Chapter 11

Early improvement in depressive symptoms and response to antidepressants: an individual patient data meta-analysis

Ymkje Anna de Vries, Annelieke M. Roest, Elske H. Bos,
J. G. M. (Hans) Burgerhof, Hanna M. van Loo, Peter de Jonge

Submitted

Abstract

Objective: To investigate whether early improvement of individual depressive symptoms during antidepressant treatment predicts response or remission.

Method: We obtained individual patient data of 2,184 placebo-treated and 6,058 antidepressant-treated participants from 30 trials for major depressive disorder (MDD). The primary outcome was response on the Hamilton Depression Rating Scale (HAM-D) at week 6; secondary outcomes were remission at week 6 and response and remission at week 12. Least absolute shrinkage and selection operator (lasso) logistic regression was used for variable selection. We compared models that only included early improvement in the total HAM-D score by week 2 ($\geq$ 20% improvement vs. <20% improvement) (total improvement model) with models that included early improvement in individual HAM-D items (item improvement model) and models with interactions among early-improving HAM-D items, age, and gender (item interactions model).

Results: The models that included individual items performed slightly, but significantly, better than the total improvement model. By week 6, 51% of all antidepressant-treated participants responded, but participants who were predicted not to respond had a 29% chance of response. By week 12, the overall probability of response was 69%; of participants who were predicted not to respond, 43% responded. In post-hoc analyses with early improvement as a continuous variable, including individual items did not enhance model performance.

Conclusions: Examining individual symptoms does not add meaningfully to the predictive ability of early improvement in the total score, particularly if improvement is seen as continuous. In addition, absence of early improvement does not rule out good outcomes. Therefore, adapting treatment because of limited improvement by week two is not supported by our findings.

Introduction

Antidepressants are first-line treatments for major depressive disorder (MDD) [73, 74, 86]. However, many patients fail to respond, with response rates averaging around 50% in clinical trials [59], and it is important to identify these patients as soon as possible to minimize the duration of ineffective treatment and the time until response.

Clinical guidelines currently recommend 4 - 8 weeks of treatment before evaluating the effects of treatment and considering a change in management for patients who show no improvement [73, 74, 86]. The evidence base for this recommendation, however, is limited. At a group level, antidepressant effects can be detected within the first week of treatment [456], and at the level of an individual patient numerous studies have found that improvement within the first one or two weeks of treatment is associated with later response or remission [87, 88, 94, 95, 457, 458, 459, 460, 461].

Despite this general consensus, these studies disagree on whether lack of early improvement is a sufficiently good predictor to justify a change in management. For instance, one study found that only 4% of participants with no improvement after two weeks reached remission by week four [87], suggesting that non-improvers have virtually no chance of good outcomes. In a different study that followed participants over a longer period of time, on the other hand, 44% of participants without early improvement still responded after twelve weeks of treatment [95].

On average, most studies indicate that at least 20% to 30% of participants without early improvement attain a good outcome after four to twelve weeks of treatment, which is reduced from the overall probability of around 50% but not negligible [461]. Conversely, many early-improvers do not achieve good outcomes. Hence, better predictive models are desirable.

One possibility to extend these models is to examine individual symptoms, rather than only the total depression score. There are meaningful differences between symptoms (e.g. regarding risk factors and disability) [96], and severity of specific symptoms has been found to be associated with prognosis [83, 462]. Previous studies have also found that response or remission can be predicted by early improvement in several specific symptoms, including depressed mood, somatic symptoms, loss of insight, and others [98, 99, 100, 463, 464, 465].

However, these studies did not investigate whether improvement in individual symptoms is more informative than improvement in the total score alone. In the current study, we therefore investigated this question. We also examined whether there are interactions between early-improving symptoms, gender, and age. Finally, we examined whether individual symptoms are differentially predictive for response to different antidepressant classes.

Methods

Data source and trial selection

We requested individual patient data (IPD) from Clinical Study Data Request [441], a data-sharing platform providing data from (among others) trials of antidepressants developed by sponsors using this platform (GlaxoSmithKline and Lilly).

We examined second-generation antidepressants (SGAs, defined as selective serotonin reuptake inhibitors (SSRIs), serotonin-norepinephrine reuptake inhibitors (SNRIs), or other antidepressants approved after 1987), since older antidepressants are considered second-line options. However, we also included trials of new chemical entities, which were never approved for MDD, if an approved SGA was used as an active comparator.

Trials were required to be randomized, placebo- or active-comparator-controlled, and double-blind, and to have a minimum duration of 6 weeks, with trial visits in which the Hamilton Depression Rating Scale (HAM-D) was administered at baseline, week 2, and either week 6 (± 1) or week 12 (± 1) (or both). We excluded trials in children (<18 years old), trials for non-MDD indications, and trials that specifically included only participants with additional symptoms (e.g. MDD with pain).

Patient population

We only included participants assigned to placebo or SGAs. No eligible trials included participants assigned to non-SSRI/SNRI SGAs (e.g. mirtazapine), so our final sample consisted of participants assigned to placebo, SSRIs, or SNRIs.

We took a complete-case approach, only including participants who had valid HAM-D scores at baseline, week 2, and week 6 or 12. Week 2 visits took place on day 14 (± 7 days), week 6 visits on day 42 (± 14 days), and week 12 visits on day 84 (± 14) days. If a participant had multiple visits within the eligible time frame, we selected the visit closest to the intended visit day or, if eligible visits were equally close to the intended visit day (e.g. day 35 and day 49), we randomly selected one of the visits.

Training and test data

We randomly split the data into an 80% training set and 20% test set, stratified by treatment group (placebo, SSRI, or SNRI). The training set was used for model discovery and cross-validation, while prediction accuracy was assessed in the test set.

Outcomes and predictors

Our primary outcome was response ($\geq$ 50% reduction in HAM-D (17-item version) score) at week 6 [466]. Secondary outcomes were remission (score of $\leq$ 7 on the HAM-D-17) at week 6, and response and remission at week 12.

Improvement in symptoms was calculated from the baseline and week 2 HAM-D items and dichotomized into "no improvement" (no change or worsening) and "improvement" (improvement in the item score of $\geq$ 1). Baseline HAM-D items were dichotomized into absent (score of 0) or present (score of 1). Early improvement on the total HAM-D score was dichotomized into no improvement (<20% improvement) or improvement ($\geq$ 20% improvement), consistent with other studies [461], while the baseline HAM-D score was standardized. As demographic variables, we included age (standardized) and gender.

Statistical analysis

Our primary analyses only included antidepressant-treated participants. For variable selection, we used least absolute shrinkage and selection operator (lasso) logistic regression [467], implemented in the glmnet package (version 2.0-5) for R (version 3.3.0).

For each outcome (response and remission at week 6 and 12), we built four models: (1) the baseline model, which included only baseline HAM-D score, HAM-D items, age, and gender; (2) a total improvement model, which included these baseline variables and early improvement in the total HAM-D score; (3) an item improvement model, which included all these variables and early improvement in the 17 HAM-D items; and (4) an item interactions model, which included all of the above and all two-way interactions among early-improving items, age, and gender.

The optimal regularization penalty (λ) was determined by ten-fold cross-validation. We favored sparser models by choosing the largest λ whose deviance was within one standard error of the minimal deviance [468]. From each lasso model, we selected all variables with non-zero coefficients to build a mixed-effects logistic regression model with a random intercept for trial, using the lme4 package (version 1.1-12). Hence, we built four separate mixed-effects models for each outcome.

Model performance

The prediction accuracy of each mixed-effects model was assessed in the test set by determining the area under the receiver operating characteristic (ROC) curve (AUC) (R package pROC, version 1.8). The model with the highest AUC was considered the best model. We also determined the accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for each model by assigning participants with

a model-predicted probability of response/remission of $\geq 50\%$ to the response/remission group.

Secondary analyses and post-hoc analyses

We performed secondary analyses in the total group of both antidepressant- and placebo-treated participants. In these analyses, we included treatment group (placebo vs. SSRI vs. SNRI) as a predictor in the lasso regressions to examine whether associations between early-improving items and outcome were dependent on antidepressant class (suggesting a drug-specific mechanism).

In our main and secondary analyses, we dichotomized early improvement, for comparability with other studies. However, we conducted additional post-hoc analyses in which baseline item scores and early improvement (change from baseline in the total score and the individual items) were included as continuous variables.

Results

Trials and patients

We requested and received data for 32 trials. However, 2 trials proved to be ineligible (no week 6 or 12 visit). The remaining 30 trials investigated duloxetine (15 trials), paroxetine (13 trials), or new chemical entities (2 trials). Thirteen trials also included other SGAs (escitalopram, fluoxetine, paroxetine, or venlafaxine).

The total number of participants in these 30 trials was 10,365, of whom 8,242 participants had a week 6 visit. The ten trials with a duration of ≥ 12 weeks included 4,487 participants, of whom 3,103 had a week 12 visit. Sample characteristics are shown in Table 11.1. Table 11.3 in the Appendix provides further details about the individual trials.

Variable selection

Detailed information about the variables selected by the lasso regressions are provided in Tables 11.4 - 11.7 in the Appendix. In brief, all improvement models selected early improvement in the total score. The item improvement models generally selected most of the early-improving HAM-D items. However, items 3 (suicide) and 15 (hypochondria) were never selected, while items 1 (depressed mood), 2 (guilt), 4 (early insomnia), 7 (work and activities), 10 (psychological anxiety), and 13 (general somatic symptoms) were always selected. Baseline HAM-D items were selected infrequently. The item interaction models generally selected a number of interactions among early-improving symptoms;

Table 11.1: *Sample characteristics*

	Week 6 sample			Week 12 sample		
	Placebo	SSRI	SNRI	Placebo	SSRI	SNRI
Sample size	2,184	3,322	2,736	652	1,270	1,181
Mean baseline HAM-D (SD)	21.5 (5)	22.1 (4)	20.8 (5)	22.2 (5)	22.8 (4)	22.1 (5)
Mean age (SD)	44 (14)	43 (14)	45 (14)	50 (16)	45 (14)	48 (15)
Female (%)	63.5	61.8	65.7	64.7	62.7	65.6
Early improvement (%)	52.7	62.9	62.3	54.4	63.1	66.3
Response (%)	38.3	52.4	49.9	53.2	69.4	67.2
Remission (%)	22.6	32.1	32.7	34.5	49.4	48.9

HAM-D: Hamilton Depression Rating Scale; SNRI: serotonin-norepinephrine reuptake inhibitor; SSRI: selective serotonin reuptake inhibitor.

only for remission at week 12 were any interactions between symptoms and age or gender selected.

Model performance

ROC curves obtained from the mixed-effects logistic regression models for response at week 6 are shown in Figure 11.1. The baseline model performed quite poorly (AUC: 0.60). The total improvement model performed significantly better (AUC: 0.73), and the item improvement and item interactions model performed similarly (AUC: 0.77) and significantly better than the total improvement model. For remission at week 6 and response and remission at week 12, the patterns were similar, although model performance was worse for the week 12 outcomes (Table 11.2).

The accuracy, sensitivity, specificity, PPV, and NPV of each model are also given in Table 11.2. There were only minor differences between the three early improvement models. At week 6, 51% of antidepressant-treated participants in the test set responded and 33% remitted. The most parsimonious model with the highest AUC, the item improvement model, predicted non-response for 46% of participants; the associated NPV was 0.71, indicating that 29% of these participants were false negatives who did actually respond by week 6. Conversely, of participants who were predicted to respond, 70% actually responded (Figure 11.2). For remission at week 6, the model identified a large majority group (78% of participants) with a slightly reduced probability of remission (24%) and a minority group with an increased probability of remission (66%) (Figure 11.4 in the Appendix).

By week 12, 69% of participants in the test set responded and 51% remitted. The item improvement model predicted non-response for 16% of participants, but these participants

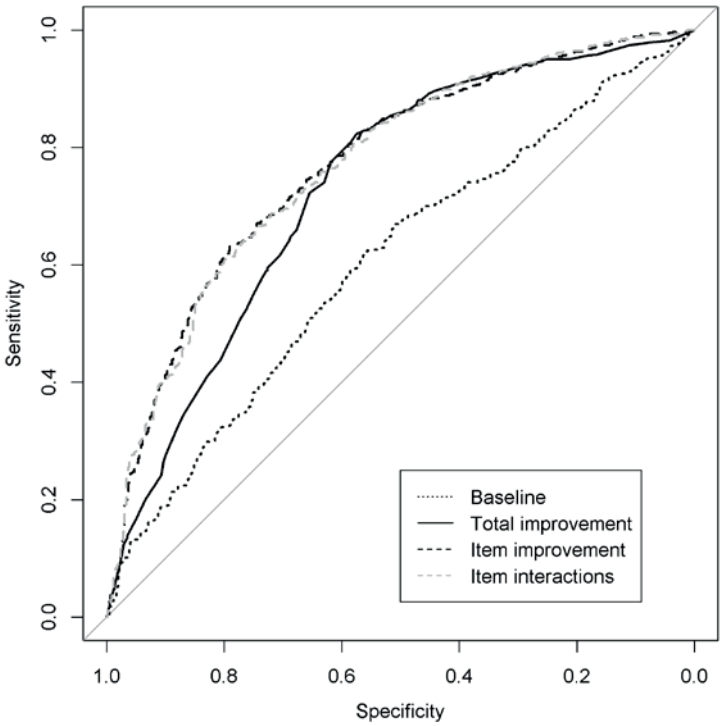

Figure 11.1: *Receiver-operating characteristic curve for the baseline, total improvement, item improvement, and item interactions model for response at week 6.*

still had a 43% probability of response. For remission, 47% of participants were predicted non-remitters, and these had a 34% probability of remission (see Figures 11.5 and 11.6 in the Appendix).

Post-hoc, we also used the predicted probability of responding or remitting to divide participants into quintiles and examined each quintile's actual probability of response or remission (Figures 11.7 - 11.10). This more fine-grained approach suggested that the improvement models could identify a risk group with poor outcomes at week 6, but prediction was less accurate and not much better than the baseline model by week 12.

Secondary analyses

Treatment group (placebo vs. SSRI vs. SNRI) was a significant predictor of response and remission at both week 6 and week 12. However, models that only included a main effect for treatment group performed as well as models with interactions between group and other variables (including gender, age, baseline score, baseline items, total improvement,

Table 11.2: *Model performance*

Week	Outcome	Model	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
6	Response	Baseline	0.60	0.59	0.67	0.51	0.59	0.59
		Total improvement	0.73	0.70	0.81	0.59	0.67	0.74
		Item improvement	0.77	0.70	0.74	0.66	0.70	0.71
		Item interactions	0.77	0.70	0.72	0.67	0.70	0.70
	Remission	Baseline	0.64	0.68	0.09	0.98	0.69	0.68
		Total improvement	0.74	0.71	0.30	0.92	0.64	0.72
		Item improvement	0.78	0.74	0.44	0.89	0.66	0.76
		Item interactions	0.78	0.74	0.43	0.89	0.66	0.76
12	Response	Baseline	0.62	0.69	1.00	0.00	0.69	N/A
		Total improvement	0.67	0.72	0.93	0.27	0.73	0.62
		Item improvement	0.71	0.71	0.90	0.29	0.73	0.57
		Item interactions	0.70	0.70	0.90	0.29	0.73	0.56
	Remission	Baseline	0.62	0.59	0.61	0.57	0.60	0.58
		Total improvement	0.68	0.64	0.69	0.59	0.64	0.64
		Item improvement	0.74	0.67	0.69	0.64	0.67	0.66
		Item interactions	0.73	0.66	0.65	0.68	0.68	0.65

AUC: Area under the (receiver operating characteristic) curve; NPV: negative predictive value; N/A: not applicable (undefined); PPV: positive predictive value.

and early-improving items), indicating no evidence for different associations between early-improving symptoms and response or remission depending on antidepressant class (Table 11.8 in the Appendix).

Post-hoc analyses

Because the total improvement model performed nearly as well as models that included individual items, we performed additional analyses using continuous change from baseline, since dichotomizing a continuous variable might affect model performance.

For response at week 6, the lasso regressions for the total improvement, item improvement, and item interaction models all selected the same variables (baseline score and change from baseline). The AUC of this model was 0.79. For remission at week 6 and response and remission at week 12, the lasso regressions did select individual items for the item improvement and/or the item interactions model. However, these models had identical or slightly (and non-significantly) worse AUCs than the (more parsimonious) total improvement model. The AUC for the total improvement model was 0.79 for remission at week 6; 0.71 for response at week 12; and 0.75 for remission at week 12.

Figure 3 depicts the probability of response or remission as a function of the percentage change from baseline in the total HAM-D score at week 2. The probability of response

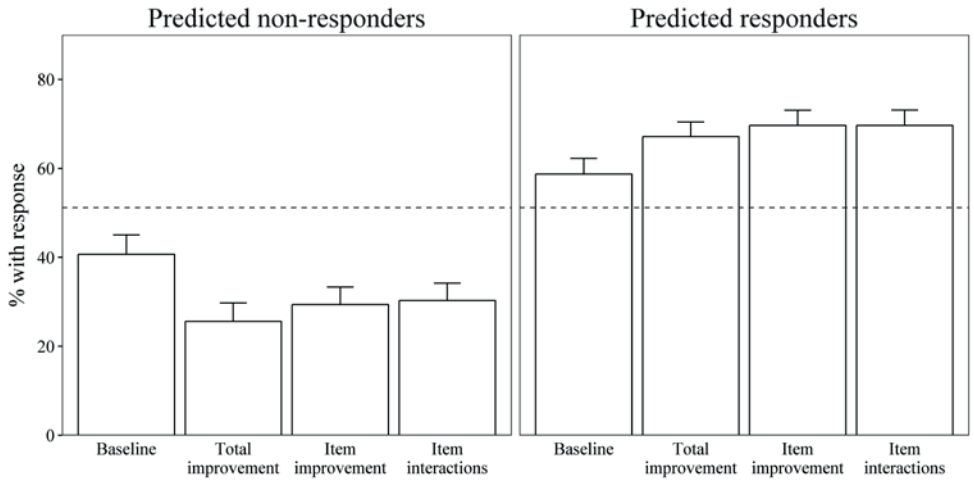

Figure 11.2: *Actual probability of response at week 6 according to participants' predicted outcome (non-response vs. response). The dashed line indicates the baseline probability of response. The models predicted non-response for 42% (baseline), 38% (total improvement), 46% (item improvement), and 47% (item interactions) of participants. Error bars indicate the 95% confidence interval.*

at week 6 was 91% for the few participants (163 (3%) of 6,058) who improved by 80% by week 2, decreasing to 17% for participants who showed any early worsening (573 (9%) participants). At week 12, however, even participants who showed early worsening still had a 39% probability of responding.

Discussion

In this individual patient data meta-analysis, we investigated whether early improvement in individual HAM-D symptoms could predict response and remission better than early improvement in the total score alone. Consistent with previous literature [461], we found that patients without early improvement were less likely to respond or remit. In our main analyses, a model with individual symptoms did perform better than a model that only included total improvement. However, the difference was relatively small, and secondary analyses examining continuous change from baseline did not confirm an added benefit from examining individual symptoms.

There was also no evidence that interactions between age, gender, and symptoms improved model performance. Our secondary analyses found no evidence that associations between early-improving symptoms and outcome differed between placebo, SSRIs, and SNRIs, since models that contained these interactions performed no better than models that did not. Taken together, our results show that early improvement is a non-specific

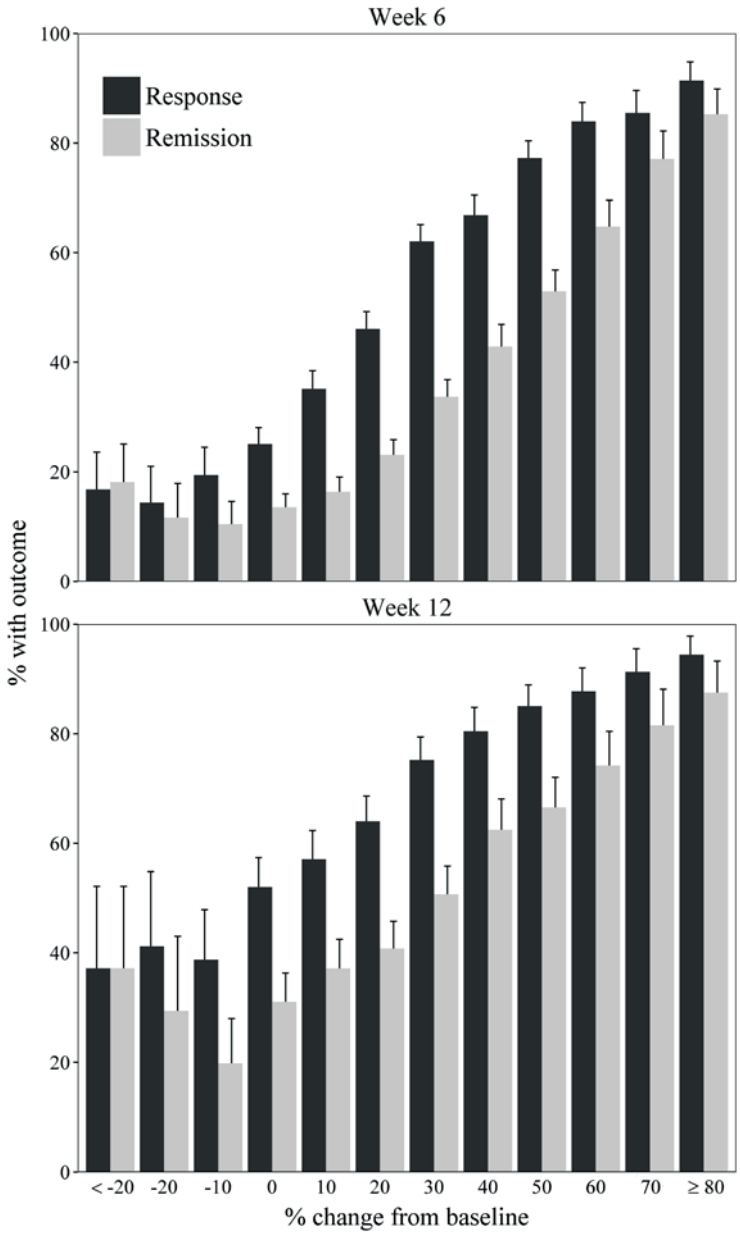

Figure 11.3: *Proportion of participants who responded or remitted according to the percentage improvement from baseline. Error bars indicate the confidence interval.*

predictor of good outcomes, regardless of treatment type.

While our results confirm the predictive value of early improvement, this value was still relatively limited, especially for longer-term outcomes. Some authors have suggested that non-improvers have virtually no chance of attaining remission and that these patients' treatment should be adapted [87], but our results indicate that these patients can still achieve good outcomes.

By 12 weeks of treatment, around 40% of patients who were predicted to have a poor outcome had achieved response and around 35% had achieved remission. These probabilities are comparable to those previously found in another large, 12-week trial (GENDEP) [95] and show that a patient's eventual outcome cannot be predicted with any certainty after two weeks of treatment. Indeed, one study found that the probability that non-responders would respond within the next two weeks was stable throughout the first twelve weeks of treatment, at around 15% [469].

A degree of caution in adapting treatment may therefore be warranted, all the more so because switching antidepressants is no more effective than continuing the same antidepressant [470]. A systematic review also found little evidence in favor of the effectiveness of early dose escalation, although escalation was clearly associated with reduced tolerability [471]. Other strategies, such as augmentation, may be more successful, but also result in decreased tolerability [472].

Such strategies might be appropriate for some patients without early improvement, for instance if a fast response is essential because of suicidality, but are likely to be premature for many patients. Given the limited predictive accuracy of models based on symptoms alone, inclusion of a broader set of predictors (e.g. psychiatric history, comorbidity, or adverse events) may be necessary to achieve better predictions.

Previous research has indicated that symptoms are not interchangeable and that the depression sum score could obscure important information [96]. Several studies have also found that early improvement in specific symptoms is associated with good outcomes [98, 99, 100, 463, 464, 465], in seeming contrast to our work. However, none of these studies included early improvement in the total score, so their findings are not directly comparable to ours. Furthermore, a variety of symptoms were associated with good outcomes, including general somatic symptoms, gastrointestinal symptoms, insomnia, depressed mood, agitation, loss of interest, feeling slowed down, and others, which also suggests that the association between early-improving symptoms and good outcomes is not particularly specific.

Our lasso regressions also tended to select most of the HAM-D items, rather than a few specific items, although some items were consistently not selected (suicidality and hypochondriasis). These results suggest that, with regard to early improvement, individual symptoms do not *add* meaningful predictive information to the sum score (especially when taken as continuous).

This may be because symptoms are actually more or less interchangeable in this regard and early improvement in any symptom is associated with good outcomes, or because symptoms are correlated and tend to improve together. However, it could also be related to the reliability of individual items. Single items are more strongly affected by random error than multi-item scales, for which the random error can balance out, which could degrade the predictive ability of a symptom. Furthermore, since our outcomes were derived from the HAM-D sum score, they are inherently dependent upon improvement in all individual items, although this would not, *a priori*, exclude differences in predictive ability, particularly if the probability or time course of improvement differs.

Our post-hoc analyses show that the association between early improvement and outcome is gradual. While a cut-off, such as $\geq 20\%$ improvement, may be easier to use in clinical practice, there is no major difference between patients on either side of this cut-off. The likelihood of response or remission does, however, seem to plateau as the percentage improvement by week 2 drops below around 10%. For instance, the likelihood of response by week 6 is only 17% for patients who deteriorate early in treatment, and the likelihood of remission is only 13%.

By week 12, however, around 39% of patients who deteriorate early in treatment have responded and 26% have remitted, which suggests that good outcomes are still possible for these patients (though less likely), if a longer period until remission can be tolerated. These results may therefore offer some guidance to clinicians who are faced with patients showing variable degrees of early improvement and need to decide between continuing, switching, or intensifying treatment.

Strengths and limitations

Among the strengths of our study is our large sample size, achieved through combining IPD. We used a rigorous approach to building predictive models, including using lasso to prevent over-fitting and using separate test data to examine model performance. We also examined multiple outcomes (response and remission) and both a short and a longer time frame (6 and 12 weeks).

A limitation of our study is that we did not take dosing schedules into account. One study has found that early improvement was more predictive when rapid, rather than slow, dose escalation was used [87]. However, there is only limited evidence for a dose-response relationship for second-generation antidepressants [381, 473, 474], and dose escalation usually also continues beyond two weeks in clinical practice.

We also took a complete-cases approach, since we were interested in predicting outcomes in patients who are receiving treatment. Our results therefore do not apply to participants who discontinue their medication and drop out of the trial. Because participants may discontinue due to lack of efficacy, the probability of response or remission may have been

overestimated somewhat. Another important reason for discontinuation was the presence of adverse events, which are also a highly relevant factor in weighing the (expected) risk-benefit ratio of treatment, but this was beyond the scope of this study.

An additional limitation is that our data were derived from clinical trials with strict inclusion and exclusion criteria. Hence, the study population represents only a subset of treatment-seeking patients, and participants may, on average, have better outcomes than patients seen in clinical practice [475]. Further research is therefore necessary to confirm that our results generalize to the broader patient population, including those with extensive comorbidity or chronic depression.

Finally, we chose the threshold of $\geq 50\%$ probability to assign participants to the response or remission category. This is a reasonable cut-off with the advantage of being independent of the data. However, a different cut-off could increase the negative predictive value (at the cost of positive predictive value). In principle, this might identify a group of participants with a lower probability of response or remission. However, because of decreasing specificity, this group would become progressively smaller as negative predictive value increases, which would limit clinical applicability.

In post-hoc analyses, we examined risk quintiles, which suggested that a small group of participants with poor outcomes at week 6 could be identified, but predictive accuracy was reduced for week 12 outcomes. Similar results were obtained while examining continuous early improvement, which suggests that this is the upper bound of predictive accuracy that can be achieved on the basis of symptoms alone.

Conclusions

Our results show that a model with only early improvement in the total score is about as predictive as models that also contain individual symptoms. Hence, clinicians need not focus on specific symptoms, but can gain as much information about the likelihood of a good outcome from improvement in the total score alone, particularly if improvement is interpreted as a continuous measure. However, the absence of early improvement does not rule out later response or remission with certainty. Therefore, adapting treatment because of limited improvement in the first two weeks would be premature for many patients.

Appendix

Table 11.3: *Supplemental table of studies*

		Duration	Dose	N (week 6)		Baseline score
Drug	Trial	(weeks)	(mg/day)	Placebo	Drug	Mean (SD)
Paroxetine	01/001 [476]	6	10 – 50	18	19	25.0 (3.2)
IR	02/001 - 004 [477]	6	10 – 50	100	112	23.5 (4.0)
	03/001 - 006 [273]	6	10 – 50	117	141	23.3 (3.7)
	7 [478]	6	10 – 60	7	8	25.1 (4.2)
	9 [479]	6	10, 20, 30, 40	31	262	22.5 (3.0)
	115 [480]	12	20	92	206	22.3 (3.6)
			Fluox 20		215	
	128 [481]	12	20	115	257	23.1 (3.8)
			Fluox 20		276	
	276 [482]	6	30	13	15	22.8 (3.9)
	279 [483]	6	30	7	14	20.8 (3.7)
Paroxetine	448 [189]	12	IR 20 – 50	94	173	23.3 (2.8)
CR			CR 25 – 62.5			
	449 [189]	12	IR 20 – 50	97	193	23.6 (3.1)
			CR 25 – 62.5			
	487 [484]	12	IR 10 – 40	97	189	22.1 (3.1)
			CR 12.5 – 50			
	810 [485]	8	12.5, 25	128	267	23.5 (3.1)
Duloxetine	HMAQ-A [486]	8	40 – 120	56	56	18.5 (4.4)
			Fluox 20		27	
	HMAQ-B [487]	8	40 – 120	60	67	18.1 (5.2)
			Fluox 20		29	
	HMAT-A [488]	8	40, 80	76	142	17.5 (5.3)
			Parox 20		73	
	HMAT-B [489]	8	40, 80	71	142	17.9 (5.2)
			Parox 20		66	
	HMAY-A [490]	8	80, 120	87	171	20.0 (3.7)
			Parox 20		79	
	HMAY-B [491]	8	80, 120	96	184	21.0 (3.6)
			Parox 20		89	
	HMBH-A [492]	9	60	102	99	21.2 (4.0)
	HMBH-B [493]	9	60	112	99	20.5 (3.4)
	HMBU [494]	12	60 - 120	-	137	23.1 (3.7)
			Ven 75 - 225		153	
	HMCQ [494]	12	60 - 120	-	133	22.3 (3.3)
			Ven 75 - 225		294	
	HMCR [495]	8	60	113	225	17.7 (5.0)
			Escit 10		237	
	HMCV [496]	8	60	-	183	21.2 (3.9)
			Parox 20		204	
	HMFA [497]	12	60	89	205	18.8 (6.3)

continued

Table 11.3: *Supplemental table of studies*

		Duration	Dose	N (week 6)		Baseline score
Drug	Trial	(weeks)	(mg/day)	Placebo	Drug	Mean (SD)
	HMFS-A [498]	8	60	103	221	22.9 (4.2)
	HMFS-B [498]	8	60	111	225	22.8 (4.7)
NCEs	NKD20006 [499]	8	Parox 20	95	83	24.5 (2.8)
	NKF100096 [500]	8	Parox 20 - 30	97	88	22.2 (5.6)

CR: controlled release; Escit: escitalopram (active comparator); Fluox: fluoxetine (active comparator); IR: immediate release; NCEs: new chemical entities; Parox: paroxetine (active comparator); Ven: venlafaxine (active comparator).

Table 11.4: *Model coefficients for response at week 6*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Intercept	-0.874 (0.457)	-0.952 (0.082)	-1.680 (0.100)	-1.271 (0.124)
Age	-0.059 (0.034)			
HAM-D item 1	0.736 (0.452)			
HAM-D item 6	0.135 (0.066)			
HAM-D item 9	0.052 (0.063)			
HAM-D item 12	0.058 (0.063)			
HAM-D item 16	0.107 (0.077)			
HAM-D item 17	0.157 (0.082)			
Total improvement		1.606 (0.067)	0.600 (0.094)	0.734 (0.099)
Improv item 1			0.351 (0.077)	0.326 (0.138)
Improv item 2			0.376 (0.069)	0.077 (0.146)
Improv item 4			0.182 (0.069)	-0.122 (0.112)
Improv item 5			0.302 (0.070)	0.219 (0.143)
Improv item 6			0.243 (0.070)	0.119 (0.093)
Improv item 7			0.197 (0.072)	-0.128 (0.125)
Improv item 8			0.148 (0.070)	-0.094 (0.117)
Improv item 9			0.220 (0.069)	0.172 (0.127)
Improv item 10			0.274 (0.070)	0.091 (0.149)
Improv item 11			0.231 (0.068)	-0.029 (0.134)
Improv item 12			0.195 (0.079)	0.066 (0.103)
Improv item 13			0.328 (0.073)	-0.044 (0.148)
Improv item 14			0.319 (0.081)	-0.013 (0.174)
Improv item 16				-0.120 (0.113)
Improv 1 × improv 2				0.034 (0.149)
Improv 1 × improv 5				0.014 (0.146)
Improv 1 × improv 9				-0.015 (0.144)
Improv 1 × improv 10				-0.025 (0.150)
Improv 1 × improv 14				0.163 (0.184)
Improv 2 × improv 8				0.220 (0.140)
Improv 2 × improv 10				0.075 (0.139)
Improv 2 × improv 11				0.202 (0.137)
Improv 2 × improv 13				0.173 (0.146)
Improv 4 × improv 7				0.333 (0.137)
Improv 5 × improv 9				0.130 (0.141)
Improv 5 × improv 10				-0.015 (0.140)
Improv 5 × improv 11				0.037 (0.141)
Improv 6 × improv 8				0.192 (0.142)
Improv 6 × improv 14				0.255 (0.171)
Improv 8 × improv 12				0.232 (0.163)
Improv 10 × improv 7				0.240 (0.142)
Improv 10 × improv 11				0.067 (0.138)
Improv 10 × improv 13				0.055 (0.149)

continued

Table 11.4: *Model coefficients for response at week 6*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Improv 11 × improv 4				0.268 (0.141)
Improv 13 × improv 7				0.142 (0.148)
Improv 13 × improv 16				0.788 (0.213)
Improv 14 × improv 13				0.217 (0.170)
Observations	4,847	4,847	4,847	4,847
Log Likelihood	-3,291.889	-2,982.251	-2,864.412	-2,834.873
AIC	6,601.778	5,970.503	5,760.824	5,749.746
BIC	6,660.153	5,989.961	5,864.602	6,009.191

AIC: Akaike Information Criterion; BIC: Bayes Information Criterion; HAM-D: Hamilton Depression Rating Scale.

Table 11.5: *Model coefficients for remission at week 6*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Intercept	-0.775 (0.071)	-1.880 (0.088)	-2.610 (0.229)	-2.170 (0.149)
Baseline score	-0.424 (0.037)	-0.493 (0.039)	-0.709 (0.047)	-0.732 (0.044)
HAM-D item 2			-0.167 (0.121)	
HAM-D item 12			0.160 (0.091)	
HAM-D item 13			-0.241 (0.165)	
HAM-D item 16			0.139 (0.166)	
Total improvement		1.591 (0.080)	0.395 (0.110)	0.692 (0.121)
Improv item 1			0.390 (0.088)	0.218 (0.170)
Improv item 2			0.539 (0.081)	0.107 (0.182)
Improv item 4			0.238 (0.073)	-0.127 (0.161)
Improv item 5			0.321 (0.074)	0.108 (0.171)
Improv item 6			0.205 (0.075)	-0.055 (0.170)
Improv item 7			0.283 (0.077)	-0.179 (0.181)
Improv item 8			0.274 (0.074)	-0.031 (0.142)
Improv item 9			0.192 (0.073)	-0.232 (0.149)
Improv item 10			0.274 (0.076)	-0.177 (0.161)
Improv item 11			0.251 (0.073)	0.046 (0.137)
Improv item 12			0.097 (0.099)	0.056 (0.107)
Improv item 13			0.442 (0.076)	0.008 (0.168)
Improv item 14			0.433 (0.081)	-0.129 (0.194)
Improv item 16			0.250 (0.181)	-0.007 (0.216)
Improv item 17			0.274 (0.125)	0.086 (0.140)
Improv 1 × improv 2				-0.197 (0.172)
Improv 1 × improv 4				0.147 (0.177)
Improv 1 × improv 5				0.106 (0.175)
Improv 1 × improv 6				-0.022 (0.175)
Improv 1 × improv 7				0.222 (0.173)
Improv 1 × improv 14				0.272 (0.203)
Improv 1 × improv 16				0.167 (0.232)
Improv 2 × improv 6				0.195 (0.149)
Improv 2 × improv 7				0.107 (0.158)
Improv 2 × improv 8				0.256 (0.150)
Improv 2 × improv 9				0.225 (0.150)
Improv 2 × improv 10				0.208 (0.153)
Improv 2 × improv 11				0.083 (0.147)
Improv 2 × improv 13				0.015 (0.154)
Improv 4 × improv 7				0.243 (0.153)
Improv 4 × improv 8				0.190 (0.150)
Improv 5 × improv 10				0.048 (0.151)
Improv 5 × improv 11				0.144 (0.145)
Improv 5 × improv 13				0.027 (0.149)
Improv 6 × improv 8				0.102 (0.149)

continued

Table 11.5: *Model coefficients for remission at week 6*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Improv 6 × improv 9				0.277 (0.147)
Improv 7 × improv 10				0.203 (0.155)
Improv 9 × improv 10				0.288 (0.154)
Improv 10 × improv 13				0.215 (0.156)
Improv 11 × improv 13				0.201 (0.147)
Improv 14 × improv 12				0.250 (0.184)
Improv 14 × improv 13				0.491 (0.167)
Improv 16 × improv 8				0.233 (0.203)
Improv 16 × improv 12				0.225 (0.210)
Improv 16 × improv 17				1.010 (0.334)
Observations	4,847	4,847	4,847	4,847
Log Likelihood	-2,947.784	-2,712.485	-2,547.361	-2,512.073
AIC	5,901.569	5,432.971	5,140.722	5,122.146
BIC	5,921.027	5,458.915	5,289.903	5,439.966

AIC: Akaike Information Criterion; BIC: Bayes Information Criterion; HAM-D: Hamilton Depression Rating Scale.

Table 11.6: *Model coefficients for response at week 12*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Intercept	0.773 (0.140)	0.081 (0.130)	0.007 (0.161)	-0.051 (0.155)
Age		-0.228 (0.065)	-0.235 (0.066)	-0.222 (0.066)
HAM-D item 4			-0.580 (0.134)	
Total improvement		1.202 (0.104)	0.394 (0.146)	0.452 (0.149)
Improv item 1			0.456 (0.126)	0.475 (0.163)
Improv item 2			0.306 (0.116)	0.010 (0.173)
Improv item 4			0.536 (0.129)	0.237 (0.201)
Improv item 7			0.168 (0.120)	0.119 (0.143)
Improv item 6				-0.416 (0.176)
Improv item 10			0.293 (0.116)	-0.331 (0.180)
Improv item 12				-0.285 (0.185)
Improv item 13			0.411 (0.124)	-0.046 (0.274)
Improv 1 × improv 4				-0.028 (0.239)
Improv 1 × improv 13				-0.042 (0.274)
Improv 2 × improv 10				0.505 (0.222)
Improv 4 × improv 13				0.195 (0.247)
Improv 10 × improv 6				0.675 (0.231)
Improv 10 × improv 12				0.793 (0.272)
Improv 13 × improv 2				0.138 (0.248)
Improv 13 × improv 6				0.494 (0.247)
Improv 13 × improv 7				0.161 (0.254)
Improv 13 × improv 10				0.127 (0.245)
Observations	1,961	1,961	1,961	1,961
Log Likelihood	-1,200.156	-1,125.716	-1,085.160	-1,076.193
AIC	2,404.312	2,259.431	2,192.319	2,196.386
BIC	2,415.474	2,281.756	2,253.712	2,319.172

AIC: Akaike Information Criterion; BIC: Bayes Information Criterion; HAM-D: Hamilton Depression Rating Scale.

Table 11.7: *Model coefficients for remission at week 12*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Intercept	0.305 (0.150)	-0.425 (0.160)	-1.034 (0.183)	-0.552 (0.226)
Age	-0.135 (0.059)	-0.142 (0.059)	-0.139 (0.061)	-0.065 (0.079)
Gender				-0.063 (0.149)
Baseline score	-0.270 (0.056)	-0.328 (0.058)	-0.398 (0.062)	-0.436 (0.066)
HAM-D item 3	-0.298 (0.102)	-0.231 (0.106)	-0.173 (0.109)	-0.160 (0.113)
HAM-D item 4	-0.322 (0.109)	-0.395 (0.113)	-0.588 (0.134)	-0.593 (0.138)
HAM-D item 16	0.355 (0.121)	0.361 (0.124)	0.430 (0.129)	0.474 (0.222)
Total improvement		1.151 (0.103)	0.096 (0.148)	0.261 (0.163)
Improv item 1			0.557 (0.125)	0.487 (0.207)
Improv item 2			0.388 (0.108)	-0.200 (0.246)
Improv item 4			0.480 (0.124)	0.194 (0.245)
Improv item 5			0.233 (0.108)	0.078 (0.151)
Improv item 6				-0.001 (0.190)
Improv item 7			0.322 (0.111)	-0.066 (0.263)
Improv item 8				-0.268 (0.158)
Improv item 9				-0.133 (0.191)
Improv item 10			0.244 (0.110)	-0.091 (0.170)
Improv item 11			0.167 (0.104)	-0.029 (0.175)
Improv item 13			0.268 (0.111)	-0.178 (0.213)
Improv item 14			0.262 (0.122)	-0.086 (0.258)
Improv item 15				0.010 (0.121)
Improv item 16				-0.294 (0.274)
Age × improv 8				-0.130 (0.107)
Age × improv 14				-0.128 (0.125)
Gender × improv 7				0.196 (0.213)
Improv 1 × improv 2				-0.015 (0.248)
Improv 1 × improv 4				-0.095 (0.241)
Improv 1 × improv 7				0.226 (0.246)
Improv 1 × improv 14				0.350 (0.294)
Improv 2 × improv 7				0.022 (0.229)
Improv 2 × improv 8				0.395 (0.216)
Improv 2 × improv 10				0.435 (0.218)
Improv 2 × improv 11				0.201 (0.210)
Improv 2 × improv 13				0.147 (0.228)
Improv 4 × improv 6				0.396 (0.217)
Improv 4 × improv 7				0.002 (0.220)
Improv 4 × improv 9				0.477 (0.215)
Improv 6 × improv 9				-0.788 (0.220)
Improv 7 × improv 9				0.451 (0.213)
Improv 10 × improv 13				0.289 (0.223)
Improv 11 × improv 5				0.251 (0.210)
Improv 13 × improv 6				0.445 (0.221)

continued

Table 11.7: *Model coefficients for remission at week 12*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Improv 15 × improv 16				0.631 (0.309)
Observations	1,961	1,961	1,961	1,961
Log Likelihood	-1,315.276	-1,249.787	-1,195.004	-1,164.499
AIC	2,644.551	2,515.574	2,424.008	2,416.998
BIC	2,683.620	2,560.224	2,518.889	2,662.571

AIC: Akaike Information Criterion; BIC: Bayes Information Criterion; HAM-D: Hamilton Depression Rating Scale.

Table 11.8: *Model performance for secondary analyses investigating interactions with treatment group*

		AUC			
		Week 6		Week 12	
Model	Interactions?	Response	Remission	Response	Remission
Baseline	No	0.61	0.66	0.63	0.62
	Yes	0.60	0.66	0.60	0.61
Total improvement	No	0.74	0.75	0.69	0.69
	Yes	0.74	0.75	0.68	0.68
Item improvement	No	0.78	0.79	0.72	0.72
	Yes	0.77	0.78	0.70	0.70
Item interactions	No	0.78	0.79	0.72	0.73
	Yes	0.77	0.78	0.69	0.72

AUC: Area under the (receiver operating characteric) curve.

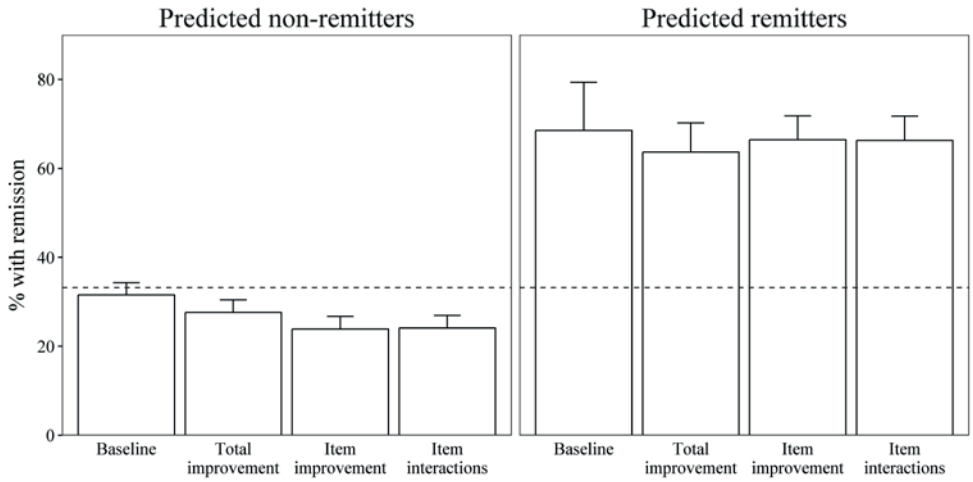

Figure 11.4: *Actual probability of remission at week 6 according to participants' predicted outcome (non-remission vs. remission). The dashed line indicates the baseline probability of remission. The models predicted non-remission for 96% (baseline), 85% (total improvement), 78% (item improvement), and 78% (item interactions) of participants.*

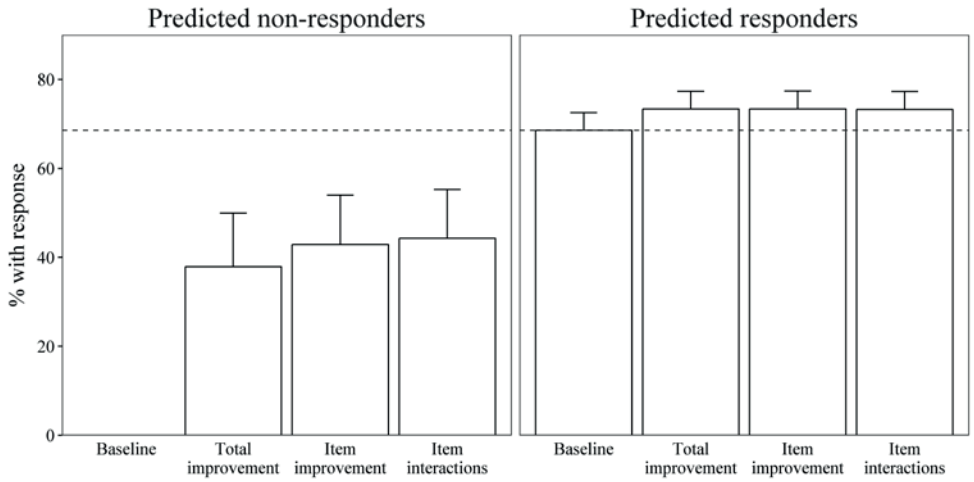

Figure 11.5: *Actual probability of response at week 12 according to participants' predicted outcome (non-response vs. response). The dashed line indicates the baseline probability of response. The models predicted non-response for 0% (baseline), 13% (total improvement), 16% (item improvement), and 16% (item interactions) of participants.*

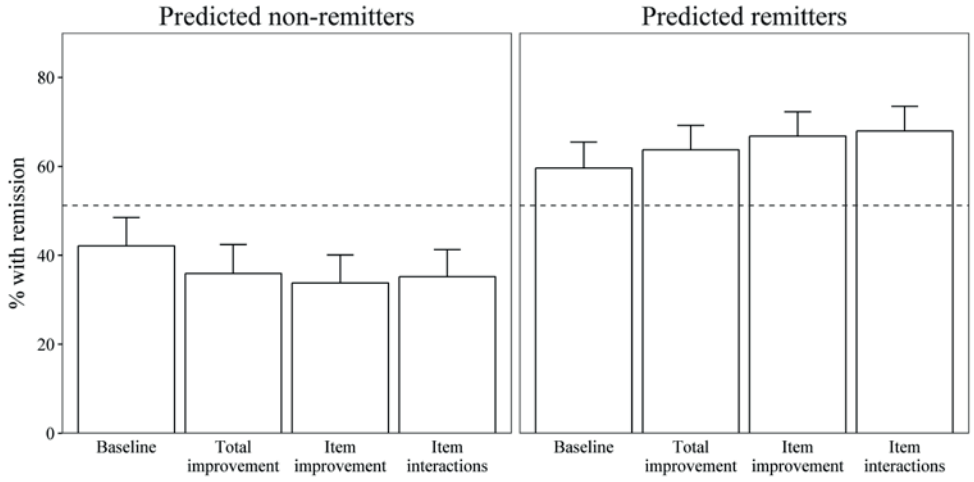

Figure 11.6: *Actual probability of remission at week 12 according to participants' predicted outcome (non-remission vs. remission). The dashed line indicates the baseline probability of remission. The models predicted non-remission for 48% (baseline), 45% (total improvement), 47% (item improvement), and 51% (item interactions) of participants.*

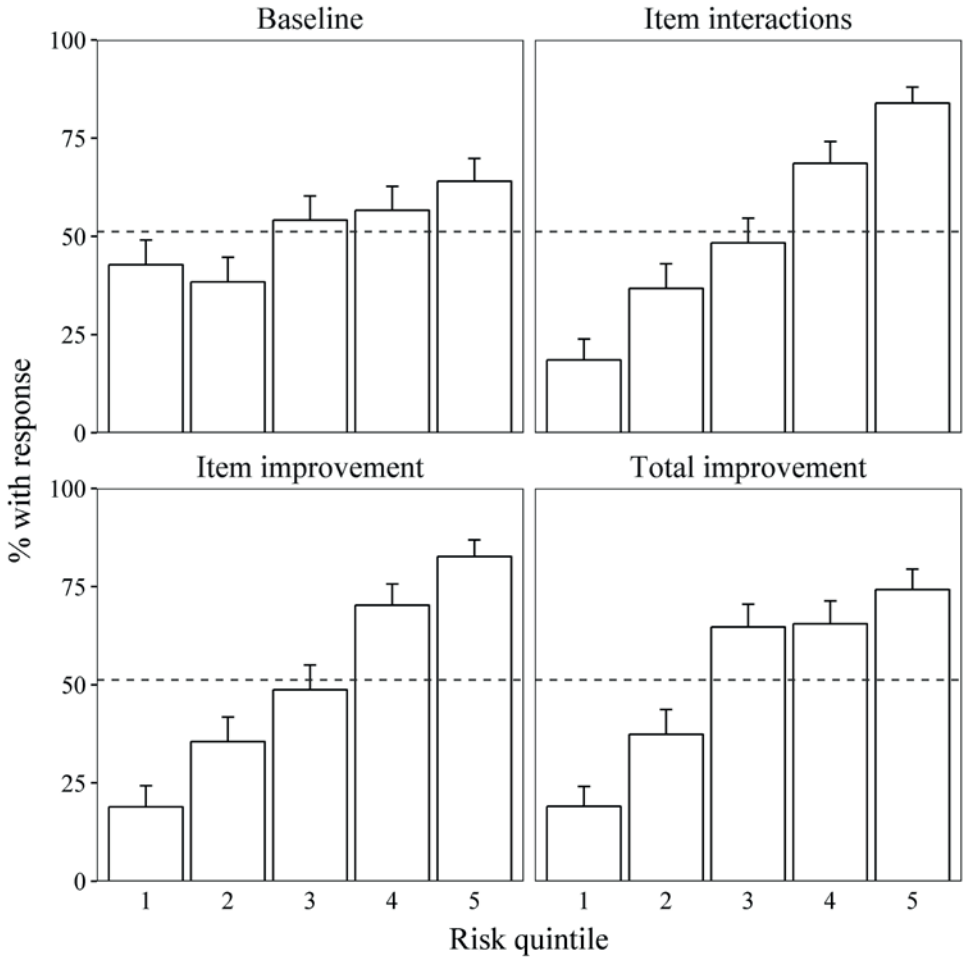

Figure 11.7: *Actual probability of response at week 6 according to risk quantiles and model. For each model, participants' predicted probability of response was used to divide participants into quintiles of "risk", with the lowest quintile having the lowest predicted probability of response.*

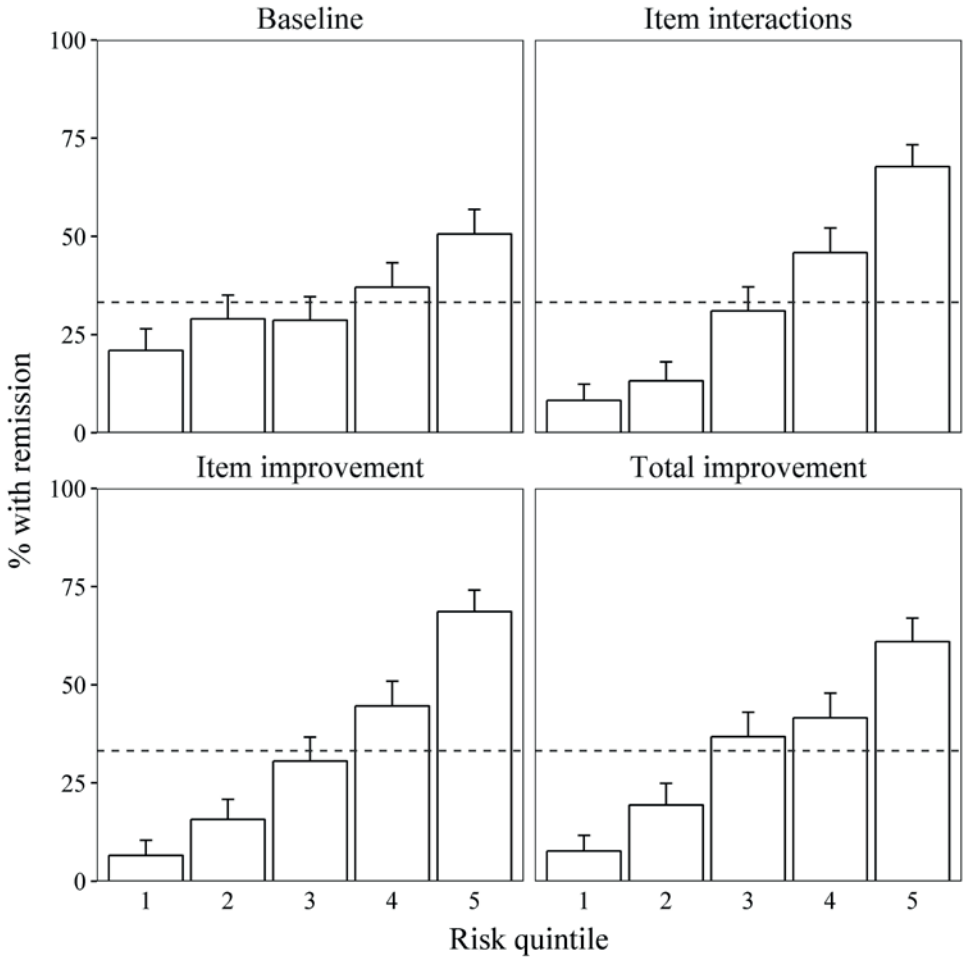

Figure 11.8: *Actual probability of remission at week 6 according to risk quantiles and model. For each model, participants' predicted probability of response was used to divide participants into quintiles of "risk", with the lowest quintile having the lowest predicted probability of response.*

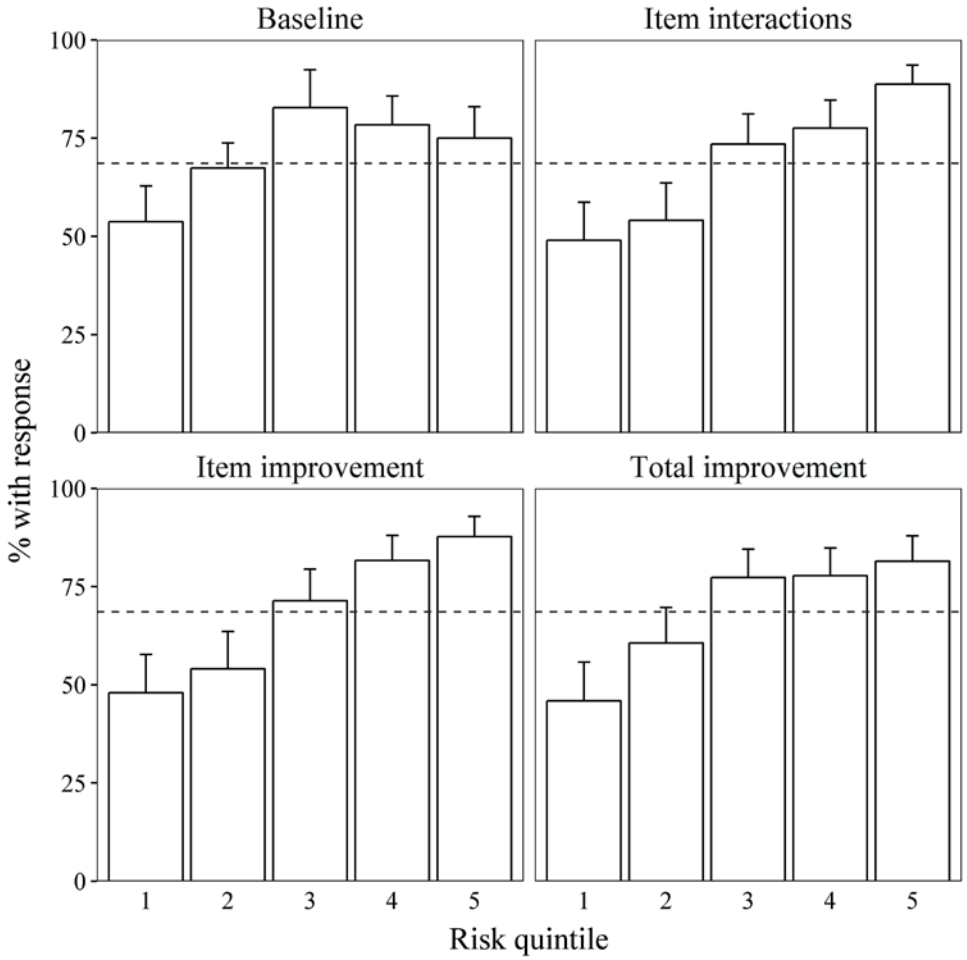

Figure 11.9: *Actual probability of response at week 12 according to risk quantiles and model. For each model, participants' predicted probability of response was used to divide participants into quintiles of "risk", with the lowest quintile having the lowest predicted probability of response.*

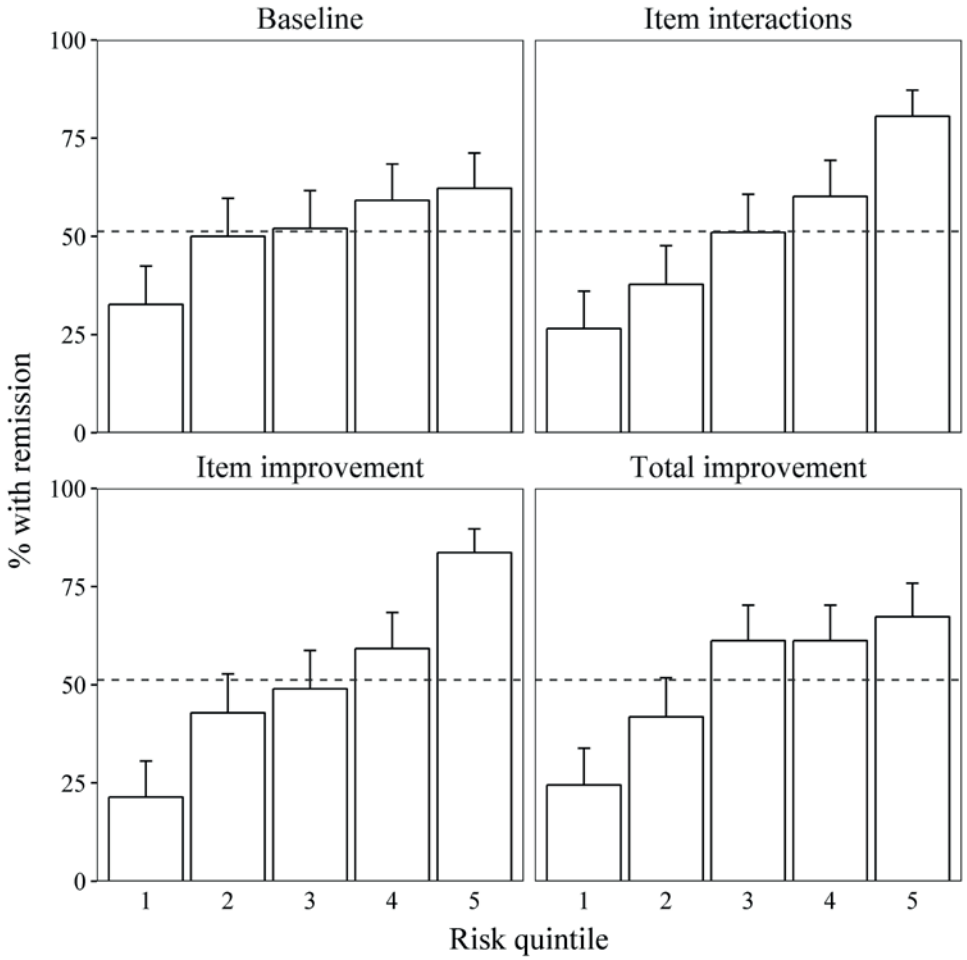

Figure 11.10: *Actual probability of remission at week 12 according to risk quantiles and model. For each model, participants' predicted probability of response was used to divide participants into quintiles of "risk", with the lowest quintile having the lowest predicted probability of response.*