Recursive Subsetting to Identify Patients in the STAR*D: A Method to Enhance the Accuracy of Early Prediction of Treatment Outcome and to Inform Personalized Care

Anthony Y. C. Kuk, PhD; Jialiang Li, PhD; and A. John Rush, MD

Objective: There are currently no clinically useful assessments that can reliably predict—early in treatment—whether a particular depressed patient will respond to a particular antidepressant. We explored the possibility of using baseline features and early symptom change to predict which patients will and which patients will not respond to treatment.

Method: Participants were 2,280 outpatients enrolled in the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study who had complete 16-item Quick Inventory of Depressive Symptomatology—self-report (QIDS-SR_{16}) records at baseline, week 2, and week 6 (primary outcome) of treatment with citalopram. *Response* was defined as a ≥ 50% reduction in QIDS-SR_{16} score by week 6. By developing a recursive subsetting algorithm, we used both baseline variables and change in QIDS-SR_{16} scores from baseline to week 2 to predict response/nonresponse to treatment for as many patients as possible with controlled accuracy, while reserving judgment for the rest.

Results: Baseline variables by themselves were not clinically useful predictors, whereas symptom change from baseline to week 2 identified 280 nonresponders, of which 227 were true nonresponders. By subsetting recursively according to both baseline features and symptom change, we were able to identify 505 nonresponders, of which 403 were true nonresponders, to achieve a clinically meaningful negative predictive value of 0.8, which was upheld in cross-validation analyses.

Conclusions: Recursive subsetting based on baseline features and early symptom change allows predictions of nonresponse that are sufficiently certain for clinicians to spare identified patients from prolonged exposure to ineffective treatment, thereby personalizing depression management and saving time and cost.

Trial Registration: clinicaltrials.gov Identifier: NCT00021528

J Clin Psychiatry 2010;71(11):1502–1508

Submitted: *April 8, 2010; accepted June 14, 2010.*
(doi:10.4088/JCP.10m06168blu).
Corresponding author: *Anthony Y. C. Kuk, PhD, Department of Statistics and Applied Probability National University of Singapore, 6 Science Drive 2, Singapore 117546 (stakuka@nus.edu.sg).*

Some patients with depression will respond to a particular medication, while others will require a different medication to achieve a response.[1–4] Unfortunately, there are no clinically useful pretreatment assessments that can reliably recommend 1 antidepressant medication over another for a particular patient.[1,5–7] Thus, the most effective antidepressant medication for each patient can presently only be identified through trial and error.[1,7–9]

During such a trial and error treatment sequence, each medication must be used for a sufficient length of time to determine whether or not it will work, an approach that may result in a prolonged sequence of several trials. If early on we could predict with a reasonably high degree of certainty that a medication will likely be ineffective for an individual patient, we could increase treatment efficiency. These patients could bypass ineffective medications and proceed to other treatments that would be more likely to be effective. Such a strategy would reduce the time and cost required to achieve treatment response and reduce patient exposure to unnecessary medication.

Attempts have been made to use the degree of depressive symptom improvement during the first few weeks of treatment to predict later outcomes (eg, response or remission after 6–12 weeks of treatment). However, these reports used receiver operating characteristics or hazard analyses to attempt to predict response/nonresponse for *all* patients rather than for individual patients.[10,11]

Some patients respond early in treatment, while others do so later.[12–14] This study was conducted to determine whether it is possible to predict early in treatment (by week 2), with a clinically meaningful level of certainty, which patients would and which patients would not respond later in treatment. The study also allowed for some patients to fall into an indeterminate group. This group was composed of patients for whom a prediction could not be made early on with a sufficient degree of certainty for a recommendation to continue or stop to be made. By separating patients into those for whom we could confidently predict response and nonresponse while reserving judgment on the remaining group (rather than making a forced call with little basis to do so), our method would seem to better mirror the realities of practice. Furthermore, the indeterminate group could also be of substantial interest for other reasons (eg, it might represent persons with a different pathophysiology or those who require a different treatment approach).

This report illustrates how such analyses can be conducted. We used data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study, which enrolled a reasonably representative sample of outpatients with depression for antidepressant treatment trials.[4,12] All participants were initially treated with citalopram for a 12-week period. Dosing was vigorous to ensure that nonresponse (or nonremission), should it occur, would not likely be due to underdosing. For this report, analyses were conducted to determine whether

response at week 6 could be predicted by either (1) baseline features or (2) baseline features plus symptom changes from baseline to week 2. The latter effort provided a basis for developing a modified classification method.

METHOD

Clinical Study Overview

The STAR*D study has been described in detail elsewhere[2,4] and is registered at clinicaltrials.gov (identifier: NCT00021528). In brief, clinical research coordinators collected standard sociodemographic information at baseline, as well as administered the Cumulative Illness Rating Scale[15] to measure current general medical conditions and the 16-item Quick Inventory of Depressive Symptomatology–self-report (QIDS-SR_{16})[16–19] forms to measure depressive symptom severity. During the initial 12-week course of treatment with citalopram, clinic visits were recommended at baseline (week 0) and at weeks 2, 4, 6, 9, and 12. For this report, the primary outcome was the QIDS-SR_{16} score collected at weeks 0, 2, and 6. We defined *response* to treatment as a ≥ 50% reduction in the total 16-item Quick Inventory of Depressive Symptomatology-Clinician Rating (QIDS-C_{16}) score by week 6 compared to baseline score.

Sample Selection

For inclusion in these analyses, participants had to have complete QIDS-C_{16} data at baseline and at weeks 2 and 6 (n = 2,280). We counted the measure as "at week 2" if the data were acquired within ± 6 days of week 2, a 13-day period. A similar rule was used for the "at week 6" data.

Statistical Analysis

On the basis of the reports of others, we selected a few baseline measurements that might predict treatment response by week 6: gender,[19] anxious features (Hamilton Depression Rating Scale anxiety/somatization factor score ≥ 7),[20] chronic episode (> 2 years),[12] and the presence/absence of 1 or more general medical conditions.[12] We first cross-tabulated these features to see whether and how well these 4 dichotomous baseline variables might predict response by week 6. For the continuous baseline variables of age and baseline QIDS-SR_{16} score, we assessed their predictive value for response at week 6 by plotting their negative predictive value (NPV), or the probability that a patient who is predicted not to respond will actually not respond, and their positive predictive value (PPV), or the probability that a patient who is predicted to respond actually does so, as a function of the cutoff point. Note that NPV and PPV are more relevant to making clinical decisions than are sensitivity and specificity. Clinicians are interested in knowing whether a person for whom they make a prediction will actually be a "responder" or a "nonresponder." In this study, we are particularly interested in NPV because if a clinician can predict early on with sufficient confidence that the current treatment is not going to work for a particular patient, he or she can terminate or modify it or recommend another treatment to improve the chance of success. This approach will not only spare identified patients from prolonged exposure to ineffective treatment but also save time and cost.

We then investigated whether we could use the change in QIDS-C_{16} score from baseline to week 2 (symptom change) to predict the response at week 6 and whether extra mileage could be gained by using both symptom change and baseline features to make a prediction. A large number of logistic regression models had been fitted using baseline features and symptom change as the predictors of response with various interaction terms included, but they failed to deliver high enough NPV/PPV and were abandoned. To achieve better predictions, we decided to modify the recursively partitioned classification tree technique,[21] which was known for its flexibility in handling complex interactions. For this purpose, we divided the participants into 6 clinically meaningful categories based on how much depressive symptom change they had shown from baseline to week 2. We defined the proportion of reduction in QIDS-SR_{16} score from baseline to week 2 (W_2) using the formula $W_2 = (S_0 - S_2)/S_0$, with S_0 representing the participant's baseline QIDS-SR_{16} score and S_2 representing the participant's QIDS-SR_{16} score at week 2. Since we defined *response* to treatment as a reduction in QIDS-SR_{16} score of ≥ 50% from baseline to week 6, then, on the basis of linear extrapolation, participants whose scores were reduced by one-sixth after week 2 would be on course to achieve a 50% reduction by week 6. Thus, 3 natural cut points are available to categorize participants based on their W_2 value: 0 (no improvement after 2 weeks), reduction by ⅙ (on course), and reduction by ½ (target already achieved). For finer classification, we somewhat arbitrarily added 2 more cut points: reduction by 1/12 (midway between 0 and ⅙) and reduction by ⅓ (midway between ⅙ and ½). Using these 5 cut points, participants could be grouped into the following 6 categories:

1. Those worse off after 2 weeks ($W_2 < 0$).
2. Those with very little improvement after 2 weeks ($0 \le W_2 < 1/12$).
3. Those who had modest symptom improvement but were still off pace to respond by week 6 ($1/12 \le W_2 < 1/6$).
4. Those right on course to respond by week 6 ($1/6 \le W_2 < 1/3$).
5. Those ahead of course to respond by week 6 ($1/3 \le W_2 < 1/2$).
6. Those who had responded by week 2 ($W_2 \ge 1/2$).

Participants in the lower-numbered categories would be expected to be less likely to respond by week 6 than those in the higher-numbered categories.

To determine whether we could gain further precision by combining baseline information with symptom change information (baseline to week 2), we proposed the following variant of the recursively partitioned classification tree technique[21] to predict which participants would *not* respond by week 6 using a prescribed NPV. This prediction would be

clinically valuable as it would enable a prediction of potential effectiveness without the need to wait the weeks it often takes for antidepressants to display their effects. It would also be valuable in that a dependable prediction of nonresponse in the longer term (in this case at week 6) would spare patients unnecessary exposure to a medication that would likely not be effective for them.

To develop a rule for our modified recursive partitioning effort, we first chose a level of certainty (say 80%, or 0.8) that we required for our prediction to be considered reliable. We thus set the NPV for nonresponse at week 6 at ≥ 0.8. We next set the following rules.

Starting with the group least likely to respond (ie, category 1), if the proportion of participants in this category (and, at the recursive stage, all other categories and subcategories previously labeled as nonresponding) who did not respond to treatment was at least 0.8, then we predicted that everyone in this category would be nonresponders and we moved to the next category. If the proportion who did not respond was < 0.8, we partitioned the category into 2 subcategories according to the values of one of the baseline variables: gender, anxious features (present/not present), chronicity (present/not present), general medical conditions (present/not present), baseline QIDS-SR_{16} score, and age. In these subcategories, subcategory 1 would be that with the higher proportion of nonresponders, subcategory 2 would be that with the lower proportion of nonresponders, and it would be required that the proportion of nonresponders in subcategory 1 together with all previously defined nonresponding categories and subcategories would be at least 0.8 cumulatively. The variables used for partitioning and its cutoff were chosen to maximize the size of subcategory 1 (ie, we wanted to be able to predict nonresponse for as many participants as possible based on the requisite NPV set at 0.8 for illustrative purposes). We then repeated the partitioning for subcategory 2. When we reached a point when a further partitioning to grow the set of participants with at least 80% actual nonrespondents was not possible, we moved to the next W_2 category and repeated the above.

Essentially, we were recursively identifying subsets of participants with high proportions of nonresponders based on their baseline measurements. Perhaps it would be more apt to call our technique "recursive subsetting" rather than recursive partitioning, since we are seeking a clinically meaningful subsample or subset rather than attempting to make a prediction for *all* participants by partitioning the whole sample.

The "training set" used for choosing the partitions and the "validation set" used to assess the predictive performance of the chosen procedures were identical, so selection bias is a valid concern. In other words, would the nominal NPV level be preserved when the classification rules constructed with the current sample are applied to a future participant sample? We addressed this concern by applying the technique of cross-validation.[22] The idea of the 10-fold cross-validation that we use later is to split the data at random into 10 subsets of equal size. We leave out the subsets 1 at a time and apply our algorithm to the remaining 9 subsets to obtain a prediction rule and assess its predictive performance over the subset left out. The predictive errors over the 10 subsets are then combined to give a more reliable assessment of the predictive performance of our procedure.

A similar recursive subsetting procedure was used to identify participants that we predicted would respond to treatment. The obvious difference from the procedure for predicting nonresponse was that we started with W_2 category 6 rather than with category 1, and we proceeded through the categories in the opposite order.

Table 1. Proportions of Participants Not Responding (Y = 0) and Responding (Y = 1) at Week 6 for Every Combination of 4 Dichotomous Baseline Variables

Chronic	Gender	General Medical Condition	Anxious	P(Y = 0)[a]	P(Y = 1)
–	Male	–	–	100/183 = 0.55	0.45
–	Male	–	+	91/177 = 0.51	0.49
–	Male	+	–	49/90 = 0.54	0.46
–	Male	+	+	121/208 = 0.58	0.42
–	Female	–	–	115/242 = 0.48	0.52
–	Female	–	+	160/340 = 0.47	0.53
–	Female	+	–	84/158 = 0.53	0.47
–	Female	+	+	172/316 = 0.54	0.46
+	Male	–	–	31/51 = 0.61	0.39
+	Male	–	+	22/49 = 0.45	0.55
+	Male	+	–	21/40 = 0.53	0.47
+	Male	+	+	57/85 = 0.67	0.33
+	Female	–	–	30/57 = 0.53	0.47
+	Female	–	+	50/79 = 0.63	0.47
+	Female	+	–	34/55 = 0.62	0.38
+	Female	+	+	85/150 = 0.57	0.43
Overall				1,222/2,280 = 0.54	0.46

[a]Denominator equals number of participants with the designated combination of baseline features. Numerator equals number of participants with nonresponse at week 6 from the group.

RESULTS

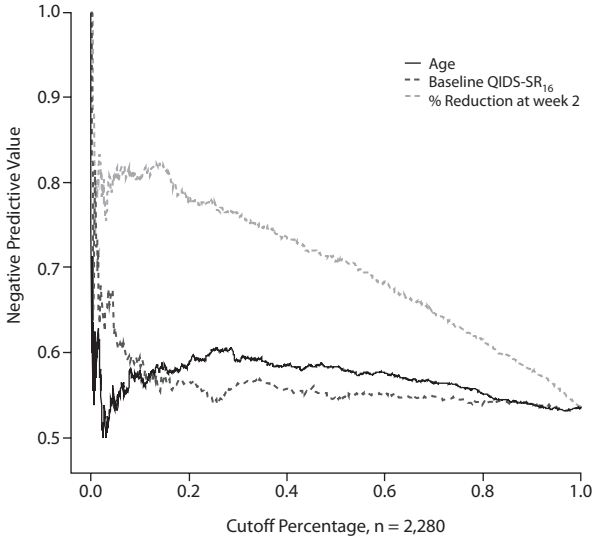
Baseline Predictions of Response

Of the 2,280 participants, 1,058 (46%) responded by week 6. Table 1 shows the proportion of participants responding and not responding to treatment at week 6 for every possible combination of the 4 dichotomous baseline variables: chronicity, gender, general medical condition, and anxious features. While patients with some of these baseline measures had a greater risk of nonresponse than patients without such features, none of the negative predictive values that could be achieved with any of the baseline parameters reached the chosen target of 0.8. For example, men with a chronic major depressive episode (MDE), anxious features, and general medical conditions (n = 85) had the highest certainty of not responding (NPV = 0.67). These baseline measures were even less predictive for responding to treatment, with the highest proportion at only 0.55.

Figure 1 shows the empirical NPV curves for both age and baseline QIDS-SR_{16} scores. Unless otherwise stated, the prediction rule being assessed is to predict nonresponse to treatment if the predictor variable being considered falls below a cutoff value. Since different predictor variables with

© Copyright 2010 Physicians Postgraduate Press, Inc.

Figure 1. Negative Predictive Value Curves to Predict Who Will Not Respond After 6 Weeks Based on Age, Baseline QIDS-SR$_{16}$ Score, and Reduction in QIDS-SR$_{16}$ Score at Week 2



Abbreviation: QIDS-SR$_{16}$ = 16-item Quick Inventory of Depressive Symptomatology–self report.

different range of values and measurement units are to be compared, we calibrate the horizontal axis of Figure 1 in terms of cutoff percentages rather than cutoff values. Reading Figure 1, we can see that the NPV is 0.56 if we predict the bottom 20% of the patients in terms of baseline score as nonresponders. For age, the upper rather than lower percentile is to be read from the horizontal axis (ie, the NPV curve shown corresponds to the rule of predicting "not responding to treatment" if patients are older rather than younger). We can see in Figure 1 that the NPV is 0.59 if we predict the oldest 20% of the patients to be nonresponders. Neither variable is satisfactory for predictive purposes since the NPV value is less than 0.7 for almost the entire range of the cutoff. Taken together, Table 1 and Figure 1 suggest that baseline information alone is not sufficient to predict—to a clinically meaningful degree—who will and who will not respond by week 6. PPV curves were also unsatisfactory (data not shown).

Postbaseline Predictors of Nonresponse

Figure 1 also shows the NPV curve for predicting who will not respond to treatment by week 6 based on the percent reduction in QIDS-SR$_{16}$ score from baseline to week 2. This NPV curve is far more informative than—and lies far above—the NPV curves based on age and baseline QIDS-SR$_{16}$.

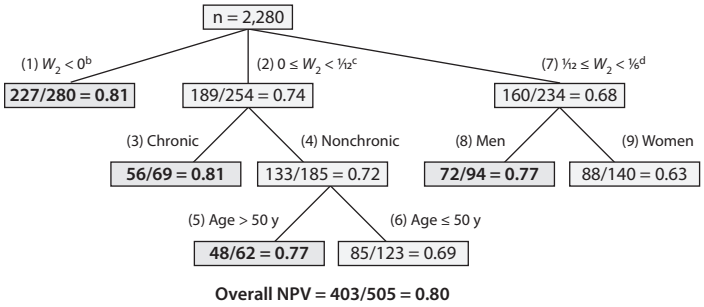
To further evaluate the usefulness of the change in QIDS-SR$_{16}$ score from baseline to week 2 for predicting nonresponse at week 6, we again used the previously described 6 categories: $W_2 < 0$, $0 \le W_2 < 1/12$, $1/12 \le W_2 < 1/6$, $1/6 \le W^2 < 1/3$, $1/3 \le W_2 < 1/2$, and $W_2 \ge 1/2$. The proportions of participants *not responding* in these 6 categories were 0.81, 0.74, 0.68, 0.61, 0.43, and 0.24, respectively. This clearly indicates that the proportion of participants not responding at week 6 decreases as the reduction in QIDS-SR$_{16}$ score over the first 2 weeks of treatment increases. If an NPV of 0.8 or above is desired, then, on the basis of the categorized W_2 alone, the best we can do is predict that participants with $W_2 < 0$ will be nonresponders. There were 280 such participants, of which 227 were true nonresponders, giving an NPV of 227/280 = 0.81.

We then incorporated baseline features into our prediction (in addition to the W_2 categories). We performed recursive subsetting (as described above) to identify subsets of participants with a high NPV within each W_2 category on the basis of baseline measurement values (Figure 2). This classification approach differs from conventional classification trees since we did not require that each individual be classified as yes/no. Rather, our rule predicted nonresponse (at week 6) for the following categories of participants: (1) those who have symptom worsening at week 2 (ie, $W_2 < 0$), (2) those with a minimal improvement in QIDS-SR$_{16}$ score after 2 weeks ($0 \le W_2 < 1/12$) *and* also have a chronic MDE, (3) those with a minimal improvement in QIDS-SR$_{16}$ score after week 2 ($0 \le W_2 < 1/12$) *and* who are nonchronic *and* who are > 50 years old, and (4) men who had modest symptom improvement but were still off pace to respond by week 6 ($1/12 \le W_2 < 1/6$). Applying this prediction rule to the sample at hand, we made a "not responding to treatment" prediction for 505 participants with an overall NPV of 403/505 = 0.80. Thus, supplementing W_2 with baseline information enabled us to identify more real nonresponders (403 vs 227 with W_2 information alone), with more or less the same NPV. For the remaining 3 W_2 categories, no further partitioning was possible while requiring an NPV of at least 0.8.

Reliability of the Recursive Subsetting

To estimate the "out of sample" performance of this recursive subsetting algorithm, we carried out cross-validation[22] and randomly divided the sample (n = 2,280) into 10 subsets (n = 228). For each subset, we carried out prediction using the rule obtained by applying our algorithm to the remaining 9 subsets as the training set. This effort resulted in 10 different scenarios. In 7 of these 10 scenarios, our algorithm yielded the same partition rules as the algorithm that was based on the full sample. In the remaining 3 scenarios, the small category of men with $1/12 \le W_2 < 1/6$ was left out of the prediction rule. Thus, a total of 471 participants (compared to the 505 without cross-validation) were predicted to not respond at 6 weeks. In fact, only 376 of the 471 were true nonresponders (cross-validated NPV of 376/471 = 0.80). This NPV value is the same as that from the non–cross-validated NPV. Thus the NPV holds up well even without cross-validation, but the number of predicted nonresponders will be overstated slightly (505 without cross-validation vs 471 with cross-validation). In other words, the ratios of true nonresponders to predicted nonresponders (NPVs) are more or less the same with or without cross-validation, but the denominator of the NPV is inflated without cross-validation.

Figure 2. Classification Tree for Predicting Who Will Not Respond to Treatment After 6 Weeks[a]



[a]Numbering indicates the partition sequence; denominator equals number of participants in the category; numerator equals number of participants in the category that did not respond. Boldface numbers indicate reliable prediction (NPV ≥ 0.8).
[b]No improvement in QIDS-SR_{16} score after 2 weeks.
[c]Very little improvement in QIDS-SR_{16} score after 2 weeks.
[d]Modest symptom improvement but still off pace to respond by week 6.
Abbreviations: NPV = negative predictive value, QIDS-SR_{16} = 16-item Quick Inventory of Depressive Symptomatology-self report.

Table 2. Responder, Nonresponder, and Indeterminate Groups Formed by Recursive Subsetting (n = 2,280)

W_2 Status[a]	Chronic[b] MDE	Age, y	Gender	n	Prediction
$W_2 < 0$				280	Nonresponse
$0 \le W_2 < 1/12$	+			69	Nonresponse
	−	> 50		62	Nonresponse
	−	≤ 50		123	Indeterminate
$1/12 \le W_2 < 1/6$			Male	94	Nonresponse
			Female	140	Indeterminate
$1/6 \le W_2 < 1/3$				531	Indeterminate
$1/3 \le W_2 < 1/2$				473	Indeterminate
$W_2 \ge 1/2$				508	Response

[a]Proportion of reduction in the 16-item Quick Inventory of Depressive Symptomatology–self-report from baseline to week 2.
[b]MDE duration > 2 years.
Abbreviation: MDE = major depressive episode.

Postbaseline Predictors of Response

We also performed recursive subsetting to predict who would respond (rather than not respond) to treatment by week 6. The best we could do was to predict that participants who had already achieved the target of at least 50% reduction in QIDS-SR_{16} from baseline by week 2 ($W_2 \ge 1/2$) would continue to do so at week 6. Even for these participants, however, the corresponding PPV was only 388/508 = 0.76. No further group could be added without decreasing the PPV to below 0.75. The fact that we could not use baseline information on top of W_2 to further pinpoint who would respond (rather than not respond) is not surprising since when we cross-classified participants according to the values of 4 baseline variables, the highest proportion that responded to treatment was 0.55 (Table 1). Thus, these variables are of very low PPV.

We used our recursive subsetting results to divide the participants into 3 groups: those predicted to *respond*, those predicted to *not respond*, and those for whom no prediction could be made—the *indeterminate* group (Table 2). As expected, the group that was predicted to not respond was concentrated in the lower-numbered W_2 categories, while the group predicted to respond coincided with the highest-numbered W_2 category.

DISCUSSION

For this report, we were interested in predicting who would or would not respond to treatment with a level of certainty that was sufficient to be clinically meaningful (ie, actionable). Note that this aim is far more ambitious and difficult than simply establishing statistically significant differences in the response rates between groups of patients as is typical for most reports. Statistical significance alone does not guarantee good enough prediction to allow clinicians to act. According to our data, 822 of 1,714 nonchronic patients responded to treatment, with a response rate of 0.48, whereas only 236 of 566 chronic patients responded to treatment, with a response rate of 0.42. The difference between the 2 response rates is highly significant statistically ($P = .01$). Despite this highly significant difference, the predictive value of chronicity was not good enough for a clinician to take action based on it. In fact, if we adopted the simple rule of predicting that nonchronic patients would respond to treatment, while chronic patients would not respond, then the empirical PPV and NPV will be 0.48 and 0.58, respectively. This predictive rule is clearly not good enough.

This study shows that baseline variables alone are not clinically useful predictors of response or nonresponse to treatment by week 6. Symptom change (QIDS-SR_{16} score) from baseline to week 2 was a more useful predictor. This finding is in line with recent findings based on randomized controlled[23] and naturalistic prospective studies[24] that have found early improvement to be a good predictor of eventual treatment outcome. However, these studies did not try to control or specify the desired PPV or NPV. They just reported them as they were. For example, in the prospective study,[24] on the basis of 795 hospitalized patients who met *DSM-IV* criteria for major depressive disorder and were treated at the discretion of the psychiatrist in charge according to the American Psychiatric Association[25] and World Federation of Societies of Biological Psychiatry[26] guidelines, the PPV was 0.88 and the NPV was 0.37 when early improvement was defined as ≥ 20% reduction in HDRS-21 score at day 14. An NPV of 0.37 means that almost five-eighths of those who showed no early improvement would respond to treatment eventually; therefore, no early improvement is not a reliable predictor of no response. The novelty of this report is to advocate for a new mind-set for predicting response/nonresponse with the PPV/NPV controlled at some prescribed levels. The desired degree of certainty (eg, ≥ 0.8) may not be attainable early on in treatment for all patients. Thus, we needed to form and allow for an indeterminate group (patients for whom predictions could not be made confidently). The ultimate aim is to make confident predictions for as many patients as possible, subject to the requisite levels of PPV/NPV, or equivalently to make the

indeterminate group as small as possible. This endeavor led us to the use of recursive subsetting and baseline variables.

While not highly predictive of treatment outcome by themselves, baseline variables in conjunction with symptom change from baseline to week 2 gave us extra mileage in predicting response/nonresponse over predictions based solely on early symptom change. By subsetting recursively on the basis of both symptom change and baseline variables, we could make a nonresponse call for 225 more patients than is possible with symptom change from baseline to week 2 alone (505 vs 280) while keeping NPV at the same value of 0.8. Cross-validation analyses suggested that the method would continue to be effective in future applications.

The numbers of patients at baseline was 4,032, at week 2 was 3,068, and at week 6 was 2,625 for this data set. We have thus far tested our procedure on the 2,280 patients with complete baseline, week 2, and week 6 information. Since our aim was to predict week 6 response using baseline variables and week 2 response, we could make predictions for 3,068 patients, but the true answer was known for 2,280 patients or 74% of the predictions only. To assess the predictive performance of our procedure over a larger set of patients, we extended the definition of nonresponse to "less than 50% reduction of baseline score at last visit" for those patients with week 2 score, no week 6 score, but at least 1 more score after week 2. With this extended definition, we had data for 2,769 patients, 489 more than before. The overall response rate under the extended definition was 1,381/2,769 = 49.7%, which is slightly lower than the previous 51.3%. Applying recursive subsetting to data from these 2,769 patients resulted in the same classification tree as in Figure 1, with an NPV of 472/617 = 0.765. It appears that the results were not affected much by dropouts.

Another novelty of this report is the reconceptualization of statistical classification to enhance the potential precision needed for an individually tailored (personalized) approach to disease management. For those patients in whom a poor outcome could be predicted early *and* with a reasonably high degree of certainty, action can be taken (ie, the treatment modified). Recursive subsetting enables us to identify such patients, and it allows us to choose the degree of certainty for our desired prediction (eg, NPV = 0.8, 0.9, etc) Recursive subsetting avoids using a forced classification or prediction for every patient by allowing for an indeterminate group (ie, patients who do not provide clear early clues as to later outcomes). As clinicians know, there are some patients for whom one cannot be sufficiently confident in predicting—early in treatment—whether they will or will not respond, and so treatment is continued. Participants in the indeterminate group may also be of interest clinically as they could have a different pathogenesis, different prognosis, or different treatment requirements. For example, one could envision adding a second drug to the initial drug early on (weeks 2–4) for the indeterminate group.

There is a clear clinical need for an early and certain prediction of whether an intervention will be effective. In some medical areas, surrogate measures (eg, biomarkers, laboratory correlates of disease processes) may aid in the selection of a treatment for a subset of patients defined by 1 or more baseline features. For depression, we lack surrogates of outcome. Therefore, symptoms are used to reflect disease processes.

Recursive subsetting enables clinicians and researchers to combine baseline and postbaseline variables and to select the degree of certainty desired for the prediction of nonresponse (or response) for particular patient groups in different treatment contexts, resulting in clear, actionable patient identification. This can be viewed as an attempt to generate so-called tailoring variables, which allow physicians to select individuals for particular treatment changes. Such tailoring variables can be prospectively tested in adaptive treatment designs.[7,27]

Naturally, the higher the level of certainty that is required the lower the number of patients that will be identified. The desired degree of certainty is affected by treatment context and treatment risks. For example, if 3 prior treatments have already failed, one might require a higher NPV for the prediction of nonresponse (eg, 0.95) before switching to another treatment. On the other hand, an NPV of 0.8 might be more appropriate when one is delivering a first treatment for a nonresistant depression, since other likely well-tolerated and effective treatments are available.

Recursive subsetting can also serve other purposes, such as the comparison of 2 treatments. If both treatments produce similar response or remission rates and both have comparable side effect risks, then we would typically regard them as equivalent. But what if treatment A produced a signal with a 90% certainty of nonresponse after 2 weeks of treatment for 50% of those who ultimately did not respond, while treatment B identified only 10% of the ultimate nonresponders with the same 90% certainty? Clinicians would clearly consider treatment A before treatment B because more patients in treatment A could avoid prolonged exposure to an ineffective medication.

This study has several limitations. Results may not be generalizable since measurement-based care was used in STAR*D,[8] which likely resulted in a more vigorous dosing of citalopram than is common in practice. In addition, treatment visits and, therefore, the measurement occasions occurred only approximately rather than precisely at the appointed time (ie, week 2, week 6), which may have reduced the precision of predictions. Neither limitation materially affects the validity of the recursive subsetting approach. Finally, these were secondary analyses, all substantive findings must be considered to be hypothesis generating.

In summary, treatment exposure can be kept to a minimum for at least some patients by selecting clinically meaningful thresholds for NPVs, by defining patient outcomes in meaningful categories, and by conducting recursive subsetting. This approach is worth further evaluation using other data sets to determine whether clinically useful recommendations can be made to assist in triage to enhance clinical outcomes of patients or to enhance the cost-efficiency of clinical trials.

Author affiliations: Department of Statistics and Applied Probability, National University of Singapore (Drs Kuk and Li) and Graduate Medical School, Duke-National University of Singapore, Singapore (Dr Rush).
Potential conflicts of interest: **Dr Rush** has been a consultant to and has received honoraria from Otsuka and has received grant/research support from the National Institute of Mental Health (NIMH) and University of Michigan. **Drs Kuk** and **Li** report no biomedical financial interests or other potential conflicts of interest.
Funding/support: The Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study was funded with federal funds from the NIMH, National Institutes of Health (contract no. N01MH90003 to University of Texas Southwestern Medical Center at Dallas [principal investigator: Dr Rush]).

REFERENCES

1. Rush AJ. Strategies and tactics in the management of maintenance treatment for depressed patients. *J Clin Psychiatry*. 1999;60(suppl 14):21–26, discussion 31–35.
2. Fava M, Rush AJ, Trivedi MH, et al. Background and rationale for the sequenced treatment alternatives to relieve depression (STAR*D) study. *Psychiatr Clin North Am*. 2003;26(2):457–494, x.
3. Rush AJ, Thase ME, Dubé S. Research issues in the study of difficult-to-treat depression. *Biol Psychiatry*. 2003;53(8):743–753.
4. Rush AJ, Fava M, Wisniewski SR, et al; STAR*D Investigators Group. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Control Clin Trials*. 2004;25(1):119–142.
5. Klein DN, Schatzberg AF, McCullough JP, et al. Age of onset in chronic major depression: relation to demographic and clinical variables, family history, and treatment response. *J Affect Disord*. 1999;55(2–3):149–157.
6. Rush AJ, Trivedi MH, Carmody TJ, et al. Response in relation to baseline anxiety levels in major depressive disorder treated with bupropion sustained release or sertraline. *Neuropsychopharmacology*. 2001;25(1):131–138.
7. Rush AJ, Wisniewski SR, Warden D, et al. Selecting among second-step antidepressant medication monotherapies: predictive value of clinical, demographic, or first-step treatment features. *Arch Gen Psychiatry*. 2008;65(8):870–880.
8. Adli M, Rush AJ, Möller H-J, et al. Algorithms for optimizing the treatment of depression: making the right decision at the right time. *Pharmacopsychiatry*. 2003;36:222–229.
9. Kemp AH, Gordon E, Rush AJ, et al. Improving the prediction of treatment response in depression: integration of clinical, cognitive, psychophysiological, neuroimaging, and genetic measures. *CNS Spectr*. 2008;13(12):1066–1086, quiz 1087–1088.
10. Nierenberg AA, McLean NE, Alpert JE, et al. Early nonresponse to fluoxetine as a predictor of poor 8-week outcome. *Am J Psychiatry*. 1995;152(10):1500–1503.
11. Quitkin FM, Petkova E, McGrath PJ, et al. When should a trial of fluoxetine for major depression be declared failed? *Am J Psychiatry*. 2003;160(4):734–740.
12. Trivedi MH, Rush AJ, Wisniewski SR, et al; STAR*D Study Team. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am J Psychiatry*. 2006;163(1):28–40.
13. Keller MB, McCullough JP, Klein DN, et al. A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. *N Engl J Med*. 2000;342(20):1462–1470.
14. Trivedi MH, Rush AJ, Pan J-Y, et al. Which depressed patients respond to nefazodone and when? *J Clin Psychiatry*. 2001;62(3):158–163.
15. Linn BS, Linn MW, Gurel L. Cumulative illness rating scale. *J Am Geriatr Soc*. 1968;16(5):622–626.
16. Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry*. 2003;54(5):573–583.
17. Trivedi MH, Rush AJ, Ibrahim HM, et al. The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. *Psychol Med*. 2004;34(1):73–82.
18. Rush AJ, Bernstein IH, Trivedi MH, et al. An evaluation of the quick inventory of depressive symptomatology and the hamilton rating scale for depression: a sequenced treatment alternatives to relieve depression trial report. *Biol Psychiatry*. 2006;59(6):493–501.
19. Kornstein SG, Sloan DM, Thase ME. Gender-specific differences in depression and treatment response. *Psychopharmacol Bull*. 2002;36(4):99–112.
20. Fava M, Rush AJ, Alpert JE, et al. Difference in treatment outcome in outpatients with anxious versus nonanxious depression: a STAR*D report. *Am J Psychiatry*. 2008;165(3):342–351.
21. Trevor H, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, NY: Springer; 2006:270–271.
22. Venables WN, Ripley BD. *Modern Applied Statistics with S-Plus*. New York, NY: Springer; 1994:343–344.
23. Szegedi A, Müller MJ, Anghelescu I, et al. Early improvement under mirtazapine and paroxetine predicts later stable response and remission with high sensitivity in patients with major depression. *J Clin Psychiatry*. 2003;64(4):413–420.
24. Henkel V, Seemüller F, Obermeier M, et al. Does early improvement triggered by antidepressants predict response/remission? analysis of data from a naturalistic study on a large sample of inpatients with major depression. *J Affect Disord*. 2009;115(3):439–449.
25. American Psychiatric Association. Practice guideline for the treatment of patients with major depressive disorder (revision). *Am J Psychiatry*. 2000;157(suppl 4):1–45.
26. Bauer M, Bschor T, Pfennig A, et al; WFSBP Task Force on Unipolar Depressive Disorders. World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of unipolar depressive disorders in primary care. *World J Biol Psychiatry*. 2007;8(2):67–104.
27. Rush AJ, Carmody TJ, Haight BR, et al. Does pretreatment insomnia or anxiety predict acute response to bupropion SR? *Ann Clin Psychiatry*. 2005;17(1):1–9.

© Copyright 2010 Physicians Postgraduate Press, Inc.