Flexible Treatment Strategies in Chronic Disease: Clinical and Research Implications

Philip W. Lavori, Ree Dawson, and A. John Rush

Multiple treatments are available for nearly all the mood disorders. This range of treatment options adds a new dimension of choice to clinical decision making. In addition to prescribing the best initial treatment, clinicians should have an algorithm for deciding if and when to make subsequent changes in treatment to take advantage of second-line treatment options when necessary. This article aims to 1) show that a wide variety of clinical decisions can be framed as choices among adaptive (within-patient) threshold-based strategies or algorithms, illustrating the generality of the concept; 2) illustrate two ways to design randomized clinical trials to compare treatment strategies with each other to decide which strategy is best; and 3) discuss some of the advantages offered by these designs, in terms of both patient acceptability and adherence to experimental protocols. Biol Psychiatry 2000;48: 605–614 © *2000 Society of Biological Psychiatry*

Key Words: Clinical trials design, treatment algorithms

Introduction

Psychopharmacologic management of the mood disorders has reached a level of development characterized by the availability of several first- and second-line treatments for nearly all of these illnesses. This adds a new dimension to clinical decision making. In addition to prescribing the best initial treatment, clinicians should have an algorithm for deciding if and when to make subsequent changes in treatment to take advantage of second-line treatment options when necessary. This article aims to 1) show that a wide variety of clinical decisions can be framed as choices among adaptive (within-patient) threshold-based strategies or algorithms, 2) illustrate two ways to design randomized clinical trials to compare treatment strategies with each other to decide which strategy is best, and 3) discuss some of the advantages offered by these designs in terms of both patient acceptability and adherence to experimental protocols. The methods described in this article apply generally to chronic disease. We illustrate them using examples chosen from the mood disorders, especially bipolar disorder.

Lithium would be widely regarded as an indicated mood stabilizer in patients who present with mania, for example. After prescribing medication, the clinician waits for the manic symptoms to remit or at least to respond. If the manic symptoms begin to return, the lithium dose would be escalated, (up to the limits defined by maximally safe therapeutic blood levels). If the patient continues to experience some manic symptoms, which might presage a full clinical relapse or cause ongoing disability, the clinician likely would consider adding a second mood stabilizer, such as divalproex (DVP). Suppose the choice of first-line treatment (lithium) and second-line treatment (add DVP) are taken as the standard-of-practice. (Note that DVP might well be the first-line treatment. We are simply illustrating with specifics, a general principle.) How should clinicians weigh the various components of the patient's response over time to the first treatment? What threshold recommends the addition of DVP to achieve optimal control over manic symptoms?

Consider a patient in the depressed phase of bipolar disorder. Adding an antidepressant agent to the mood stabilizer might carry the risk of inducing manic relapse. Thus, the initial step may be to wait, possibly raising the mood stabilizer dose. Suppose, however, that the patient's depressive symptoms continue or worsen over time. Clinicians must again weigh the burden of ongoing depression against the risk of mania to define a threshold for adding an antidepressant medication (AD). If some patients would recover from the depressed phase without an AD, they could be spared the exposure to the risk of mania by setting the threshold for adding the AD at a higher level of severity or duration of depressive symptoms. Depending on the balance of benefit and risk, the optimal algorithm for adding an AD might call for waiting until a sufficiently severe, lasting, and disabling depression has been established, such that the risk of a manic relapse is a chance worth taking. If the risk were great, the optimal strategy might be to set a high threshold, either effectively avoiding or rarely using AD. If the risk were more modest or minimal, the threshold should logically be lower.

Such threshold decisions abound in a mature field of

From the Department of Veterans Affairs Cooperative Studies Program, Palo Alto (PWL), and the Department of Health Research and Policy, Stanford University, Stanford (PWL), California; Frontier Science Research Foundation, Boston, Massachusetts (RD); and the Department of Psychiatry, University of Texas Southwestern Medical Center, Dallas (AJR).

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

Received March 22, 2000; revised June 1, 2000; accepted June 2, 2000.

0006-3223/00/$20.00
PII S0006-3223(00)00946-X

chronic disease management such as psychopharmacology in which there are a number of treatment options. Yet, there is often little or no empirical evidence to guide the choice of threshold, especially if one demands evidence from well-controlled, randomized trials. The standard two- or three-group drug trial in major depressive or bipolar disorders is specifically designed to distinguish among fixed treatment options. Our understanding of the value of individual drugs is based largely on such trials. But the standard trial does not answer the critical question, "How should the clinician guide the ongoing treatment of a patient, given current outcomes with that patient?"

Adaptive Strategies in Psychopharmacology

Suppose that the first and second line treatments have been settled. This lets us focus on the new ideas. We will also limit our discussion to the simplest kind of adaptive strategies, or algorithms, that we define by example.

Bipolar Disorder, in Remission

Continuing the first example discussed above, suppose that the reference population consists of newly remitted patients, following a first-break manic episode, and that the goal is to avoid major clinical relapses. We postulate that the initial treatment in all patients is lithium, and the second-line therapy is the addition of DVP. Whether and when to move from lithium alone to lithium plus DVP depends on the course of the patient's response to the lithium. The best strategy is one that minimizes the probability of relapse, among all the possible strategies that could be applied to this patient group.

The first requirement for a usable strategy is a data-based summary of the current and past history of symptoms. Suppose, for example, that we sum up the number of weeks (since the remission started) that the patient has had subsyndromal symptoms (e.g., of hypomania) every day of that week, and subtract the number of weeks that the patient has been free of such symptoms every day of that week. (For discussion purposes, we will ignore mixed weeks). Let us call this score S. We chose this method of scoring because it is easy to describe and because related measures have been shown to predict relapse (Keller et al 1992). Note that an increasing S reflects continuing illness, whereas a decreasing S reflects continuing wellness. The score S can be calculated on a regular basis (e.g., at every weekly visit) to give a sequence of measurements S(1), S(2), and so on. Figure 1A illustrates the scores over time for a hypothetical patient X, who is continuously subsyndromal for 2 weeks and then continuously well, all the while being maintained on lithium alone (the first treatment). Such a patient might be described as a "delayed responder" to lithium, perhaps because the right dose needs to be found.

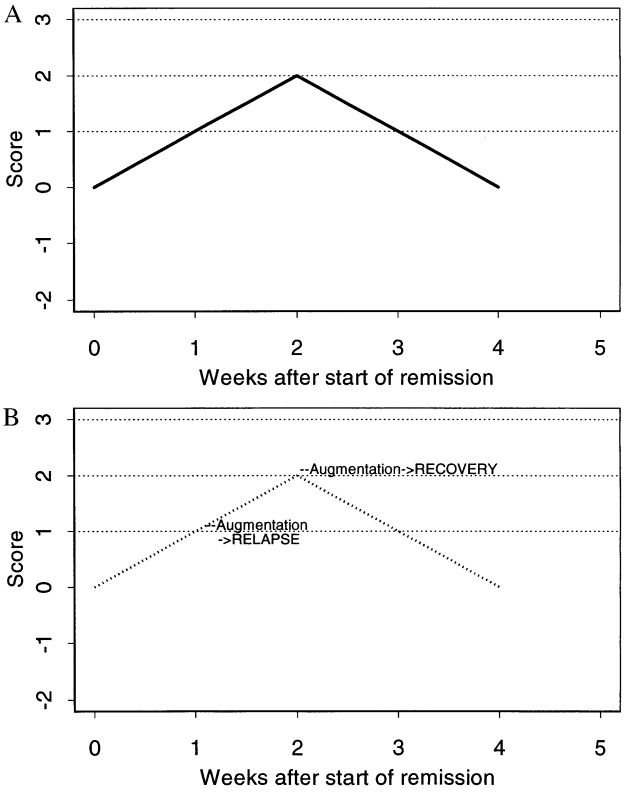

Figure 1. **(A)** Cumulative summary symptom score on continuous lithium therapy for a hypothetical patient "X." **(B)** Outcomes of hypothetical patient X under strategies "augment at score = 1" and "augment at score = 2."

Recall that a particular strategy consists of a choice of a threshold for moving from lithium monotherapy to the combined or augmented treatment (lithium plus DVP). The dotted lines in Figure 1A indicate the possible thresholds; their intersection with the scores S indicate the resulting potential augmentation times for patient X. For the sake of illustration, suppose that patient X moves to augmentation at week 1, and that the results are bad, perhaps because residual symptoms and new side effects conspire to make X give up treatment altogether, and, therefore, relapse during the second week. Suppose that if augmentation is delayed until week 2, the results are much better. For instance, X's symptoms abate, as they would have anyway, even without augmentation (Figure 1A), so he is tolerant of the side effects of augmented therapy. Figure 1B illustrates patient X's course with augmentation at either week, superimposed on the dotted line that indicates patient X's course during initial treatment (as in Figure 1A).

Suppose we adopt the strategy "augment at threshold S = 1." Under this strategy, patient X would receive

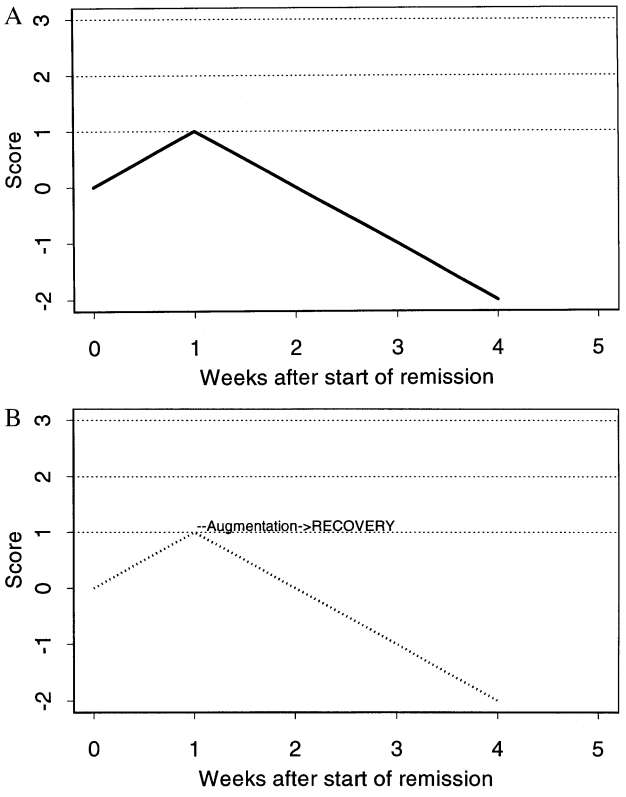

Figure 2. **(A)** Cumulative summary symptom score on continuous lithium therapy for a hypothetical patient "Y." **(B)** Outcomes of hypothetical patient Y under strategies "augment at score = 1" and "augment at score = 2."

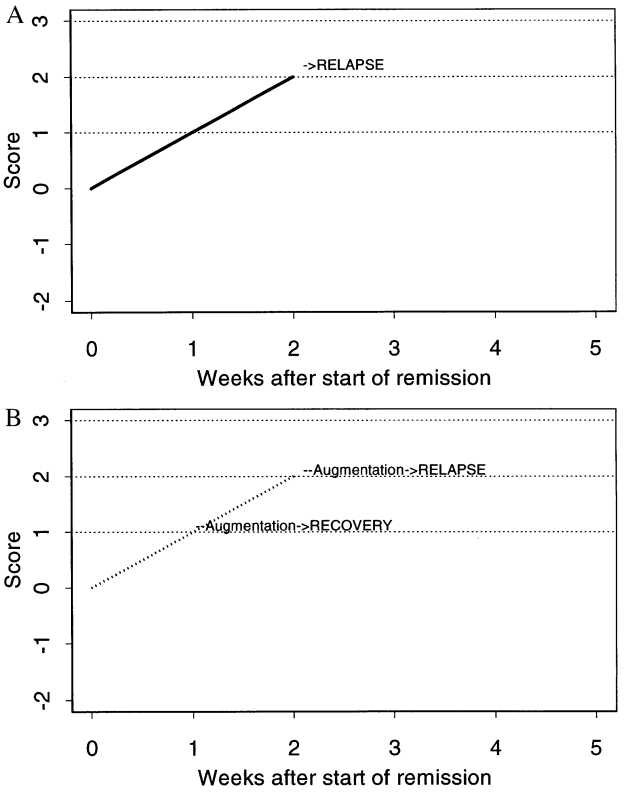

Figure 3. **(A)** Cumulative summary symptom score on continuous lithium therapy for a hypothetical patient "Z." **(B)** Outcomes of hypothetical patient Z under strategies "augment at score = 1" and "augment at score = 2."

augmentation at the end of week 1, when S first reaches 1 (the horizontal dotted line at S = 1 in Figure 1B). The strategy "augment at threshold S = 2" would leave a patient on the original "lithium alone" treatment until the score reached 2; for patient X, this occurs after the second week. Therefore, X is a patient who does best (and indistinguishably, at least in terms of relapse) under strategies defined by thresholds of 2 or more and worst under strategy augment at S = 1. If we knew the contents of Figure 1B in advance, we could decide which strategy would be best for this patient.

Now consider a different patient (Y), whose response to lithium is depicted in Figure 2A. Patient Y has symptoms for 1 week, then remains well, while continuously on lithium alone. Under strategy "threshold = 1," patient Y would receive augmentation after week 1, just the same as did patient X. But because patient Y never reaches S = 2 under lithium treatment, all strategies defined by thresholds of 2 or higher leave patient Y permanently on lithium alone . Now suppose that patient Y does well (no relapse) throughout the augmentation period (Figure 2B). Then, this patient has the same benign course of illness under all strategies, even though strategy threshold = 1 would augment Y's lithium with DVP at week 1, whereas other strategies would never lead to augmentation.

Finally, consider patient Z (Figure 3A), who has a bad course on lithium alone, with subsyndromal symptoms persisting for 2 weeks and then experiences a clinical relapse during the third week. Patient Z would receive augmentation at week 1 on the threshold = 1 strategy, at week 2 on the threshold = 2 strategy, and would relapse before augmentation is provided under the strategy threshold = 3 (or higher). Now suppose that Z is a patient who should receive augmentation as soon as possible (Figure 3B). Under strategy threshold = 1, Z receives augmentation at week 1, thereby avoiding relapse. Under strategy threshold = 2, augmentation begins at week 2, but Z still goes on to relapse, augmentation having arrived too late.

Notice that a single *strategy* (say, threshold = 2) may lead to different patterns of *treatment* (i.e., augmentation times) in different patients. Because the timing of augmentation under a strategy depends on the ups and downs of the course of illness on the initial treatment, these are adaptive strategies. This idea of adaptive strategies also

generalizes to any change in treatments, such as "switching" treatments (i.e., stopping one treatment and starting another).

Thus, what works for X fails for Z, and vice versa. The optimal strategy is defined by the threshold for augmentation that would produce the lowest *overall* relapse rate, if all patients were managed according to the strategy. The best threshold will depend on the mix of patients of these (and other) types. Note that if the second treatment step (in this case, augmentation of lithium with divalproex) is uniformly better, the best threshold is low. Conversely, if the second treatment step has no effect, the best threshold is high. We are interested, in the case where the best is in between, because in the former case, augmentation should be the first-line treatment, whereas in the latter case one should never proceed to the second-line treatment. Note also that once the patient relapses (whether on the initial or subsequent treatment), this cycle of decision is over. A new plan of treatment must be provided for the new episode. The avoidance of relapse defines success for the augmentation strategy.

Bipolar Disorder, Depressed Phase

The second example introduced above concerns patients with bipolar disorder who develop a major depressive episode (MDE) while on the first line mood stabilizer. Suppose (as discussed above) that the main goal is to achieve remission of the depression while avoiding a manic relapse and that the first-line treatment is to "continue on the mood stabilizer (MS)." The possible second treatment is to add an antidepressant (AD) to the MS. We might score the patient's symptom severity by cumulating the Hamilton Depression Score (HAM-D, measured weekly) from the onset of the depressive episode. We can make the score better suited for defining strategies by subtracting a "target level" (say, HAM-D = 12) from the HAM-D before cumulating it, however. Figure 4 shows this calculation for a hypothetical patient whose HAM-D rises from 12 to 15 and then declines to 9, over the 10 week period of interest. This score is the "detrended area under the curve" method. It will be sensitive either to large departures or to long-lasting departures from the target level of 12.

We cue the next step (adding AD to MS) by thresholds in the detrended area under the curve. A high threshold requires that a patient demonstrate a worse course of depression (on MS alone) before adding the AD, whereas a low threshold calls for adding the AD sooner. Note that for a patient with the course on MS alone depicted in Figure 4, a threshold of 5 would call for an early addition of AD, whereas a threshold of 8 would delay the AD for a few weeks. A threshold of 10 would leave such a patient

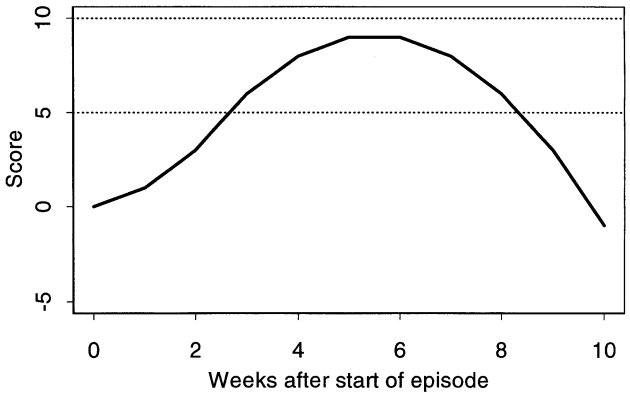

Figure 4. Possible area under the curve score, based on detrended Hamilton Depression score, reflecting initial worsening and then some improvement.

on the MS alone, for the entire duration of the treatment period (here, 10 weeks).

To even hypothetically evaluate and compare strategies, we must weigh the depressive symptoms and manic relapse against each other. For simplicity, suppose a strategy is a "success" if it results in a final (10-week) score S below 0 and no manic relapse during this period. Otherwise, we call it a "failure" (i.e., a manic relapse occurs or the depression continues). Such a score weights the depression and mania equally. Then the criterion for judging a strategy is the probability of success for that strategy, in the reference population. Note that the score used to determine the change in treatment need not be the same as the criterion for evaluating different strategies, although in most instances they will likely be related to each other.

Nonbipolar Depressive Disorder

The idea of basing the change in treatment on the "area under the curve" may help to define treatment for nonbipolar depressions. Suppose clinicians adopt fluoxetine as the first-line treatment, and the next step is a switch to bupropion. At each week after beginning the first-step treatment with fluoxetine, let S cumulate the adjusted HAM-D scores up through that week (as discussed above, $-$S is the detrended area under the curve). For each threshold choice, clinicians have a possible strategy: "switch to bupropion if and when S reaches the threshold." In either case, they continue until some fixed time has elapsed. One natural overall summary of outcome (for comparing strategies) is the score S at the last time point.

Instead of the HAM-D, one could also use any function of the history of symptoms and side effects. The function could include individual patient-specific preferences if they can be elicited or could be based on population preferences. The cumulative score can be defined to ignore

all but the most recent observations ("windowing"), or it might give more weight to recent observations ("tapering"). Each such choice reflects a belief about the proper role of current and past observations in the current decision to continue the first-line treatment or switch to the alternative.

Placebo Trials and Watchful Waiting

Threshold-based strategies provide a potentially useful way to capture the clinical concept of an initial trial of a placebo or other "nonspecific" treatment (e.g., "watchful waiting"). We use the word *trial* here in its original, clinical sense, not to refer to an "experiment." If the placebo is the initial treatment and the second-line treatment is an active agent, clinicians can use the cumulative scores discussed above to define when to start a specific active treatment. This corresponds to conservative management of mildly symptomatic patients, motivated by a desire to avoid unnecessary exposure to active treatments, for patients who will improve anyway. The threshold choice sets the level of ongoing symptoms to trigger the switch to active treatment. Such a threshold-based treatment plan is a specific, well-defined implementation of watchful waiting. Any of the scores discussed above could be used, with an appropriate set of thresholds from which to choose. The criterion for success might be taken as the cumulative score at the final time, perhaps with a penalty for the occurrence of side effects or other poor outcomes not captured by the score. Then the strategies range from a low threshold (always treat right away) to a very high threshold (never treat), with alternatives in between that call for treatment if patients continue to be depressed.

Experimental Comparisons of Strategies

As described above, the right threshold depends on an unknown mix of patient types in the population to be treated. Because we cannot know in advance what each individual patient's best threshold might be, it is necessary to define the optimal threshold in terms of average hypothetical responses to the threshold-based strategies. Furthermore, to make inferences about these averages requires experiment, with randomization providing the strongest basis for such inference. We describe two kinds of randomized designs that can be used for inference about adaptive strategies: baseline randomization (BR) and biased-coin adaptive within subjects (BCAWS).

Baseline Randomization

One way to compare alternative thresholds is to randomize each subject at the outset to one of several thresholds, to define if and when to implement the next step of treatment. (We refer to "subjects" in experiments, to distinguish them from patients in the world of clinical practice. This distinction will become important in a subsequent section.) As the experiment progresses, the subject's scores are updated after each measurement visit. If a subject's score exceeds the threshold to which that subject was assigned initially, the investigator takes the next step (i.e., moves to the second-line treatment from the first-line treatment).

Suppose we consider the augmentation strategies discussed in the first example, maintenance treatment for patients with bipolar disorder who have remitted after one manic episode. All study subjects begin on lithium alone. The "subsyndromal symptom score" S is tracked. Then if a subject is randomized to augment at threshold 1, as soon as that subject's score S reaches 1, the investigator adds DVP. Subjects randomized to different thresholds (2, 3, etc.) are managed accordingly. These changes in treatment should be managed to achieve single or double masking of current treatment, with dummy pills, unmasked case managers, and so forth. The requirement for masking applies to adaptive designs (of all kinds), as well as to fixed trials.

In addition to the data collection required to calculate the subsyndromal symptom score, investigators keep track of the subject's relapse status, as the primary outcome. If the subject relapses, whether on lithium or on the combined treatment (lithium + DVP), the experiment is over for that individual. As usual, post-protocol treatments for relapse are provided outside the study proper. So for each patient/subject, investigators record the outcome (relapse or not) and the (randomized) threshold. The usual methods of analysis suffice. There is no difference caused by the dynamic, adaptive nature of the strategies. Referring back to the bipolar maintenance example, the randomization to thresholds would on average balance the proportions of subjects of type X, Y, Z, and so forth across the threshold-assignment groups. Therefore, the baseline randomization design provides unbiased contrasts of the overall relapse rates produced by each strategy if applied to the whole reference population of the trial.

The VANQWISH study (Boden et al 1998; Ferry et al 1998) is an example of such a trial in a nonpsychiatric illness. Subjects with a new myocardial infarction, without Q-waves in the electrocardiogram, who were initially stable, were randomized to one of two strategies. The "invasive" group received immediate diagnostic catheterization and angiography, followed by revascularization if indicated. The "conservative" group received noninvasive testing and monitoring of ischemic changes; only if and when such ischemic changes manifested did subjects receive cardiac catheterization and angiography, followed by revascularization if indicated. Here, the score S is the current ischemic status of the subject (S = 1 if yes, 0 if no), and the thresholds were either 0 or 1. The outcome for

this trial was either death or a new MI at 1 year of follow-up.

The VANQWISH study was designed to compare two different thresholds of ischemia used to trigger the invasive diagnostic procedures and subsequent procedures. In principle it would have been possible to have more than two threshold levels. Investigators could have had a group whose catheterization was delayed until the ischemia was even more strongly manifested. When VANQWISH was designed, this was not seen as a possible choice, given the prevalent views of the risks of delaying catheterization, so the study proceeded with two thresholds. As it happened, subjects randomized to the conservative (noninvasive) strategy (corresponding to the higher threshold of ischemia) had a significantly lower mortality and no greater risk of new MI than those randomized to the invasive strategy (corresponding to the lower threshold). Furthermore, the conservative strategy successfully postponed catheterization in nearly two thirds of patients for at least 1 year (Boden et al 1998). The conservative strategy was more "cost effective" than the invasive strategy, perhaps setting the stage for a subsequent study exploring even higher thresholds for catheterization.

As in VANQWISH, direct baseline randomization to strategies is attractive if the number of distinct strategies to be compared is small and if they are different enough such that a difference is expected to emerge. The relapse-avoidance example discussed above provides a chance to employ these ideas. The threshold for adding DVP to lithium could be set at "low" (add as soon as subsyndromal symptoms appear), "medium" (add if subsyndromal symptoms persist for more than one week), or "high" (add if these symptoms persist for at least one month, despite adjustments to the lithium level). Such an experiment does not involve new statistical methodology, and it may be a useful addition to current designs.

Biased-Coin Adaptive Within-Subject Randomization

We recently proposed a new class of randomized designs to compare adaptive strategies that we call the biased-coin adaptive within-subject (BCAWS) designs (Lavori and Dawson 2000). Basically, the BCAWS design introduces randomization at each time that the subject is eligible to change from the first-line to the second-line treatment. Compare this with the baseline randomization, which randomly allocates subjects to thresholds at the start but then changes treatments deterministically depending on the subjects' responses (if a subject reaches the threshold while on the initial treatment, then treatment is changed). In the BCAWS design, the subject may move on from the first-line to the second-line treatment at any time, with a probability that depends on the value of a score at that time (such as any one of the scores S described above). The variation in the probability of changing treatments is what makes this a "biased coin" design. The design is called "within-subject" and "adaptive" because the bias toward or against changing at any time depends on each subject's history of responses up to that time. If the "biased coin" is chosen such that the change probability rises as the score S indicates worse outcomes, then the resulting design tends to change subjects' treatments when they are doing poorly. A baseline randomization to a set of strategies defined by a particular set of thresholds in a particular score S, can support inferences *only* among strategies in that set. In contrast, it is possible to create a BCAWS design that also provides estimates of the results of some strategies that are outside the original set.

For illustration, suppose that the original set of strategies was defined as in the bipolar maintenance example. The score S is the balance of subsyndromally ill and well weeks, and the strategies are of the form "add DVP to lithium as soon as the score reaches the threshold" for various threshold choices. Suppose that the BR experiment came to a conclusion that the best strategy was to augment at threshold $S = 3$. Now suppose that a question arises as to whether it might be better to discount early symptoms to "give lithium a chance to work." This might be described by a set of "delayed" strategies: "continue lithium alone for D months, regardless of symptoms; after that, add DVP when S reaches the threshold" (again, for various thresholds). These delayed strategies are motivated by the idea that at the start of the remission, one might want to adapt to symptoms by adjusting the lithium level instead of adding DVP, whereas after some time D, one uses the thresholds to decide whether to augment. It turns out that inference to the delayed strategies of this form will be covered by a version of the BCAWS design that also yields inferences about the "undelayed" strategies. Under some circumstances, one can make inference about both the choice of delay D and the threshold.

The Basis for Inferences about Outcomes from BCAWS

To be clear about the possible inferences from a BCAWS design, it is necessary to emphasize the distinction between the "real world of actual clinical strategies" and the "world of the BCAWS experiment." The world of clinical strategies is defined by a particular patient population, a clinical context of first-line and second-line treatments, a scoring method (S) that clinicians use to rate the history of the patient's response to the first-line treatment, and a set of thresholds that determine adaptive strategies (change from initial to second treatment as soon as S reaches the

threshold). The examples given at the beginning of this paper take on reality in the world of the clinical strategies.

Once clinicians adopt a score S and a particular threshold, treatment decisions are determined by the patient's responses, and thus, so are the outcomes of the strategies. There is no randomness in any patient's experience of the strategy, although there is *a priori* ignorance about both the treatment sequence (change times) and the outcomes observed in a particular patient under a particular strategy. This ignorance is the reason we resort to experiments, such as BR or BCAWS, to infer something about the potential outcomes of the different strategies, only one of which will ever be observed in any patient.

In a BCAWS experiment, the subject's exposure to treatment is partly determined by the randomness of the (biased) coin flip. This gives rise to observations of the results of certain strategies in that patient, whereas the results of other strategies are not observed. If subjects do not change from first- to second-line treatment when scores first reach 3, then we do not get to observe potential responses under strategy "change treatment at threshold = 3"; if subsequent changes occur when they reach a score of 4, they give complete information about their responses to strategy change at threshold = 4. We cannot observe which is the better threshold for an individual subject. Instead, we try to pull together information on strategies such that we can compare the average responses of the entire reference patient population. The way we do this for BCAWS exploits a remarkable correspondence between the concept of treatment effect and the mechanisms of missing data; the modern version of this idea is attributable to Rubin (1974).

As noted above, the outcome-dependent treatment changes in the BCAWS design create a pattern of missing data on strategy outcomes. Specifically, when subjects are changed from first- to second-line treatment, they no longer provide observations on the outcomes that would have been observed from that point onward under all strategies that would have left them on the first-line treatment at that point (strategies corresponding to higher thresholds than the value of the score S at the point of change). Such subjects provide full outcome data on the strategy corresponding to the threshold equal to the score S at the time that their treatment was changed. Conversely, when subjects do not change from first- to second-line treatment, such subjects no longer provide observations on the outcomes of the strategy corresponding to the threshold equal to the value of the score S at the time the subjects do not change. Any subject who continues on first-line treatment to the end of the trial provides complete observations on outcomes for all strategies corresponding to thresholds greater than the highest score experienced by that subject.

This mechanism of missing data is not "missing completely at random" (Little and Rubin 1987) because "missingness" depends on outcomes; however, it is "missing at random" (MAR) because missingness only depends on the values of the *observed* outcomes. We know the probabilities of changing treatments at all times in all subjects (by design). Imputation for MAR data can be used to infer the distributions of outcomes and to compare them across thresholds, as described in Lavori and Dawson (2000) and Dawson and Lavori (unpublished data).

The imputation for a missing response, which would have been observed under a change in treatment at a particular time for a subject who does not change at a particular threshold, is as follows: a "donor subject" is drawn at random from the subjects who have the same history of scores up to that time, but who did change treatments. The entire future from that time on is imputed all at once. The imputation for missing responses that would have been observed under higher thresholds for subjects who do change at a particular threshold at a particular time (t) proceeds by single time steps. A donor subject is chosen, matched on current history, from those subjects who did not change at time t. That donor provides data for the recipient subject's unobserved response at time $t + 1$ for all thresholds higher than the one at which the recipient subject actually changed. The process of imputation is repeated inductively up to the designed end of the experiment. Statistical inference for the observed and imputed data proceeds as described in Lavori and Dawson (2000) and Dawson and Lavori (unpublished data).

Science, Ethics, and Informed Consent

At least three points of view must be considered in evaluating the two trial designs (BR and BCAWS) in comparison with a standard RCT with fixed treatments. These viewpoints belong to the research scientist, the patient, and the Institutional Review Board (the ethics panel). We discuss each perspective in the context of the first example, described above, of the treatment of patients with bipolar disorder, newly remitted from a manic episode. The first-line treatment is lithium, and the second-line treatment is augmentation with DVP.

What potential advantages and disadvantages do such designs pose for the research scientist? The standard RCT of the fixed treatments involves randomizing subjects at the outset to either lithium alone or combination lithium and DVP and maintaining these controlled treatments up to either full clinical relapse or study exit. Any departure from these fixed treatments (in response to emergent symptoms of hypomania, for example) threatens the power or validity of the study. The decision to compare fixed treatments in this way amounts to a strong hypothesis that the intercurrent symptoms should not play a role in clinical

decision making. But if such symptoms emerge in substantial numbers of subjects, the consumer of the research may view study results as a comparison of options that are not realistic. If clinicians prefer to use an adaptive strategy, starting with lithium and augmenting with DVP at some symptom threshold, they will be frustrated by designs that compare fixed treatments, instead of informing the choice of threshold. Such study results will have little impact on practice and may even be regarded as irrelevant. Thus, the researchers should consider the choice of fixed or adaptive design in the context of current practice.

Of course, research scientists (and others involved in a BR or BCAWS study) should be in equipoise about the choice of threshold, so as to have an ethical experiment. Here, *equipoise* means that there should be considerable uncertainty about the right level of symptoms to trigger augmentation. Importantly, such equipoise persists even as subjects experience symptoms because it expresses the lack of certainty as to the practical significance of the symptoms and how they should influence treatment decisions.

In any experiment, as well as any clinical algorithm, there must be allowance for the possibility of bad outcomes that would trigger a "rescue treatment" even if the score has not reached the threshold. These safety boundaries need to be considered in advance of experiment, and they will have an effect on inference and interpretation. Part of the strength of BR and BCAWS comes from the explicit way that such boundaries can be incorporated into the experiment, making them, in effect, part of the strategies. In BR, one can make sure that the score S is constructed in such a way that any patient who reaches a safety boundary would also reach the maximal score. In BCAWS, the bias of the coin can be set to switch any patient who reaches the boundary.

The scientific choice between BR and BCAWS is both more subtle and less well understood at this point. The tradeoff is between the increased simplicity and power of the BR (comparing among one set of strategies) and the more complex BCAWS design, which trades away some power for the ability to explore more general sets of strategies. A useful analogy is to the choice between a single-factor design, which holds other factors fixed, and a factorial design, which randomizes two or more factors simultaneously. In the BR design, once the scoring method is fixed, only the threshold parameters can be compared with each other with randomization-based inference. The BCAWS design offers the chance to see if the optimal threshold depends sensitively on the choice of scoring method. This may help to allay concerns that the truly optimal combination of scoring method and threshold has been left out of the experiment.

The increased flexibility of BCAWS offers the following attractive possibility. Select the principal set of strategies by choosing a score S and a set of thresholds. Construct a BCAWS design that is similar enough to the BR design to have adequate power to discriminate among principal strategies (Dawson and Lavori, unpublished data). After the study is over, use the methods in Lavori and Dawson (2000) to extend inference to related strategies (delays, different scoring methods, etc.) as far as data permit. These exploratory (but randomization-based) analyses set the stage for further trials and provide an idea of the "robustness" of the principal strategies. Taking full advantage of the possibilities of BCAWS requires further exploration and understanding of the related-but-different strategies estimable after a BCAWS experiment. The new feature of the BCAWS design is that the inferences it supports depend on the actual outcomes of patients. Much more work needs to be done to understand this tradeoff.

What are the advantages and disadvantages of the three kinds of experiments to the subject? In a fixed treatment design, the subjects and the clinicians know at the outset that the study treatments will not adapt to changes in symptoms. This knowledge may increase the tendency to withdraw from study treatment as soon as symptoms worsen because no change in treatment is possible within the fixed study treatment protocol. In turn, such withdrawals lead to a dilemma for scientists, who must decide between two unattractive alternatives. If the scientists truncate follow-up once the subject withdraws from the study treatment, the resulting loss of data disables the intent-to-treat analysis. If researchers follow the strict intent-to-treat principle and collect the subjects' postadherence outcomes, the power of the study may be compromised. This dilemma puts the scientific validity of the study in direct conflict with the interests of the subject.

In contrast, patients and clinicians taking part in a BR or BCAWS study will have in mind the possibility that they are still on the first-line treatment (lithium alone) and may augment with DVP in the future. A rational discussion about withdrawal may go as follows:

> You have developed some symptoms of hypomania. If you are currently on lithium alone, you may at any time have your treatment augmented with DVP, as we discussed. Because we still do not know the level of symptoms that should trigger augmentation, it makes sense to let the study continue to drive the treatment. As long as your symptoms persist, you will have a chance to have your treatment augmented. If you have already had augmentation with DVP, it may have been very recent and may not have had a chance to work. In any case, augmentation with DVP is the treatment that we would recommend for you if we knew you would not be protected from relapse by lithium alone. Therefore, it makes sense to stay the course.

Subjects know that by design, they will start on the most favored initial treatment. Subjects also know that if symp-

tom scores deteriorate over time, clinicians are more likely to change the treatment. There is a chance that changing treatment at the current score is not the best thing to do. These points apply both to the BR and BCAWS. Indeed, from the subject perspective, there may be little difference between the two adaptive designs. It may be important to learn more about how patients see the randomization process; some may view an ongoing lottery differently than a single coin-flip that determines the future treatments.

How would an institutional review board (IRB) view the design? Two issues should occupy the IRB in deciding whether a BR or BCAWS design is ethical. First, the IRB should be convinced that the balance of expert opinion favors equipoise about the thresholds for augmentation (as discussed above). Indeed, if the IRB believes that the fixed treatments represent unrealistic options, then the fixed RCT is not an appropriate ethical choice because it does not lead to useful information. The second issue concerns informed consent. Because the adaptive designs come closer to standard clinical care than the traditional fixed treatment trial, special care must be taken to avoid misunderstanding the distinction between a clinical trial and standard clinical care. It seems easier to explain randomization to fixed treatments at the outset (the fixed RCT), and perhaps a little less easy to explain randomization to thresholds of symptoms (BR). Explaining an ongoing lottery (BCAWS) may be appreciably more difficult.

Discussion

We have described the threshold strategies, which formalize a large part of clinical decision making. We have described two methods for research and inference about such strategies, one classical and the other based on modern ideas and methods of missing data. The former is appropriate when the number of options is small and well understood. The latter may help in the more common situation in which there is considerable unknown data. Finally, we have begun to describe the possible responses of patients, clinical researchers, and IRB members. Future work will provide more details on optimal design, power, sample size requirements, and other practical details.

The usual discrete times of follow-up in clinical care (weekly or monthly follow-up, for example) make it sensible to adopt a set of discrete occasions for reevaluation of S and decision to switch based on the chosen threshold. Because the symptom measures are coarse, we do not lose much generality if we boil down the basic measurements at each occasion to $+1$ (for "illness") or -1 (for "wellness") and limit ourselves to thresholds such as 0, 1, 2, and so forth. This already takes us far in adding flexibility to fixed strategies. In the same spirit of generality, we pass lightly over the difficult issues of measurement and of weighing different outcomes against each other (such as symptoms and side effects). We do not think these are more or less important in this context than in the fixed treatments case.

In contrast, we are particular about the idea that the outcomes used to compare the strategies in the experiment (or in the abstract) should cover the whole time span of the experiment. In particular, patients who do not respond to the first-line treatment but are moved to the second-line treatment early because of a low threshold and then go on to have a good outcome under the second-line treatment may be considered a "success" overall and tend to influence the optimal threshold accordingly.

The examples convey the considerable generality of the idea of threshold strategies, and the sense that they are closer to what clinicians do than the fixed treatments that are typically compared in most clinical trials. It may seem that the restriction to thresholds in a single score do not adequately sum up all the considerations a clinician brings to the treatment decision. Even with perfect knowledge of treatment effects, however, one often has to weigh benefits in one domain against costs in another, so as to rank order the strategies by the desirability of their outcomes. Similarly, it is possible to be ingenious in defining scores and to include several domains of symptoms and side effects. Ultimately, the value of a strategy depends on its behavior on average in patients to whom it is applied.

Up to now, the design of clinical trials has been dominated by the needs of the pharmaceutical industry to satisfy regulatory requirements for proof of efficacy. This agenda leads to trials that efficiently compare highly separable treatments (such as a new selective serotonin reuptake inhibitor vs. a placebo). But such studies shed less light on the correct use of the currently broad choice among agents with "better than placebo" performance. For this illumination, we need designs such as BR and BCAWS that compare adaptive strategies with randomization-based inferential strength.

Supported in part by the Department of Veterans Affairs Cooperative Studies Program; National Institute of Mental Health Grant No. R01-MH51481 to Stanford University; National Institute of Mental Health Contract No. N01-MH-90003; and Mental Health Connections, a partnership between Dallas County Mental Health and Mental Retardation and the Department of Psychiatry of the University of Texas Southwestern Medical Center that is funded by the Texas State Legislature and Dallas County Hospital District.

Aspects of this work were presented at the conference "Bipolar Disorder: From Preclinical to Clinical, Facing the New Millennium," January 19–21, 2000, Scottsdale, Arizona. The conference was sponsored by the Society of Biological Psychiatry through an unrestricted educational grant provided by Eli Lilly and Company.

References

Boden WE, O'Rourke RA, Crawford MH, Blaustein AS, Deedwania PC, Zoble RG, et al (1998): Outcomes in patients with acute non-q-wave myocardial infarction randomly assigned to an invasive as compared with a conservative strategy. *N Engl J Med* 338:1785–1792.

Dawson R, Lavori PW (in submission): Comparison of designs for adaptive treatment strategies: Classical vs. adaptive randomization.

Ferry D, O'Rourke R, Blaustein A, Crawford M, Deedwania P, Carson P, et al (1998): Design and baseline characteristics of the veterans affairs non-q-wave infarction strategies in-hospital (VANQWISH) trial. *J Am Coll Cardiol* 31:312–320.

Keller MB, Lavori PW, Kane JM, Gelenberg AJ, Rosenbaum JF, Walzer EA, Baker LA (1992): Subsyndromal symptoms in bipolar disorder: A comparison of standard and low serum levels of lithium. *Arch Gen Psychiatry* 49:371–376.

Lavori PW, Dawson R (2000): A design for testing clinical strategies: Biased-coin adaptive within-subject randomization. *J R Stat Soc A* 163:29–38.

Little R, Rubin D (1987): *Statistical Analysis with Missing Data.* New York: Wiley.

Rubin D (1974): Estimating causal effects of treatments in randomized and non-randomized studies. *J Educ Psychol* 66:688–701.