Strengthening Clinical Effectiveness Trials: Equipoise-Stratified Randomization

Philip W. Lavori, A. John Rush, Stephen R. Wisniewski, Jonathan Alpert, Maurizio Fava, David J. Kupfer, Andrew Nierenberg, Frederic M. Quitkin, Harold A. Sackeim, Michael E. Thase, and Madhukar Trivedi

As psychiatric practice patterns evolve to take advantage of the growing list of treatments with proven efficacy, research studies with broader aims will become increasingly important. Randomized trials may need to accommodate multiple treatment options. In completely randomized designs, patients are assigned at random to one of the options, requiring that patients and clinicians find each of the options acceptable. In "clinician's choice" designs, patients are randomized to a small number of broad strategies and the choice of specific option within the broad strategy is left up to the clinician. The clinician's choice design permits some scope to patient and clinician preferences, but sacrifices the ability to make randomization-based comparisons of specific options. We describe a new approach, which we call the "equipoise stratified" design, that merges the advantages and avoids the disadvantages of the other two designs for clinical trials. The three designs are contrasted, using the National Institute of Mental Health Sequenced Treatment Alternatives to Relieve Depression trial as an example. Biol Psychiatry 2001; 50:792–801 *© 2001 Society of Biological Psychiatry*

Key Words: Equipoise, clinical trials, design, methodology, treatment, statistics

Introduction

The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study (Rush 2001) aims to define the optimal sequence of treatment options in patients whose major depressive disorder has not sufficiently improved with a vigorous course of a selective serotonin reuptake inhibitor (SSRI), citalopram. As psychiatric practice patterns evolve to take advantage of the growing list of efficacious treatments, research studies with broader aims will become increasingly important. Likewise, in other chronic disease areas that have benefited from an increase in available treatment options, clinicians face similar questions of treatment sequencing in cases of inadequate or incomplete response to front-line therapy.

Treatment research, however, is a victim of its own success at developing new medications, as well as psychosocial and other kinds of interventions. In STAR*D, the final list of treatment options for patients with incomplete response to citalopram included seven contenders at the "next step" (see Table 1). This number could have easily risen to more than a dozen, were it not for practical limits on the feasible number of study treatment arms. The process of winnowing the list down to the final choices sparked lively and contentious discussions because there are so many different treatment options with some claim on the researchers' attention. Even limited to seven treatments, the design raises many issues of feasibility, statistical power, and complexity. One of the most difficult issues concerns the way that the randomization will accommodate the multiplicity of treatment options defined in Table 1. Broadly speaking, there are two design templates in common use.

The simplest and scientifically most straightforward method is to recruit patients whose treatment can be chosen at random from the full set of seven options. In this "completely randomized" (CR) design, a patient for whom *any* of the treatments is not acceptable must be excluded from the study, or will choose not to enter, as might any patient for whom one of the options is clearly preferred. Thus, only those patients able to pass through this intake sieve would be assigned at random to one of the seven treatments, yielding a completely randomized design, but with possible constraints on generalizability due to sample selection bias.

An alternative design combines treatment options into larger groups of strategies. Table 1 suggests one such aggregation into "switch" or "augment" treatment strategy groups. Then each patient is assigned at random to one of the two strategies (switch or augment), and the choice of

From the Department of Veterans Affairs Cooperative Studies Program, Palo Alto VA, Palo Alto, California and Stanford University Department of Health Research and Policy (PWL); Department of Psychiatry, University of Texas Southwestern Medical Center at Dallas (AJR, MT); University of Pittsburgh Departments of Epidemiology and Psychiatry (SRW); Clinical Psychopharmacology Unit, Massachusetts General Hospital, Boston, Massachusetts (MF, AN, JA); Department of Psychiatry, UPMC Health System, Western Psychiatric Institute and Clinic (DJK, MET); and New York State Psychiatric Institute, Columbia University, New York, New York (HAS, FMQ).

Address reprint requests to Dr. P.W. Lavori, Stanford University School of Medicine, Department of Health Research and Policy, M/C 5405, Stanford, CA 94305.

Received March 2, 2001; revised June 5, 2001; accepted June 13, 2001.

0006-3223/01/$20.00
PII S0006-3223(01)01223-9

Table 1. Individual Treatment Options in STAR*D after Failure of Citalopram

Switch Options:
● Bupropion (BUP) ● Sertraline (SER) ● Venlafaxine (VEN) ● Cognitive Therapy (CT)
Augmentation options:
● Citalopram + Bupropion (+BUP) ● Citalopram + Buspirone (+BUS) ● Citalopram + Cognitive Therapy (+CT)

the specific option within the strategy is left to the judgment of the clinician. The "clinician's choice" (CC) design offers a way to permit flexibility of treatment, while still providing a comparison of the broad strategies. One of the obvious advantages of the CC design is that it permits a much wider population of patients to be studied, including patients for whom as few as one option in each strategy is acceptable. An example of such a design is described by the Atrial Fibrillation Follow-Up Investigation of Rhythm Management (AFFIRM investigators 1997), in the context of atrial fibrillation, comparing rate and rhythm management, but leaving the specific choice of drug up to the clinician.

In this article, we discuss the strengths and weaknesses of the CR and CC designs, and describe an alternative "equipoise stratified" (ES) design that was developed and specifically adopted for the STAR*D project. We describe statistical estimation and hypothesis testing for the new design, as well as power and sample size considerations. We identify determinants of feasibility of the new design, and briefly touch on the particular concerns about research ethics and informed consent that this new design provokes.

Critique of the Completely Randomized Design

The simplicity of the CR design has much to recommend it. Statistical inference from a CR design is straightforward, since it is just a more ambitious version of the standard two or three group design used in nearly all clinical efficacy research; however, the design may suffer scientifically (e.g., in terms of generalizability) and practically (e.g., in terms of slow recruitment), because of the exclusions and preferences in the patient reference population. To consider randomizing a STAR*D patient to the CR design, both the patient and the clinician must be in the state of "equipoise" with respect to all seven of the options in Table 1. Equipoise refers to the absence of conviction that one or more of the treatment options are unacceptable to the particular patient, or that one or more of the options are clearly superior. Thus, even if the patient and clinician are indifferent to the choice among six options, if the seventh option is either unacceptable or clearly superior for this patient, randomization among the seven options would not be warranted and the patient must be excluded.

In the context of STAR*D, whole categories of patients (such as those who are not interested in psychotherapy or who do not want to take more than one medication) would be excluded. Furthermore, some patients may have a response without remission following a course of citalopram. Such patients (or their clinicians) may be reluctant to abandon these initial gains by switching to another treatment altogether. The preferences and attitudes of individual clinicians about the unsuitability of certain treatments for particular patients will also limit participation. For example, if a patient has received one of the study medications in the past with disappointing results or dose-limiting side effects, the clinician may not be comfortable with the possibility of randomization to that option.

The STAR*D planners estimated that only a small minority—perhaps less than 10%—of otherwise eligible and consenting patients would fall into the stratum defined by equipoise among all seven options. Such a low eligibility rate would be unacceptable, given the intent to compare the effectiveness of treatment alternatives in a broadly defined population. The resulting intake sample would represent a small and distinct proportion of the affected population, severely limiting the generalizability of study results. Generalizing CR study findings to the excluded populations requires an unverifiable assumption; namely, that sample characteristics are unrelated to the efficacy of treatments. Even if one were prepared to make such assumptions, the low efficiency of recruitment would cause the cost and duration of a seven-group CR design to rise to unacceptable levels in the pursuit of adequate statistical power.

Switch or Augment: The Clinician's Choice Design

As shown in Table 1, the natural structure of the options for treatment points toward a two-strategy clinician's choice (CC) design. In a CC design, subjects are randomized between the two broad strategies, switch or augment, with the particular treatment option chosen by the clinician from the corresponding list in Table 1. For a patient to be eligible for the CC design, there must be at least two treatments, not all from one strategy, that the clinician and patient would accept. The reference population would be broadened by the CC design, owing to the much less restrictive condition for equipoise. The STAR*D planners thought that about half of all otherwise eligible, consenting subjects would satisfy the requirement for the CC design. A five-fold increase in recruitment efficiency, compared to the CR design, would reduce the cost and duration of the

CC study. The CC design offers the clinician greater flexibility, within the broad strategy chosen by the randomization, and, therefore, decreases the likelihood that the clinician will experience unacceptable restrictions on his or her judgment. Such flexibility is a major inducement to clinicians to participate, and contributes to the "ecological validity" of the study by its light hand on the clinician's arm.

Flexibility does come at a cost; namely, the loss of ability to compare specific options. The protections against bias provided by the randomization by strategy grouping do not extend to the contrasts of specific treatment options within or across the two strategies, such as switch to bupropion (BUP) versus switch to venlafaxine (VEN) or the addition of buspirone (+BUS). Contrasts among options are confounded by the patient characteristics that influence the choice of option, given the (randomized) assignment of strategy. Furthermore, suppose it is discovered in subsequent studies that one of the options (e.g., SER) is really less effective than the other options in its (switch) class, perhaps because failure to respond to citalopram selects patients who would not benefit from another SSRI. Then, if the SER option was a popular choice among clinicians in the CC study (for patients randomized to switch), the whole study is undermined. This is because there is no way to excise the SER outcomes from the analysis and still retain the randomization inferences. In contrast, the CR design allows contrasts among all pairs of treatments.

An argument in support of CC begins with the likely premise that the clinician knows something about the patient and can, therefore, optimize the choice of a specific option within the broad class. Then one obtains from CC a comparison of putatively optimized strategies, rather than a comparison of options chosen without the full benefit of the clinicians' specific insights. For this argument to have force, it must be true of most patients that the clinician has a clear patient-specific preference among options *within* a strategy, but not *between* strategies. Thus, it must often occur that the clinician strongly prefers exactly one option from one or both strategies, and that this preference varies among patients. (To be a meaningful preference, the choice must be reliable, such that the same patient presenting to different clinicians should elicit similar preferences; otherwise the goal of optimization is not achievable.)

The randomization-based inference is confined to the contrast of broad strategies, whose interpretation depends on the completely unspecified mix of clinician judgments and preferences that prevailed in the study. The more heterogeneity in the choices among clinicians, the less well the study informs the individual clinician's decisions. In the worst case, if the choice among "switch" options is made poorly while the choice among "augment" options is made well, the augment strategy may win the horse race even if in each patient the optimal switch option would uniformly beat the optimal augment option. Nevertheless, the "optimization" argument has a strong attraction based on a germ of truth, and we will return to it below.

The above discussion emphasizes the importance of the choice of strategy (augment or switch). It might indeed turn out that one or the other strategy overall is a better guide to decision, especially if the individualized decisions on specific options add value. However, the inability to extend inference to the individual options suggests that the design of studies should not be based on such generalized comparisons unless the scenario of the preceding paragraph obtains: clear, reliable, and valid patient-specific options within strategy.

Next, we describe another approach to the problem—one that makes it possible to make broad comparisons among strategies *when it is clinically sensible to do so*, while also preserving the ability to make randomization-based statistical inferences at the level of specific treatments. The method is designed to preserve the strengths of the CC design while extending randomization-based inference to cover wider possibilities.

The Equipoise-Stratified Randomized Design

Equipoise is part of the ethical basis of experimentation in humans. To be in equipoise with respect to a set of prospective treatment options is to regard them as approximately equal in terms of likelihood of success. Lack of definitive scientific evidence creates controversy among experts, who may agree that there is uncertainty about the right treatment in a particular context. A specific clinician and patient, after reviewing what is known about the effects of the alternative treatments, including the fact that expert opinion is not settled, may agree that all of the options presented in a trial are of roughly equal potential benefit. To consider entering a patient into any study, the clinician and patient must be able (in principle) to define the list of specific study treatments that are acceptable and of rough parity. We call this list the "equipoise stratum" (ES or e-stratum). Furthermore, there must not be a treatment available to the patient that is known to be better than any of the treatments on the list. The list of acceptable treatments might depend on the patient's own history of response to treatment, including side effects as well as efficacy. It might depend on the clinician's knowledge of relevant patient characteristics, which have been shown to predict response in other patients. It might reflect the patient's own preferences with regard to particular known side effects or mode of action. Note also that a patient's equipoise-stratum might change over time, in response to subsequent experience with treatments.

The ES design randomly assigns each patient to a specific option within that patient's e-stratum. Then, individual pairwise contrasts of treatments are built up by statistical pooling of within-stratum contrasts. Because the contrasts within ES are based on randomization, they are not confounded as in the CC design. In addition, the ES design retains the desirable features of the CC design by allowing the patient and clinician to exclude certain options. This exclusion limits the generalization, of course, to subject groups that accepted randomization to some options that are being compared. Essentially, the ES design converts the clinician's judgment from an unspecified, postrandomization confounder in the CC design to a fully observed, prerandomization factor that can be balanced explicitly, and, therefore, statistically controlled. In the following, we describe the ES configuration expected to obtain in the STAR*D design, and then discuss the statistical process involved in making specific contrasts among options, using both within and across strategy examples.

The ES design is related to a more complex design suggested by Kadane and Sedransk (1980) and Kadane (1986). They proposed allowing accruing data in the trial to influence the acceptability of treatment options and provided a Bayesian framework for inference. Korn and Baumrind (1991; 1998) reviewed issues related to clinician preferences for treatment, and the resulting difficulties with inference about treatment effects from observational studies as well as generalizability from randomized trials. They proposed a design that allows randomization of consenting patients to a clinician who favors treatment A or a clinician who favors B, as long as the panel of clinicians is not unanimous in their preferences for A or B. The commentaries by Zelen, Freedman, Ashby, and Harrison contained in the Korn and Baumrind (1998) review are also relevant to the issues raised here.

*What Equipoise Strata are Expected in STAR*D?*

In the abstract, there are 120 possible equipoise strata in STAR*D, since there are 128 (2 to the seventh power) distinct subsets of the seven treatment options at Level 2, of which eight subsets are irrelevant (the empty set and the seven singletons). Because, however, of regularities in the strata (Rush 2001), it appears likely that many of the potential ES will be unpopulated. In the opinion of the STAR*D planners, over 95% of patients are expected to fall into one of the seven largest equipoise strata. In Table 2, we list those seven strata, along with the estimated proportion of patients expected to fall into each stratum. The small residual minority may be included arbitrarily into a "near neighbor," more restrictive large stratum. For example, fewer than 1% of patients are expected in the ES defined by the options {+BUP, +BUS, CT, +CT}. These patients who cannot accept a *medication* switch might be accommodated either in the ES {+BUS, +BUP, +CT} (any augmentation) or in the ES {CT, +CT} (switch to or augment with CT). By dividing the proportion of patients expected to fall into each stratum by the number of options in the stratum, we obtain the proportion of patients assigned to each option, in each stratum (see Table 2, column labeled "option percent"). Assuming a total sample size of 2000 entering Level 2 (incomplete response to citalopram), we can estimate the expected number of patients assigned to each option (see Table 2, last column). For example, stratum 2 (Any Switch) is expected to be acceptable to about 12% of patients, and since there are four treatment options that are "switches," each option is expected to receive 3% of the total sample, or 60 patients, from this stratum. The reader will note that we do not believe that many patients would occupy strata such as {SER,VEN}, in which one of the switch medications (BUP) has been ruled out, but the other two are acceptable. Such strata accommodate patients who have tried that medication (BUP) in the past, without success. Note that STAR*D excludes at intake patients who have had an unsuccessful, full-strength course of any study medication during the index episode, or who have demonstrated clear intolerance or have a medical contraindication to any study medicine. We believe that few patients will have experienced a "full-strength" trial of the study medications in the current episode, and, therefore, nearly all will be willing to accept randomization to all three medication switch alternatives. This assumption is, of course, subject to empirical test.

Table 2. What ES are There?

Stratum Name (Option list)	ES %	Option %	*n* Per Option
1. Universal Donor (SER, BUP, VEN, CT, +BUS, +BUP, +CT)	8.4	1.2	24
2. Any Switch (SER, BUP, VEN, CT)	12.0	3.0	60
3. Any Augment (+BUS, +BUP, +CT)	12.0	4.0	80
4. Any Medication (SER, BUP, VEN, +BUS, +BUP)	25.0	5.0	100
5. Any Medication Switch (SER, BUP, VEN)	14.0	4.7	94
6. Any Medication Augment (+BUS, +BUP)	14.0	7.0	140
7. No New Medication (CT, +CT)	11.0	5.5	110
Total	96.4%		

Statistical Analysis for the Equipoise Stratified Design

Consider the contrasts among medication switch options (SER, BUP, and VEN). From Table 2, one can read off the four ES that contribute to this contrast (they are the first, second, fourth, and fifth strata). These four ES contribute

Table 3. Pairwise Comparisons among Four Options

Pair	Contributing ES
Medication Switch vs Medication Augment	1 and 4
Medication Switch vs CT Switch	1 and 2
Medication Augment vs CT Augment	1 and 3
CT Switch vs CT Augment	1 and 7

(respectively) 24 + 60 + 100 + 94 = 278 patients to each medication switch option (see the last column of Table 2). To compare the three medication switch options, one can proceed as if the design were a three-group, stratified design. For example, if the outcome is binary (such as full recovery), a Mantel-Haenszel test can be applied to the four contingency tables, each tabulating the (three) groups by (two) outcomes results. Or, symptom-based scores can be analyzed by a standard F test for treatment effect applied to the two-way layout of strata and groups. In either of these two approaches, the inclusion of the ES as a stratifying variable in the analysis accounts for possible differences in the patients associated with the decision about which treatment options are acceptable (main effect of ES). The three pairwise contrasts of options (such as BUP vs. VEN) can be analyzed as usual, by *t* test or *z* test, given a significant omnibus F- or chi-square test. The "medication augmentation" test has 24 + 80 + 100+ 140 = 344 (from the first, third, fourth, and sixth ES listed in Table 2), in each of two groups (+BUS, +BUP). Note that the comparison of the three options within the medication switch strategy is completely independent of the comparison of the two medication augment options. This is because no patient contributes to both kinds of comparison. Therefore the analyses can proceed as if the data for switch and augment strategies were collected in two separate studies.

If *either* the "medication switch" or "medication augment" null hypothesis is rejected (i.e., one or more treatments found to be especially effective), one would not want to combine the corresponding results into a broad strategy contrast, since that would lump "winners and losers." Instead, we identify the surviving representatives (superior treatments) in both the medication switch and medication augment options. Specifically, if the medication switch null hypothesis is rejected, the switch representative is the option with the best result; otherwise it is the pooled set of options. The same procedure applies to the medication augment options. Then, the two representatives contribute to tests among the four options of medication switch, medication augment, CT switch, CT augment. Table 3 tracks the ES contributions to each pairwise contrast of the treatment strategies. Table 3 shows that the samples overlap due to the patients in ES 1, so they are not quite independent (as they would be if there were no patients in common); however, ES 1 is expected to be small, so the correlation of the statistical tests is also small. Note that the two pairwise tests to which *only* the universal donor ES contributes (ES 1) are not listed, since the sample size and statistical power are too low. These two comparisons are "Medication Switch versus CT Augment" and "Medication Augment versus CT Switch."

As discussed above, one or both of the initial medication switch or augment tests may reject the null hypothesis; that is, there may be a significant difference among the three medication switch options, or between the two medication augment options. Then only the "winning" option(s) would go forward to the tests against each other and the two CT options.

This approach justifies the ES design, because such detection and comparison cannot be conducted with a CC design in an unconfounded manner. For example, suppose there is one option (say, "switch to BUP") that is much better than the others. The CC design may have low power to detect the strategy that includes a single winning option with the randomized comparison of "augment versus switch" (barring a fortuitous pileup of clinician's choices on the option "switch to BUP"), because the benefits of "switch to BUP" would be diluted by the results of the less effective switch options. The contrast among all seven individual treatment groups might have more power to reject the null hyothesis, since "switch to BUP" would stand alone, unmixed with other options; however, recall that the CC design randomizes the patient to a strategy (such as "augment" or "switch") but it leaves the specific choice of option within strategy (which specific augmentation or switch) to the clinician. Therefore, CC does not randomize patients among options, but only among strategies. Even a highly significant result favoring one option cannot rule out the possibility that the result is due to the confounding of patient characteristics with clinician's choice (in the CC design).

Interaction of Equipoise Strata With Treatment Options

For analyses of both categorical and continuous variable outcomes, the treatment group by stratum interaction is of interest, testing the important question of whether the effect of treatment depends on the ES. The STAR*D design and power calculation are based on the assumption of a negligible stratum-by-treatment interaction, but it is useful to have a test available, albeit with limited power. We expect that the "Universal Donor Stratum" will contribute only 24 patients per treatment to this contrast, so that the "global" comparison among treatments in that stratum has very low power. Other strata contribute in the neighborhood of 100 patients per treatment. Thus, certain components of the interaction contrast are tested at acceptable power. For example, consider the test of the null interaction hypothesis that the difference in outcomes between SER and BUP switching options does not depend on whether an augmentation drug treatment is acceptable. This interaction contrast is built up by pooling the SER versus BUP differences from ES 2 and 5 (which allow SER vs. BUP but disallow any medication augmentation) and comparing that effect to the corresponding SER versus BUP effect from ES 1 and 4 (which allow SER vs. BUP, and also medication augmentation options). A total of 308 patients would contribute to the estimate of the SER versus BUP effect in "non augmentable" patients, and 248 to the SER versus BUP effect in "augmentable" patients. (From Table 2, ES 2 and 5 contribute 60 + 94 = 154 to each of SER and BUP, for a total of 308, while ES 1 and 4 contribute 24 + 24 + 100 = 124 to each option, for a total of 248.)

Refinement of the Equipoise Strata Analysis to Account for Other Stratifiers

Other prognostic factors might be used as stratifying variables. In particular, the clinical site is a customary stratifier. The equipoise strata can be defined within each site, giving rise to site-specific ES. This introduces no difficulties of a fundamental nature, but suggests an alternative line of analysis based on combining randomization-based tests across site-specific ES.

Within each site, there is a version of the ES distribution described in Table 2. The analyses described above can be conducted separately in each site, but since the sample sizes within site-specific ES may be small, it may be desirable and feasible to compute the *p* value for a given contrast within each ES by a randomization (so-called "exact") test. For example, consider ES 3, in which remission rates among the medication augment options are compared. With 80 patients per group, from as many as 30 sites, a single site might have as few as two or three patients per group. In that case, it might be possible to enumerate all permutations of the treatment group labels that preserve the site/ES totals. One would calculate the Mantel-Haenszel stratified chi-square test over site/ES micro-strata for each random permutation of treatment group labels. In the simplest case of one patient in each of two groups, there are only two re-randomizations, including the observed one. Across 30 sites, there would be about one billion different re-randomizations, and it would be feasible to calculate the MH chi-square test for each one. Then it would be possible to tabulate the full randomization distribution of the chi-squared statistic (under the null hypothesis) and calculate the tail area to the right of the observed statistic, which would be a model-free test of the null hypothesis. For cases where enumeration of the randomization distribution is not feasible, it would be possible to sample it, using Monte-Carlo methods. To the extent that the ES controls some of the variation in outcome, the use of randomization tests in this way can "tighten up" the trial, as described by Tukey (1993).

Model-based Analyses

If the investigator is prepared to write down a statistical model for the effects of the factors (including the ES), along with treatment, then the possibilities widen. For example, a linear (additive) regression model for the effects of treatment and strata expresses the expected outcome as a simple function of stratum membership and treatment received, with coefficients estimated from the observed data. The coefficients for the treatment indicators can be interpreted as causal effects of the treatment, given the randomization and the correct specification of the model. The coefficients can be tested for statistical significance using the error terms in the model, as usual in the analysis of variance. To use such a model one must assume that the effects of treatments are the same in each stratum, i.e., there is no treatment-by-stratum interaction. Under such an assumed model for the effects of treatment, it is possible (for example) to use data from ES 2 (any switch) and ES 7 (no new medication) to contribute to inference about the relative merits of medication switch versus CT augmentation, even though neither stratum has a head-to-head comparison of these two options.

For example, if CT is better than SER in ES 2, and +CT is better than CT in ES 7, it is attractive to conclude that +CT is better than SER. One must be clear about what group to which the latter statement applies. If the ES truly have patient-based reality then, strictly speaking, it does not apply to ES 2 or ES 7, since the patients in these groups do not have a well-defined effect of SER versus +CT. Indeed, it would only apply to ES 1, where there is a direct (but low N) comparison available.

This is a classic "bias/variance" tradeoff: the estimate of SER versus CT+ (to be applied to patients like those in ES 1) obtained indirectly (using ES 2 and ES 7) may be biased, while the direct estimate is unbiased. But the former estimate will have lower sampling variance due to the larger sample size. Minimizing the mean squared error (the sum of the variance and the squared bias) leads to "shrinkage" estimators of the Stein type (weighted averages of the direct and indirect estimators, where the weights depend on the between and within ES variation in the effects). If the indirect and direct estimates are far apart, relative to the sampling error of the direct estimates, the indirect estimates will be down-weighted. So ES 1 is useful in checking the assumptions of the model of equal effects.

Clinician's Choice: Equivalence or Optimization?

There are two different interpretations of the uses of the CC design. First, some of the options within a strategy may be regarded as equivalent, across a large number of patients. In that case, the clinician's choice represents inessential variation. The more interesting position, referred to above, is that the clinician's choice represents the best option, based on patient characteristics. The STAR*D study takes place against a background of sparse information about the relative merits of the various options (Rush 2001). Therefore, the STAR*D investigators assume that most clinicians *should* start out with equipoise among most of the treatment options in Table 1, for the generic patient.

Moving from the generic to the specific patient, there are two kinds of patient-specific information that come into play. These are the patient's prior experience with one or more specific *options*, before the citalopram trial, and the patient's a priori rejection of certain entire *strategies*, based on preferences without direct experiential support. As discussed above, patients who have experienced robust trials with one or more of the study treatments in the index episode will be excluded. We believe that most patients will not reject a possible aggressive retrial of a treatment they have experienced (without success) in previous episodes. This, too, will require empirical test.

In contrast, we suspect that a priori rejection of strategies may play the major role in determining the ES. For example, a patient who has experienced a partial remission on CIT may not be an appropriate candidate for a switch option. A patient who has little interest in "talk therapy" may rule out any strategies involving CT. Rejection of entire strategies leaves unspecified the choice among treatment options within the acceptable strategies. As a result, neither the patient's nor the clinician's choice is likely to contribute to optimizing the treatment to a significant degree. We note that comparisons of acceptability are of interest to the field, and contribute new information in and of themselves.

A broad strategy is a reliable guide to treatment only if the basis for the within-strategy variations in options are well understood. A broad strategy may work because the options are interchangeable and equivalet for the bulk of patients. This would be useful to know. The ES design builds in the ability to detect gross failures of equivalence, as described above. If one of the broad strategies (e.g., switching) has on average a poor result because only one of the particular switches actually works well in most patients, or a popular switch option works especially poorly, then the *clinical* idea of the strategy fails, independently of how it is studied in an experiment. The ES design offers the ability to detect the failure of specific options and broader strategies.

Practical Comparison of Clinician's Choice and Equipoise Stratified Designs

The clinician intending to randomize patients into the CC design must get information of the same general kind as needed for the determination of the ES. The difference is purely a matter of making the judgements explicit, before the randomization. At any event, the clinician's choice must become definite as soon as possible after randomization. Advancing the decision to just before randomization seems feasible. The salient difference is that in the CC design the clinician *apparently* need only think through the options with the patient for the particular strategy that has been randomly selected. But this is only apparently true, since the clinician must go at least as far as to determine that there is an acceptable option in either strategy before randomization.

We believe that the careful, joint review of options by patient and clinician forms a solid basis for the informed consent process. Moreover, the process of picking research treatments parallels decision-making in clinical practice. For these reasons, we think that the ES design has a clear advantage over either CC or CR in terms of ethical concerns. For example, the CR design puts a great deal of pressure on the need to be in full equipoise. A patient who wants very much to participate in the research might be tempted to "game" the consent process, by hoping for a desired outcome of randomization. In ES, the patient's preferences are explicitly accommodated, excluding only patients who have a specific option in mind. The CC design may encourage vagueness in the discussion of treatment options, since the clinician does not have to decide on the option until after randomization to broad strategy. In contrast, the ES design is built on specific discussion of options, elicitation of preferences, and identification of the patient's individual e-stratum.

We can examine what would happen to patients in each ES, during a CC trial, and compare the ES and CC designs, in terms of the generalizability, lack of bias, and precision of statistical estimates of treatment effects. Patients in ES 2,3,5, and 6 cannot be randomized into Clinician's Choice, between "add" and "switch," because they are only eligible for one or the other strategy. Thus, the CC design leaves these patients out of the experiment, despite the presumption that they account for over half the patients whose treatment is in question. Such exclusions decrease generalizability and increase the number of patients that need to be screened to achieve a sample size that maintains adequate precision of treatment comparisons. If the patient is in ES 1,4, or 7, there is no reason not to fully randomize the treatments within the respective ES, since the clinician and patient are in equipoise. That randomization would convert the CC design into an ES design. In this sense, the CC design is merely an inferentially weaker (because of possible treatment selection bias) and less efficient version of the ES design.

It is of course possible that the actual ES distribution will deviate dramatically from that expected (see Table 2). If this happens, the ES design must adapt to the true distribution. For example, consider patients who have no interest in CT and, therefore, would be in ES 4 (any medication), the most populous stratum we expect; however, suppose many of these patients have tried one of the study medications in the past, with poor results, and suppose that we have been mistaken in our assumption that they will agree to a possible retrial of that medication. Such patients do not fit into ES 4 because they refuse one of the options. For example, consider the patient who rejects switch to VEN. Then the ES design as discussed above would place the patient into the nearest neighbor ES, namely, ES 6 {+BUP,+BUS}. Then the comparison of the two medication augment options gains in precision by virtue of the addition of that patient to ES 6, but the comparison of the SER and BUP switch options loses precision because of the loss of the patient from ES 4, even though in principle the patient is randomizable to either SER or BUP.

Alternatively, such patients could be accommodated in one of five possible strata, (SER, BUP, +BUS, +BUP), (BUP, VEN, +BUS, +BUP), etc., each leaving out the single rejected drug. The analysis described above becomes more complex, but can still be guided by the principle that patients contribute to all pairwise contrasts of specific options to which they were "exposed" by the randomization in their ES. This is not a serious problem for the ES study; but the proliferation of strata raises practical issues. It seems to make sense to start with ES that reflect the simplifying assumptions described above, and then reevaluate the decision as the true ES distribution becomes clear. In a sense, the need to change to a more detailed stratification would reflect the failure of the global strategy as an organizing principle.

Clinician-level Causal Effects

The only difference between the equipoise stratum and any other fixed, prerandomization characteristic of patients is that the ES relies on the clinician's and patient's judgments of acceptability of treatments. Thus, it is a property of the clinician/patient dyad, and not strictly speaking, a patient factor. If all clinicians would agree on the e-stratum of any particular patient, then an e-stratum is a fixed effect, just like age or gender. The difficulty arises when clinicians disagree on the distribution of the same patients into e-strata. There are two distinct possibilities. Part of clinicians' disagreement on ES may arise from unreliable or idiosyncratic estimation of patient characteristics. This disagreement has little effect on the design or interpretation of the ES study. A more interesting disagreement is due to different ideas about what is appropriate, including varying degrees of comfort with specific medications. This is not a trivial issue: practice variation in the choice of specific option within a strategy, that is not explainable by specific patient variation, is commonplace in medicine.

If the investigators were able to codify the rules for establishing e-strata (e.g., by history of nonresponse, expected side effects, or elicitation of patient preferences), then e-strata may be regarded as a fixed, crossed patient effect. Such effects can be incorporated into design and analysis much as any other important prognostic factor. Moreover, tests of interaction would be readily interpretable. If such codification is not considered feasible, then the e-stratum can be considered a nested (within clinician) effect. Either way, it would appear that analyses that take account of the clinician/patient effect would be sensible, and these could be based on the randomization if "clinician" were used as a blocking factor.

Programmatic Considerations

From the point of view of a program of research, the ES design offers some advantages. To keep the example simple, consider the task of defining the preference order among three treatments, such as BUP, +BUP, and +CT. One way to proceed is to screen patients and exclude all but the universal donors, and employ a CR design. From the programmatic point of view, the CR design leaves unanswered the question "What would the right treatment choices be in the other 3 e-strata"? If these strata are substantial in the population, that would be a good question to answer. On the other hand, one can think of answering the questions in 3 different studies, each comparing 2 treatments. Then, for example, the "BUP, +BUP" study contrasts treatments "switch to or augment with BUP" in a mixed sample consisting of equipoise strata (BUP, +BUP) and (BUP, +BUP, +CT), the "+BUP, +CT" study contrasts "augment with BUP or CT" in a mixture of (+BUP, +CT) and (BUP, +BUP, +CT) patients, etc. From the programmatic point of view, it is inefficient to run three disconnected studies, if only because of duplication of screening efforts as well as other infrastructure costs. Furthermore, it is unlikely that three independent studies would capitalize on the opportunity to link the treatment comparisons through the universal donor e-stratum. If this is a numerous stratum, then finding the overall winner in that stratum (if there is one) is important. But the three pairwise experiments do not necessarily define a best treatment in that stratum.

By contrast, the ES design offers maximal efficiency, giving opportunities to participate to all patients for whom there is a researchable choice of treatments (an e-stratum). This, of course, only makes sense when the resources match the scope of the problem. If the resource constraints force the investigator to cut the problem into inefficient pieces, the program director may well lament, "I did not have the resources to be thrifty!" In addition, the ES design is most advantageous when there are several populous e-strata; otherwise, there is no better alternative than to randomize among the treatments in the single large e-stratum.

Table 4. Comparison of Three Designs

Conceptual Features	CR	CC	ES
Unbiased comparison of individual treatment options?	√		√
Test of stratum-by-option interaction?			√[a]
Robustness to loss of options in the future?	√		√
Estimable preference distribution?			√
Generalize to broad segment of patient population?		√[b]	√
Readily interpretable results?	√[c]		√
Practical Features			
Recruitment open to most patients?		√[c]	√
Efficient recruitment process?		√	√
Simplicity of consent process?		√[d]	
Adaptation to patient and clinician preferences?		√	√
Robustness to prior assumptions of preferences?		√[b]	√
Simple design and analysis?	√	√	

[a] possibly low power

[b] unless many patients would exclude an entire strategy (e.g., partial responders unwilling to switch)

[c] but restricted to "universal donor" sub-population of unknown proportion of the patient population

[d] but perhaps less specifically informative

Conclusions

If investigators eventually want to know which of a large number of treatment options is truly the best, and which treatments to use when not all are appropriate, then the most efficient program of research is one that compares the treatments all at once. In the preceding, we showed that any other program is less efficient, other things equal. This line of thinking becomes even more compelling when one considers that each option may not be equally appropriate to every patient. For example, options that involve addition of (or switch to) a particular psychosocial treatment may be unacceptable to some patients, or may be a poor fit (e.g., family therapy for patients without cooperative families).

We described the ES design, a method for stratifying and randomizing patients, that allows the greatest number of patients to contribute data to as many comparisons as possible. Taken together with some care in the data analysis, the ES method offers the investigator the ability to optimize the available recruitment resources, mobilizing them to test the entire ensemble of important hypotheses.

The analysis of ES designs is very similar to the analysis of multi-site data by "pooling." It also offers a side benefit of increased power and efficiency, due to the fact that any between-stratum variation in outcome is controlled. But the most important feature of the pooled analysis is that it is based in the actual randomization, and makes no unnecessary assumptions about the stratum effects and stratum-by-treatment interactions. In fact, a true randomization analysis is very attractive.

Table 4 summarizes the features of the three designs discussed above. The CR design has conceptual strengths, owing to its solid basis in the randomization of all patients among all options. The CC design suffers conceptually from its lack of full specification of the treatments, but does well in practical terms. The practical problems that afflict the CR design increase with the number of different options. In contrast, the ES design appears to offer the advantages of each of the other designs, while avoiding their specific weaknesses. The STAR*D trial will provide useful experience with the ES design, no doubt revealing some unanticipated problems of implementation. A substantial side-benefit will be the elicitation and tabulation of patient and clinician preferences for options in treatment-resistant depression.

Supported in part by a Contract and Grants from the National Institute of Mental Health (N01-MH-90003 and MH-30915), and by Mental Health Connections, a Texas State Legislature and Dallas County Hospital District funded partnership between Dallas County Mental Health and Mental Retardation and the Department of Psychiatry of the University of Texas Southwestern Medical Center.

References

AFFIRM Investigators (1997): Atrial fibrillation follow-up investigation of rhythm management - the AFFIRM study design. The plannning and steering committees of the AFFIRM study for the NHLBI AFFIRM investigators. *Am J Cardiol* 79:1198–1202.

Kadane JB (1986): Progress toward a more ethical method for clinical trials. *J Med Philos* 11:385–404.

Kadane JB, Sedransk N (1980): Toward a more ethical clinical trial. In: Bernardo JM et al, editors. *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*. Valencia: Valencia Univ. Press, pp 329–338.

Korn EL, Baumrind S (1991): Randomized clinical trials with clinician-preferred treatment. *Lancet* 337:149–152.

Korn EL, Baumrind S. (1998): Clinician Preference and the estimation of causal treatment differences. *Statistical Science* 13(3):209–235.

Rush AJ (2001): Sequenced Treatment Alternatives to Relieve Depression (STAR*D). In: Syllabus and Proceedings Summary, American Psychiatric Association 154th Annual Meeting, New Orleans, LA, May 5–10, 2001, p 182.

Tukey JW (1993): Tightening the clinical trial. *Control Clin Trials* 14(4):266–85.