Journal of Biopharmaceutical Statistics, 15: 419–436, 2005

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1081/BIP-200056525

Taylor & Francis
Taylor & Francis Group

COST-EFFECTIVENESS INFERENCES FROM BOOTSTRAP QUADRANT CONFIDENCE LEVELS: THREE DEGREES OF DOMINANCE

Robert L. Obenchain, Rebecca L. Robinson, and Ralph W. Swindle
U.S. Medical Outcomes Research, Eli Lilly and Company, Indianapolis, Indiana, USA

When with at least 95% confidence a new treatment is shown to be not only less costly (LC), but also more effective (ME), than a current treatment, that new treatment can be said to "strictly dominate" the current treatment statistically. But what can be said when head-to-head treatment comparisons turn out to be less clear-cut than this? Here, we propose two additional sets of specific LC and/or ME confidence thresholds to define the concepts of "some dominance" and "much dominance." Confidence levels associated with entire quadrants of the incremental cost–effectiveness (ICE) plane are easily computed using the same bootstrapping techniques used to estimate an "acceptability curve." Our two proposed additional "degrees" of dominance, although less stringent than strict dominance, are nevertheless more stringent than commonly accepted approaches using ICE ratio or net benefit calculations.

To illustrate analysis concepts, we use data from a randomized, double-blind, placebo- and active comparator-controlled clinical registration trial for treatment of major depressive disorder (MDD). As is typical, our case study is rather small and short term, providing outcome information for a total of only 264 patients during their initial 8 weeks of acute-phase MDD treatment. Thus, we focus attention on sensitivity analyses, showing that the bootstrap distribution of cost-effectiveness uncertainty is robust across two alternative ways of measuring overall effectiveness and three alternative ways of imputing missing values.

Evaluation of the balance between cost and benefit is particularly difficult when a new pharmacological treatment is first introduced, yet information of this sort is highly desired by decision makers. We show that, even with only a relatively modest amount of clinical trial information, sensitivity analyses can still confirm that cost-effectiveness comparisons are being made in a consistent fashion. In contrast, extensive follow-up comparisons using data from actual clinical practice will almost always ultimately be needed to better inform health policy makers.

Key Words: Bootstrap uncertainty distribution; Cost-effectiveness; Depression; Duloxetine; Paroxetine; Resource utilization; Sensitivity analysis.

Received March 12, 2004; Accepted November 8, 2004
Address correspondence to Robert L. Obenchain, U.S. Medical Outcomes Research, Eli Lilly and Company, Indianapolis, Indiana, USA; E-mail: Obenchain_Robert_L@lilly.com

1.0. INTRODUCTION

The tutorial by Briggs and Fenn (1998) provides a good summary of the many methods for statistical inference in cost-effectiveness (cost–benefit) analysis that have been proposed in the health economics literature during the last 20 or so years. The most simple techniques compare only two treatments and are known as incremental cost-effectiveness (ICE) methods (Black, 1990). In these head-to-head comparisons, the unknown true differences in cost and effectiveness between treatments are denoted by

ΔC = Difference in overall expected cost, "new" minus "standard"

ΔE = Difference in overall expected effectiveness, "new" minus "standard"

By convention, the ICE plane is depicted with its horizontal axis representing ΔE with a larger-is-more-favorable to "new" interpretation. Meanwhile, the vertical axis of the ICE plane depicts ΔC with a more negative or smaller-is-more-favorable to "new" interpretation.

1.1. Confidence Interval Approaches

In many approaches to ICE inference, attention is focused on computing a confidence interval for a scalar-valued ICE summary statistic. The two most commonly used statistics are the ICE ratio (ICER) $= \Delta C/\Delta E$, and net (monetary) benefit, (NB) $= \lambda \times \Delta E - \Delta C$, where λ is some stated value for society's shadow price (or willingness to pay) expressed in dollars per unit of health care improvement (Stinnett and Mullahy, 1998). Laska et al. (1999) argued that making multiple, pairwise treatment comparisons using either ICER or NB statistics leads to identical, optimal overall health care allocations when uncertainty is ignored.

When uncertainty is sufficiently high, ICER point estimates become almost useless. Specifically, when neither estimated treatment difference (neither ΔC nor ΔE) is significantly different from 0 at roughly 95% confidence, the closed-form solution of Chaudhary and Stearns (1996) for Fieller's theorem ICER confidence limits are typically imaginary. In other words, although a study with this much uncertainty does produce a point estimate of the true ICER, there is no realistic way to estimate the precision of that ICER from the estimated precision, skewness, and correlation between the observed ΔC and ΔE treatment differences. The proper interpretation of this result is that the unknown, true ICER comparing these two treatments could be almost any positive or negative numeric value.

1.2. Bootstrap Distribution of ICE Uncertainty

Even in high uncertainty situations, we demonstrate that ICE outcomes can still be much more favorable to one of the two treatments than to the other. As explained here, evidence for differential cost effectiveness can still be provided by computing ICE quadrant "confidence levels" from the bootstrap distribution of uncertainty about the unknown true location of (ΔE, ΔC) within the ICE plane. Many authors have proposed applying bootstrap methods (resampling with replacement) to ICE statistical inference (see Briggs et al., 1997; Obenchain, 1997,

1999; Polsky et al., 1997; Stinnett, 1996; Tambour and Zethraeus, 1998; and Van Hout et al., 1994). Because methods that simply rank-order ICER resamples (without keeping track of the ICE quadrant that generated the outcome) can yield badly biased confidence intervals in high-uncertainty situations, the computing algorithms suggested by Obenchain (1997, 2003) and Cook and Heyse (2000) are much more reliable. Specifically, they use polar coordinates to form "circular" order statistics (rank on angle) to identify highly directional, wedge-shaped confidence regions.

Here, the simulated probability content within any region of the ICE plane generated using the bootstrap distribution of ICE uncertainty will be called that region's "confidence level." Van Hout et al. (1994) called this same probability content a region's "acceptability." Furthermore, interest focuses primarily on bootstrap confidence levels within the four individual quadrants of the ICE plane, not on the content of all possible half-planes below and/or to the right of a rotating line of slope λ, as in Van Hout et al. (1994).

Because of the extremely weak conditions needed to apply the central limit theorem, estimates of ΔC and ΔE in head-to-head treatment studies are typically asymptotically normal (see Chaudhary and Stearns, 1996; O'Brien et al., 1994). Thus, it is certainly not surprising that the bootstrap distribution of ICE uncertainty is almost invariably observed to be quite close to bivariate normality, at least when treatment group sizes are not very small ($\geq$50) and the number of resamples is large (e.g., the default of 25,000 in the ICEplane software of Obenchain, 2003). Although our proposed methods for establishing degrees of dominance using the bootstrap distribution of ICE uncertainty do not technically require that these distributions be normal or even asymptotically normal, we nevertheless consider some simple bivariate normal numeric approximations to the bootstrap distribution of ICE uncertainty here. After all, outcomes researchers can be highly confident that such approximations are both relevant and realistic.

Our experience with the ICEplane software is that estimates of ΔC and ΔE derived from actual samples of cost-effectiveness outcomes on treated patients (from clinical trials, nonrandomized studies, or retrospective database analyses, as averse to predictions from decision-theoretic tree models) commonly tend to be almost uncorrelated. Quite possibly, this may be due to the additional uncertainty implied by the presence of comorbid conditions in realistic data. Nevertheless, we consider the possibility of correlation between outcome differences (ΔE, ΔC) simply to explore the robustness of our proposed sets of confidence thresholds for determining degrees of dominance.

1.3. Case Study Example

To illustrate inferences about dominance, we use outcomes from a registration trial that compared duloxetine, a new serotonin and norepinephrine reuptake inhibitor, with paroxetine, a standard selective serotonin reuptake inhibitor, for treatment of major depressive disorder (MDD). The primary efficacy and safety analyses for this study, along with details of its somewhat innovative design characteristics, are reported in Goldstein et al. (2004).

The measures of effectiveness considered are derived from blind, clinical assessments using the Hamilton Depression Rating Scale (HAMD-17), (Hamilton,

1967), and made during acute-phase treatment of MDD at weeks 0 (baseline), 1, 2, 4, 6, and 8. The study randomized $N = 86$ adult patients to duloxetine 40 mg/d (20 mg BID), $N = 91$ to duloxetine 80 mg/d (40 mg BID) and $N = 87$ to paroxetine 20 mg/d. Because outcomes for the $N = 87$ patients randomized to placebo were relatively poor in this study, analyses reported focus exclusively on the $N = 264$ patients randomized to an active treatment for MDD.

Patients self-reported all health care resources used beyond those dictated by the study protocol via a modified version of the Resource Utilization Survey (Copley-Merriman et al., 1992). Costs were then computed by rounding the 1998 prices reported by Schoenbaum et al. (2001) to the nearest multiple of $50. Thus, each visit to a psychiatrist, psychologist/therapist, or "other mental health care worker" was assigned a standardized cost of $100. An emergency room visit was assigned a standardized cost of $450, whereas all other visits to health care professionals were assigned a standardized cost of $50. In this study, the most commonly reported types of resource utilization in the "other" category were visits to dentists, chiropractors, and physical therapists.

During the study, patients were asked to report resource utilization "since the last visit." Unfortunately, there was considerable variation in actual time between visits. For example, the final visit varied from a minimum of 47 days to a maximum of 115 days after baseline, with a mean of 57.8 days. Therefore, "cost per week" was calculated here by multiplying total accumulated cost for a patient by 7 and then dividing by the total days of cost accumulation for that patient. For patients who discontinued early from the study, this calculation corresponds to average value carried forward (AVCF) imputation.

2.0. THREE DEGREES OF DOMINANCE IN ICE INFERENCE

2.1. Expected ICE Quadrant Confidence Under Equivalence

When the two treatments being compared are actually equivalent on effectiveness, exactly 50% of the bootstrap distribution of ICE uncertainty is expected to fall into the two right-hand ICE quadrants ($\Delta E > 0$), where the new treatment is more effective (ME) than the standard. Similarly, exactly 50% of the bootstrap distribution of ICE uncertainty would be expected to fall into the two lower ICE quadrants ($\Delta C < 0$), where the new treatment is less costly (LC) than the standard, when, in reality, the two treatments are actually equivalent on cost. Finally, when the two treatments are not only equivalent on cost and effectiveness, but also these cost and effectiveness outcomes are uncorrelated measures, exactly 25% of the bootstrap distribution of ICE uncertainty is expected to fall into each of the four quadrants of the ICE plane.

Correlation between the ΔC and ΔE estimates for two treatments that actually are equivalent can cause the probability of observing a "LC and ME" outcome ($\Delta C < 0$ and $\Delta E > 0$) to vary all the way from 0 to 0.5. Again, this probability is .25 when the correlation between ΔC and ΔE estimates is zero, but it approaches either 0 as the correlation approaches $+1$ or .5 as the correlation approaches -1. For example, numeric integration for the bivariate normal approximation yields Pr(LC and ME) = .167 or .333 when the $(\Delta E, \Delta C)$ correlation is $\rho = \pm 0.5$, and .115 or .385 when this correlation is $\rho = \pm 0.75$.

In view of these scenarios, the simple fact that the *point estimate* of (ΔE, ΔC) from a study happens to fall within the South-East ICE quadrant has no clear implication. This observation certainly does not provide sufficient evidence to conclude, with any reasonable confidence, that the "new" treatment actually does dominate the "standard."

2.2. ICE Quadrant Confidence for Strict Dominance

In sharp contrast to the previous hypothetical situation where the two treatments being compared are actually "equivalent" on cost and effectiveness, the new treatment might reasonably be considered to be statistically "strictly dominant" over the standard if, say, 95% or more of the bootstrap distribution of ICE uncertainty falls within the single, lower-right (South-East) ICE quadrant. In actual practice, it is unlikely that a fair comparison of ICE outcomes between two active treatments would be this extreme in favor of either treatment unless sample sizes were very large indeed.

2.3. ICE Quadrant Confidence in Actual Practice

The sort of situation one might realistically find when attempting to differentiate between two active treatments using relatively small samples from a registration trial might be described as follows. Suppose the bootstrap distribution of ICE uncertainty within all four quadrants of the ICE plane has been simulated. Furthermore, regardless of the observed correlation between cost and effectiveness outcome differences, suppose the observed bootstrap ICE uncertainty quadrant confidence levels surpass a pair of thresholds.

Confidence Threshold Pairing for "Some Dominance":

[1] at least 50% confidence falls within the South-East (LC and ME) quadrant, plus
[2] at least 90% confidence falls within the three (LC or ME, inclusive) quadrants.

Confidence Threshold Pairing for "Much Dominance":

[1] at least 65% confidence falls within the South-East (LC and ME) quadrant, plus
[2] at least 95% confidence falls within the three (LC or ME, inclusive) quadrants.

Confidence Threshold for "Strict Dominance":

at least 95% confidence falls within the South-East (LC and ME) quadrant.

A study with bootstrap ICE quadrant confidence levels that surpass these specific thresholds could certainly be said to favor the "new" treatment over the "standard" treatment. In fact, the first two pairings provide our proposed definitions for "some dominance" and "much dominance" in ICE inference.

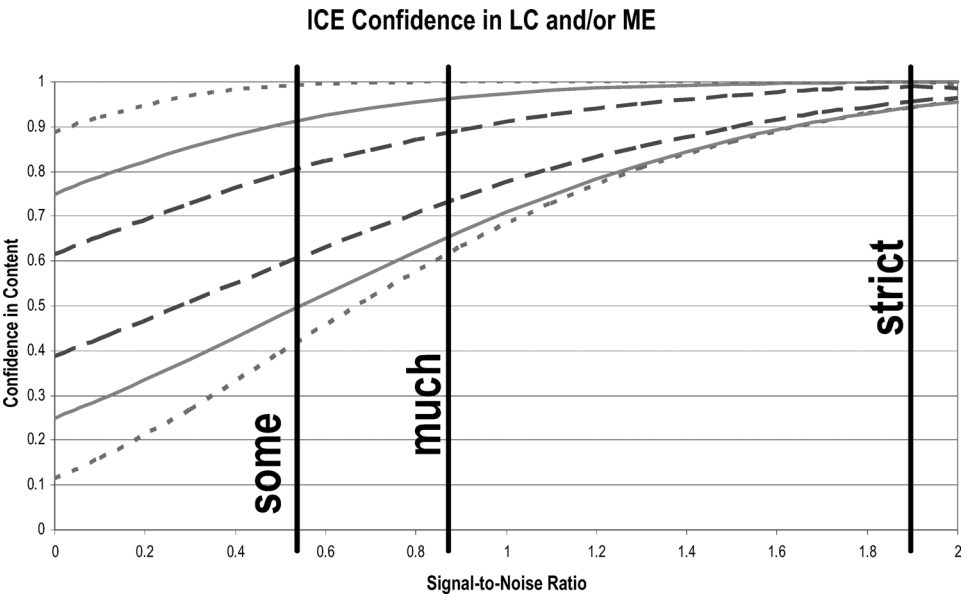

Figure 1 Expected ICE quadrant confidence levels for "threshold pairings" from bivariate normal approximations with $\mathrm{SNR}(\Delta E) = -\mathrm{SNR}(\Delta C) = \mu/\sigma$. Highest to lowest confidence level curves: dotted: LC or ME for $\rho = -0.75$; solid: LC or ME for $\rho = 0.0$; dashed: LC or ME for $\rho = +0.75$; dashed: LC and ME for $\rho = +0.75$; solid: LC and ME for $\rho = 0.0$; dotted: LC and ME for $\rho = -0.75$.

2.4. Proposed Threshold Pairings Work Well Together

Figure 1 depicts expected ICE quadrant confidence levels calculated for three special cases where estimates of the true treatment differences (ΔE, ΔC) are correlated with $\rho = -0.75$, 0, or $+0.75$. Results displayed are based on numeric integration with a bivariate normal approximation to the bootstrap distribution of ICE uncertainty and use signal-to-noise ratios (SNRs) that are equal but opposite in sign: $\mu/\sigma = \mathrm{SNR}(\Delta E) = -\mathrm{SNR}(\Delta C)$. In other words, as μ increases (relative to σ) from left to right in Fig. 1, the true value of (ΔE, ΔC) moves down the diagonal of the South-East (LC and ME) quadrant of the ICE plane. The three lower curves in Fig. 1 depict expected confidence percentages within the single "LC and ME" ICE quadrant. The three upper curves in Fig. 1 depict expected confidence percentages summed over the three "LC or ME" (inclusive) ICE quadrants.

First, consider the two solid curves that depict the case where (ΔE, ΔC) are uncorrelated. Note that the "new" treatment reaches our proposed "and/or" threshold pairing of 50% and 90% confidence for "some dominance" in Fig. 1 when the SNR in the uncorrelated case reaches $\mu/\sigma = 0.544$. More precisely, the "and/or" confidence percentages achieved are actually 50.0% and 91.4%. Similarly, the new treatment reaches our proposed "and/or" threshold pairing of 65% and 95% for "much dominance" in Fig. 1 when the SNR for the uncorrelated case reaches $\mu/\sigma = 0.863$. More precisely, the "and/or" confidence percentages achieved there are 65.0% and 96.2%. Finally, the new treatment becomes statistically "strictly dominant" in Fig. 1 when the SNR for the uncorrelated case reaches $\mu/\sigma = 1.951$. More precisely, the actual "and/or" confidence percentages are then 95.0% and 99.9%.

Next, consider the two middle dashed curves in Fig. 1 that depict the case where the (ΔE, ΔC) correlation of $\rho = +0.75$ is strongly positive. Note, in particular, that although it is easier to meet "LC and ME" thresholds in this positively correlated case, it is also much more difficult to meet the inclusive "LC or ME" thresholds. Specifically, the SNR required to achieve "some dominance" in Fig. 1 increases from about 0.5 to $\mu/\sigma = 0.9$, whereas the SNR for "much dominance" increases from less than 0.9 to $\mu/\sigma = 1.4$. Also, the new treatment still becomes statistically "strictly dominant" in Fig. 1 when $\rho = +0.75$ as the SNR approaches $\mu/\sigma = 2.0$.

Finally, consider the two most extreme dotted curves in Fig. 1 that depict the case where the (ΔE, ΔC) correlation of $\rho = -0.75$ is strongly negative. Note, in particular, that although it is now easier to meet the inclusive "LC or ME" thresholds in this negatively correlated case, it is also more difficult to meet the "LC and ME" thresholds. Specifically, the SNR required to achieve "some dominance" in Fig. 1 increases from about 0.5 to $\mu/\sigma = 0.7$, whereas the SNR for "much dominance" increases from about 0.9 to $\mu/\sigma = 1.0$. Finally, the new treatment still becomes statistically "strictly dominant" in Fig. 1 when $\rho = -0.75$ as the SNR approaches $\mu/\sigma = 2.0$.

Rather than considering only cases with $\mu/\sigma = \text{SNR}(\Delta\text{E}) = -\text{SNR}(\Delta\text{C})$, let us now also consider some extreme cases where only one of these two SNRs needs to be relatively large in absolute value. For example, when $\text{SNR}(\Delta\text{E}) \geq 2.6$, one's confidence that the new treatment is ME always exceeds 99.5%, at least according to the realistic normal approximation for the asymptotic distribution of ΔE. In these cases, our "some dominance" threshold is then met whenever $\text{SNR}(\Delta\text{C}) \leq 0.0$ and our "much dominance" threshold is also met when $\text{SNR}(\Delta\text{C}) \leq -0.4$, regardless of any ($\Delta$E, ΔC) correlation. Similarly, $\text{SNR}(\Delta\text{C}) \leq -2.6$ would essentially assure LC, but $\text{SNR}(\Delta\text{E}) \geq 0.0$ or $\text{SNR}(\Delta\text{E}) \geq +0.4$ would still be required for "some dominance" or "much dominance," respectively.

In summary, our proposed threshold pairings are most easily achieved (smallest, equal SNR absolute values) when the bootstrap distribution of ICE uncertainty reveals approximately uncorrelated outcome differences. In particular, our proposed threshold pairings work well together in the sense that any correlation making one of the two thresholds easier to achieve also makes the other threshold in the same pair harder to achieve. For example, our proposed "LC or ME" inclusive threshold will be most stringent (conservative) when outcome differences are highly positively correlated. In other words, after exploring several types of alternative true ICE scenarios, we believe our proposed "and/or" confidence threshold pairings are mutually supportive, consistent, intuitive, and sufficiently modest (relative to strict dominance) to be demonstrated in typical cost-effectiveness inference situations, using data either from clinical trials or from naturalistic studies.

2.5. The Role of Sensitivity Analysis in ICE Statistical Inference

As with most registration trials, the case study we describe here was not powered to detect differences between active treatments on the secondary cost measure needed in ICE statistical inference. Thus, although the results reported here cannot be definitive, their credibility is unquestionably enhanced by using sensitivity

analyses to demonstrate their robustness across various alternative ways to quantify outcomes and impute missing values.

The data that forms the basis for treatment comparisons are paired (cost, effectiveness) outcomes observed on individual patients within each of the two separate treatment cohorts. The bivariate cost-effectiveness outcome for each patient thus consists of his or her dollars-per-week measure of all resource utilization above and beyond what is dictated by the study protocol plus his or her corresponding summary measure of clinical efficacy (over the entire 8-week period of active treatment for MDD).

Each patient's cost measure (average resource utilization in dollars per week) is a normalized measure of area under the cumulative cost curve, and clearly has the required numerically smaller-is-more-favorable to "new" interpretation. No monotone transformation of cost, such as a logarithm, is used in the analyses reported here. Logarithms could not be taken here because roughly one-third of all patients reported zero resource utilization beyond that specified in the 8-week study protocol. Furthermore, the observed distribution of strictly positive dollars per week values was not highly skewed. Thus, the sensitivity analyses of cost described here are confined to only two alternatives: either imputing all missing values via AVCF or else using only patients with "relatively complete" data in the sense described in section 2.6.

The measures of effectiveness considered here must have numerically larger-is-better interpretations. HAMD-17 scores, themselves, have a smaller-is-better interpretation. In contrast, any "decrease in HAMD-17 score" from baseline to endpoint, denoted here by DB, does have the required interpretation. A patient's DB is his or her HAMD-17 score at baseline minus his or her HAMD-17 score at endpoint; this is the efficacy measure most commonly seen in U.S. registration trials for MDD. The alternative effectiveness measure considered here is "integrated" decrease in HAMD-17 score from baseline to endpoint, denoted by integrated decreases from baseline (IDB). IDB is the total (signed) area lying not only (1) above the curve resulting from connecting repeated HAMD-17 measures with straight line segments, but also (2) below the horizontal line defined by that patient's baseline HAMD-17 score. See Table 1 for two example calculations of DB and IDB.

To have been effectively treated for MDD, both the DB and the IDB value for a patient would need to be strictly positive. For two patients with the same numerical value for DB, the patient with the larger value of IDB could be said to have experienced an earlier or stronger onset of action in treatment for MDD. Our sensitivity analyses of effectiveness use either DB or IDB in combination with any one of the three methods for imputing missing HAMD-17 values described in section 2.6.

2.6. Imputation of Missing Values

Early study discontinuation averaged 28% by week 4 and 36% by week 8, and was not significantly different ($p \geq .05$) across treatment groups; see Goldstein et al. (2004) for more details. Some form of missing data imputation is thus required to produce patient-level summaries of overall cost and effectiveness. In addition to imputing missing values for all patients, we also performed an alternative "relatively complete" data analysis using only patients who completed five or more visits. No data from patients who missed four or more visits were used in this case.

Table 1 Trapezoid weights and example calculations of IDB on HAMD-17 total score

			Example 1		Example 2	
Weeks on drug	Visits	Trapeziod weight	HAMD-17	Decrease from baseline	HAMD-17	Decrease from baseline
0	3	0.5	32	0[a]	25	0[a]
1	4	1.0	31	1	24	1
2	5	1.5	31	1	23	2
4	6	2.0	30	2	25	0
6	7	2.0	30	2	27	−2
8	8	1.0	29	3	30	−5
	IDB			13.5		−5[b]
	DB			3		−5[b]

[a]Because visit three is baseline, change from baseline is always zero.

[b]A negative IDB or DB is an increase in depression severity.

Technical Note: The contribution to IDB from any two consecutive visits is defined, via the "Trapezoid Rule," to be the between-visit time "width" (of 1 or 2 weeks) times the average "height" of the area between the horizontal HAMD-17 baseline for the patient and the line segment connecting the consecutive observed or imputed HAMD-17 values for that patient. This signed "height" is the baseline HAMD-17 minus the average of the two consecutive visit HAMD-17 scores and will be strictly positive whenever an actual decrease from baseline has persisted between the two consecutive visits. The overall IBD for a patient is the sum of five trapezoidal areas and corresponds to weighting the six observed or imputed decrease from baseline values for active treatment weeks 1 through 8 as shown in Table 1. Note that the weights sum to 8 because the study protocol called for a total of 8 weeks of therapy for MDD.

This "relatively complete" data approach reduces the sample size of duloxetine 40 mg/d treated patients from 86 to 73, duloxetine 80 mg/d treated patients from 91 to 71, and paroxetine 20 mg/d treated patients from 87 to 70.

Mixed models, which include random and fixed effects, have been widely used in analyses of clinical data that contain missing values since their introduction by Laird and Ware (1982). The mixed models repeated measures (MMRM) method of Mallinckrodt et al. (2001a,b) is particularly attractive because it can be used as a primary analysis; the MMRM approach does not require imputation of missing values. Furthermore, MMRM assumes data are missing at random (MAR) is fairly robust in the sense that it treats visit (time) as a qualitative factor and is the primary method of longitudinal analysis for the current study reported in Goldstein et al. (2004).

In all effectiveness analyses reported here, either last observation carried forward (LOCF) or MMRM prediction is used to impute missing values before calculating each patient's summary DB or IDB measure.

All attempts to model longitudinal variation in expected cost ($/week) at the individual patient level failed; there appears to be no sound basis for concluding that expected cost per week is anything but constant over the 8-week period studied. Thus, all missing values of cost were imputed by AVCF.

The majority of statistical analyses reported here were performed using JMP® version 4.0.4, SAS Institute Inc.; but MMRM analyses were performed using SAS release 8.02; and bootstrap ICE quadrant confidence levels were simulated using version 2001.08 of ICEplane Obenchain (2003).

Table 2 Patient baseline characteristics by treatment group

Patient characteristic	Duloxetine 40 mg/d[a] $N = 86$	Duloxetine 80 mg/d[b] $N = 91$	Paroxetine 20 mg/d $N = 87$
Gender (% female)	56	62	64
Mean age in years (SD)	41 (10)	41 (12)	40 (11)
Ethnicity (%)			
Caucasian	84	85	74
Hispanic	10	10	14
African descent	5	5	10
Other	1	0	2
Mean HAMD-17 (SD)	18.74 (5.97)	17.86 (4.66)	17.83 (5.19)
Resource utilization			
Cost per week in $ (SD)	11.92 (4.33)	10.44 (2.47)	16.67 (5.13)
No utilization (%)	76	78	78

[a]Administered as 20 BID.
[b]Administered as 40 BID.

2.7. Characteristics of the Study Sample

The three active treatment groups did not differ significantly ($p \geq .05$) in baseline characteristics (Table. 2). Study participants were 60.6% female and had an average age of 40.6 (± 0.68) years. Somewhat fewer paroxetine recipients were Caucasian ($p = .12$). Participants had moderately severe depression, with average HAMD-17 scores of 18.1 (± 0.33) at baseline. Baseline average cost per week for health care resource utilization outside protocol was \$12.97 ($\pm$\$2.36) per week. There were also no significant differences ($p \geq .05$) in baseline resource utilization across treatment groups. The majority of patients (77.3%) used no additional resources in the 2 weeks immediately prior to their randomization to active treatment.

2.8. Effectiveness

As shown in Table 3, alternative methods of computing overall effectiveness resulted in the same numeric ordering of treatment groups. The greatest IDB were experienced by patients receiving duloxetine 80 mg/d followed by duloxetine 40 mg/d and, last by paroxetine 20 mg/d. The magnitude of HAMD-17 decrease across all treatment arms was greatest with MMRM imputation, especially when sample sizes were restricted to the patients with "relatively complete" data for five or more visits. No significant treatment differences ($p \geq .05$) were found, except that duloxetine 80 mg/d was more effective than paroxetine 20 mg/d in the DB/MMRM imputation analysis using all patients ($p = .04$.)

Alternative approaches yield the rather wide range of ΔE estimates displayed in Table 3. Note that LOCF imputation introduces serious downward bias into estimation of IDB, and the DB estimates are not directly comparable to the IDB estimates.

Table 3 Average and incremental analyses of clinical efficacy, costs, and cost-effectiveness by treatment group

	Duloxetine		Paroxetine
	40 mg/d[a] $N = 86$	80 mg/d[b] $N = 91$	20 mg/d $N = 87$
Average effectiveness HAMD-17 (SD)			
IDB/MMRM	127.52 (63.58)	128.84 (55.14)	122.68 (54.06)
IDB/LOCF	41.52 (49.17)	45.51 (46.64)	41.23 (44.91)
IDB/"rel. complete"	128.33 (65.38)	131.10 (55.47)	124.10 (57.95)
DB/MMRM	8.66 (6.98)	9.52 (6.26)	7.75 (7.22)
AVCF	9.60 (18.54)	5.75 (12.42)	8.65 (26.21)
"rel. complete"	10.18 (18.46)	5.64 (11.39)	7.10 (18.18)
Comparisons with paroxetine (SD)	$\Delta C/\Delta E = ICER$	$\Delta C/\Delta E = ICER$	
IDB/MMRM	0.95 (32.10)/4.84 (83.46) = 0.196	−2.90 (29.00)/6.15 (77.22) = −0.471	—
IDB/LOCF	0.95 (32.10)/0.29 (66.20) = 3.256	−2.90 (29.00)/4.28 (64.75) = −0.677	—
IDB/"rel. complete"	3.08 (25.91)/4.23 (87.36) = 0.727	−1.47 (21.46)/7.00 (80.22) = −0.471	—
DB/MMRM	0.95 (32.10)/0.92 (10.04) = 1.038	−2.90 (29.00)/1.78 (9.56) = −1.633	—

IDB, integrated decrease from baseline to end point; DB, decrease from baseline to end point; AVCF, average value carried forward; MMRM, mixed model repeated measure; LOCF, last observation carried forward.

[a]Administered as 20 BID.

[b]Administered as 40 BID.

2.9. Cost

Average costs of additional resources used during the trial were least for patients receiving duloxetine 80 mg/d, greater for paroxetine 20 mg/d, and greatest for patients receiving duloxetine 40 mg/d. When analyses were restricted to only patients with "relatively complete" data for five or more visits, average costs were further reduced for patients treated with duloxetine 80 mg/d ($5.64 ± $11.39 per week) or paroxetine 20 mg/d ($7.10 ± $18.18 per week). No significant treatment differences in cost ($p \geq .05$) were found.

Similar to baseline resource utilization within the 2 weeks immediately preceding randomization, the majority of patients did not report using any health care resources other than the services specified in the study protocol between baseline and endpoint. Percentages of patients reporting $0 per week after randomization did not differ by treatment ($p = .17$): 56% of patients receiving duloxetine 40 mg/d, 66% of patients receiving duloxetine 80 mg/d, and 69% of patients receiving paroxetine 20 mg/d.

2.10. Incremental Cost-Effectiveness

Details of alternative ICE analyses comparing duloxetine at 40 mg/d or 80 mg/d with paroxetine 20 mg/d are summarized in Tables 3 and 4. Figures 2A and 2B depict bootstrap distributions of ICE uncertainty; each is a simulated bivariate "scatter" of resampled ICE outcomes for the first 1000 of 25,000 bootstrap replications for one of our eight alternative analyses. Specifically, Fig. 2A (which compares the lower dose of duloxetine with the standard dose of paroxetine) and

Table 4 Bootstrap ICE quadrant confidence levels and dominance thresholds

	Duloxetine 40 mg/d[a] vs. Paroxetine 20 mg/d		Duloxetine 80 mg/d[b] vs. Paroxetine 20 mg/d	
	LC and ME	LC or ME	LC and ME	LC or ME
Effectiveness measure/imputation method				
IDB/MMRM	27%	81%	64%	96%
IDB/LOCF	20%	70%	61%	95%
IDB/"rel.complete"	10%	71%	55%	93%
DB/MMRM	31%	88%	79%	99%
Dominance thresholds (minimum confidence levels)				
Uncorrelated, equivalent treatments	25%	75%	25%	75%
Some dominance	≥50%	≥90%	≥50%	≥90%
Much dominance	≥65%	≥95%	≥65%	≥95%
Strict dominance	≥95%		≥95%	

For abbreviations, see footnote to Table 3.

In all four alternative analyses, the cost measure was dollars per week with imputation of missing values by AVCF.

[a]Administered as 20 BID.

[b]Administered as 40 BID.

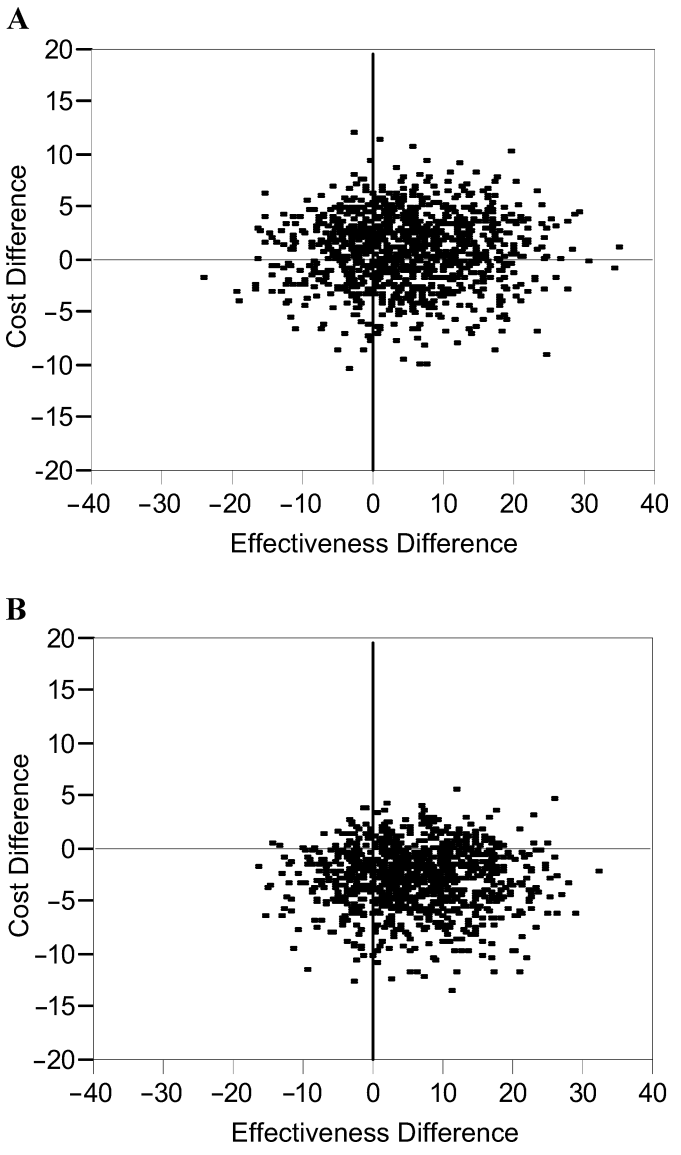

Figure 2 Bootstrap distributions of ICE uncertainty when the effectiveness measure is IDB with MMRM imputation: first 1000 of 25,000 total replications. (A) Duloxetine 40 mg/d[a] minus paroxetine 20 mg/d. (B) Duloxetine 80 mg/d[b] minus paroxetine 20 mg/d. [a]Administered as 20 BID; [b]administered as 40 BID.

2B (which compares the higher dose of duloxetine with the standard dose of paroxetine) depict results from the analyses that used all patients and MMRM imputation to define IDB for HAMD-17. As is typical in empirical practice, cost and effectiveness outcomes were essentially uncorrelated in our eight alternative bootstrap simulations.

Note that ICE uncertainty is clearly quite high in these analyses, primarily due to the limited number of patients in the current trial. For example, 95% confidence limits from Fieller's theorem do not exist for any of the eight ICERs listed in Table 3. In contrast, although most observed differences are too small to be declared significant at traditional confidence levels, interesting qualitative (numerical) comparisons have emerged.

Note that estimated ICE quadrant confidence levels for comparing duloxetine 80 mg/d as the new treatment with paroxetine 20 mg/d as the standard treatment ranged from 55% to 79% within the dominant ICE quadrant (LC and ME). This range of percentages is clearly higher than the minimum (25%) expected when comparing uncorrelated outcomes on two equivalent treatments. Similarly, the total ICE quadrant confidence levels summed over the three (LC or ME) inclusive ICE quadrants for duloxetine 80 mg/d relative to paroxetine 20 mg/d ranged from 93% to 99%, which is again higher than the 75% that would be expected from uncorrelated outcomes on equivalent treatments. In fact, using our proposed confidence thresholds (repeated in the last three rows of Table 4), duloxetine 80 mg/d displays not only "some dominance" over paroxetine 20 mg/d in all four sensitivity analyses, but also "much dominance" in the DB/MMRM imputation analysis.

Potential advantages of treating MDD with duloxetine 40 mg/d relative to paroxetine 20 mg/d were much less straightforward. All alternative methods attributed more than 50% confidence to duloxetine 40 mg/d being ME than paroxetine 20 mg/d. But duloxetine 40 mg/d may be more costly in patient-reported resource utilization than paroxetine 20 mg/d.

Of the four alternative types of ICE dominance analyses attempted, the traditional DB analyses in row four of Table 4 resulted in the highest ICE quadrant confidence levels in favor of duloxetine 80 mg/d or duloxetine 40 mg/d relative to paroxetine 20 mg/d. The ICE quadrant confidence levels least favorable to duloxetine came from the analyses of patients with "relatively complete" data for five or more visits, where sample sizes were smaller, but fewer missing values needed to be imputed than in other analyses.

3.0. SUMMARY AND DISCUSSION

3.1. Case Study Example

Much attention has been focused here on the numeric example because it illustrates some quite typical shortcomings when using data from clinical trials to address questions about head-to-head cost-effectiveness comparisons. Specifically, the numeric example illustrates the sort of careful attention needed to address the dual problems of small samples and missing values. Unlike clinical trials, studies designed to more realistically estimate real world costs tend to use much larger treatment groups and do not allow patients to discontinue from the cost accumulation phase of the study.

Like most registration trials, the study described here was powered only to detect anticipated duloxetine advantages over placebo and its noninferiority relative to a standard SSRI treatment. The study was not powered to detect differences in secondary cost measures needed to determine cost effectiveness.

Although methodology for powering studies to make cost comparisons is available (O'Brien et al., 1994; Willan and O'Brien, 1999), sample sizes this large are rarely found in U.S. registration trials where value comparisons are not mandated.

Our analyses using the 80 mg/d dose of duloxetine as the new treatment suggest that this higher dose of duloxetine displays, in our "and/or" threshold pairings sense, at least "some dominance" and possibly "much dominance" over paroxetine 20 mg/d in acute-phase treatment of MDD. Although duloxetine 80 mg/d does not appear to statistically "strictly dominate" paroxetine 20 mg/d (i.e., at least 95% confidence in LC and ME), evidence that this dose of duloxetine is indeed LC and ME than paroxetine at its standard dose is nevertheless much higher (55% to 79% confidence) than would be expected if these two treatments were equivalent and outcomes remained uncorrelated (25% confidence).

Because the incremental cost-effectiveness comparison between the lower dose of duloxetine (40 mg/d) and the standard dose of paroxetine (20 mg/d) is less favorable, the possibility of a dose–response relationship in acute-phase treatment of MDD with duloxetine is suggested. Interestingly, the primary effect of doubling the daily dose of duloxetine appears more as a reduction in resource utilization costs rather than as an additional improvement in clinical MDD effectiveness. For example, this cost reduction is quite clear when comparing Figs. 2A and 2B. In contrast, assuming linear pricing, this doubling of the daily dose would also double the acquisition cost of duloxetine.

The ICE comparisons made here should be updated after the price of duloxetine for treatment of MDD has been determined. As of this writing, we anticipate that the recommended starting and maintenance dose of duloxetine for treatment of MDD will be 60 mg/d, which is midway between the two doses considered in this study. We believe the relevance of Figs. 2A and 2B is enhanced by the fact that drug acquisition costs were not included in our computations. After all, the vertical scale in these figures is measured in dollars per week ($/week). That way, once a difference in acquisition cost per week is determined from announced prices (or changes later due to generic competition), it would be a simple matter to shift the origin of the vertical axis up or down in Figs. 2A and 2B to "visualize" the resulting ICE quadrant confidence levels.

Trial findings summarized here also have the usual limitations of registration studies of focusing on a clinical population with limited generalizability due to entry criteria restrictions imposed by FDA regulations. Because almost all anticipated resource utilization was scheduled within the trial protocol, most of the patients in this study had no additional resource use in their 8 weeks of active treatment for MDD. Also, standardized prices were used instead of actual costs, and concomitant medication costs, as well as costs associated with hospital and community services, may not have been fully captured. Larger and more naturalistic ICE comparisons that include patients with the full spectrum of medical comorbidities are needed to better inform decisions between alternative MDD treatments. For examples, see Obenchain et al. (1997) and Lave et al. (1998).

The growth of managed care and pharmacy benefit management has resulted in increasing reliance on economic evaluation of pharmaceuticals. These evaluations are particularly difficult when assessing new medications. Here, we present sensitivity analyses to begin to address this important but complex issue,

using data from a clinical registration trial. Longitudinal, naturalistic trials and/or observational studies are needed to support the results discussed.

3.2. Assumptions of the "Threshold Pairings" Approach

Our proposed threshold pairings are quite realistic when treatment differences on cost and effectiveness are approximately uncorrelated within the bootstrap distribution of ICE uncertainty. Instead, outcomes may be highly positively correlated in situations where, say, death is viewed as reducing cost to zero, but effective treatment preserves life and therefore increases both short- and long-term cost. In these cases, the "LC or ME" inclusive threshold may be the most difficult to achieve. Alternatively, outcomes could be highly negatively correlated in situations where effective treatment avoids a serious adverse event (such as a hip fracture) that triggers large end-of-life costs; the "LC and ME" threshold would be most demanding in these cases. In contrast, whenever cost and effectiveness measures are genuinely highly correlated, it is really only necessary to analyze one of these two outcomes; an analysis of either cost or effectiveness alone would then tell much of the whole story!

Our proposed confidence threshold pairings are most appropriate when cost and effectiveness outcomes are measured separately and independently. In other words, our "and/or" approach appears to be most realistic in truly two-dimensional cost-effectiveness comparisons, when both thresholds are equally stringent and neither outcome alone can provide a reasonable overall summary.

3.3. Advantages of the "Threshold Pairings" Approach

Conceptual Simplicity: Our proposal is actually quite easy to understand, explain, and apply in practice. Furthermore, due to large-sample statistical theory yielding asymptotic normality, the bootstrap approach to quantifying ICE uncertainty has considerable intuitive appeal. More important, we demonstrate that our threshold pairings are highly realistic in the sense that they are actually achieved by bivariate normal distributions for uncorrelated (nonredundant) outcomes. Finally, ICE quadrant confidence levels can be easily visualized using highly versatile and informative scatter plots, or simply tabulated.

Outcome Standardization Unnecessary: Arbitrary rescalings of cost measures and/or effectiveness measures leave ICE quadrant confidence levels unchanged. In other words, our threshold pairings proposal is invariant to specification of society's (unknown) true willingness to pay for health care, λ. Furthermore, there is no need to convert one's effectiveness measure into quality adjusted life years. These are truly major advantages.

Highly Discriminating: Surprisingly, simply by stressing relatively high confidence levels (of at least either 50%, 65%, or 95%) that new is not only LC, but also ME than standard, all three of our dominance proposals also turn out to be highly discriminating. For example, conventional approaches, such as the ICER and NMB methods based on willingness to pay, can label a new treatment "cost effective" when, in reality, it is clearly more costly. A new treatment of this supposedly well-worth-its-additional-cost type will fail to meet our first "LC and

ME" threshold. In other words, a new treatment that is dominant in any of our proposed senses is almost surely cost effective in all the traditional senses, but not vice versa.

ACKNOWLEDGMENTS

The authors thank the associate editor and referees for comments that greatly improved the logic and readability of this article.

REFERENCES

Black, W. C. (1990). The CE plane: A graphic representation of cost-effectiveness. *Med. Decis. Making* 10:212–214.

Briggs, A. H., Fenn, P. (1998). Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane (Student Corner). *Health Econ.* 7:723–740.

Briggs, A. H., Wonderling, D. E., Mooney, C. Z. (1997). Pulling cost-effectiveness analysis up by its bootstraps: A non-parametric approach to confidence interval estimation. *Health Econ.* 6:327–340.

Chaudhary, M. A., Stearns, S. C. (1996). Estimating confidence intervals for cost-effectiveness ratios: An example from a randomized trial. *Stat. Med.* 15:1447–1458.

Cook, J. R., Heyse, J. F. (2000). Use of an angular transformation for ratio estimation in cost-effectiveness analysis. *Stat. Med.* 19:2989–3003.

Copley-Merriman, C., Egbuonu-Davis, L., Kotsanos, J. G., Conforti, P., Franson, T., Gordon, G. (1992). Clinical economics: A method for prospective health resource data collection. *Pharmacoeconomics* 1(5):370–376.

Goldstein, D. J., Lu, Y., Detke, M. J., Wiltse, C., Mallincrodt, C., Demitrack, M. A. (2004). Duloxetine in the treatment of depression—A double-blind, placebo-controlled comparison with paroxetine. *J. Clin. Psychopharmacol.* 24:389–399.

Hamilton, M. (1967). Development of a rating scale for primary depressive illness. *British J. Soc. Clin. Psychol.* 6:278–296.

Laird, N. M., Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 38:963–974.

Laska, E. M., Meisner, M., Siegel, C., Stinnett, A. A. (1999). Ratio-based and net benefit-based approaches to health care resource allocation: Proofs of optimality and equivalence. *Health Econ.* 8:171–174.

Lave, J. R., Frank, R. G., Schulberg, H. C., Kamlet, M. S. (1998). Cost-effectiveness of treatments for major depression in primary care practice. *Arch. Gen. Psychiatry* 55(7):645–651.

Mallinckrodt, C. H., Clark, W. S., David, S. R. (2001a). Accounting for dropout bias using mixed-effects models. *J. Biopharm. Stat.* 11:1–13.

Mallinckrodt, C. H., Clark, W. S., David, S. R. (2001b). Type I error rates from mixed effects model repeated measures versus fixed effects ANOVA with missing values imputed via last observation carried forward. *Drug Info. J.* 35:1215–1225.

O'Brien, B. J., Drummond, M. F., Labelle, R. J., Willan, A. R. (1994). In search of power and significance: Issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Med. Care* 30:231–243.

Obenchain, R. L. (1997). Issues and algorithms in cost-effectiveness inference. *Biopharm. Rep.* 5:1–7.

Obenchain, R. L. (1999). Resampling and multiplicity in cost-effectiveness inference. *J. Biopharm. Stat.* 9(4):563–582.

Obenchain, R. L. (2003). *ICEplane: Microsoft Windows Software for Incremental Cost-Effectiveness Confidence Regions, Tolerance Regions, Acceptability and Net Benefit via Bootstrapping*. Copyright © Pharmaceutical Research and Manufacturers of America (PhRMA). http://www.math.iupui.edu/∼indyasa.

Obenchain, R. L., Melfi, C. A., Croghan, T. W., Buesching, D. P. (1997). Bootstrap analysis of cost-effectiveness in antidepressant pharmacotherapy. *Pharmacoeconomics* 11(5):464–472.

Polsky, D., Glick, H., Willke, R., Shulman, K. (1997). Confidence intervals for cost effectiveness ratios: A comparison of four methods. *Health Econ.* 6:243–252.

Schoenbaum, M., Unutzer, J., Sherbourne, C., Duan, N., Rubenstein, L. V., Miranda, J., Meredith, L. S., Carney, M. F., Wells, K. (2001). Cost-effectiveness of practice-initiated quality improvement for depression: Results of a randomized controlled trial. *JAMA* 286(11):1325–1330.

Stinnett, A. A. (1996). Adjusting for bias in C/E ratio estimates. *Health Econ.* 5:470–472.

Stinnett, A. A., Mullahy, J. (1998). Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Med. Decis. Making* (Special Issue on Pharmacoeconomics) 18:S68–S80.

Tambour, M., Zethraeus, N. (1998). Bootstrap confidence intervals for cost-effectiveness ratios: Some simulation results. *Health Econ.* 7:143–147.

Van Hout, B. A., Al, M. J., Gordon, G. S., Rutten, F. F. H. (1994). Costs, effects and C/E ratios alongside a clinical trial. *Health Econ.* 3:309–319.

Willan, A. R., O'Brien, B. J. (1999). Sample size and power issues in estimating incremental cost-effectiveness ratios from clinical trials data. *Health Econ.* 8:203–211.