

Original Article

Cite this article: Furukawa TA, Kato T, Shinagawa Y, Miki K, Fujita H, Tsujino N, Kondo M, Inagaki M, Yamada M (2019). Prediction of remission in pharmacotherapy of untreated major depression: development and validation of multivariable prediction models. *Psychological Medicine* **49**, 2405–2413. <https://doi.org/10.1017/S0033291718003331>

Received: 11 June 2018

Revised: 29 September 2018

Accepted: 15 October 2018

First published online: 15 November 2018

Key words:

Major depression; pharmacotherapy; prediction model; remission

Author for correspondence:

Toshi A. Furukawa, E-mail: furukawa@med.nagoya-cu.ac.jp

Prediction of remission in pharmacotherapy of untreated major depression: development and validation of multivariable prediction models

Toshi A. Furukawa¹, Tadashi Kato², Yoshihiro Shinagawa³, Kazuhira Miki⁴, Hirokazu Fujita⁵, Naohisa Tsujino⁶, Masaki Kondo¹, Masatoshi Inagaki⁷ and Mitsuhiko Yamada⁸

¹Department of Psychiatry and Cognitive-Behavioral Medicine, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan; ²Aratama Kokorono Clinic, Nagoya, Japan; ³Shiki Clinic, Nagoya, Japan; ⁴Miki Mental Clinic, Yokohama, Japan; ⁵Center to Promote Creativity in Medical Education, Kochi Medical School, Kochi University, Nankoku, Japan; ⁶Department of Neuropsychiatry, Toho University School of Medicine, Tokyo, Japan; ⁷Department of Psychiatry, Shimane University Faculty of Medicine, Izumo, Japan and ⁸Department of Neuropsychopharmacology, National Center of Neurology and Psychiatry, Tokyo, Japan

Abstract

Background. Depression is increasingly recognized as a chronic and relapsing disorder. However, an important minority of patients who start treatment for their major depressive episode recover to euthymia. It is clinically important to be able to predict such individuals. **Methods.** The study is a secondary analysis of a recently completed pragmatic megatrial examining first- and second-line treatments for hitherto untreated episodes of non-psychotic unipolar major depression ($n = 2011$). Using the first half of the cohort as the derivation set, we applied multiply-imputed stepwise logistic regression with backward selection to build a prediction model to predict remission, defined as scoring 4 or less on the Patient Health Questionnaire-9 at week 9. We used three successively richer sets of predictors at baseline only, up to week 1, and up to week 3. We examined the external validity of the derived prediction models with the second half of the cohort.

Results. In total, 37.0% (95% confidence interval 34.8–39.1%) were in remission at week 9. Only the models using data up to week 1 or 3 showed reasonable performance. Age, education, length of episode and depression severity remained in the multivariable prediction models. In the validation set, the discrimination of the prediction model was satisfactory with the area under the curve of 0.73 (0.70–0.77) and 0.82 (0.79–0.85), while the calibration was excellent with non-significant goodness-of-fit χ^2 values ($p = 0.41$ and $p = 0.29$), respectively.

Conclusions. Patients and clinicians can use these prediction models to estimate their predicted probability of achieving remission after acute antidepressant therapy.

Introduction

It has been increasingly recognized that major depressive disorder is more often than not a chronic and relapsing disorder (Furukawa *et al.*, 2000; Kanai *et al.*, 2003; Furukawa *et al.*, 2009). As a result there have been many attempts to determine the characteristics of the patients who do not respond to treatment (Bagby *et al.*, 2002).

However, in the real world, the illness course of major depression is highly variable and a substantial minority of the patients do show complete remission from a major depressive episode (Kessler *et al.*, 2017). It will help the practicing clinicians a great deal if they know the baseline demographic and clinical characteristics of such patients and if they can indeed discern such patients with satisfactory confidence at an early stage of the treatment. Remission to a completely euthymic state, rather than response and improvement of the depression severity, has now been proposed to be a desirable and achievable goal of the treatment of patients with major depression (Nierenberg and Wright, 1999; Keller, 2003; Nierenberg, 2013).

Studies of the course of major depression have identified, although often inconsistently, the following demographic, clinical or psychosocial predictors of poor response: older age, unemployment, low education, unmarried status, high baseline severity, longer duration of episode, greater number of previous episodes, younger age of onset, comorbid personality disorder, comorbid anxiety disorder, comorbid substance use disorder, poor social support and poor physical functioning among others (Bagby *et al.*, 2002; Kessler *et al.*, 2017). Observations early in the course of the treatment can also be informative: early improvement within 1–3 weeks of treatment has been found repeatedly to be a predictor of good outcomes (Katz *et al.*, 2004; Henkel *et al.*, 2009; Szegedi *et al.*, 2009; Tadic *et al.*, 2010).

© Cambridge University Press 2018. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is unaltered and is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use or in order to create a derivative work.

However, it is not known if any combination of these factors has enough discriminatory power to be used in clinical practices: even when each predictor is statistically significant, if the positive predictive value (PPV) of the positive predictions is, for example, 50% or even lower, then such a prediction model cannot be used in the clinical practice. Unfortunately, as far as the current authors are aware, there has been no study which has built, and examined the performance of, a prediction model of remission using appropriate psychometric methodology. There are a number of salient weaknesses in the available literature. First, most if not all studies suffer from a substantial loss to follow-up and these dropouts are often simply ignored in the complete case analyses or handled inappropriately with the last-observation-carried-forward method (Little and Rubin, 2002). It is very important in the studies of disease prognosis to limit the loss to follow-up as much as possible and, in the case of unavoidable missing data, to use appropriate imputation methods such as multiple imputation (MI) (Sterne *et al.*, 2009). Second, science of prediction has seen much advance and refinement in the past decade so that we now have a consensus methodology to appropriately design the study, collect the data, analyse the dataset and report the results (Collins *et al.*, 2015a; Debray *et al.*, 2017). We now have growing consensus that the model must be developed using the multivariable analyses and that it must be examined for external validation. Such properly developed prediction models are expected to play greater roles in informing decision making at various stages in the clinical pathway (Rabar *et al.*, 2012; Goff *et al.*, 2014).

We have conducted a pragmatic megatrial examining the first- and second-line treatments for untreated non-psychotic major depression that involved 2011 patients and followed them up to 25 weeks with the follow-up rate of 95.0%. This study is a secondary analysis of this dataset to delineate the demographic and clinical characteristics of the remitters to acute phase antidepressant treatment and to examine if and how we can predict them based on such variables. The prediction model will be built and examined using the recommended methodology.

Methods

Study and the participants

SUN☺D is a 25-week, multi-centre, parallel-group, assessor-blinded, pragmatic megatrial. The details of the study procedure and the results are reported elsewhere (Furukawa *et al.*, 2011; Kato *et al.*, 2018). In brief, it involved two randomizations: the first was a cluster-randomization by site at week 1 between the initial strategy to titrate the first-line treatment with sertraline up to the minimum or the maximum of the licensed dosage. The second was an individual randomization to allocate the participants who had not remitted by week 3 to continue sertraline, to augment it with mirtazapine, or to switch it to mirtazapine. The primary outcome was the score of the Patient Health Questionnaire-9 (PHQ-9) at week 9.

This study is a secondary analysis of the course of the patients participating in the SUN☺D pragmatic trial. In this study, we focus on those who show complete remission after the acute phase treatment.

Participants were eligible when (i) they suffered from non-psychotic unipolar major depression according to DSM-IV in the past month as ascertained by the clinician with the use of the semi-structured interview, Primary Care Evaluation of Mental Disorders (PRIME-MD) (Spitzer *et al.*, 1994), (ii) of either

sex, aged between 25 and 75, (iii) had not been treated with antidepressants, antipsychotics or mood stabilizers in the past month, (iv) were deemed suitable to start the treatment with sertraline by the clinician, and (v) had provided informed consent. Exclusion criteria included comorbidity with psychotic disorders, personality disorders and substance dependence. More details of the inclusion as well as exclusion criteria are provided in the protocol (Furukawa *et al.*, 2011).

Interventions

The first-line treatment consisted of sertraline started with 25 mg/day, then titrated to either 50 or 100 mg/day, according to the cluster randomization by site, by week 3. Those who remitted by week 3 (defined as scoring 4 or less on PHQ-9 at week 3) continued with their allocated first-line treatment. Those who had not remitted were randomized 1:1:1 to continue sertraline, to augment sertraline with mirtazapine, or to switch to mirtazapine at week 3. These second-line treatments were continued up to week 9. After week 9, the treatment was at discretion by the physicians and the final assessment was made at week 25.

Co-administration of non-protocol antidepressants, antipsychotics or mood stabilizers was prohibited up to week 9; anxiolytics and hypnotics were permitted. After week 9, the treatments were at the study physicians' discretion and there were no prohibited treatments.

Assessments

Before entry to the study, the physicians gathered information about the baseline demographic as well as clinical characteristics of the patients. After entry into the study, trained interviewers assessed the participants with the PHQ-9 and Frequency, Intensity, and Burden of Side Effects Rating (FIBSER) by telephone at weeks 1, 3, 9 and 25. The inter-rater reliability of the assessors as well as the success of blinding of the assessors have been ascertained (Shimodera *et al.*, 2012; Kato *et al.*, 2018).

Patient Health Questionnaire-9

PHQ-9 consists of the nine diagnostic criteria items of a major depressive episode of the DSM-IV. Each item is rated between 0 = 'Not at all' through 3 = 'Nearly every day', and the total score ranges between 0 and 27. The scores are interpreted clinically (Kroenke *et al.*, 2001) as indicating

- 0–4: No depression
- 5–9: Mild depression
- 10–14: Moderate depression
- 15–19: Moderately severe depression
- 20–: Severe depression

Good reliability, validity as well as sensitivity to change have been documented (Furukawa, 2010).

Beck Depression Inventory, Second edition

The participants were also asked to fill in the Beck Depression Inventory, second edition (BDI-II) on a bi-weekly basis when they visited the clinicians. BDI-II is a 21-item self-report measure of depression severity. The total score ranges between 0 and 63. Two subscales based on 'cognitive' and 'non-cognitive' factors have been proposed (Beck *et al.*, 1996). Excellent reliability,

validity as well as sensitivity to change have been reported (Furukawa, 2010).

Frequency, Intensity, and Burden of Side Effects Rating

FIBSER was originally used in a large NIMH-funded depression trial as a global rating scale for side effects which assesses the frequency, intensity and burden of side effects, each on a seven-point scale between 1 and 7. The total score therefore ranges between 3 and 21, with higher ratings indicating greater severity (Rush *et al.*, 2006).

Adherence

Adherence was measured as the number of days that the patient reported having taken the study medication.

Statistical analyses

We defined remitters as scoring 4 or less on PHQ-9, which was the primary outcome measure in the original megatrial. We first compared remitters *v.* non-remitters at week 9 with regard to the baseline demographic as well as clinical characteristics. Missing data were imputed by way of MI, using chained equations under the assumption that data were missing at random. Fifty multiply imputed datasets were created, using sex, age, education, employment, marital status, age of onset for depression, number of depressive episodes, length of index episode, PHQ-9, BDI-II, BDI-II subscales (Beck *et al.*, 1996), FIBSER and adherence as predictors. Rubin's rules were used to pool the regression coefficient estimates from the imputed datasets (Rubin, 1987). The association was expressed as odds ratio (OR) and its 95% confidence intervals (CIs).

Secondly, we examined whether we could predict remitters at week 9 by entering all the predictors into one model. As predictors, we used three successively richer sets, namely (i) all the demographic and clinical variables at baseline as listed above, (ii) plus the clinical variables by week 1 and treatment allocation at week 1, and (iii) plus the clinical variables by week 3 and treatment allocation at week 3. We then used the manual MI-stepwise logistic regression with backward selection method with *p* to leave set at 0.10 (Wood *et al.*, 2008; Chen and Wang, 2013), while also considering the clinical importance, clinical convenience and collinearity. We calculated variance inflation factors (VIFs) in order to ascertain that the obtained models did not suffer from multi-collinearity.

In order to avoid overfitting the data to the sample and to ascertain the external validity of the prediction model thus obtained, we split the sample by the median date of enrolment (temporal validation) (Collins *et al.*, 2015a). We used the first half of the total cohort as the derivation set to build the prediction model, and then examined its prediction performance on the second half of the cohort as a validation set.

Because the clinical focus of prediction was to see if the screening-positive population would eventually turn out to be true remitters, in building and assessing the prediction model, we tried to maximize the PPV, while not unduly sacrificing the total number of screening-positive population (assessed by the sensitivity of the prediction model) or the overall discrimination [assessed by the area under the receiver-operating characteristics curve (AUC)] and calibration (assessed by calibration plots and goodness-of-fit statistics).

In the validation sample, we examined AUC, goodness-of-fit statistics, calibration plots, PPV, negative predictive value (NPV),

sensitivity and specificity of the prediction model against the remitters at week 9 as well as those at week 25.

We conducted all statistical analyses with STATA Version 15.1 (College Station, TX, USA).

Results

Participants, interventions and assessments

Figure 1 shows the screening, randomization and follow-up of the study participants. Between December 2010 and March 2015, 56261 first-visit patients to the participating 48 clinics and hospitals in Japan underwent eligibility assessment, of whom 7895 suffered from untreated unipolar major depressive episodes. Of these, 2011 patients satisfied eligibility criteria, provided informed consent and were enrolled into SUN☺D.

At week 1, 970 participants were allocated to the 50 mg/day and 1041 to the 100 mg/day arms by cluster randomization. In the 50 mg/day arm, 91.7% had been prescribed 50 mg/day, 0.1% 37.5 mg/day, 1.3% 25 mg/day and 0.1% 75 mg/day by week 3; in the 100 mg/day arm, 82.0% had reached 100 mg/day, 5.3% 75 mg/day, 6.7% 50 mg/day and 0.9% 25 mg/day. In the 50 mg/day arm, 6.8% had stopped treatment as had 5.1% in the 100 mg/day arm.

Of all enrolled patients, 1953 (97.1%) completed telephone assessment at week 3, at which point 230 had remitted and continued on their allocated sertraline dose. Of those who had not remitted, 551 were randomized to continue sertraline ($n = 551$), augment sertraline with mirtazapine ($n = 538$) or switch to mirtazapine ($n = 558$). Of the initial cohort randomized at week 1, 1927 (95.8%) and 1910 (95.0%) were successfully followed-up at weeks 9 and 25, respectively.

Univariable predictors

In total, 37.0% (95% CI 34.8–39.1%) of the original cohort were remitted at week 9.

Table 1 shows the ORs for the association between the baseline predictors and the remission status at week 9. Older age, longer education, married status, older age at onset, shorter length of index episode as well as lower depression severity at weeks 0, 1 and 3 and less adverse effects at weeks 1 and 3 were significantly associated with remission.

Prediction models in the derivation set

We next constructed prediction models applying MI-stepwise logistic regression to the derivation set ($n = 1009$).

The prediction model based on the baseline, week 0 data only did not perform satisfactorily (Table 2a). The model included age, education, length of index episode and depression severity at baseline but the overall AUC was only 0.69. Even when the cut-off post-test probability for positive prediction was set at 0.70, PPV was 0.67; moreover, only 1% of the true remitters at week 9 were predicted to be so at baseline.

The prediction model based on data up to week 1 performed better (Table 2b). The final model included age, education, length of index episode, depression severity at baseline and at week 1, and the total burden of side effects at week 1. The overall AUC was 0.75, and PPV reached 0.80 when the cut-off post-test probability was set at 0.70. However, it was possible to identify only 16% of the true remitters as such.

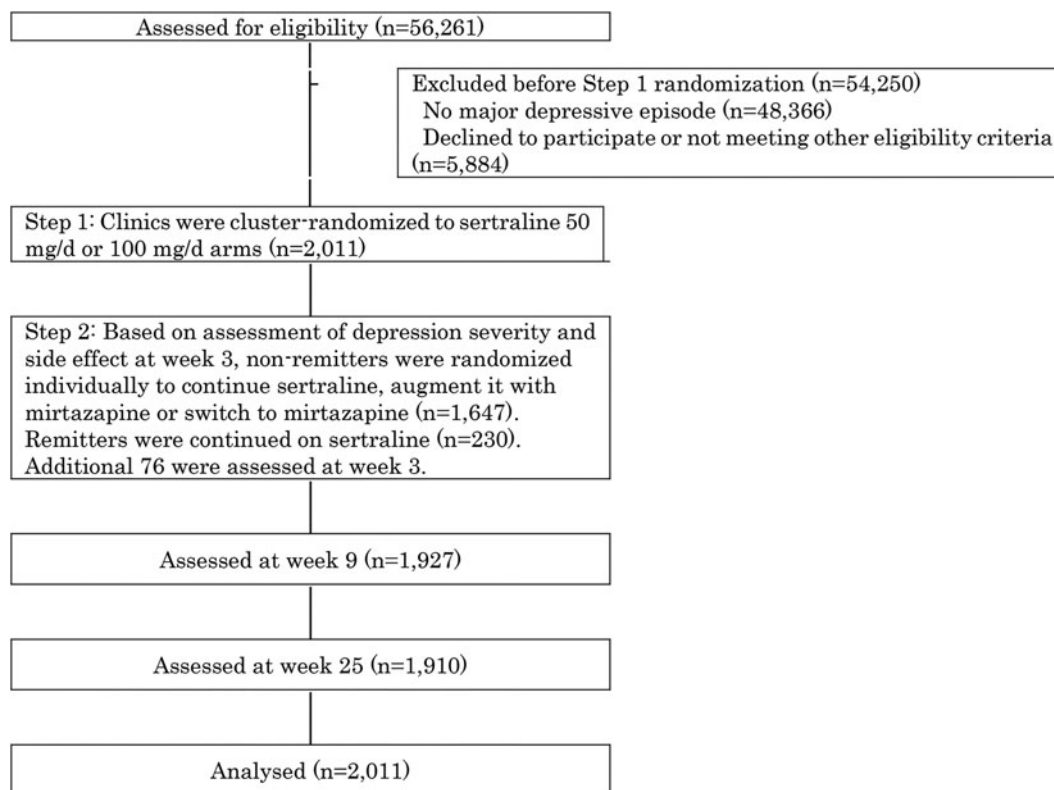


Fig. 1. Participants flow.

When we included data up to week 3, the prediction performance improved further (Table 2c). The final model included age, education, length of index episode as before but now only depression severity measures at week 3. The AUC was now 0.85; at the cut-off post-test probability of 0.70, PPV was 0.83, allowing 40% of the final remitters to be identified.

None of the VIFs in the three models was >5, suggesting that the obtained models did not present with a problem in multi-collinearity.

External validation of the prediction model in the validation set

Only the models using data up to week 1 or to week 3 were tested for external validity (Table 3). When these two prediction models were applied to the validation set, discrimination was still satisfactory, with AUC of 0.73 (95% CI 0.70–0.77) and 0.82 (0.79–0.85) for the models up to week 1 and up to week 3, respectively. Figure 2 shows calibration plots for the two models: the predicted and the observed matched closely, with no statistically significant Hosmer–Lemeshow statistics ($p = 0.41$ and $p = 0.29$, respectively) for this large validation set ($n = 1002$).

Setting the threshold for positive prediction at 0.70, the models showed similar performance to predict remitters as in the derivation set. Using the data up to week 1, PPV (i.e. proportion of true remitters among the positively predicted) remained at 0.74 (0.64–0.83); using the data up to week 3, PPV improved to 0.83 (0.76–0.88). The sensitivity (i.e. proportion of positive prediction among the true remitters) was 0.17 (0.13–0.21) and 0.36 (0.31–0.42).

The models can be used to predict non-remitters as well. At the cut-off post-test probability of 0.30, the model with data up

to week 1 showed NPV (i.e. proportion of true non-remitters among the negatively predicted) of 0.81 (0.77–0.85); that with data up to week 3 NPV of 0.84 (0.81–0.87). The specificity (i.e. proportion of negative prediction among the true non-remitters) was 0.55 (0.51–0.59) and 0.68 (0.65–0.72).

It is interesting to note that, using the same prediction model, we can predict the remitters at week 25 (remission rate: 51.8%, 95% CI 49.5–54.0%) with similar accuracy, with AUC of 0.69 (0.66–0.72) and 0.75 (0.72–0.78) for the models up to week 1 and up to week 3, respectively. The PPV for remission at week 25 based on the data up to week 1 was 0.87, and that based on data up to week 3 was 0.86. In other words, if the predictions based on age, education, length of episode and depression severity were positive, we can be fairly confident that such patients would remit by week 9 or, at least eventually, by week 25. The calibration for predicting week 25 remission was poor, mainly because more participants reached remission at week 25 than predicted by the models predicting remission at week 9 (eTable 1 and eFigure 1 in the online Supplementary material).

The prediction models

The final prediction models based on data up to week 1 and on data up to week 3 were as follows:

- logit by week 1 data = $-0.841 + 0.059 \times \text{PHQ9 at week 0} + 0.028 \times \text{age} + 0.087 \times \text{education (years)} - 0.045 \times \text{length of episode (months)} - 0.076 \times \text{PHQ9 at week 1} - 0.056 \times \text{BDI2 at week 1} - 0.056 \times \text{FIBSER at week 1}$
- logit by week 3 data = $0.343 + 0.029 \times \text{age} + 0.080 \times \text{education (years)} - 0.037 \times \text{length of episode (months)} - 0.176 \times \text{PHQ9 at week 3} - 0.061 \times \text{BDI2 at week 3}$

Table 1. Univariate prediction of complete remission at week 9

Variable	Complete remitters at week 9 (n = 717)	Non-remitters at week 9 (n = 1210)	OR
<i>Demographic and clinical characteristics</i>			
Sex (men %)	48% (44–52%)	46% (44–49%)	0.93 (0.78–1.12, <i>p</i> = 0.47)
Age (years)	44.8 (43.8–45.7)	41.1 (40.4–41.7)	1.29 (1.19–1.39, <i>p</i> < 0.001) for every 10-year increase in age
Education (years)	14.2 (14.0–14.4)	13.8 (13.7–14.0)	1.33 (1.13–1.56, <i>p</i> < 0.001) for every 4-year increase in education
Employment (employed)	62% (59–66%)	58% (56–61%)	1.17 (0.97–1.41, <i>p</i> = 0.11)
Marital status (married)	60% (56–64%)	49% (46–52%)	1.58 (1.31–1.90, <i>p</i> < 0.001)
Age of onset for depression	40.3 (39.2–41.3)	36.3 (35.6–37.1)	1.25 (1.16–1.34, <i>p</i> < 0.001) for every 10-year increase in age of onset
No of previous depressive episodes	2.3 (2.0–2.6)	2.3 (2.1–2.4)	1.00 (0.98–1.03, <i>p</i> = 0.76)
Length of index episode (months)	4.2 (3.6–4.8)	7.1 (6.1–8.0)	0.97 (0.96–0.99, <i>p</i> < 0.001) for every 1-month increase in length
<i>Baseline</i>			
PHQ-9	17.7 (17.5–18.0)	18.9 (18.6–19.1)	0.93 (0.91–0.95, <i>p</i> < 0.001) for every one-point increase in PHQ-9
BDI-II	28.2 (27.5–28.9)	33.4 (32.9–34.0)	0.94 (0.93–0.95, <i>p</i> < 0.001) for every one-point increase in BDI-II
<i>Week 1</i>			
PHQ-9	13.1 (12.7–13.4)	16.6 (16.4–16.9)	0.87 (0.85–0.88, <i>p</i> < 0.001)
BDI-II	22.6 (21.9–23.4)	30.4 (29.8–31.0)	0.93 (0.92–0.94, <i>p</i> < 0.001)
FIBSER	6.1 (5.9–6.4)	7.0 (6.7–7.2)	0.95 (0.92–0.97, <i>p</i> < 0.001)
Adherence (days/week)	6.0 (5.9–6.1)	6.1 (6.0–6.2)	0.97 (0.91–1.03, <i>p</i> = 0.29)
<i>Week 3</i>			
PHQ-9	7.6 (7.2–7.)	13.8 (13.5–14.1)	0.78 (0.76–0.80, <i>p</i> < 0.001)
BDI-II	15.1 (14.4–15.7)	26.7 (26.1–27.3)	0.88 (0.87–0.89, <i>p</i> < 0.001)
FIBSER	6.0 (5.8–6.3)	7.2 (6.9–7.4)	0.93 (0.91–0.95, <i>p</i> < 0.001)
Adherence (days/week)	6.5 (6.4–6.6)	6.4 (6.4–6.5)	1.02 (0.95–1.10, <i>p</i> = 0.54)

Numbers are mean (95% confidence interval) or percentage (95% confidence interval).

BDI-II, Beck Depression Inventory, second edition; FIBSER, Frequency, Intensity, and Burden of Side Effects Rating; PHQ-9, Patient Health Questionnaire-9.

where post-test probability is obtained by $\exp(\text{logit})/[1 + \exp(\text{logit})]$. The Excel spreadsheet to calculate the post-test probability is provided on our department homepage at <http://ebmh.med.kyoto-u.ac.jp/toolbox.html> and also attached to this article as an electronic supplement.

Discussion

The biggest inception cohort to date to study the outcome of patients undergoing antidepressant therapy for an untreated episode of major depression revealed that, if we include observations up to week 1 or 3 after commencement of therapy, we can have reasonably satisfactory and usable prediction models to predict remission at the end of acute phase treatment. The same models were able to predict remission after 25 weeks as well.

Older age, higher education, married status, shorter duration of episode, older age of onset and milder initial depression severity were associated with remission in our univariable analyses. Older age has often been associated with poorer prognosis

(Bagby *et al.*, 2002; Kessler *et al.*, 2017); in our cohort of patients with major depression without comorbidities, older age predicted better prognosis. When built into a multivariable prediction model, age, education and length of episode along with depression severity emerged as independent predictors. In other words, marital status and age of onset may have been confounded by these factors.

The performance of the prediction models improved when we included depression severity in the early course of treatment. The added predictive value of depression severity in the first 1–3 weeks of treatment is in line with the literature (Katz *et al.*, 2004; Henkel *et al.*, 2009; Szegedi *et al.*, 2009; Tadic *et al.*, 2010). Indeed only the models incorporating data up to week 1 or 3 demonstrated satisfactory performance in the development set. In the external validation set, the discrimination of these models was good with AUC between 0.73 and 0.82 and the calibration was excellent as shown in the calibration plots (Fig. 2). When the model prediction is positive after 1–3 weeks of initial treatment, one can be 70–80% sure that the patient would remit within 9 or, at least,

Table 2. Final prediction models using the derivation set ($n = 1009$)

Predictor	OR	<i>p</i> value	VIF
(a) Based on the baseline data only			
Baseline PHQ-9	0.95 (0.92–0.98)	0.004	1.01
Age	1.04 (1.03–1.05)	<0.001	1.10
Education (years)	1.11 (1.04–1.18)	0.001	1.10
Length of index episode (months)	0.96 (0.93–0.98)	<0.001	1.01
Constant	0.09 (0.02–0.32)	<0.001	

AUC = 0.66, PPV = 0.67 at cut-off post-test probability of 0.70 with sensitivity of 0.01 (i.e. with 1% of the final remitters correctly identified).

Predictor	OR	<i>p</i> value	VIF
(b) Using data up to week 1			
Baseline PHQ-9	1.06 (1.02–1.11)	0.006	1.50
Age	1.03 (1.02–1.04)	<0.001	1.15
Education (years)	1.09 (1.02–1.17)	0.010	1.12
Length of index episode (months)	0.96 (0.93–0.98)	<0.001	1.02
PHQ-9 at week 1	0.95 (0.91–0.98)	0.007	1.07
BDI-II at week 1	0.93 (0.89–0.97)	0.001	2.77
FIBSER at week 1	0.95 (0.93–0.96)	<0.001	2.38
Constant	0.42 (0.10–1.85)	0.258	

AUC = 0.75, PPV = 0.80 at cut-off post-test probability of 0.70 with sensitivity of 0.16 (i.e. 16% of the final remitters correctly identified).

Predictor	OR	<i>p</i> value	VIF
(c) Using data up to week 3			
Age	1.03 (1.01–1.04)	<0.001	1.13
Education (years)	1.08 (1.00–1.17)	0.037	1.11
Length of index episode (months)	0.96 (0.94–0.99)	0.003	1.03
PHQ-9 at week 3	0.84 (0.80–0.88)	<0.001	3.09
BDI-II at week 3	0.94 (0.92–0.97)	<0.001	3.12
Constant	1.41 (0.33–6.10)	0.647	

AUC = 0.85, PPV = 0.83 at cut-off post-test probability of 0.70 with sensitivity of 0.40 (i.e. with 40% of the final remitters correctly identified).

AUC, area under the curve; BDI-II, Beck Depression Inventory, second edition; FIBSER, Frequency, Intensity, and Burden of Side Effects Rating; PHQ-9, Patient Health Questionnaire-9; PPV, positive predictive value; VIF, variance inflation factor (95% CI in parentheses).

by 25 weeks. Such information will be very encouraging both for the patients and the clinician in the actual practices.

The treatments did not emerge as strong predictors. In the model using data up to week 1, when patients were randomized to either 50 mg or 100 mg/day of sertraline, the treatment allocation did not emerge as a significant predictor. This finding is in line with

the results of the original randomized controlled trial (RCT), which found that there was no difference in PHQ-9 scores at week 9 between these two arms (Kato *et al.*, 2018). In the model using data up to week 3, when non-remitted patients were randomized to continue sertraline, augment it with mirtazapine or switch to mirtazapine, the original RCT found small but statistically significant superiority of the augmentation or switching strategies over the continuation in terms of the PHQ-9 scores at week 9 among the non-remitters (Kato *et al.*, 2018). In the current analyses, augmentation or switching emerged as significant predictors in initial steps of variable selection: however, when PHQ-9 scores at week 3 were included, they were no longer statistically significant. In other words, PHQ-9 at week 3 was a stronger predictor of remission at week 9 over changing the treatments among the non-remitters.

The study has some limitations. First, although the model showed good overall discrimination and satisfactory PPV when the post-test threshold of positive prediction was set at 0.70, it was only able to identify a minority (30–40%) of the actual remitters. This limitation is well illustrated by Fig. 2: the probability of accurate prediction is high to the right of the 7th decile; however, there are always patients who are less likely to remit but still do remit to the left of the 7th decile. The users of the model using this threshold need to be aware that there are patients who still remit even when they are negatively predicted below this threshold. Second, we were unable to examine variables that were used as exclusion criteria or not measured originally in the SUN☺D trial. Among such were personality disorders, substance use disorders, anxiety disorders, social support and social functioning. The prediction performance may have improved had we measured these variables at the baseline. However, the set of variables in the current study represents the minimum set clinicians would be measuring in daily practices and serve to indicate which variables to look for in the case of non-complicated major depression. Third, the findings would apply to chronic or non-chronic drug-naïve patients without significant comorbidities, and possibly not to treatment-refractory populations or in the context of salient psychiatric or physical comorbidities. We need further research to build prediction models for such difficult-to-treat depression and examine if similar variables would be at play. Fourth, it is not known whether the obtained models will be applicable when treatments other than the ones used in this megatrial are administered. The final models did not include treatment variables. However, when different drugs and different therapies are used, including psychotherapies or physical therapies, different factors might emerge as important predictors. Finally, although using the latter half of the sample as a validation set is considered a form of external validation (Collins *et al.*, 2015b), the validity coefficients thus obtained could have been higher than using a dataset from completely new settings. Performance of the obtained models need to be assessed with further validation samples from different settings and broader types of participants.

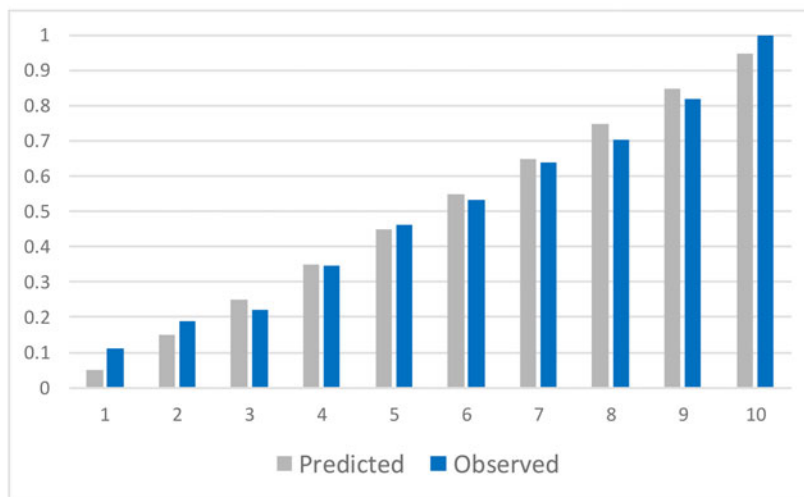
However, this study possesses several unique strengths. This is the largest cohort of patients with hitherto untreated episodes of major depression, treated with step-wise antidepressant pharmacotherapy. The participants were recruited in 48 clinics and hospitals across Japan. The dropout rates were <5% up to week 25, and the appropriate imputation method was applied for the missing data. The sample size allowed two datasets, each comprising approximately 1000 patients, one for derivation of the models and another for external validation of the models. The development of the prediction models followed the most recent guideline and used the one-step multivariable procedure (Collins *et al.*, 2015a). The

Table 3. Predicting remission at week 9 in the validation set (*n* = 1002)

		Using data up to week 1	Using data up to week 3
AUC		0.73 (0.70–0.77)	0.82 (0.79–0.85)
Hosmer–Lemeshow statistics		$\chi^2_{df=10} = 10.34, p = 0.41$	$\chi^2_{df=10} = 11.96, p = 0.29$
Cut-off post-test prob = 0.70	PPV	0.74 (0.64–0.83)	0.83 (0.76–0.88)
	NPV	0.66 (0.63–0.69)	0.72 (0.68–0.75)
	Sensitivity	0.17 (0.13–0.21)	0.36 (0.31–0.42)
	Specificity	0.97 (0.95–0.98)	0.95 (0.94–0.97)
Cut-off post-test prob = 0.30	PPV	0.51 (0.47–0.53)	0.60 (0.55–0.64)
	NPV	0.81 (0.77–0.85)	0.84 (0.81–0.87)
	Sensitivity	0.79 (0.74–0.83)	0.79 (0.74–0.83)
	Specificity	0.55 (0.51–0.59)	0.68 (0.65–0.72)

AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value (95% CI in parentheses).

(a) Prediction model based on data up to week 1 (Hosmer-Lemeshow goodness-of-fit test *p*=0.41)



(b) Prediction model based on data up to week 3 (Hosmer-Lemeshow goodness-of-fit test *p*=0.29)

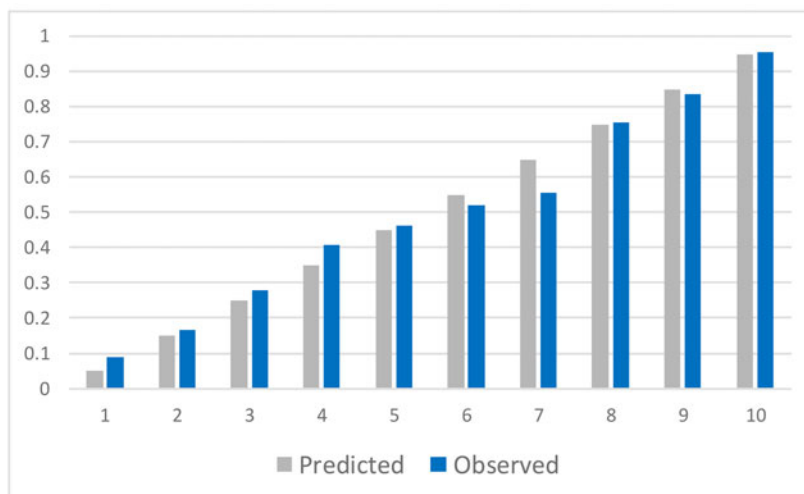



Fig. 2. Predicted v. observed by decile in the validation set.

performance of the obtained models in the validation set was satisfactory. The study focused on remission, which is clearly the most desirable outcome of acute phase depression treatment (Nierenberg and Wright, 1999; Keller, 2003; Nierenberg, 2013).

We have provided the whole prediction models as Excel spreadsheets as an online Supplementary material. Patients and clinicians can enter their age, education, length of episode, PHQ-9 and BDI-II scores to obtain predicted probabilities of achieving remission at week 9 and at week 25. We hope that clinically informed, judicious use of this tool will help the patients and clinicians make better informed decisions.

Author ORCIDs.  Toshi A. Furukawa 0000-0003-2159-3776

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/S0033291718003331>

Competing interests. TAF has received lecture fees from Meiji, Mitsubishi-Tanabe, MSD and Pfizer. He has received research support from Mitsubishi-Tanabe. TK has received lecture fees from Eli Lilly and Mitsubishi-Tanabe, and has contracted research with GSK, MSD and Mitsubishi-Tanabe. He has received royalties from Kyowa Yakuhin. YS has received lecture fees from Janssen, Kyowa-Yakuhin, Meiji, MSD, Otsuka and Mitsubishi-Tanabe. KM has received lecture fees from Eisai, GSK, Kyowa Yakuhin, Meiji, MSD, Otsuka, Pfizer, Eli Lilly, Mochida, Yoshitomi, Dainippon-Sumitomo, Takeda and Shionogi. HF has received lecture fees from Mochida and Tsumura. NT has received lecture fees from Astellas, Eisai, Shionogi, Novartis, Fujifilm RI Pharma, Meiji, Mochida, MSD, Janssen, Eli Lilly and Dainippon-Sumitomo. MK has received lecture fees from Yoshitomi and a research grant from Novartis. MI has received a grant from Novartis Pharma. He has received lecture fees from Meiji Seika Pharma, Mochida and Takeda. MY has contracted research with Nippon Chemipharm.

Funding. The study was funded by the Ministry of Health, Labor and Welfare, Japan (H-22-Seishin-Ippan-008) from April 2010 through March 2012 to TAF (<http://www.mhlw.go.jp/english/>), and thereafter by the Japan Foundation for Neuroscience and Mental Health (JFNMH) to TAF (<http://www.jfnm.or.jp/>). The JFNMH received donations from Asahi Kasei, Eli Lilly, GSK, Janssen, MSD, Meiji, Mochida, Otsuka, Pfizer, Shionogi, Taisho, and Mitsubishi-Tanabe. The study is partly supported by Japan Agency for Medical Research and Development (18dk0307072) and the Ministry of Health, Labour and Welfare (H29-ICT-Ippan-010). The funders of the study had no role in study design, data collection, data analysis, data interpretation or writing of the report.

Authors' contributions. TAF is the principal investigator and had overall responsibility for the management of the study. TAF had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analyses. TAF conceived and designed this study. TAF and MY obtained the funding. YS, KM, HF, NT, MK, MY, MI and TK acquired, analysed or interpreted the data. TAF conducted the statistical analyses. TAF drafted the manuscript, and YS, KM, HF, NT, MK, MY, MI and TK contributed critical revision of the manuscript. All authors contributed to and approved the final manuscript.

References

- Bagby RM, Ryder AG and Crispien C (2002) Psychosocial and clinical predictors of response to pharmacotherapy for depression. *Journal of Psychiatry and Neuroscience* 27, 250–257.
- Beck AT, Steer RA and Brown GK (1996) *BDI-II: Beck Depression Inventory*, 2nd Edn. Manual. San Antonio: The Psychological Corporation.
- Chen Q and Wang S (2013) Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine* 32, 3646–3659.
- Collins GS, Reitsma JB, Altman DG and Moons KG (2015a) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Medicine* 13, 1.
- Collins GS, Reitsma JB, Altman DG and Moons KG (2015b) Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Annals of Internal Medicine* 162, 55–63.
- Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, Riley RD and Moons KG (2017) A guide to systematic review and meta-analysis of prediction model performance. *British Medical Journal (Clinical Research Ed.)* 356, i6460.
- Furukawa TA (2010) Assessment of mood: guides for clinicians. *Journal of Psychosomatic Research* 68, 581–589.
- Furukawa T, Konno W, Morinobu S, Harai H, Kitamura T and Takahashi K (2000) Course and outcome of depressive episodes: comparison between bipolar, unipolar and subthreshold depression. *Psychiatry Research* 96, 211–220.
- Furukawa TA, Yoshimura R, Harai H, Imaizumi T, Takeuchi H, Kitamura T and Takahashi K (2009) How many well vs. unwell days can you expect over 10 years, once you become depressed? *Acta Psychiatrica Scandinavica* 119, 290–297.
- Furukawa TA, Akechi T, Shimodera S, Yamada M, Miki K, Watanabe N, Inagaki M and Yonemoto N (2011) Strategic use of new generation antidepressants for depression: SUN☺D study protocol. *Trials* 12, 116.
- Goff Jr. DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith Jr. SC, Sorlie P, Stone NJ, Wilson PW, Jordan HS, Nevo L, Wnek J, Anderson JL, Halperin JL, Albert NM, Bozkurt B, Brindis RG, Curtis LH, DeMets D, Hochman JS, Kovacs RJ, Ohman EM, Pressler SJ, Sellke FW, Shen WK, Smith Jr. SC, Tomaselli GF and American College of Cardiology/American Heart Association Task Force on Practice, G. (2014) 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 129, S49–S73.
- Henkel V, Seemuller F, Obermeier M, Adli M, Bauer M, Mundt C, Brieger P, Laux G, Bender W, Heuser I, Zeiler J, Gabel W, Mayr A, Moller HJ and Riedel M (2009) 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. *Journal of Affective Disorders* 115, 439–449.
- Kanai T, Takeuchi H, Furukawa TA, Yoshimura R, Imaizumi T, Kitamura T and Takahashi K (2003) Time to recurrence after recovery from major depressive episodes and its predictors. *Psychological Medicine* 33, 839–845.
- Kato T, Furukawa TA, Mantani A, Kurata K, Kubouchi H, Hirota S, Sato H, Sugishita K, Chino B, Itoh K, Ikeda Y, Shinagawa Y, Kondo M, Okamoto Y, Fujita H, Suga M, Yasumoto S, Tsujino N, Inoue T, Fujise N, Akechi T, Yamada M, Shimodera S, Watanabe N, Inagaki M, Miki K, Ogawa Y, Takeshima N, Hayasaka Y, Tajika A, Shinohara K, Yonemoto N, Tanaka S, Zhou Q, Guyatt GH and for the SUN☺D Investigators (2018) Optimising first- and second-line treatment strategies for untreated major depressive disorder – the SUND study: a pragmatic, multi-centre, assessor-blinded randomised controlled trial. *BMC Medicine* 16, 103.
- Katz MM, Tekell JL, Bowden CL, Brannan S, Houston JP, Berman N and Frazer A (2004) Onset and early behavioral effects of pharmacologically different antidepressants and placebo in depression. *Neuropsychopharmacology* 29, 566–579.
- Keller MB (2003) Past, present, and future directions for defining optimal treatment outcome in depression: remission and beyond. *JAMA* 289, 3152–3160.
- Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Ebert DD, de Jonge P, Nierenberg AA, Rosellini AJ, Sampson NA, Schoevers RA, Wilcox MA and Zaslavsky AM (2017) Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences* 26, 22–36.
- Kroenke K, Spitzer RL and Williams JB (2001) The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine* 16, 606–613.
- Little RJ and Rubin DB (2002) *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

- Nierenberg AA** (2013) Strategies for achieving full remission when first-line antidepressants are not enough. *Journal of Clinical Psychiatry* **74**, e26.
- Nierenberg AA and Wright EC** (1999) Evolution of remission as the new standard in the treatment of depression. *Journal of Clinical Psychiatry* **60** (Suppl 22), 7–11.
- Rabar S, Lau R, O'Flynn N, Li L, Barry P and Guideline Development, G** (2012) Risk assessment of fragility fractures: summary of NICE guidance. *British Medical Journal (Clinical Research Ed.)* **345**, e3698.
- Rubin DB** (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Rush AJ, Trivedi MH, Wisniewski SR, Stewart JW, Nierenberg AA, Thase ME, Ritz L, Biggs MM, Warden D, Luther JF, Shores-Wilson K, Niederehe G and Fava M** (2006) Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. *New England Journal of Medicine* **354**, 1231–1242.
- Shimodera S, Kato T, Sato H, Miki K, Shinagawa Y, Kondo M, Fujita H, Morokuma I, Ikeda Y, Akechi T, Watanabe N, Yamada M, Inagaki M, Yonemoto N and Furukawa TA** (2012) The first 100 patients in the SUN[©]D trial (strategic use of new generation antidepressants for depression): examination of feasibility and adherence during the pilot phase. *Trials* **13**, 80.
- Spitzer RL, Williams JB, Kroenke K, Linzer M, deGruy III FV, Hahn SR, Brody D and Johnson JG** (1994) Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. *JAMA* **272**, 1749–1756.
- Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM and Carpenter JR** (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal (Clinical Research Ed.)* **338**, b2393.
- Szegedi A, Jansen WT, van Willigenburg AP, van der Meulen E, Stassen HH and Thase ME** (2009) Early improvement in the first 2 weeks as a predictor of treatment outcome in patients with major depressive disorder: a meta-analysis including 6562 patients. *Journal of Clinical Psychiatry* **70**, 344–353.
- Tadic A, Helmreich I, Mergl R, Hautzinger M, Kohnen R, Henkel V and Hegerl U** (2010) Early improvement is a predictor of treatment outcome in patients with mild major, minor or subsyndromal depression. *Journal of Affective Disorders* **120**, 86–93.
- Wood AM, White IR and Royston P** (2008) How should variable selection be performed with multiply imputed data? *Statistics in Medicine* **27**, 3227–3246.