in which primary and perhaps secondary hypotheses of interest are a priori organized into ordered sets, and testing in the current set depends on results of previous sets.\textsuperscript{7,8,35} Using these procedures in perioperative medicine studies would force investigators to prioritize hypotheses and outcomes in the design phase, and would thus likely reduce the number of outcomes included in studies. Gatekeeping designs are attractive because in practice there is often some natural hierarchy of importance across outcomes of interest. Also, groups of outcomes often fall into natural “sets,” which can be tested together. Another attractive feature is that if no effect is found on any of the primary outcome(s), secondary outcomes would not be analyzed. This has long been considered an important principle for clinical trials, and seems just as valid for most observational studies.\textsuperscript{31} As with the simpler joint hypothesis testing, adherence to the planned design for a gatekeeping procedure is necessary for valid interpretation of results and maintaining type I error at the planned level.

A potential limitation to implementing these methods is that there is no CONSORT (Consolidated Standards of Reporting Trials) statement extension for reporting on a joint hypothesis testing design. However, there is a detailed extension for noninferiority and equivalence trials that stresses the need to clearly state whether the design of the study is noninferiority or equivalence, and the importance of prespecifying the noninferiority \( \delta.\textsuperscript{36} \) For the designs we have discussed, similar guidelines would apply. For example, researchers could state that “the study design is a noninferiority–superiority trial in which both pain and opioid consumption were jointly tested. In order for one intervention to be claimed more effective than the other, it needed to be significantly noninferior on both outcomes and superior on at least one of them.” In line with the CONSORT statement one also needs to prespecify all primary and secondary outcomes, as well as details of how the analyses will be conducted, including whether or not a global test will be required and methods for testing individual outcomes. One needs to specify clearly what the decision-making process will be with regard to the multiple endpoints. Results could be reported as we have done in Figure 1 and Table 2.

In conclusion, joint hypothesis testing and gatekeeping procedures are straightforward to implement and should be considered in any study design with multiple outcomes of interest. They can markedly improve the organization, interpretation, and efficiency of both randomized and nonrandomized studies in comparison with considering multiple outcomes in isolation.

APPENDIX 1: CORRELATION BETWEEN PRIMARY OUTCOMES

The amount and the direction of the correlation between outcomes being assessed in a joint hypothesis setting, for
example, pain and opioid consumption, may have implications for the interpretation of the results and conclusions, as well as affecting the power of the study.

Independent of the mean pain and opioid consumption differences between groups, a strong negative correlation within an intervention group may imply an effective treatment: patients with higher opioid consumption tend to have lower pain, and vice versa. A strong positive correlation might imply a not-so-effective treatment: patients with higher opioid consumption still tend to have high pain, and vice versa. In our practice we have observed correlations ranging from 0 to about 0.50 (Table 3), all positive and in the mild-to-moderate range. Figure 5 shows the mildly positive correlation between pain and opioid consumption through 72 hours for our nicotine study, with Spearman (Pearson) correlations of 0.21 (0.04) for nicotine patients and 0.40 (0.51) for placebo.

### APPENDIX 2: DETAIL ON TESTING FOR NONINFERIORITY

#### Constructing the Noninferiority Test

Assuming that smaller outcome values are desirable (e.g., lower pain better than higher), noninferiority testing for the ith outcome is most intuitively done by simply observing whether the upper limit of the confidence interval (e.g., 95% CI for \( \alpha = 0.025 \)) for the difference in means lies below the noninferiority \( \delta \). In addition, a test with a \( P \) value is often desired. We proceed by rearranging terms and expressing each alternative hypothesis \( H_{A1} \), in equation (4) as \( H_{A1}: \mu_{Ei} - \mu_{Si} - \delta_i < 0 \), and constructing a \( t \) test statistic to assess whether the difference in means is below the given \( \delta \) as

\[
T = \frac{\hat{\mu}_E - \hat{\mu}_S - \delta_i}{\sqrt{\frac{s_{\hat{\mu}}^2}{n_E} + \frac{s_{\hat{\mu}}^2}{n_S}}} \tag{9}
\]

where \( n_E \) and \( n_S \) are sample sizes for Experimental and Standard groups, \( s_{\hat{\mu}}^2 \) is the pooled variance

\[
s_{\hat{\mu}}^2 = \frac{(n_E - 1)s_E^2 + (n_S - 1)s_S^2}{n_E + n_S - 2}, \quad \text{and} \quad s_E^2 \quad \text{and} \quad s_S^2 \quad \text{are} \quad \text{group} \quad \text{variances (SD squared). Noninferiority for the ith outcome is claimed if } T \text{ is far enough below zero (i.e., smaller than the value of } T \text{ from a } t \text{ distribution with } n_E - n_S - 2 \text{ degrees of freedom at } 1 - \alpha).}

#### Applying Noninferiority Test to Nicotine Study

Pain score mean (SD) was observed for the nicotine study at 1.45 (0.9) and 1.12 (0.8) for the nicotine and placebo patients, respectively. With a noninferiority \( \delta \) of 1.0, and \( n = 43 \) and 42 patients per group, the \( t \) test statistic for noninferiority in equation (9) is

\[
T = \frac{4.4 - 1.45 - 1}{\sqrt{\frac{0.85^2(1/43 + 1/42)}}} = -0.67
\]

Because -3.63 is smaller than -1.66, the value of \( T \) from a \( t \) distribution with 43 degrees of freedom at \( \alpha = 0.025 \), we reject the null hypothesis and claim noninferiority at the 2.5% significance level (\( P = 0.002 \), i.e., nicotine is no more than 1 point worse than placebo. Correspondingly, the upper limit of the 95% confidence interval (-0.05 to 0.71) for the difference in means is below the noninferiority \( \delta \) of 1 (Table 2). A 95% confidence interval (difference in means \pm 1.96 standard errors) is used because the test is 1 tailed, with our noninferiority \( \alpha = 0.025 \) in the upper tail (because lower values are desirable).

We test noninferiority of nicotine to placebo on opioid consumption using a noninferiority \( \delta \) of 1.2 for the ratio of geometric means (or equivalently, the ratio of medians) because the data appear to be log-normal. After log-transforming the data (log base 2), the null, and alternative hypotheses are

\[
H_{A1}, (null): \mu_E - \mu_S \geq \log_{2}(1.2) = 0.263
\]

versus

\[
H_{A1}, (alt): \mu_E - \mu_S < \log_{2}(1.2) = 0.263 \tag{10}
\]

where \( \mu_E \) and \( \mu_S \) are the means of log-transformed opioid consumption for the Experimental (nicotine) and Standard (placebo) groups, respectively, and 0.263 is log(base 2) of 1.2. Applying the \( t \) test statistic as in equation (9) we have

\[
T = \frac{4.4 - 5.3 - 0.263}{\sqrt{\frac{0.69^2(1/43 + 1/42)}}} = -0.663 = -4.4. \text{ Because -4.4 is much more extreme than -1.66, we claim noninferiority of nicotine to placebo at the 0.025 level. Correspondingly, the upper limit of the 90% confidence interval for the ratio of geometric means is <1.2, at (0.64, 0.97), calculated by back-transforming the confidence interval for } \log_{2}(\text{opioids}), \text{ Table 2.}

### APPENDIX 3: TESTS FOR GLOBAL SUPERIORITY WITH ALTERNATIVE DATA TYPES

For a set of continuous or ordinal outcomes, or mixed continuous/ordinal, we recommended O’Brien’s rank sum test for assessing global superiority (see the section “Global Superiority Testing”). When continuous outcomes follow a multivariate normal distribution, O’Brien’s ordinary least squares (OLS) test\(^{19} \) may sometimes be more powerful than the rank sum test. The OLS test is equivalent to a \( t \) test on the sum of standardized scores across the outcomes. Each patient outcome is first standardized by subtracting the pooled (across groups) mean and dividing by the pooled SD for that outcome, which helps prevent overweighting by any one endpoint. The standardized outcome scores are then summed within subject.
and groups compared with t test or analysis of variance. However, because the OLS test assumes multivariate normality, the rank sum test is often more attractive in practice.

For binary outcomes, global test options include 2 multivariate tests that adjust for the within-subject correlation among outcomes: the common effect generalized estimating equation (GEE) test\(^\text{37}\) and the average relative effect GEE test\(^\text{38}\); a frequently chosen option is to compare groups on the collapsed composite (any event versus none) with, say, a chi-square test. The common effect method estimates a common treatment effect across the outcomes, but is still powerful given moderate treatment effect heterogeneity. In the average relative effect test, we first estimate the individual log-odds ratio for each outcome and then test the average log-odds ratio against zero. This test avoids being driven by the most frequent components, a problematic feature of other methods when components differ on both incidences and treatment effects.\(^\text{39}\) For survival outcomes, options for a 1-tailed global test include the method of Wei, Lin, and Weissfeld,\(^\text{39}\) which incorporates the correlation across outcomes within subject and is based on the Cox proportional hazards model.

**APPENDIX 4: EXTENSIONS OF JOINT HYPOTHESIS TESTING TO MULTIPLE TREATMENTS AND UNSPECIFIED DIRECTIONS**

In an ongoing randomized clinical trial at Cleveland Clinic, investigators are comparing 3 femoral nerve catheter insertion techniques during ultrasound-guided femoral nerve block during total knee replacement (ClinicalTrials.gov, identifier: NCT00927368): A = stimulation needle and stimulating catheter; B = stimulation needle, but nonstimulating catheter; or C = nonstimulating catheter. Groups will be compared on average 48-hour postoperative VAS pain score and total 48-hour postoperative opioid requirement in morphine equivalents. Joint hypothesis testing will be used to conclude one treatment more effective than another if it can be shown to be noninferior (NI) on both pain and opioid consumption and also superior on at least 1 of the 2 outcomes.

This joint hypothesis testing design is complex because (1) 3 treatments are being compared and (2) because there is no specified direction for the comparisons: A might be noninferior to B (and perhaps superior), or vice versa, and similarly for the A–C and B–C comparisons. For each outcome there are thus 6 comparisons of interest for noninferiority testing (3 groups \(\times \) 2 directions each). An \(\alpha\) of 0.025 for noninferiority testing overall was chosen a priori. Because noninferiority is required on both pain and opioid consumption, each outcome will be evaluated at 0.025 (i.e., no adjustment made for multiple outcomes), as in equation (4); however, within each outcome we will use the Holm–Bonferroni procedure to control \(\alpha\) at 0.025 for the 6 comparisons. If any intervention is noninferiority to another on both outcomes, that comparison (say, A is noninferior to B) proceeds to superiority testing in step 2.

Superiority on each of opioid consumption and pain score will then be assessed for each of the \(W\) comparisons that passed noninferiority testing in step 1 (\(W = \) at most 3 of 6, because the groups will have an ordering, e.g., A \(\leq\) C \(\leq\) B), using 1-tailed tests in the same direction as found in noninferiority testing. No global superiority test will be used because superiority on either outcome is sufficient in this study. To maintain an overall significance level of 0.025 across the 2 outcomes tested for superiority (pain, opioids), one would test each of the \(W\) comparisons passing noninferiority testing at the 0.025/(\(W \times 2\)) significance level. For example, if \(W = 2\) comparisons showed noninferiority (A to B and B to C), then each superiority comparison would be tested with a significance criterion of 0.025/4 = 0.00625.

**APPENDIX 5: DETAIL ON PARALLEL GATEKEEPING**

Overall type I error is protected in parallel gatekeeping procedures by reducing the significance level for each sequential set of tests according to a rejection gain factor, \(\rho_{ij}\), which reflects the cumulative proportion of hypotheses rejected in previous sets. Each set of hypotheses is thus tested at the \(\rho_{ij} \times \alpha\) significance level. If all previous hypotheses have been rejected, the current set would be tested at \(\alpha\), because the rejection gain factor would be 1.

Specifically, when each hypothesis in each set is given the same weight (equal to 1/number of tests in a set), the rejection gain factors, \(\rho_1 = \rho_2 = \cdots = \rho_M\), are simply the product of the rejection proportions for the previously tested sets, such as

\[
\rho_1 = \rho_2 = \cdots = \rho_M = \prod_{j=1}^{m} \left( \frac{r_j}{n_i} \right), \quad m = 2, \ldots, M, \tag{11}
\]

where \(r_j\) is the number of rejections and \(n_i\) the number of tests in the \(i\)th set. For example, with overall \(\alpha\) of 0.05, if 1 of 2 outcomes in the first set is rejected, the rejection gain factor for the second set would be 0.5 (i.e., 1/2). The second set would thus be tested at the 0.5\(\alpha\) level, or 0.025. If then only 1 of the 4 hypotheses in set 2 were rejected, the significance level for the third ordered set would be 0.5 \((0.25)(0.05) = 0.00625\).

The procedure can be made more flexible by assigning differing a priori weights to the hypotheses in a set, with all weights in a set summing to 1. Then the rejection proportion \(r_j/n_i\) for the \(j\)th set in equation (11) would be replaced by a sum of weights,

\[
\sum_{j=1}^{p_i} r_j w_{ij},
\]

where \(w_{ij}\) is the weight and \(r_j\) the rejection status (1 = rejected, 0 = not) for the \(j\)th hypothesis in the \(i\)th set. For example, with weights of 0.2, 0.3, and 0.5 and rejection status of 1, 0, and 1 for outcomes 1, 2, and 3 in the \(i\)th set, the rejection proportion would be 1(0.2) + 0(0.3) + 1(0.5) = 0.7.

**APPENDIX 6: POWER OF JOINT HYPOTHESIS TESTS AS A FUNCTION OF CORRELATION AMONG OUTCOMES**

We used simulations to assess the relative powers of 2 different intersection–union joint hypothesis tests of noninferiority on both outcomes (pain and opioid consumption) and superiority on at least one as a function of the correlation between the outcomes. Results show, for example, that requiring global superiority is a more powerful strategy when one intervention is superior to the other on both outcomes (Fig. 4B), while not requiring that global superiority (and thus using a multiple testing adjustment) is more powerful when the intervention is only superior on one outcome (Fig. 4A).

Figure 4A shows that when an intervention is superior on only 1 of 2 outcomes, not requiring global superiority (red line) is substantially more powerful (in rejecting the joint null hypothesis.
of noninferiority on both and superiority on at least one) for all positive correlations than with no global test requirement (red line). However, Figure 4B shows that when an intervention is superior on both outcomes, requiring the global test may be more powerful for correlations less than 0.3, and otherwise only somewhat less powerful.

When only one outcome is superior (Fig. 4A), power requiring NI and superiority without the global test increases with correlation (red), and power for NI/global/superiority decreases with increasing correlation (green). In Figure 4B, power for both of these tests decreases with increasing correlation.

DISCLOSURES
Name: Edward J. Mascha, PhD.
Contribution: This author helped design the study, conduct the study, analyze the data, and write the manuscript.
Attestation: Alparslan Turan approved the final manuscript.
Name: Alparslan Turan, MD.
Contribution: This author helped write the manuscript.
Attestation: Alparslan Turan approved the final manuscript.
This manuscript was handled by: Franklin Dexter, MD, PhD.

REFERENCES
4. Rohmel JGC, Benda N, Lauter J. On testing simultaneously non-inferiority in two primary endpoints and superiority in at least one of them. Biom J 2006;48:916–33
35. Dmitrienko A, Tamhane AC. Gatekeeping procedures with priority in at least one of them. Biom J 2006;48:916–33