Psychological Medicine (2012), **42**, 309–316. © Cambridge University Press 2011
doi:10.1017/S0033291711001280

ORIGINAL ARTICLE

A practical approach to the early identification of antidepressant medication non-responders

J. Li[1,2]*, A. Y. C. Kuk[1] and A. J. Rush[2]

[1] *Department of Statistics and Applied Probability, National University of Singapore, Singapore*
[2] *Duke-National University of Singapore, Graduate Medical School, Singapore*

Background. The aim of the present study was to determine whether a combination of baseline features and early post-baseline depressive symptom changes have clinical value in predicting out-patient non-response in depressed out-patients after 8 weeks of medication treatment.

Method. We analysed data from the Combining Medications to Enhance Depression Outcomes study for 447 participants with complete 16-item Quick Inventory of Depressive Symptomatology – Self-Report (QIDS-SR_{16}) ratings at baseline and at treatment weeks 2, 4 and 8. We used a multi-time point, recursive subsetting approach that included baseline features and changes in QIDS-SR_{16} scores from baseline to weeks 2 and 4, to identify non-responders (<50% reduction in QIDS-SR_{16}) at week 8 with a pre-specified accuracy level.

Results. Pretreatment clinical features alone were not clinically useful predictors of non-response after 8 weeks of treatment. Baseline to week 2 symptom change identified 48 non-responders (of which 36 were true non-responders). This approach gave a clinically meaningful negative predictive value of 0.75. Symptom change from baseline to week 4 identified 79 non-responders (of which 60 were true non-responders), achieving the same accuracy. Symptom change at both weeks 2 and 4 identified 87 participants (almost 20% of the sample) as non-responders with the same accuracy. More participants with chronic than non-chronic index episodes could be accurately identified by week 4.

Conclusions. Specific baseline clinical features combined with symptom changes by weeks 2–4 can provide clinically actionable results, enhancing the efficiency of care by personalizing the treatment of depression.

Received 24 April 2011; Revised 14 June 2011; Accepted 25 June 2011; First published online 25 July 2011

Key words: Antidepressant medication, cross-validation, diagnostic medicine, negative predictive value, recursive subsetting.

Introduction

Major depressive disorder is a serious, disabling, life-shortening illness with high lifetime risks. Present practice entails a trial-and-error approach with beginning a medication, adjusting doses upwards (side effects permitting), and then deciding (after 2–8 weeks) to either stop or alter the treatment if response has not occurred (APA, 2010).

If we could reliably predict that non-response will be present at 8 weeks, within a few weeks into treatment, we could reduce exposure of selected patients' treatment that is very likely to not succeed. Such a prediction could save patients and care systems the time, side-effect risk and costs associated with a more prolonged treatment trial that in turn could hasten treatment revisions for such patients.

* Address for correspondence: J. Li, Ph.D., Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore 117546.
(Email: stalj@nus.edu.sg)

While several efforts to forecast response or non-response at 8–12 weeks using pretreatment features alone or with post-baseline symptom-change measures, none has been of sufficient value to become part of routine care (Nierenberg *et al*. 1995; Quitkin *et al*. 2003). The statistical procedures used to date have not directly led to a clinical decision because they have not aimed to control the positive predictive value (PPV) (the probability that a patient who is predicted to respond actually does so) or the negative predictive value (NPV) (the probability that a patient who is predicted to not respond actual does so). NPV and PPV are more relevant to clinical decision making than sensitivity and specificity because clinicians have to try to decide for each individual whether to continue or stop (alter) the treatment. They must have a reasonably high degree of certainty about the decision before they act.

However, unlike traditional statistical approaches such as logistic regression and classification trees (Venables & Ripley, 1994; Trevor *et al*. 2006), recursive subsetting allows binary predictions (will or will not

respond later) in which the clinician can set the degree of certainty needed for treatment. Specifically, both logistic regression and classification trees cannot allow investigators to control the NPV and PPV directly. These approaches are thus not helpful for the current goal. Typically, researchers will just report PPVs and NPVs as they are rather than trying to control them at some prescribed values. For example, Henkel *et al*. (2009) reported an NPV of 0.37 for one of their classification rules, which is clearly not good enough. Since the full benefit of treatment takes 2–3 months to realize (Trivedi *et al*. 2006), we could use recursive subsetting with a predefined reasonably high NPV (e.g. 0.75, 0.80, 0.90) so that the clinician can act.

We have previously found that recursive subsetting allows clinicians to select the degree of certainty that he/she needs to identify individuals who will do poorly (e.g. not respond) at 6 weeks using baseline features and one post-baseline symptom-change measure (Kuk *et al*. 2010). In practice, however, multiple post-baseline treatment visits are held at which additional post-baseline symptom-change measures can be obtained. This report uses a recursive subsetting approach with a new dataset in which we evaluate the utility of baseline feature measurements and one or two follow-up time points (at weeks 2 and 4) to predict non-response at 8 weeks. If successful, then efforts to automate this approach for clinical use would be justified.

Specifically, this report addresses three questions:

(1) Can recursive subsetting produce clinically useful and actionable information in a new dataset of representative out-patients with major depressive disorder?
(2) Does the best actionable information depend on both baseline and post-baseline findings?
(3) What are the advantages of using multiple time points over a single time point in predicting patient response? How do we make use of the time course/trajectory of early symptom change to predict eventual outcome?

Method

Study overview

This report used data from the Combining Medications to Enhance Depression Outcomes (CO-MED) trial (www.co-med.org;; Rush *et al*. 2011), a 7-month single-blinded, randomized, placebo-controlled trial that compared the efficacy of each of two different antidepressant medication combinations (bupropion-sustained release + escitalopram and venlafaxine-extended release + mirtazapine) *versus* escitalopram + placebo in 12 weeks of first-step medication treatment for major depressive disorder. Study details and results are available elsewhere (Rush *et al*. 2011). In brief, clinical research coordinators collected standard baseline sociodemographic and clinical information, as well as the 16-item Quick Inventory of Depressive Symptomatology – Self-Report (QIDS-SR_{16}) (Kornstein *et al*. 2002; Rush *et al*. 2003, 2006; Trivedi *et al*. 2004) to measure depressive symptom severity. During the initial 12 weeks of acute treatment, clinic visits were recommended at baseline (week 0) and at weeks 1, 2, 4, 6, 8, 10 and 12. For this report, we defined response as a $\geqslant 50\%$ reduction from baseline in the total QIDS-SR_{16} score by week 8. (Note: we chose the week 8 time point because CO-MED allowed participants with minimal response at that time to exit.)

Sample selection

For inclusion in these analyses, participants had to have complete QIDS-SR_{16} data at baseline and at week 2, 4 and 8. The analysis included 447 participants. After participants were provided with a complete description of the study, written informed consent was obtained.

Statistical analysis

For these analyses, we used the recursive subsetting approach (Kuk *et al*. 2010) and we chose an initial level of 75% NPV as a requisite for a clinically useful, reliable prediction.

We selected three dichotomous baseline features based on previous reports (Kornstein *et al*. 2002; Trivedi *et al*. 2006; Fava *et al*. 2008) that might predict treatment non-response: gender, anxious features (Hamilton anxiety/somatization factor score $\geqslant 7$) (Rush *et al*. 2004) and chronic index episode (episode duration >2 years) and two continuous variables (baseline severity based on QIDS-SR_{16} and age). (Note that a variety of other baseline variables could also have been used.)

We used change in QIDS-SR_{16} from baseline to week 2 and baseline to week 4 to predict non-response at week 8. Specifically, 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 being the participant's baseline QIDS-SR_{16} score and S_2 being the participant's QIDS-SR_{16} score at week 2. Similarly, we defined the proportion of reduction in QIDS-SR_{16} score from baseline to week 4 (W_4) using the formula $W_4 = (S_0 - S_4)/S_0$, with S_4 being the participant's QIDS-SR_{16} score at week 4.

For W_2, we specified six non-overlapping ascending intervals: $W_2 < 0$, $0 \leqslant W_2 < \frac{1}{16}$, $\frac{1}{16} \leqslant W_2 < \frac{1}{8}$, $\frac{1}{8} \leqslant W_2 < \frac{1}{4}$, $\frac{1}{4} \leqslant W_2 < \frac{1}{2}$ and $W_2 \geqslant \frac{1}{2}$. These cut-offs were selected on the basis of linear extrapolation of response and

https://doi.org/10.1017/S0033291711001280 Published online by Cambridge University Press

resulted in six meaningful categories of participants, namely those who were worse off after 2 weeks ($W_2 < 0$), those with very little improvement after 2 weeks ($0 \leqslant W_2 < \frac{1}{16}$), those who had modest symptom improvement but were still off pace to respond by week 8 ($\frac{1}{16} \leqslant W_2 < \frac{1}{8}$), those right on course to respond by week 8 ($\frac{1}{8} \leqslant W_2 < \frac{1}{4}$), those ahead of course to respond by week 8 ($\frac{1}{4} \leqslant W_2 < \frac{1}{2}$), and, those who had already responded by week 2 ($W_2 \geqslant \frac{1}{2}$). We expected that lower values of W_2 would be more closely associated with a non-response. Therefore, there might be a decreasing trend in the probability of not responding to treatment from the 1st to the 6th group. We started with the group least likely to respond (category 1) and partitioned the participants into subsets according to baseline information such that a satisfactory NPV could be achieved while the sample size of the subset is maximized. The same procedure was applied to the following groups sequentially. We stopped the partitioning when the overall NPV for the subsets reached 75% and further subsetting did not increase the NPV (Kuk *et al*. 2010).

We also used appropriately modified cut-offs for W_4 to define six intervals of W_4: $W_4 < 0$, $0 \leqslant W_4 < \frac{1}{8}$, $\frac{1}{8} \leqslant W_4 < \frac{1}{4}$, $\frac{1}{4} \leqslant W_4 < \frac{3}{8}$, $\frac{3}{8} \leqslant W_4 < \frac{1}{2}$, and $W_4 \geqslant \frac{1}{2}$ with the same interpretation as before. We could then conduct the same recursive subsetting program described above for W_2. Since week 4 is temporally closer to week 8 than to week 2, we expected that week 4 changes might be more predictive of the final outcome. We used W_4 information in three different ways: (i) using W_4 alone; (ii) using W_2 alone in the first subsetting procedure and then using W_4 alone in a second subsetting on those for whom we could not predict with the first subsetting; (iii) using W_4 and W_2 together in a single subsetting procedure.

We further used the recursive subsetting method on two different subgroups (severe/non-severe groups and chronic/non-chronic). For these subgroup analyses, we used an NPV of 80% to indicate a clinically useful reliable prediction.

One important statistical issue is to assess the external validity of our procedure. We consider the cross-validation approach in this paper. By splitting the whole data into 10 subsets at random, we formed separate training samples and test samples and evaluated the accuracy of the test samples by using the prediction rule constructed from the training samples. The cross-validated NPV provides a reasonable estimate for the prediction accuracy for future analysis.

Results

Predictions of non-response with baseline data alone

Of the 447 participants, 250 (56%) responded by week 8. Table 1 shows the proportion of participants who did and did not respond by week 8 for every possible combination of the three dichotomous baseline variables: chronicity, gender and anxious features. None of these baseline measures were clinically useful in predicting either response or non-response at 8 weeks because none of the PPVs and NPVs reached the pre-specified required level of accuracy of 75%.

Fig. 1 shows the empirical NPV curves for age and baseline QIDS-SR_{16} scores, each a continuous function of the percentiles of the continuous predictor. Neither age nor severity was sufficiently predictive for clinical purposes since the NPV value is <0.7 for almost the entire range of the cut-off. Taken together, Table 1 and Fig. 1 indicate that baseline information alone is not sufficient to predict non-response by week 8.

Prediction of non-response using week 2 data

Fig. 1 also shows that the NPV curve for predicting non-response by week 8 based on percentage reduction in QIDS-SR_{16} score from baseline to week 2 (W_2) is above the age and baseline QIDS-SR_{16} curves for almost the entire percentage range. Thus, percentage reduction from baseline to week 2 (W_2) in QIDS-SR_{16} score is far more informative about week 8 outcome than either age or baseline severity.

To further evaluate the predictive usefulness of W_2, we formed the six categories noted above: $W_2 < 0$, $0 \leqslant W_2 < \frac{1}{16}$, $\frac{1}{16} \leqslant W_2 < \frac{1}{8}$, $\frac{1}{8} \leqslant W_2 < \frac{1}{4}$, $\frac{1}{4} \leqslant W_2 < \frac{1}{2}$, and $W_2 \geqslant \frac{1}{2}$. The proportion of participants who did not respond by 8 weeks in each category was 0.64, 0.68, 0.64, 0.60, 0.38 and 0.22, respectively. While NPV was relatively higher when W_2 was small, each NPV value was still <0.75. Consequently, the clinician cannot decide to stop the medication because the certainty is insufficient.

We then incorporated baseline features into the W_2 categories and conducted recursive subsetting to identify subsets of participants with a high NPV within each W_2 category. One subset did achieve the NPV threshold of 0.75. Specifically, participants whose QIDS-SR_{16} had not dropped by at least 1/8 ($W_2 < \frac{1}{8}$) by week 2 and who were in a chronic index episode were predicted to not respond by week 8. In fact, 48 participants were in this group, and 36 of them were true non-responders at week 8, which yielded an overall NPV of $36/48 = 0.75$.

Prediction of non-response using week 4 data

Fig. 2 displays the NPV curve for predicting who will not respond to treatment by week 8 based on the percentage reduction in QIDS-SR_{16} score from baseline to week 2 and from baseline to week 4. The NPV curve

https://doi.org/10.1017/S0033291711001280 Published online by Cambridge University Press

Table 1. *Proportions of participants not responding and responding to antidepressant medication at week 8 for every combination of three dichotomous baseline variables*

Baseline variables			Proportions of participants			
			Not responding		Responding	
Chronic	Gender	Anxious	Proportion	Numbers[a]	Proportion	Numbers[b]
Yes			0.46	113/244	0.54	131/244
No			0.41	84/203	0.59	119/203
	Male		0.42	129/303	0.58	174/303
	Female		0.47	68/144	0.53	76/144
		Yes	0.42	48/114	0.58	66/114
		No	0.45	149/333	0.55	184/333
Yes	Male	Yes	0.41	14/34	0.59	20/34
Yes	Male	No	0.44	49/112	0.56	63/112
Yes	Female	Yes	0.33	9/27	0.67	18/27
Yes	Female	No	0.58	41/71	0.42	30/71
No	Male	Yes	0.50	17/34	0.50	17/34
No	Male	No	0.39	49/123	0.61	74/123
No	Female	Yes	0.42	8/19	0.58	11/19
No	Female	No	0.37	10/27	0.63	17/27
Overall			0.44	197/447	0.56	250/447

[a] Denominator = number of participants with the designated combination of baseline features. Numerator = number of participants with non-response at week 8 from the group.

[b] Denominator = number of participants with the designated combination of baseline features. Numerator = number of participants with response at week 8 from the group.

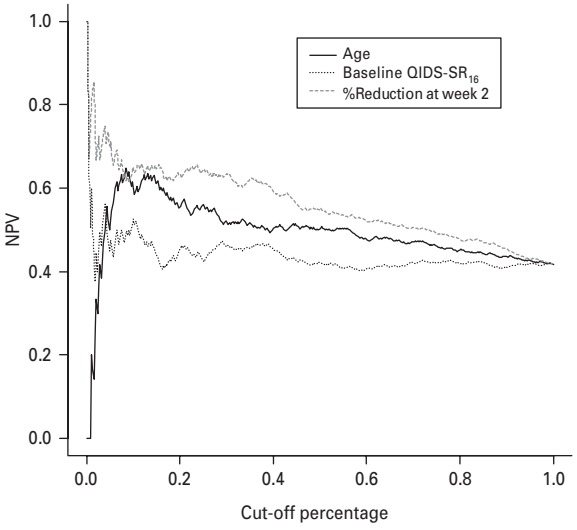

Fig. 1. Negative predictive value (NPV) curves for age, baseline 16-item Quick Inventory of Depressive Symptomatology – self-rated (QIDS-SR_{16}) score, and percentage reduction in QIDS-SR_{16} score at week 2.

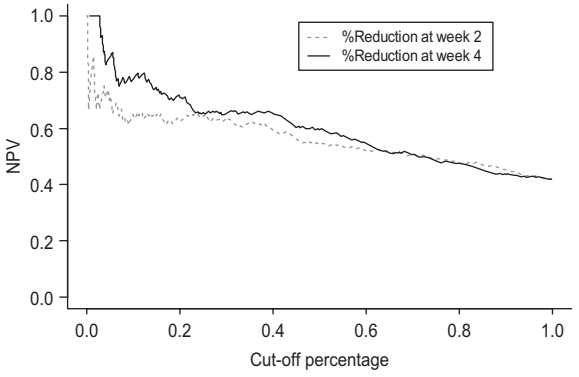

Fig. 2. Negative predictive value (NPV) curves for percentage reduction in 16-item Quick Inventory of Depressive Symptomatology – self-rated (QIDS-SR_{16}) score at week 2 and at week 4.

for W_4 is above the W_2 curve for most of its range, which indicates better predictive power for W_4.

We then considered the six categories of W_4 described earlier: $W_4 < 0$, $0 \leqslant W_4 < \frac{1}{8}$, $\frac{1}{8} \leqslant W_4 < \frac{1}{4}$, $\frac{1}{4} \leqslant W_4 < \frac{3}{8}$, $\frac{3}{8} \leqslant W_4 < \frac{1}{2}$, and $W_4 \geqslant \frac{1}{2}$. The proportion of participants who did not respond in each category was 0.85, 0.67, 0.57, 0.63, 0.35 and 0.22, respectively. The first category achieved the required NPV.

We then conducted recursive subsetting to seek further subsets of the remaining five categories and arrived at two subsets: (i) $W_4 < 0$ (i.e. worsening by week 4); (ii) $0 \leqslant W_4 < \frac{1}{4}$ and being in a chronic index episode. We could then make a 'will not respond'

https://doi.org/10.1017/S0033291711001280 Published online by Cambridge University Press

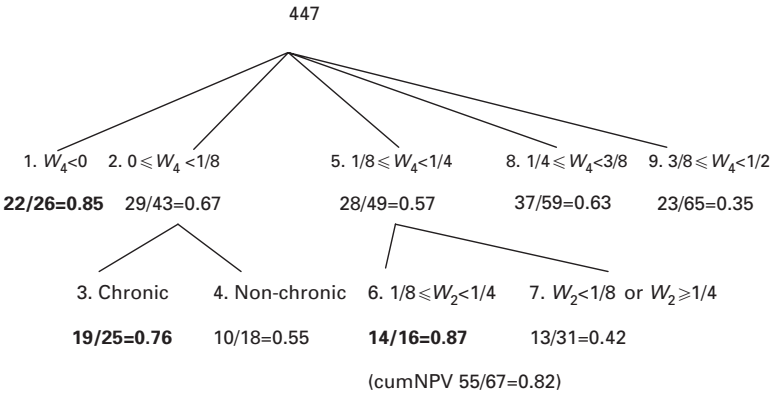

Fig. 3. Recursive subsetting algorithm using week 2 and week 4 data with a target negative predictive value (NPV) of 0.8. The numbering of subsets is according to our previous report (Kuk *et al.* 2010), indicating the sequence of the subsets generation. The NPV of each selected subset is shown in bold and CumNPV is the cumulative NPV for all the selected subsets.

prediction for 79 participants, 60 of whom were true non-responders giving an NPV of 76% (60/79). Using baseline and W_4 data allowed us to make actionable predictions for more participants than relying only on baseline and W_2 data. Of these 79 participants, 50 were not predicted by using the baseline and W_2 information. On the other hand, there were 19 participants predicted by W_2 but not by W_4. The fact that W_2 and W_4 predicted quite different individuals suggests that it is better to use both W_2 and W_4 to make predictions than to use just one of them.

Prediction of non-response using baseline, week 4 and week 2 data

To further evaluate this idea, we incorporated a two-step sequential approach. For step 1, we predicted 48 participants would not respond by using W_2 and baseline data as described in the preceding subsection. For step 2, we used W_4 and baseline data for the remaining 399 (447–48) participants. For this group, we predicted that another 39 would not respond. They belonged to the following subsets: (i) $W_4<0$; (ii) $0 \leqslant W_4 <\frac{1}{8}$ and anxious features. Thus, overall we identified 87 participants who we predicted would not respond by week 8, of whom 66 were true non-responders (NPV of 75%). Thus, this combination could predict poor outcomes for nearly 20% of the sample (87/447). Using both W_2 and W_4 data in a two-step sequence, along with baseline information, provided predictions on more participants than using either one alone.

If a higher NPV of 80% is desired, the first stage of the sequential procedure based on W_2 and baseline features is not able to predict any non-responder with the required certainty of 80% or more. Thus no action can be taken at week 2. At the second stage at week 4, both W_4 and W_2 become available, and we can use both of them as well as baseline features to predict who are the non-responders. As shown in Fig. 3, we are able to predict 67 non-responders (approximately 16% of the sample) from the following categories: (i) $W_4<0$ (26 participants, of whom 22 are true non-responders); (ii) $0 \leqslant W_4 <\frac{1}{8}$ and chronic episode (25 participants, of whom 19 are true non-responders); (iii) $\frac{1}{8} \leqslant W_4 <\frac{1}{4}$ and $\frac{1}{8} \leqslant W_2 <\frac{1}{4}$ (16 participants, of whom 14 are true non-responders). Using W_4 alone, recursive subsetting identified only the first two categories. Thus making use of both W_2 and W_4 allowed us to uncover the third subset of 16 participants as likely non-responders. There is an interesting way to describe this group of participants. They are participants who had not kept up their progress. While '$\frac{1}{8} \leqslant W_2 <\frac{1}{4}$' is considered 'on course' in week 2, if there is no further reduction in QIDS-SR$_{16}$ score and the participant is '$\frac{1}{8} \leqslant W_4 <\frac{1}{4}$' after 4 weeks, then this stagnation of progress is a sign that a patient may not respond to treatment. This is an example of how using both W_2 and W_4 gives us extra mileage in predicting patient response than using either alone. In this case, the cross-validated NPV from a 10-fold cross-validation procedure is 0.809. Results of cross-validation in all other cases show that the test samples maintain the desired NPV and suggest that our results are applicable for future data.

Application of recursive subsetting to clinically defined subgroups

We evaluated two different subgroups, defined by baseline features, to determine if the approach might be particularly useful in one or another subgroup.

We divided the participants into severe (baseline QIDS-SR$_{16}$ >15) and non-severe (baseline QIDS-SR$_{16}$ $\leqslant 15$) groups, fixed the NPV level at 80%, and conducted recursive subsetting for these two groups using W_2, W_4 and baseline measures together. For the

https://doi.org/10.1017/S0033291711001280 Published online by Cambridge University Press

severe group ($n=216$), we could predict 35 non-responders (16% of the sample) who were from the following subsets: (i) $W_4<0$ ($n=8$); (ii) $0\leqslant W_4<\frac{1}{8}$ and chronic episode ($n=11$); (iii) $\frac{1}{8}\leqslant W_4<\frac{1}{4}$ and chronic episode ($n=16$). There were 30 true non-responders, giving an NPV of 85%. For the non-severe group ($n=231$), we could predict 26 non-responders (11% of the sample) who were from the following subsets: (i) $W_4<0$ ($n=18$); (ii) $\frac{1}{8}\leqslant W_4<\frac{1}{4}$ and $\frac{1}{8}\leqslant W_2<\frac{1}{4}$ ($n=8$). There were 22 true non-responders, giving an NPV of 84%. The results from the two subgroups were not that different.

We also conducted separate analyses for the chronic/non-chronic groups, with an NPV level at 80% using recursive subsetting. For the chronic group ($n=244$), we could predict 50 non-responders (20% of the sample) who were from the following categories: (i) $W_4<0$ ($n=15$); (ii) $0\leqslant W_4<\frac{1}{8}$ and anxious features ($n=18$); (iii) $\frac{1}{8}\leqslant W_4<\frac{1}{4}$ and $W_2<\frac{1}{4}$ ($n=17$). There were 42 true non-responders, giving an NPV of 84%. For the non-chronic group ($n=203$), we could only predict 18 participants (only 7.2% of the sample) who were from the following categories: (i) $W_4<0$ ($n=11$); (ii) $\frac{1}{8}\leqslant W_4<\frac{1}{4}$ and $\frac{1}{8}\leqslant W_2<\frac{1}{4}$ ($n=7$). There were six true non-responders, giving an NPV of 86%. In this case, we could make reliable predictions for more chronic than non-chronic participants.

Discussion

The present results replicate and extend our initial work (Kuk *et al*. 2010) with recursive subsetting. We found that one can make clinically useful and actionable early predictions about individual patients who would not respond later (in this case at 8 weeks) to antidepressant medication treatment. This report used 75% as the requisite NPV for the whole sample and 80% for the subgroup analyses because we thought that such degrees of certainty would be clinically meaningful. We found that this approach did identify a clinically meaningful proportion of participants early in treatment with sufficient certainty that clinicians could take action (i.e. with a 75–80% chance of non-response, treatment could be changed). This report also found that the combination of both baseline and post-baseline (i.e. depressive symptom change) information provided reliable predictions of who would not respond, while baseline information alone was not clinically useful (i.e. NPV less than 75%) so medication would not be stopped. Further, two post-baseline assessments of symptom change (weeks 2 and 4) enhanced the number of participants for whom actionable predictions could be made as compared with only baseline plus week 2 or baseline plus week 4 information. In particular, we demonstrated that stagnation of progress after some initial improvement is a potentially useful indicator of subsequent non-response to treatment.

This report focused on NPV because we wished to predict non-response. This procedure can be used to predict positive outcomes (e.g. response) via the PPV.

In addition, the recursive subsetting approach enables clinicians to specify the requisite accuracy level (e.g. NPV $\geqslant$75% or 0.75). The choice of the accuracy level may be affected by practical considerations such as side effect risk, history of treatment resistance, cost, etc. For example, a higher NPV may be preferred if there are few treatment options, so only the patients who have less than a 10% or even 5% chance of responding would be identified to have their treatment changed.

The application of recursive subsetting is new. There are no reports other than our own (Kuk *et al*. 2010) by which to compare directly our findings. As in our initial report (Kuk *et al*. 2010), we used cross-validation to assess the out-of-sample performance of our procedure and observed that the stated NPV holds up well. In retrospect, this is not surprising since the recursive subsetting procedure only maximizes the number of predictions made subject to the chosen NPV rather than maximizing the NPV itself. A rigorous proof that the selected subset achieves the stated NPV with probability approaching 1 as sample size increases has been found and will be included in a more mathematical paper, together with other theoretical results. Study limitations include the nature of the sample (selected to have recurrent or chronic depression), the choice of categories to describe post-baseline changes, and the limited number of baseline measures evaluated. As well, response was defined by self-report and was based on symptom change. We would suggest, however, that other approaches such as sensitivity and specificity are not appropriate for clinical decision making since they can only evaluate overall misclassification performance. That is, they do not provide sufficiently specific predictions for individual participants so that clinicians can act.

It should be noted that the choices of cut-off points for W_2 and W_4 in our subsetting procedure were based on an assumption that symptom change follows a linear pattern. One can, however, use other (non-linear) patterns, which would lead to other cut-off values that may be more (or less) predictive. A grid search of optimal cut-off values may be employed to refine our algorithm.

Further, it should be noted that recursive subsetting can be adapted to any treatment or disease in which (*a*) the ultimate results/outcomes are removed in time from baseline (treatment initiation), (*b*) there is heterogeneity of response (i.e. some people respond

https://doi.org/10.1017/S0033291711001280 Published online by Cambridge University Press

and others do not), and (*c*) there may be a post-baseline indicator of effect – whether symptoms, laboratory test values or other indicators. In addition, the approach can be used for situations in which one is concerned about the development of longer-term side effects, and in which individual predictions are also needed (e.g. weight gain over time with selected antidepressants).

We stress that decision tree analytic techniques have also appeared in similar clinical contexts (see Mulsant *et al*. 2006; Andreescu *et al*. 2008). These findings converge with ours, as both document the importance of early symptom change during treatment of depression as a clinically actionable signal about the likelihood of non-response by 12 weeks. Taken together, our adult sample and their older adult sample (Mulsant *et al*. 2006; Andreescu *et al*. 2008) provide ample empirical evidence of the clinical utility of statistical techniques such as recursive subsetting and decision tree analysis as dynamic tools for the clinicians.

To bring this methodology into practice, an automated algorithm is needed in which a range of baseline and post-baseline variables can be provided and the requisite NPV and PPV levels can be chosen. In healthcare systems in which such measures are obtained, the deployment of such a tool would then need an evaluation as to clinical utility and cost-effectiveness.

In summary, recursive subsetting is an approach that can define and identify early individual depressed patients who are sufficiently unlikely to respond, so that clinicians can modify the treatment early in a meaningful proportion of patients. A combination of baseline (sociodemographic or clinical features) and early post-baseline depressive symptom changes provide a sufficiently accurate prediction that from 1 in 5 to 1 in 7 depressed out-patients can be spared at least 4 weeks of an ultimately ineffective treatment, which is a clinically and fiscally meaningful result.

Acknowledgements

This project was funded by the National Institute of Mental Health (NIMH) under contract N01MH90003 to UT Southwestern Medical Center at Dallas, TX, USA (principal investigator A.J.R.). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government. The NIMH had no role in the drafting or review of the manuscript, nor in the collection or analysis of the data.

We appreciate the support of Forest Pharmaceuticals Inc., GlaxoSmithKline, Organon Inc., and Wyeth Pharmaceuticals in providing medications at no cost for this trial.

The authors thank the CO-MED research team and the Depression Trials Networks for providing the data for this report. We also acknowledge the editorial support of Jon Kilner, MS, MA (Pittsburgh, PA, USA).

[The study's ClinicalTrials.gov registration no. is NCT00590863 (http://www.clinicaltrials.gov/ct2/show/NCT00590863?term=CO-MED&rank=1).]

Declaration of Interest

A.J.R. has received consulting fees from Otsuka, University of Michigan and Brain Resource; consultant/speaker fees from Otsuka and Bristol–Myers Squibb and author royalties from Guilford Publications, Healthcare Technology Systems and the University of Texas Southwestern Medical Center and has received research support from the National Institute of Mental Health.

References

Andreescu C, Mulsant BH, Houck PR, Whyte EM, Mazumdar S, Dombrovski AY, Pollock BG, Reynolds CF (2008). Empirically derived decision trees for the treatment of late-life depression. *American Journal of Psychiatry* **165**, 855–862.

APA (2010). Practice Guideline for the Treatment of Patients with Major Depressive Disorder, 3rd ed. American Psychiatric Publishing: Arlington, VA, pp. 17–18 (http://www.psychiatryonline.com/pracGuide/PracticePDFs/PG_Depression3rdEd.pdf). Accessed 25 January 2010.

Fava M, Rush AJ, Alpert JE, Balasubramani GK, Wisniewski SR, Carmin CN, Biggs MM, Zisook S, Leuchter A, Howland R, Warden D, Trivedi MH (2008). Difference in treatment outcome in outpatients with anxious *versus* nonanxious depression: a STAR*D report. *American Journal of Psychiatry* **165**, 342–351.

Henkel V, Seemüller F, Obermeier M, Adli M, Bauer M, Mundt C, Brieger P, Laux G, Bender W, Heuser I, Zeiler J, Gaebel W, Mayr A, Möller HJ, Riedel M (2009). Does early improvement triggered by antidepressant predict response/remission? Analysis of data from a naturalistic study on a large sample of inpatients with major depression. *Journal of Affective Disorders* **115**, 439–449.

Kornstein SG, Sloan DM, Thase ME (2002). Gender-specific differences in depression and treatment response. *Psychopharmacology Bulletin* **36**, 99–112.

Kuk AYC, Li J, Rush JA (2010). 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. *Journal of Clinical Psychiatry* **71**, 1502–1508.

Mulsant BH, Houck PR, Gildengers AG, Andreescu C, Dew MA, Pollock BG, Miller MD, Stack JA, Mazumdar S, Reynolds CF (2006). What is the optimal duration of a short-term antidepressant trial when treating geriatric depression? *Journal of Clinical Psychopharmacology* **26**, 113–120.

Nierenberg AA, McLean NE, Alpert JE, Worthington JJ, Rosenbaum JF, Fava M (1995). Early nonresponse to fluoxetine as a predictor of poor 8-week outcome. *American Journal of Psychiatry* **152**, 1500–1503.

Quitkin FM, Petkova E, McGrath PJ, Taylor B, Beasley C, Stewart J, Amsterdam J, Fava M, Rosenbaum J, Reimherr F, Fawcett J, Chen Y, Klein D (2003). When should a trial of fluoxetine for major depression be declared failed? *American Journal of Psychiatry* **160**, 734–740.

Rush AJ, Bernstein IH, Trivedi MH, Carmody TJ, Wisniewski S, Mundt JC, Shores-Wilson K, Biggs MM, Woo A, Nierenberg AA, Fava M (2006). An evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression: a STAR*D report. *Biological Psychiatry* **59**, 493–501.

Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, Thase ME, Nierenberg AA, Quitkin FM, Kashner TM (2004). Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Controlled Clinical Trials* **25**, 119–142.

Rush AJ, Kraemer HC, Sackeim HA, Fava M, Trivedi MH, Frank E, Ninan PT, Thase ME, Gelenberg AJ, Kupfer DJ, Regier DA, Rosenbaum JF, Ray O, Schatzberg AF (2006). Report by the ACNP Task Force on Response and Remission in Major Depressive Disorder. *Neuropsychopharmacology* **31**, 1841–1853.

Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC, Ninan PT, Kornstein S, Manber R, Thase ME, Kocsis JH, Keller MB (2003). The 16-item Quick Inventory of Depressive Symptomatology (QIDS) Clinician Rating (QIDS-SR) and Self-Report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry* **54**, 573–583.

Rush AJ, Trivedi MH, Stewart JW, Nierenberg AA, Fava M, Kurian BT, Warden D, Morris DW, Luther JF, Husain MM, Cook IA, Shelton RC, Lesser IM, Kornstein SG, Wisniewski SR (2011). Combining medications to enhance depression outcomes (CO-MED): acute and long-term outcomes: a single-blind randomized study. *American Journal of Psychiatry*. Published online: 2 May 2011. doi:10.1176/appi.ajp.2011.10111645.

Trevor H, Tibshirani R, Friedman J (2006). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, pp. 270–271. Springer: New York.

Trivedi MH, Rush AJ, Ibrahim HM, Carmody TJ, Biggs MM, Suppes T, Crismon ML, Shores-Wilson K, Toprac MG, Dennehy EB, Witte B, Kashner TM (2004). The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-SR) and Self-Report (QIDS-SR) in public sector patients with mood disorders, a psychometric evaluation. *Psychological Medicine* **34**, 73–82.

Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L, Norquist G, Howland RH, Lebowitz B, McGrath PJ, Shores-Wilson K, Biggs MM, Balasubramani GK, Fava M (2006). STAR*D Study Team, for the STAR*D Study Team: Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *American Journal of Psychiatry* **163**, 28–40.

Venables WN, Ripley BD (1994). *Modern Applied Statistics with S-Plus*, pp. 343–344. Springer: New York.