Accepted author's manuscript. Published in final edited form as: Journal of Affective Disorders 2020; 274: 690-697. Publisher DOI: 10.1016/j.jad.2020.05.141

5131 words
4 tables
1 figure

Can personalized treatment prediction improve the outcomes, compared with the group average approach, in a randomized trial? Developing and validating a multivariable prediction model in a pragmatic megatrial of acute treatment for major depression

Toshi A. Furukawa, MD, PhD*
Departments of Health Promotion and Human Behavior and of Clinical Epidemiology, Kyoto University Graduate School of Medicine / School of Public Health, Kyoto, Japan
furukawa@kuhp.kyoto-u.ac.jp

Thomas Debray, PhD
Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, The Netherlands,
T.Debray@umcutrecht.nl

Tatsuo Akechi, MD, PhD
Department of Psychiatry and Cognitive-Behavioral Medicine, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan
takechi@med.nagoya-cu.ac.jp

Mitsuhiko Yamada, MD, PhD
Department of Neuropsychopharmacology, National Institute of Mental Health, National Center of Neurology and Psychiatry, Tokyo, Japan
mitsuhiko_yamada@ncnp.go.jp

Tadashi Kato, MD
Aratama Kokorono Clinic, Nagoya, Japan
aratama8177@yahoo.co.jp

Michael Seo, MSc
Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland
swj8874@gmail.com

Orestis Efthimiou, PhD
Institute of Social and Preventive Medicine, University of Bern, Switzerland
oremiou@gmail.com

* Correspondence
Departments of Health Promotion and Human Behavior and of Clinical Epidemiology, Kyoto University Graduate School of Medicine / School of Public Health
Yoshida Konoe-cho, Sakyo-ku, Kyoto, Japan
Tel: +81-75-753-9491, Email: furukawa@kuhp.kyoto-u.ac.jp

ABSTRACT

Background: Clinical trials have traditionally been analysed at the aggregate level, assuming that the group average would be applicable to all eligible and similar patients. We re-analyzed a mega-trial of antidepressant therapy for major depression to explore whether a multivariable prediction model may lead to different treatment recommendations for individual participants.

Methods: The trial compared the second-line treatment strategies of continuing sertraline, combining it with mirtazapine or switching to mirtazapine after initial failure to remit on sertraline among 1,544 patients with major depression. The outcome was the Personal Health Questionnaire-9 (PHQ-9) at week 9: the original analyses showed that both combining and switching resulted in greater reduction in PHQ-9 by 1.0 point than continuing. We considered several models of penalized regression or machine learning.

Results: Models using support vector machines (SVMs) provided the best performance. Using SVMs, continuing sertraline was predicted to be the best treatment for 123 patients, combining for 696 patients, and switching for 725 patients. In the last two subgroups, both combining and switching were equally superior to continuing by 1.2 to 1.4 points, resulting in the same treatment recommendations as with the original aggregate data level analyses; in the first subgroup, however, switching was substantively inferior to combining (-3.1, 95%CI: -5.4 to -0.5).

Limitations: Stronger predictors are needed to make more precise predictions.

Conclusions: The multivariable prediction models led to improved recommendations for a minority of participants than the group average approach in a megatrial.

Keywords:

Depressive disorder; Antidepressive agents; Precision medicine; Prediction model

1 INTRODUCTION

Personalized medicine or precision medicine aims at finding the right treatment for the right person at the right time, based on each patient's unique clinical, genetic and environmental characteristics (Wium-Andersen et al., 2017). It is expected to have an important role in the future, or even replace the modus operandi of today's psychiatry, in which we can hope to find the more effective and less harmful treatment for a particular patient only through trials-and-errors (Cuijpers and Christensen, 2017).

Despite this promise, there are to date only a few applications of personalized medicine in psychiatry. The few that exist are limited to pharmacogenetic test kits (e.g. FDA-approved AmpliChip CYP450 or GeneSight), whose cost-effectiveness is debated (Perna et al., 2018). Even less evidence exists on whether any of the available algorithms for personalized medicine results in more benefit and/or less harm for the whole population to whom such algorithms can be applied than the current trial-and-error approach (Cuijpers and Christensen, 2017; Kent et al., 2018).

The quality and potential impact of personalized medicine may be improved in various ways. First, access to larger samples can improve the development and validation of personalized prediction models. In this era of big data, individual participant data from randomized controlled trials (RCTs) (Institute of Medicine Committee on Strategies for Responsible Sharing of Clinical Trial Data, 2015), large observational studies (Kessler et al., 2019), electronic healthcare registries and wearable devices (Debray et al., 2015a) (Perna et al., 2018) are often available. The resulting datasets can then be utilized to develop more accurate personalized prediction models and to obtain more detailed insight into their performance across different settings and populations (Riley et al., 2016).

Secondly, the application of more flexible modelling methods may help to better account for the complex nature of treatment-related prognosis in psychiatry. Modern methods of analysis (often termed "machine learning") in particular are capable of dealing with non-linear effects, treatment-covariate interactions and other complexities without the need to pre-specify these a priori (Kessler et al., 2019).

In this study, we sought to develop a prediction model for individualizing the choice of antidepressants for patients with major depression who have not remitted on their first line treatment. Our data came from a previously conducted, pragmatic megatrial of first- and second-line treatments for hitherto untreated episodes of major depression, which had a very large sample size (n=2,011) and a very high follow-up rate (95.8% for the primary endpoint at 9 weeks) (Furukawa et al., 2011; Kato et al., 2018). Megatrials offer the additional advantage of providing a harmonized set of covariates, in contrast to the usual situation found in big data assembled from separate RCTs or observational studies (Noma et al., 2019). In developing our personalized treatment model, we explored a wide range of statistical methods as well as machine learning approaches. After developing the model, we examined whether

its use for personalizing treatment can lead to improved patient outcomes, compared with the usual group-average, arm-level interpretation of RCT results. Finally we provide a web app to predict depression severity from a large set of baseline predictors using the model developed and validated in this study.

2 METHODS

The following report was prepared in accordance with the TRIPOD statement (Collins et al., 2015).

2.1 Design and major findings of the SUN ☺ D trial

SUN ☺ D (Strategic Use of New generation anti-depressants in Depression) was a 25-week pragmatic, multi-centre, assessor-blinded trial which involved two randomisations, aiming to examine the first- and second-line treatment strategies for hitherto untreated unipolar major depressive episodes. The study took place in 48 psychiatric clinics and hospitals across Japan between December 2010 and September 2015. The study protocol (Furukawa et al., 2011) and the results of the main analyses (Kato et al., 2018) have been published elsewhere.

Briefly, the main eligibility criteria for entry into the SUN ☺ D trial were as follows: men and women aged between 25 and 75 years, with a primary diagnosis of non-psychotic unipolar major depressive episode according to DSM-IV, and having received no antidepressant, antipsychotic or mood stabiliser in the previous month. At Step 1 all participants received sertraline, and were randomized to titration up to the minimum or the maximum of its licensed dose range, i.e. 50-100 mg in Japan. At Step 2, the participants who had not achieved remission by week 3 (defined as scoring 4 or less on the Patient Health Questionnaire-9 (PHQ-9)) were randomized to continue sertraline, combine it with mirtazapine, or to switch to mirtazapine. The primary outcomes were measured at week 9, after which the treatments were at the discretion of the physicians and the patients. The final follow-up was at week 25.

With regard to the Step 1 randomisation, analyses provided no evidence of difference between the 50 mg/day or 100 mg/day arms in efficacy or side effects at weeks 9 or 25. With regard to the Step 2 randomisation, adding mirtazapine achieved a greater reduction in PHQ-9 scores by 1.0 points (95% confidence interval [CI]: 0.4 to 1.6) and switching by 1.0 points (0.5 to 1.6) at week 9, both in comparison with continuing sertraline. The superior efficacy of the combination and the switch arm disappeared by week 25.

2.2 Intervention alternatives to be analysed in this study

In this analysis we focused on the participants in Step 2 randomisation of the SUN☺D trial, i.e. those who had entered the trial at week 0, received sertraline 50 or 100 mg/day for three weeks but did not remit by week 3.

The three intervention arms at Step 2 randomisation were as follows:

1) In the continuing sertraline arm, sertraline was administered as 50 or 100 mg/day according to Step 1 randomisation, up to week 9.
2) In the combination arm, sertraline was continued as above, and mirtazapine was added between 7.5 to 45 mg/day.
3) In the switching arm, mirtazapine between 7.5 and 45 mg/day was administered, while sertraline was gradually tapered off by week 7.

2.3 Primary and secondary outcomes in this study

The primary outcome of this study is PHQ-9 (Kroenke *et al.*, 2001) at week 9. PHQ-9 measures the severity of depression by rating each of the nine diagnostic criterion symptoms in four grades between 0='Not at all' and 3='Nearly every day' for the past two weeks. The total score therefore ranges between 0 and 27.

As a secondary outcome we also used PHQ-9 at week 25. PHQ-9 was rated by central telephone assessors blinded to the allocated treatments, whose reliability had been established (Shimodera et al., 2012).

2.4 Predictors

We included all the following baseline predictor variables.

1) Socio-demographic variables: age, sex, education (years), employment status (full time/part-time/on sick leave/housewife/student/retired/not employed), and marital status (single (never married)/divorced/widowed/married)
2) Baseline clinical characteristics: age at onset of depression, number of previous depressive episodes, length of index episode (months), concurrent physical illness (present vs absent)
3) Depression characteristics by week 3: individual item scores of PHQ-9 for the index episode, at week 1 and at week 3; individual item scores of the Beck Depression Inventory-II (BDI-II) at week 1 and week 3; individual item scores of the Frequency Intensity and Burden of Side Effects Rating (FIBSER) at week 1 and week 3; adherence to pharmacotherapy

BDI-II is a self-report measure of depression severity, consisting of 21 items, each of which is rated between 0 and 3 (Beck et al., 1996). In the current prediction model, the two items tapping appetite and sleep were divided into two items each, one evaluating increase and another evaluating decrease.

FIBSER is a self-report of overall drug side effects in terms of frequency, intensity and burden, each on a 7-point scale (Rush et al., 2006). We also measured adherence to pharmacotherapy in seven grades between 1= 'never or only taken the medication one day' and 7= 'taken the medication on all seven days of the past week.'

2.5 Statistical analyses

The overall aim of the analyses was to develop a treatment prediction model for the PHQ-9 scores at week 9, given patient-level socio-demographic and clinical characteristics measured before week 3 and treatment allocation at week 3. Since missing outcome values and missing covariate values occurred only for very few participants (2% and 4%, respectively, of the randomized population), and also assuming that covariate data were missing completely at random, we performed a complete case analysis.

2.5.1 Development of a personalized prediction model

We considered five methods for prediction model development: (i) penalized linear regression models using LASSO, (ii) penalized linear regression models using the ridge penalty, (iii) support vector machines (SVM) with a polynomial or radial kernel, and (iv) artificial neural networks with one hidden layer, 3 or 4 nodes. We fitted all models, with all predictors described in section 2.5 and treatment indicator as independent variables and the PHQ-9 score at week 9 as outcome variable. There were in total 88 patient covariates included in the models. All hyper-parameters (such as the penalty term of LASSO and ridge regression) were optimized using repeated 10-fold cross-validation. More specifically, we randomly split the data in 10 folds, and used 9 of the folds to train the model and the hold-out fold to test the model. We repeated this procedure 5 times. In each iteration, we fit each model using a range of tuning parameters. Via this procedure we developed a range of different models, using the four general methods listed above (i.e. LASSO, ridge, SVM, neural networks). The final model was chosen via an internal cross-validation, i.e. by comparing the mean squared error of the predictions (predicted vs. observed outcomes) for all candidate models.

All analyses were performed in R, using the *glmnet* (Friedman et al., 2010) and *caret* (Kuhn, 2008) packages.

2.5.2 Leave-one-patient-out cross-validation (Internal cross-validation)

In order to obtain an honest assessment of the performance of the model selected for the full dataset (see Section 2.6.1), we implemented a leave-one-patient-out cross-validation. More specifically, we took one patient out of the dataset at a time, and we developed the model using the remaining patients. For model development we used a 10-fold cross validation repeated 3 times. Note that this 10-fold

cross-validation was nested within the leave-one-patient-out cross validation. We then applied the developed model on the left-out patient, and we obtained a prediction of the patient's outcome under the three different treatments. By comparing the observed versus the predicted outcome (for the treatment actually received), we calculated the absolute difference between observation and prediction. We then summarized absolute differences across all patients. The leave-one-out validation procedure checks whether the model is valid and useful to use in new patients from the same target population, i.e. it assesses the model's reproducibility (Debray et al., 2015b).

Using this approach, for each patient we obtained a prediction of the outcome under all three different treatments. This allowed us to determine the treatment that was predicted to lead to the lowest PHQ-9 at week 9. Using the information on the best predicted treatment, we split the patients into three subgroups: patients for whom continuing sertraline was the optimal predicted treatment strategy (Group 1); combining it with mirtazapine was optimal (Group 2); or switching to mirtazapine was optimal (Group 3). After splitting the patients in the three subgroups (using only baseline information), we estimated the relative treatment effects within each subgroup, using the observed outcomes and the known treatment assignment for each patient. E.g. patients of Group 1 (for whom continuing sertraline was predicted to be the best) were randomized to either continue, combine or switch. Then the ones that were randomized to continue should, according to our algorithm, have on average a better outcome than those randomized to combine or switch. Thus, this strategy allowed us to utilize randomization in order to assess the usefulness of our prediction model. Within each of the three identified subgroups, we estimated all relative treatment effects (i.e. continue vs. combine; continue vs. switch; combine vs. switch) by estimating mean differences in PHQ-9 scores at week 9 and corresponding standard errors.

As secondary outcomes, we compared PHQ-9 scores among treatments at week 25.

2.5.3 Leave-one-site-out cross-validation (Internal-external cross-validation)

In order to get a sense of how well our model is expected to perform in a new (but similar) setting, we followed an internal-external cross-validation method (Debray et al., 2013; Royston et al., 2004; Steyerberg and Harrell, 2016). Patients in the dataset were clustered according to the clinic/hospital they were treated. We merged small clinics with less than 40 patients into one group ("small clinics"). In total, we grouped the patients in 12 different sites. With this information, we used the following iterative approach:

1) We removed one site from the dataset.
2) We developed our model in the patients of the remaining 11 sites. In this step we did not explore different types of models (i.e. LASSO/ridge/SVM/neural networks); we used the type of model that

was selected for the full dataset, as described in Section 2.6.1. Otherwise the modelling procedure was as previously described.

3) We applied the developed model to the hold-out sample. For each patient we obtained a prediction of the outcome. We then compared the prediction with the observed outcome, and calculated the absolute prediction error, i.e. the absolute difference between prediction and observation. Using all patients from the left-out clinic we calculated the MAE.
4) We repeated step 1-3 until each site was removed exactly once. This approach resulted in 12 site-specific MAE and corresponding standard errors.
5) We performed a random-effects meta-analysis of the 12 MAEs, and we estimated the 95% confidence interval, as well as an approximate 95% prediction interval of the pooled MAE (Debray et al., 2019; Debray et al., 2017).

In order to obtain an estimate of the case-mix of the 12 different sites, i.e. to assess how similar they are with each other with respect to the patients they included, we fit a set of membership models (Steyerberg et al., 2019). More specifically, using the complete dataset, for each different site we fit a ridge regression model where the outcome for each patient was "1" if a patient belonged to this site, "0" otherwise. All patient characteristics as well as the outcome at week 9 were used as predictors. The 12 models (one model per site) were separately fit, using 10-fold cross-validation. After fitting, we estimated the c-statistic for membership. A low c-statistic (i.e. close to 0.5) would mean that the different sites included similar patients. Higher c-statistic would indicate a broader case-mix, i.e. differences in baseline characteristics and outcomes of the patients among sites.

2.6 Web app

In order to facilitate the application of the obtained treatment prediction model, we have constructed a web app implementing the derived model.

3 RESULTS

3.1 Participants

A total of 2,011 patients were randomized at Step 1. Among them, a total of 1,646 patients did not remit by week 3 and were subsequently randomized at Step 2. For 33 of these patients (2%) we did not have information on the outcome on week 9, and were subsequently removed from all analyses. For 69 (4.2%) patients there were missing data for some of the predictor variables and were also excluded from the following analyses. Thus, the total sample we used amounted to 1,544 patients.

Altogether, 512 patients were allocated to continue sertraline, 502 to combine mirtazapine with sertraline, and 530 to switch to mirtazapine. Table 1 shows the baseline socio-demographic and

clinical characteristics of the total sample. Typically, patients were in their 30s and 40s, slightly more than half were women, the median number of depressive episodes was one including the index episode, and the median length of the index episode was 2.5 months.

The three treatment groups had at week 9, on average, achieved PHQ-9 scores of 9.2 (SD 5.9), 8.1 (6.0) and 8.3 (5.9) after continuing on sertraline, combining sertraline with mirtazapine and switching to mirtazapine, respectively. The evidence of differences between continuing and combining or between continuing and switching was strong (both p=0.001) but there was no evidence of a difference between combining and switching (p=0.94). (Kato et al., 2018)

3.2 Choosing the analytical method and developing the full model

After initial exploratory analyses we decided to use for the analysis three different support vector machines (SVMs) with a radial kernel, one SVM per treatment arm. This was because the initial analyses identified that this strategy produced the lower prediction error in both internal and internal-external cross validation (see Appendix 1, eFigures 1-4 for details for MAEs in internal-external cross-validation). After developing the three models, we calculated in-sample performance, i.e. we compared predictions from the models with actual observations. Overall, the MAE was 1.5 points on the PHQ-9 score. This was 1.7/2.3/1.1 for patients in the continue/combine/switch treatment arm. In eFigure 5 of the Appendix 2 we show a scatterplot of the observed versus predicted outcomes, and in eFigure 6 a histogram of the absolute prediction errors.

3.3 Leave-one-patient-out cross-validation

Next, we repeated the analysis of Section 3.2 after taking one patient out of the sample at a time. Thus, we developed three separate SVMs per each patient, using the data from the rest of the patients. We then used these models to make predictions on the left-out patient. The MAE of our predictions across all patients was 2.8 points on PHQ-9. This was 3.0/2.9/2.5 for the three randomization strata. This suggested that the in-sample MAEs presented in the previous paragraph were overoptimistic. eFigure 7 in Appendix 3 shows the scatterplot of the observed vs. the predicted scores and eFigure 8 a histogram of the absolute prediction errors.

Using the models we developed in each step of the leave-one-out procedure, we identified the best (predicted) treatment for each patient as the treatment for which our models predicted the lowest PHQ-9 scores at week 9. Out of the 1,544 patients, for 123 we predicted continuing sertraline to be the best treatment (Group 1); for 696 combination was predicted to be the best treatment (Group 2); for 725 it was switching to mirtazapine (Group 3). Table 1 summarizes the baseline characteristics of the three subgroups: item-level scores for PHQ-9, BDI-II and FIBSER of the three groups are provided

in Appendix 4. Group 1 tended to include younger patients, relatively more women, and patients with more depressive symptoms at week 3.

Table 2 shows the observed PHQ-9 scores for each subgroup. Among patients in Group 1, continuing sertraline was indeed better than switching to mirtazapine (the difference in PHQ-9 was -2.0, with 95% CI -4.7 to 0.6), but worse than combining, although the uncertainty was large (difference 1.0, -1.6 to 3.5). Among these patients, those randomized to combining rather than switching had much better outcomes, difference 3.1, (95% CI 0.5 to 5.5). Among patients in Group 2, there was in truth no difference between combining and switching, but continuing sertraline was the worse, mean difference vs. combining 1.3 (95% CI 0.2 to 2.4). Among patients in Group 3 , we observed no important difference vs. combining, but there were clinically important differences vs. continuing sertraline, mean difference 1.6 (0.5 to 2.3).

Next, we assessed whether the three subgroups we identified using the outcome at week 9 also performed similarly at week 25 (Table 3). A total of 1,512 patients provided information for PHQ-9 at week 25.. Overall we found that the differential effectiveness we observed in group 1 at week 9 carried over at week 25, i.e. for patients for which continuing was identified to be best, differences observed at week 9 became more pronounced at week 25: patients randomized to switching had on average 3.0 (95% CI -0.1 to 6.0) or 4.0 (95% CI 1.2, 7.0) higher PHQ-9 as compared to those randomized to continue sertraline or to combine it with mirtazapine, respectively.

3.4 Leave-one-site-out cross-validation (Internal-external cross-validation)

Table 4 summarizes the 12 different sites, and the results from the membership models. Based on the membership c-statistics and the mean PHQ-9 at week 9, we concluded that the patient covariates and outcomes were quite heterogeneous across sites.

Then, we proceeded with the leave-one-site-out cross-validation. Figure 1 shows the MAE (predicted vs observed) for each different site. A visual inspection of the forest plot suggested little variation in the MAEs among the different institutions. I-squared was 23% and the estimate for τ (the standard deviation of random effects) was quite low. The MAE of our leave-one-patient-out (2.8 points) was quite similar to the pooled MAE of the internal-external cross-validation at the site level (2.9 points).

3.5 Web app

We implemented the treatment prediction model based on the entire sample as a web app (https://cinema.ispm.unibe.ch/shinies/sund/.). When patient data are entered, the app predicts the outcome of PHQ-9 under the three treatments and determines which subgroup each individual would belong to. The web app currently requires manually entering all the many item-level data listed in

2.4; as such it is currently built mainly for illustrative and exploratory uses. Appendix 5 provides snapshots from the web app.

4 DISCUSSION

We used a sample of 1,544 participants with major depression who, after 3 weeks of receiving sertraline without remitting, were randomized to continue sertraline, combine with mirtazapine or switch to mirtazapine. Our aim was to predict the outcomes of these patients 6 weeks after randomization. To this aim, we built three prediction models that used support vector machines.

The traditional (group-level) analysis of this megatrial indicated that either combination or switching would be more efficacious than continuing, and that there was no difference in efficacy between combining and switching (Kato et al., 2018). Using our models, we identified three subgroups of patients: patients for whom the most efficacious treatment would be to continue sertraline, to combine with mirtazapine or to switch to mirtazapine. Subsequently, we used a leave-one-patient-out cross-validation approach, aiming to assess the robustness of these findings. This analysis showed that for the first of the identified subgroups (i.e. patients for whom continuing sertraline was predicted the best; n=123, 8% of the total population of patients), either continuing or combining would be recommended, with the latter clearly outperforming switching by 3.1 points on PHQ-9 (95% CI: 0.5 to 5.5). For the second subgroup (combination predicted best; n=696, 45%) as well as for the third (switching best; n=725, 47%), combination and switching were equally more efficacious than continuation. In other words, our validated treatment recommendation based on the results of our prediction models and the leave-one-out analysis was different from the treatment recommendation based on the overall results from the megatrial for the first subgroup, while it was similar for the second and the third subgroups.

The superiority of continuing or combining over switching that we predicted for the first subgroup persisted – or even widened – in week 25, as shown by the leave-one-patient-out cross-validation. In this analysis the benefit in terms of PHQ-9 score between continue vs. switch was 3.0 (-0.1 to 6.0) and between combine vs. switch 4.1 (1.2 to 7.0), for patients in the first subgroup. Differences between the three treatments for patients in the second and third subgroups virtually disappeared at week 25. The latter was also the case for the overall results in the original analyses of the megatrial (Kato et al., 2018). Note that information on the outcome at week 25 was not included at any stage in the development of our model. Thus, we conclude that our models can be used to identify a subgroup of patients for which switching should be avoided. For these patients, our cross-validation analysis showed that the optimal treatment would be to combine.

The leave-one-site-out cross-validation showed that the performance of the prediction model was homogeneous across the different sites, despite the fact that, as indicated by the membership models, the sites differed with respect to their case-mix distributions. This heterogeneity of patients across sites may have been partly due to cluster randomisation used at step 1 between 50 mg and 100 mg as initial dosage of sertraline. In the current analysis, the heterogeneity has contributed to increased expectation for a similar performance of the model when the model is applied to new but similar clinics or hospitals.

It was clinically counter-intuitive that patients for whom changing the treatment was recommended (Groups 2 and 3) had improved more from baseline to week 3 on initial sertraline, as compared to those for whom continuation was recommended (Group 1) (Table 1). Table 2 shows, however, that for Groups 2 and 3 patients, combining with or switching to mirtazapine did lead to even better outcomes than continuing the initially apparently effective treatment. By contrast, for Group 1 patients, even though they had improved the least on initial sertraline up to 3 weeks, either continuing on sertraline or combining it with mirtazapine achieved greater improvement by week 9 than switching to mirtazapine. It may be the case that for Group 1 patients drugs (both sertraline and mirtazapine) may take more time to show their effects than for Groups 2 or 3 patients and therefore continuing sertraline up to 9 weeks was beneficial with or without mirtazapine and prescribing mirtazapine for 6 weeks by switching was not beneficial. The overall poor outcomes of Group 1 patients, in comparison with Groups 2 or 3 patients whatever the treatment choice at week 3 was, would be compatible with this speculation that Group 1 patients need more time to improve.

Another salient difference among the three groups was proportion of women: it was very high in Group 1 but very low in Group 2, with Group 3 coming in-between. The differences in side effects (e.g. sexual dysfunction or weight gain) or efficacy on anxiety between sertraline and mirtazapine may have contributed to these differences but it is hard to know this in the current study because we did not measure specific side effects or anxiety.

Individual participant data meta-analysis of randomized controlled trials is one way to assemble a large dataset. When conducted as a network meta-analysis, it can lead to differential treatment prediction models, matching patient characteristics with treatment choices (Furukawa et al., 2018). There is, however, one big potential caveat in such endeavours, common to all studies using existing datasets, be they experimental or observational. Different studies often measure different sets of variables, and in different ways (e.g. with different scales or with different categorizations): in such cases it is very difficult to harmonize the dataset and the number of covariates usable in the meta-analysis may often diminish rapidly. Although imputation methods have been recently proposed to address this limitation (Audigier et al., 2018), their implementation is not straightforward in studies with few participants and/or clusters. By contrast, if the dataset comes from a single large trial, as in

the current study, the covariates are usually uniformly measured, while at the same time the amount of missing information is limited.

There are several weaknesses to our study. First, MAEs were relatively large (around 3 points on PHQ-9) in comparison with expected differences among the treatments. This may be due to inherent heterogeneity among patients diagnosed with the current criteria; or to an unavoidable noise in measuring and predicting depression severity. It may be reduced in future studies if we measure additional and informative covariates, such as genetic factors. Second, given the pragmatic nature of the original megatrial, we did not have any measurements of biological markers nor any detailed assessments of psychiatric comorbidities or psychosocial history of the participants. For example, several studies in the literature have suggested that comorbid anxiety disorders (Papakostas et al., 2008), substance use disorders (Rush et al., 2008) and history of adverse childhood experiences (Nemeroff et al., 2003) may be important effect modifiers. It is up to the future studies to replicate and hone the current prediction model with such covariates. Third, although we confirmed the internal-external validation of the model through the leaving-one-site-out method, thus establishing that we can expect similar performance in a new but similar setting, it is still unknown how the model would perform in completely new settings (e.g. clinics in neighbouring or remote countries). The model therefore needs to be further tested and cross-validated. For instance, this could be performed in a new cohort. However, in such a case, the validation of our algorithm might not be straightforward, because our algorithm uses treatment as a moderating factor, and thus it will be hard to assess its performance in a setting where treatment is not randomized, due to confounding. Thus, the superiority of the approach using the developed model over the approach based on the recommendations from the aggregate level analyses needs to be ideally demonstrated in an RCT.

By contrast, the major strengths of the current study may be summarized as follows. It used the largest single trial dataset to date in antidepressant treatment of depression, with consistent measurements and little missingness. The patients suffered from hitherto untreated episodes of major depression, with median length of the index episode of 2-3 months, unlike some other megatrials of antidepressants which recruited much more chronic populations (Trivedi et al., 2006). The current findings would therefore apply well to those who are starting on treatments for their major depressive episode but have not responded to their first antidepressant. Our models utilized information on a wide list of possibly predictive patient covariates. Such modelling approach is suspected to be prone to overfitting in the absence of true effect modifications (Kent et al., 2018; van Klaveren et al., 2019). However, the relatively robust internal-external validity in the current study suggests that we have been able to include some genuine effect modifiers.

In summary, the newly developed precision treatment prediction model was able to discriminate between patients who would benefit from continuing or combining rather than switching (treatment

recommendations different from those based on the traditional RCT analyses) and those who would benefit from combining or switching rather than continuing (the same recommendations as in the original RCT). The clinical implication of the current findings may be as follows: For individual patients possibly belonging to Group 1 (i.e. for patients for whom continue sertraline was predicted to be best), the treatment differences are substantive. Clinicians are advised to use the attached web app to determine if a patient belongs to Group 1 and, if so, plan the treatment accordingly. By contrast, research implications may be as follows: Group 1 constituted 8% of the entire cohort, for whom the differentiation is clinically substantive: however, the average outcome of the entire group may not improve substantially when this model is applied to the entire population of patients. If we hope to create models that will make a difference on the population level, we need to find stronger predictors, possibly including currently unknown genetic and other biomarkers, and also using datasets of large observational as well as randomised studies (Tomlinson et al., in press). Such predictors may lead in the future to more accurate prediction models; without them, precision treatment prediction models in psychiatry may never find wider acceptance.

Contributors

TAF and OE conceived and designed the study. TAF, TA, MY and TK conducted the study and acquired the data. MS and OE conducted the statistical analyses. TAF, TD and OE interpreted the data. TAF and OE drafted the manuscript and all the authors critically contributed to its revisions. All authors read and approved the final manuscript.

Role of the Funding Sources

This work has been supported in part by JSPS Grant-in-Aid for Scientific Research (Grant number 17K19808) to TAF. TD has been supported by The Netherlands Organisation for Health Research and Development (grant 91617050).

The original SUN ☺ D 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 funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Declaration of interests

TAF reports personal fees from Mitsubishi-Tanabe, MSD and Shionogi, and a grant from Mitsubishi-Tanabe, outside the submitted work; TAF has a patent 2018-177688 pending.

TA has received lectures fees from Astellas, AstraZeneca, Daiichi-Sankyo, Dainippon-Sumitomo, Eisai, Hisamitsu, Janssen, Kyowa-hakko Kirin, Kyowa, Lilly, MSD, Meiji-seika Pharma, Mochida, Mundipharma, Nipro, Otsuka, Pfizer, Shionogi, Terumo, and Tsumura. He has received research funds from Daiichi-Sankyo, Eisai, FUJIFILM RI Pharma, Lilly, MSD, Novartis, Otsuka, Shionogi, and Tanabe-Mitsubishi.

TK has received lectures fees from Eli Lilly, Mitsubishi-Tanabe and Pfizer. TK has received contracted research funds from GSK, MSD, and Mitsubishi-Tanabe.

All the other authors declare no conflict of interest.

References

Audigier, V., White, I.R., Jolani, S., Debray, T.P.A., Quartagno, M., Carpenter, J., van Buuren, S., Resche-Rigon, M., 2018. Multiple Imputation for Multilevel Data with Continuous and Binary Variables. Statist. Sci. 33, 160-183.

Beck, A.T., Steer, R.A., Brown, G.K., 1996. BDI-II: Beck Depression Inventory, Second Edition, Manual. The Psychological Corporation, San Antonia.

Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G., 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann. Intern. Med. 162, 55-63.

Cuijpers, P., Christensen, H., 2017. Are personalised treatments of adult depression finally within reach? Epidemiol Psychiatr Sci 26, 40-42.

Debray, T.P., Damen, J.A., Riley, R.D., Snell, K., Reitsma, J.B., Hooft, L., Collins, G.S., Moons, K.G., 2019. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat. Methods Med. Res. 28, 2768-2786.

Debray, T.P., Damen, J.A., Snell, K.I., Ensor, J., Hooft, L., Reitsma, J.B., Riley, R.D., Moons, K.G., 2017. A guide to systematic review and meta-analysis of prediction model performance. BMJ 356, i6460.

Debray, T.P., Moons, K.G., Ahmed, I., Koffijberg, H., Riley, R.D., 2013. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat. Med. 32, 3158-3180.

Debray, T.P., Riley, R.D., Rovers, M.M., Reitsma, J.B., Moons, K.G., 2015a. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLoS Med. 12, e1001886.

Debray, T.P., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E.W., Moons, K.G., 2015b. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J. Clin. Epidemiol. 68, 279-289.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33, 1-22.

Furukawa, T.A., Akechi, T., Shimodera, S., Yamada, M., Miki, K., Watanabe, N., Inagaki, M., Yonemoto, N., 2011. Strategic Use of New generation antidepressants for Depression: SUN(^_^)D study protocol. Trials 12, 116.

Furukawa, T.A., Efthimiou, O., Weitz, E.S., Cipriani, A., Keller, M.B., Kocsis, J.H., Klein, D.N., Michalak, J., Salanti, G., Cuijpers, P., Schramm, E., 2018. Cognitive-Behavioral Analysis System of Psychotherapy, drug, or their combination for persistent depressive disorder: Personalizing the treatment choice using individual participant data network metaregression. Psychother. Psychosom. 87, 140-153.

Institute of Medicine Committee on Strategies for Responsible Sharing of Clinical Trial Data, 2015. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risks. National Academic Press, Washington, D.C.
Kato, T., Furukawa, T.A., 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, G.H., for the SUN(^_^)D Investigators, 2018. Optimising first- and second-line treatment strategies for untreated major depressive disorder - the SUN☺D study: a pragmatic, multi-centre, assessor-blinded randomised controlled trial. BMC Med. 16, 103.
Kent, D.M., Steyerberg, E., van Klaveren, D., 2018. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. Bmj 363, k4245.
Kessler, R.C., Bossarte, R.M., Luedtke, A., Zaslavsky, A.M., Zubizarreta, J.R., 2019. Machine learning methods for developing precision treatment rules with observational data. Behav. Res. Ther. 120, 103412.
Kuhn, M., 2008. Building predictive models in R using the caret package. Journal of Statistical Software 28, 1-26.
Nemeroff, C.B., Heim, C.M., Thase, M.E., Klein, D.N., Rush, A.J., Schatzberg, A.F., Ninan, P.T., McCullough, J.P., Jr., Weiss, P.M., Dunner, D.L., Rothbaum, B.O., Kornstein, S., Keitner, G., Keller, M.B., 2003. Differential responses to psychotherapy versus pharmacotherapy in patients with chronic forms of major depression and childhood trauma. Proc. Natl. Acad. Sci. U. S. A. 100, 14293-14296.
Noma, H., Furukawa, T.A., Maruo, K., Imai, H., Shinohara, K., Tanaka, S., Ikeda, K., Yamawaki, S., Cipriani, A., 2019. Exploratory analyses of effect modifiers in the antidepressant treatment of major depression: Individual-participant data meta-analysis of 2803 participants in seven placebo-controlled randomized trials. J. Affect. Disord. 250, 419-424.
Papakostas, G.I., Stahl, S.M., Krishen, A., Seifert, C.A., Tucker, V.L., Goodale, E.P., Fava, M., 2008. Efficacy of bupropion and the selective serotonin reuptake inhibitors in the treatment of major depressive disorder with high levels of anxiety (anxious depression): a pooled analysis of 10 studies. J. Clin. Psychiatry 69, 1287-1292.
Perna, G., Grassi, M., Caldirola, D., Nemeroff, C.B., 2018. The revolution of personalized psychiatry: will technology make it happen sooner? Psychol. Med. 48, 705-713.
Riley, R.D., Ensor, J., Snell, K.I., Debray, T.P., Altman, D.G., Moons, K.G., Collins, G.S., 2016. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 353, i3140.
Royston, P., Parmar, M.K., Sylvester, R., 2004. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Stat. Med. 23, 907-926.
Rush, A.J., Trivedi, M.H., Wisniewski, S.R., Stewart, J.W., Nierenberg, A.A., Thase, M.E., Ritz, L., Biggs, M.M., Warden, D., Luther, J.F., Shores-Wilson, K., Niederehe, G., Fava, M., 2006. Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. N. Engl. J. Med. 354, 1231-1242.
Rush, A.J., Wisniewski, S.R., Warden, D., Luther, J.F., Davis, L.L., Fava, M., Nierenberg, A.A., Trivedi, M.H., 2008. Selecting among second-step antidepressant medication monotherapies: predictive value of clinical, demographic, or first-step treatment features. Arch. Gen. Psychiatry 65, 870-880.
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., Furukawa, T.A., 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.
Steyerberg, E.W., Harrell, F.E., Jr., 2016. Prediction models need appropriate internal, internal-external, and external validation. J. Clin. Epidemiol. 69, 245-247.
Steyerberg, E.W., Nieboer, D., Debray, T.P.A., van Houwelingen, H.C., 2019. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. Stat. Med.
Tomlinson, A., Furukawa, T.A., Efthimiou, O., Salanti, G., De Crescenzo, F., Singh, I., Cipriani, A., in press. Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. Evid. Based Ment. Health.
Trivedi, M.H., Rush, A.J., Wisniewski, S.R., Nierenberg, A.A., Warden, D., Ritz, L., Norquist, G., Howland, R.H., Lebowitz, B., McGrath, P.J., Shores-Wilson, K., Biggs, M.M., Balasubramani, G.K., Fava, M., 2006.

Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am. J. Psychiatry 163, 28-40.
van Klaveren, D., Balan, T.A., Steyerberg, E.W., Kent, D.M., 2019. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. J. Clin. Epidemiol. 114, 72-83.
Wium-Andersen, I.K., Vinberg, M., Kessing, L.V., McIntyre, R.S., 2017. Personalized medicine in psychiatry. Nord. J. Psychiatry 71, 12-19.

Table 1. Baseline socio-demographic and clinical characteristics of the whole sample, as well as the three identified subgroups, according to leave-one-patient-out analysis.

Socio-demographic characteristics	**Total sample (n=1544)**	**Group 1:** Continuing sertraline was the best (n=123)	**Group 2:** Combining with mirtazapine was the best (n=696)	**Group 3:** Switching to mirtazapine was the best (n=725)
Age in years, mean (SD)	41.5 (11.5)	36.1 (8.0)	41.1 (11.0)	42.8 (12.1)
Female sex, n (%)	803 (52%)	87 (70.1%)	260 (37.5%)	394 (54.3%)
Education in years, mean (SD)	14.0 (2.3)	14.2 (2.3)	13.8 (2.2)	14.1 (2.4)
Employment status, n (%)				
Full time	612 (39.6%)	50 (40.7%)	289 (41.5%)	273 (37.7%)
Part-time	133 (8.6%)	1 (0.8%)	72 (10.3%)	60 (8.3%)
On sick leave	427 (27.7%)	56 (45.5%)	176 (25.3%)	195 (26.9%)
Housewife	154 (10.0%)	5 (4.1%)	92 (13.2%)	57 (7.8%)
Student	14 (0.9%)	0 (0.0%)	6 (0.9%)	8 (1.1%)
Retired	15 (1%)	0 (0.0)	2 (0.3%)	13 (1.8%)
Not employed	189 (12.2%)	11 (8.9%)	59 (8.4%)	119 (16.4%)
Marital status, n (%)				
Single (never married)	493 (31.9%)	30 (24.4%)	218 (31.3%)	245 (33.4%)
Divorced	221 (14.3%)	15 (12.2%)	75 (10.8%)	131 (18.1%)
Widowed	32 (2.0%)	0 (0.0%)	19 (2.7%)	13 (1.8%)
Married	798 (51.7%)	78 (63.4%)	384 (55.2%)	336 (46.3%)
Clinical characteristics				
Age (yrs) at onset of first episode, years mean (SD)	36.9 (12.6)	32.5 (9.6)	35.7 (12.0)	38.8 (13.2)
Number of depressive episodes, mean (SD), median, range	2.3 (3.7), median: 1, range: 1-50	1.8 (1.5), median: 1, range: 1-10.	2.6 (3.6), median: 2, range: 1-30	2.2 (3.9), median: 1, range: 1-50
Length of current depressive episode (months), mean (SD), median, range	6.3 (15.2), median: 2.5, range: 0.5-276	7.6 (18.3) median: 3.0, range: 0.5-144	6.8 (15.7) median: 3.0, range: 0.5-240	5.6 (14.0) median: 2.0, range: 0.5-276

Comorbid physical illness, n (%)	509 (33.0%)	19 (15.4%)	249 (35.8%)	241 (33.2%)
PHQ-9 total score				
Baseline	18.7 (3.8)	19.6 (3.5)	18.8 (3.9)	18.4 (3.8)
Week 1	15.9 (4.7)	16.9 (4.2)	16.9 (4.7)	14.8 (4.6)
Week 3	12.7 (5.2)	14.9 (5.5)	13.9 (5.2)	11.2 (4.6)
BDI-II total score				
Week 1	28.7 (10.2)	33.8 (9.6)	31.2 (10.0)	25.5 (9.5)
Week 3	24.3 (10.8)	31.7 (11.0)	27.0 (10.6)	20.4 (9.3)
FIBSER total score				
Week 1	6.8 (3.9)	7.9 (3.8)	7.1 (4.1)	6.2 (3.7)
Week 3	7.1 (4.0)	6.0 (3.6)	8.1 (4.3)	6.3 (3.6)
Adherence				
Week 1	6.1 (1.4)	6.1 (1.4)	6.0 (1.4)	6.1 (1.3)
Week 3	6.6 (1.0)	6.5 (1.1)	6.6 (0.9)	6.6 (1.1)

CI: confidence interval. SD: standard deviation.

BDI-II: Beck Depression Inventory, 2nd edition. FIBSER: Frequency, Intensity and Burden of Side Effects Rating. PHQ-9: Patient Health Questionnaire-9

Table 2. Observed PHQ-9 scores at week 9 for the three identified subgroups according to the leave-one-patient-out cross-validation.

Best treatment according to the treatment prediction model	Total no of patients	Actually continued sertraline (SD)	Actually combined with mirtazapine (SD)	Actually switched to mirtazapine (SD)	Comparison between continue vs combine (95% CI)	Comparison between continue vs switch (95% CI)	Comparison between combine vs switch (95% CI)
Continue sertraline	123	10.1 (6.2)	9.1 (5.3)	12.1 (6.1)	1.0 (-1.6 to 3.5)	-2.0 (-4.7 to 0.6)	-3.1 (-5.4 to -0.5)
Combine with mirtazapine	696	10.5 (6.0)	9.2 (6.2)	9.3 (6.2)	1.3 (0.2 to 2.4)	1.2 (0.1 to 2.3)	-0.1 (-1.2 to 1.0)
Switch to mirtazapine	725	8.0 (5.5)	6.6 (5.5)	6.6 (5.0)	1.3 (0.3 to 2.3)	1.4 (0.5 to 2.3)	0.1 (-0.9 to 1.0)
Overall	1544	n=512 9.2 (5.9)	n=502 8.1 (6.0)	n=530 8.3 (5.9)	1.1 (0.4, 1.9)	0.9 (0.2, 1.7)	-0.2 (-0.9, 0.6)

CI: confidence interval. SD: standard deviation.

A comparison between two treatments A and B is provided as mean difference, where a value smaller than zero favours treatment A (i.e. corresponds to a lower score).

Table 3. PHQ-9 at week 25 and FIBSER at weeks 9 and 25 for the three identified subgroups according to the leave-one-patient-out cross-validation.

PHQ-9 at week 25							
Best treatment according to the treatment prediction model	Total no of patients	Actually continued sertraline (SD)	Actually combined with mirtazapine (SD)	Actually switched to mirtazapine (SD)	Comparison between continue vs combine (95% CI)	Comparison between continue vs switch (95% CI)	Comparison between combine vs switch (95% CI)
Continue sertraline	118	7.4 (6.4)	6.4 (5.7)	10.4 (7.0)	1.0 (-1.7, 3.4)	-3.0 (-6.0, 0.1)	-4.1 (-7.0, -1.2)
Combine with mirtazapine	681	7.4 (5.9)	7.2 (6.5)	7.6 (6.3)	0.2 (-0.9, 1.4)	0.2 (-0.9, 1.3)	-0.4 (-1.6, 0.8)
Switch to mirtazapine	713	5.4 (5.1)	5.3 (5.2)	5.0 (5.2)	0.1 (-0.8, 1.1)	0.4 (-0.5, 1.4)	0.3 (-0.7, 1.2)
Overall	1512	n=505 6.4 (5.6)	n=491 6.3 (5.9)	n=516 6.6 (6.1)	0.1 (-0.6, 0.9)	-0.2 (-0.9, 0.6)	-0.3 (-1.0, 0.4)

CI: confidence interval. SD: standard deviation.

A comparison between two treatments A and B is provided as mean difference, where a value smaller than zero favours treatment A (i.e. corresponds to a lower score).

Table 4. Summary of the 12 sites (clinics/hospitals) and results from the membership models.

#	Name	Number of patients	Mean PHQ-9 at week 9 (SD)	c-statistic for membership
1	D06	44	8.3 (6)	0.81
2	N06	45	10.1 (6.3)	0.95
3	K05	50	8.8 (7)	0.76
4	T02	51	8.5 (6.1)	0.79
5	A03	53	6.5 (4.2)	0.80
6	K06	62	8.5 (6.6)	0.82
7	R05	68	8.3 (5.7)	0.93
8	K09	96	8.6 (6.3)	0.95
9	H03	101	9.8 (6.4)	0.86
10	H05	115	8.8 (6)	0.83
11	N03	368	7.7 (5.4)	0.82
12	SC (small clinics/hospitals)	491	8.9 (6)	0.73

Figure 1. Internal-external cross-validation: median absolute errors by clinic/hospital. Clinics with less than 40 patients were grouped. MAE: Median absolute error. se: standard error