Application of Item Response Theory to Model Disease Progression and Agomelatine Effect in Patients with Major Depressive Disorder

Marc Cerou, Sophie Peigné, Emmanuelle Comets, Marylore Chenel

▶ **To cite this version:**

Marc Cerou, Sophie Peigné, Emmanuelle Comets, Marylore Chenel. Application of Item Response Theory to Model Disease Progression and Agomelatine Effect in Patients with Major Depressive Disorder. AAPS Journal, 2020, 22 (1), pp.4. 10.1208/s12248-019-0379-x . hal-02394266

HAL Id: hal-02394266

https://univ-rennes.hal.science/hal-02394266

Submitted on 17 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Application of item response theory to model disease progression and agomelatine effect in patients with major depressive disorder

M. Cerou [*1,3], S. Peigné [3], E. Comets [†1,2] and M. Chenel [†3]

[1]Université de Paris, IAME, INSERM, F-75018, Paris, France
[2]Université Rennes-1, CIC1414, 35700 Rennes, France
[3]Division of Clinical Pharmacokinetics and Pharmacometrics, Institut de Recherches Internationales Servier, Suresnes, France

Abstract

Introduction: In this paper, we studied the effect over time of agomelatine, an antidepressant drug administered in patient with major depressive disorder, through item response theory (IRT), taking into account a strong placebo effect and missing not at random data. We also assessed the informativeness of the HAMD-17 scale's item.

Materials and Methods: The data includes five phase III clinical trials sponsored by Servier Institute, totalling 1549 patients followed during a maximum of 1 year. At each observation, individual scores for the 17 items of the HAMD scale were recorded. The probability for each score was modelled with IRT. A non-linear mixed effects model was used to describe the evolution of the disease and was coupled with a time to event model to predict dropout. Clinical trial simulations were then used to compare placebo and active treatment. Informativeness of each item was evaluated using the Fisher information theory.

Results: The best model combined an IRT model, a longitudinal model for underlying depression which describes the remission and then a possible relapse, and a hazard model for dropout depending on the evolution from baseline. The drug effect was best modelled as an effect on the remission and the relapse phases. The median predicted drop in HAMD between baseline and 6 weeks was 8.8 (90% PI, 8.3–9.2) when on placebo and 13.1 (90% PI, 12.8–13.4) when treated. Nine items were found to be the most informative.

Conclusion: The IRT framework allowed to characterise the evolution of depression with time and estimate the effect of agomelatine, as well as the link between symptoms and disease.

Keywords— Major Depressive Disorder, IRT, Agomelatine, MNAR, NLMEM

*Corresponding author: marc.cerou@inserm.fr
†Co-last author

Accepted Manuscript

Introduction

Major depressive disorder (MDD), also known simply as depression, is a common but serious mental disorder. According to the World Health organisation [1] (WHO), MDD is one of the leading causes of disability worldwide, and affects more than 300 million people. It is characterised by mood, cognitive, vegetative and psychotic symptoms which can persist through most of the day, nearly every day during at least two weeks. Major depression is often thought of as an episodic disorder. Individuals who have experienced earlier episodes will have an increased risk for a new episode of 16% [2]. Epidemiologic data [3] showed that depression may strike at any time, but is most common later than the age of onset which is in median in the early to mid 20s. Individuals who are separated or divorced [4] and women (two-fold increase compared to men) [5] are at higher risk of major depression. The prevalence of the disease decreases with age [4] in Western countries. According to the National Institute of Mental Health in USA, depression is caused by combination of genetic, environmental, and psychological factors, which include personal or family history of depression, major life changes, trauma, stress, as well as certain physical illnesses and medications. WHO recommends psychotherapy or/and medications for the treatment of depression [1], but symptoms often persist despite pharmacological treatment [6] showing the urgent need to find better therapeutic options.

Antidepressants drugs aim to correct imbalances of neurotransmitters in the brain that are described to be responsible of mood and behaviour changes. The antidepressant agomelatine was marketed in Europe in 2009 for the treatment of major depressive episodes in adults [7]. It is thought to act through a combination of antagonist activity at 5HT2C receptors and agonist activity at melatonergic MT1/MT2 receptors [7, 8]. Agomelatine has shown efficacy in MDD [9] while retaining a favourable tolerance profile, attributed in part to a lack of interference with the reuptake of neurotransmitters like serotonin and dopamine [10]. Recent meta-analyses [9, 10] have shown efficacy of agomelatine in MDD while considering published and unpublished non-positive studies.

Clinical trials in depression face a number of challenges, reflecting the complex etiology and chronic nature of the disease, the large subject to subject variability in expression and resilience to depression, and the natural course of disease progression. Moreover a large placebo response is generally observed in trials evaluating novel antidepressant treatments [11, 12]. The impact of depression on sleep, anxiety, mood, somatic and psychotic components is captured through a multi-item symptom inventory composed of 17 items, the Hamilton depression rating scale [13] (HAMD-17). HAMD-17 is used as the primary clinical outcome in clinical trials on depression for its ability to cover the spectrum of depressive symptoms. However, this multidimensionality also complicates the interpretation and detection of drug effects [14, 15], and alternative scales more sensitive to a change due to the drug effect have been proposed [16, 17, 18, 19, 20]. Using an appropriate scale, as well as extracting all possible information from the different items, improves the power of clinical trials.

Recent publications have demonstrated, in a nonlinear mixed effect model (NLMEM) context, that the use of the item response theory (IRT) framework allows an increase in the precision of composite score estimation, and thus increases the power to detect a drug effect [21, 22]. IRT was first proposed in the 50s for educational and achievement tests [23]. IRT is often used in the design of questionnaires by tailoring tests through item selection. It is an important tool in psychometrics and is more and more used in healthcare field. Combined with NLMEM, this methodology was first used in the context of Alzheimer disease [21, 24, 25, 26] and also in other diseases such as multiple sclerosis [27], cognition [28], in mechanically ventilated preterm neonates [29], Parkinson [30], schizophrenia [31], acute ischemic stroke [32] and migraine patients [33]. IRT is a statistical framework aimed at characterising the relation between each component of a scale and an individual ability/disability. In our study, each symptom in the HAMD-17 scale is used as a surrogate measure of a latent variable measuring the depression. Combining IRT and NLMEM provides a useful framework to characterise mathematically the relation between each symptom and the latent variable, model

Accepted Manuscript

the evolution of the latent variable over time, and capture more precisely the multidimensionality of the scale [34].

A further issue in chronic diseases, and particularly in clinical trials in depression, is how to appropriately handle data from patients dropping out of the study. Dropout rates can typically range from 20%-40% for short studies [35] to 50% in studies lasting over a year [36]. Patients may leave the study because they suffered from an adverse event or because they feel no benefit from the treatment [37], as well as for non-medical reasons. Interestingly, adverse events can also occur in the placebo arms, a fact known as nocebo effect and related to negative expectations of the medication [38]. In the literature and nicely summarised in [39], the dropout process is usually classified into Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR) depending on the relationship between the risk of leaving the study and hidden or observed variables. MNAR is also called informative dropout [40, 41]. In longitudinal studies, the need to integrate the dropout process in the model to obtain unbiased estimates in case of MNAR has been demonstrated ([41]. More recently, Björnsson et al. [42] verified and quantified the magnitude of the bias depending on different estimation methods. In the context of MDD and NLMEM, Gomeni et al. [35] showed that modelling the dropout process is necessary to avoid biased inferences.

In the present study, we built a model to take into account the multidimensionality of the scale, the occurrence of dropout events, and further complexities such as an open label period for one of the studies, and dose adjustments for non-responsive patients in clinical trials involving agomelatine. This work aimed to (1) combine IRT, NLMEM and dropout model to characterise the disease progression of MDD and drug effect of agomelatine, (2) use the IRT methodology to identify the most informative items. The data used and the modelling approach are presented in Materials and Methods. The resulting model and its validation are outlined in the Results section. The most informative items are described at the end of this section. We then discuss the strengths and limitations of this work, and the perspectives for the future.

Materials and Methods

Data

The data for this paper was collected in five phase III ([36, 43, 44, 45, 46]), multi-centre, placebo-controlled, flexible dose, double-blind studies, as shown in Table I. Trials were selected based on a similar clinical management of the placebo arms, as well as the availability of individual HAMD scores to allow longitudinal modelling of the evolution of depression with IRT. All studies included men and women aged between 18 and 65 years, except [Heun2013] where patients were over 65 years. Inclusion criteria was a moderate to severe major depressive episode according to DSM-IV-TR criteria, characterised by an HAM-D score of at least 22 requiring an antidepressant treatment. Additionally for some studies the Clinical Global Impression (CGI)- severity of illness (CGI-S) needed to be at least 4 with a duration of the current depressive episode of 4 to 8 weeks.

Detailed information on each study is given in Supplementary material. Briefly, all trials included a compulsory treatment period and an optional period where the patients could continue receiving the treatment they had been randomised to, summarised in table I. Patients were randomised to treated or placebo groups at entry, except in [Goodwin2009] where patients were first enrolled in an open-treatment (TO) period for 8 to 10 weeks. Patients were seen every 2 weeks during the mandatory phase and every 4 to 9 weeks during the optional phase. The main clinical endpoint was the Hamilton Depression Rating Scale, a questionnaire composed of 17 items (HAMD-17), which was taken at each visit, along with a CGI scale determined by the physician. Patients could continue on to the optional phase subject to their and/or the investigator's agreement. Additionally, the CGI Improvement score (CGI-I) was used to determine withdrawal at specific visits for some studies. Patients were withdrawn from study if the CGI-I was greater than 3 at week 10

Accepted Manuscript

in [Kennedy2006], greater than 3 at week 6 and greater than 2 at week 10 in [Kennedy2014], and greater than 2 at week 12 in [Heun2013]. All studies compared a dose of 25 mg of agomelatine to placebo, except [Kennedy2014] which included a 10 mg and a 25 mg fixed dose arms. For all patients except in two arms (10 mg and 25 mg fixed dose) of [Kennedy2014], the response was evaluated after 2 weeks, and if deemed insufficient according to an external group of experts, the dose was doubled.

Fig. 1 shows the evolution of the HAMD-17 score over time stratified by study and arm. Overall, 1549 patients were followed over time with a median follow-up of 25 weeks. Demographic characteristics can be found in table II.

Model building

Item-response theory and non-linear mixed effect models were used to analyse the longitudinal measurements of the HAMD scale. The individual scores were modelled through IRT as a function of a latent variable D representing depression. A model of the evolution of D under placebo was developed to describe the natural evolution of depression in the clinical trials. Agomelatine concentrations were predicted through a K-PD model [47] based on individual patient regimen, and found to act on parameters governing time to remission and relapse. A joint model was developed to account for dropout related to D.

To develop this complex model, we proceeded sequentially. The dataset was split in a building dataset (N=1088) and an evaluation dataset (N=461), stratifying on study and treatment group. First, an initial estimation of the population-specific parameters of the IRT model for each item was obtained using only the placebo and baseline data. Second, a structural model for the evolution of D under placebo and for the effect of agomelatine was developed, and coupled with a model for dropout, investigating the link between D and the probability to leave the study. Standard model building and evaluation methods were used. The model was further extended to account for specific features such as dose increase after 2 weeks in poor responders and an additional effect of open treatment. Third, the final model was evaluated on both the building and the evaluation datasets with 1000 trial replicates.

Model application

The model was then used to predict the difference between agomelatine and placebo on the level of the HAMD-17 scores over time. A measure of efficacy is the standardised mean difference (SMD) at 6 weeks, also known as effect size. This was computed with Cohen's d-statistic [48] first in the context of clinical trial simulation (CTS) using the last prediction carried forward, and second using predictions which enable the computation of a dropout-corrected SMD. We reported the median SMD and dropout-corrected SMD with their 90% prediction interval (5th and 95th percentiles). We also used information theory to assess which items are most sensitive to changes in depression levels.

Detailed information on the equations for different models tested, model building and evaluation, and model use are reported in Supplementary Materials. The main components of the final model are described in Results.

Accepted Manuscript

Results

Details concerning the different steps of model building are reported in the Supplementary material. Below we describe the main components of the final model.

Model building

IRT Model

HAMD-17 is a composite score assessed by a questionnaire constituted of 17 items exploring several symptoms of the disease. Some questions had three possible answers while others had five. Instead of modelling directly the total score, we consider these items as a surrogate measure of the depression level and modelled the impact of disease progression on each item. Let observation Y_{ijs} represent the value of the score for question s of the HAMD-17 scale for patient i at time t_{ij}. For each item, the response was modelled by an ordered categorical model [49, 50]. The probability of being equal to K depends on item-specific parameters, which were assumed to be characteristic of the population, and an individual latent variable D_{ij} representing the level of the depression for subject i at time t_{ij}:

$$P(Y_{ijs} \geq K) = \frac{e^{a_s \cdot (D_{ij} - b_{sK})}}{1+e^{a_s \cdot (D_{ij} - b_{sK})}} \quad , K = 1,\cdots,Kmax \ (Kmax = 2 \text{ or } 4) \tag{1}$$

$$\begin{cases} P(Y_{ijs} = K) = P(Y_{ijs} \geq K) - P(Y_{ijs} \geq K+1) \quad , K = 0,\cdots,Kmax-1 \ (Kmax = 2 \text{ or } 4) \\ P(Y_{ijs} = Kmax) = P(Y_{ijs} \geq Kmax) \\ \sum_0^{Kmax} P(Y_{ijs} = K) = 1 \end{cases} \tag{2}$$

where a_s is the slope (or discrimination) and b_{sK} the difficulty parameter (or item location) for item s and value K of the score. Difficulty parameters associated to each score are in the scale of the latent variable. Higher values indicate high scores are only likely in the presence of severe depression. The difficulty parameter was constrained to be non-decreasing ($b_{s,K+1} \geq b_{s,K}$). For simplicity and readability, the indices are removed.

Longitudinal model of depression status

Clinical trials on depression characteristically exhibit a marked placebo response, as patients will experience a phase of remission (decreasing severity of the disease) and possibly a phase of relapse (increasing severity of the disease). Clinically, remission is characterised as achieving a total HAMD-17 score of less than 6 while relapse is defined by a total HAMD-17 score that increases and exceeds 16 points. If patients relapsed, they were withdrawn from the study so only the initial part of the relapse phase was available. The assumption is that all patients return in time to a score close to their baseline. Several models accounting for the evolution of depression have been proposed in the literature [51, 52, 53, 54]. Here, we used a flexible model which can capture the remission and relapse phase, combining a Weibull model to describe the remission with flexibility as presented by Gomeni et al, and an exponential model that tends towards the baseline value to describe the relapse (equation 3).

This model assumes that under placebo, depression represented by the variable D changes with time t according to the following equation:

$$D(t) = D_0 - Drem \cdot \left(e^{-Krel(t) \cdot t} - e^{-(Krem(t) \cdot t)^{\gamma}} \right) \tag{3}$$

Accepted Manuscript

where D_0 denotes the value of the latent variable at baseline and the magnitude of the placebo effect is characterised by $Drem$ (latent variable unit). The median time to remission (year) is $Trem_0$ and the scale parameter ($Krem_0$) is calculated as $\frac{1}{Trem_0} \times log(2)^{\frac{1}{\gamma}}$. The half-life (year) of relapse is $Trel_0$ and the associated rate ($Krel_0$) is $\frac{1}{Trel_0} \times log(2)$. The shape parameter is γ and controls the sigmoidicity (curvature), reflecting how steeply depression changes over time. When $\gamma = 1$, this model reduces to the inverse Bateman function.

The individual parameter for subject i was described as $\theta_i = \mu * \exp(\eta_i)$ for Drem, Trem, Trel and γ, and $\theta_i = \mu + \eta_i$ for D_0, where μ denotes fixed effects and η_i denotes random effects assumed to be distributed as $\eta_i \sim N(O, \Omega)$ with variance-covariance matrix Ω.

Drug effect

Individual dose regimens were recorded for all studies, but there were no blood samples collected. In order to take into account agomelatine pharmacokinetics and their influence on the evolution of depression, we used the K-PD model proposed by Jacqmin et al. [47]. This model assumes that the kinetics of the drug can be summarised in a virtual compartment, called biophase, in equilibrium with the dynamics of the PD marker it acts upon. The elimination rate from the biophase can be inferred from the dynamics of the marker. Assuming a bolus input, the virtual concentration CE in the effect compartment can be modelled using the following differential equation:

$$\frac{dCE}{dt} = -\frac{log(2)}{Teq} \times CE \tag{4}$$

This model has a single parameter, Teq, representing the half-life of elimination from the biophase.

The best model of the drug effect assumed a linear model of the virtual concentration, with the compound acting on the remission scale (slope: α_{Trem}) and relapse half-life (slope: α_{Trel}) as follows:

$$Krem(t) = Krem_0 \left(1 + \alpha_{Trem} \times CE\left(t\right)\right) \tag{5}$$

$$Krel(t) = Krel_0 \left(1 - \alpha_{Trel} \times CE\left(t\right)\right) \tag{6}$$

Dropout model

The baseline hazard model was modelled by a Weibull function, and the link between the latent variable and the hazard was best described using a weighted difference between the current and baseline value of the latent variable.

$$h_i(t|\theta_i) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \times \exp\left(\beta \cdot f\left(t|\theta_i\right)\right) \tag{7}$$

with $f\left(t|\theta_i\right) = \left(1 - \frac{D(t|\theta_i) - D_0}{Drem}\right)$.

Refining the model

Goodness of fit plots, including Visual Predictive Checks (VPC), were produced to evaluate the intermediate models. The un-stratified total score VPC showed a strong underestimation of the effect, both in the remission and relapse phases. The plots were then stratified by treatment sequence and dose adjustment (at week 2). They showed overestimation of HAMD-scores during the open treatment period in [Goodwin2009]. Also, a delay was apparent in the HAMD decrease for subjects who required a dose adjustment after two weeks. These factors were included in the model as follows.

First, the open treatment period was modelled as a symptomatic effect with a temporary improvement:

Accepted Manuscript

$$
\begin{aligned}
D(t) &= D(t) - \beta_{open}\left(1 - \exp\left(-\frac{log(2)}{T_{open}} \cdot t\right)\right) && \text{if } t \leq t_{rand}, \\
D(t) &= D(t) - \beta_{open}\left(1 - \exp\left(-\frac{log(2)}{T_{open}} \cdot t_{rand}\right)\right) \times \exp\left(-\frac{log(2)}{T_{open}} \cdot (t - t_{rand})\right) && \text{else}
\end{aligned} \tag{8}
$$

where t_{rand}, corresponding to the randomisation time, was a design variable (8 or 10 weeks). Two parameters were estimated in this model: T_{open} which represents the half-life of improvement and β_{open} the maximal improvement from $D(t)$. Moreover, the primary outcome in [Goodwin2009] study was the relapse so patients who dropped out in the open treatment period were not included in the study. Thus for this specific study, the risk to dropout was null during this period.

Second, the dose adjustment after two weeks was handled by defining a binary covariate with value 1 for the subjects who required a dose change. As these patients did not experience an improvement in the level of depression by the end of the second week, we modelled this by adding a fixed lag-time of two weeks, and introducing a covariate effect on the remission scale, as follows:

$$
Krem(t) = Krem(t) \times (1 - \beta_{adj}) \tag{9}
$$

Parameter estimates

Parameters of the main models and their relative standard errors are reported in table III. The parameter estimates for each item of the IRT model can be found in supplementary material (Table SI). All parameters were precisely estimated with very small estimation errors reported by NONMEM [55].

Model evaluation

Fig. 2 shows the item characteristic curves for the 17 items of the HAMD scale. They are compared to a GAM with cross-validated cubic spline which makes no assumption on the shape of the models. The probability for a score for each item over depression level is similarly well described by the smooth regression and the IRT model, which supports the use of this model.

Model adequacy was assessed both in the building and in the evaluation dataset. The predictive performances on the total HAMD score of the longitudinal model are shown in Fig. 3. The plots are stratified by dose adjustment (rows) and treatment sequence (column). In each panel, we see that the 5th, median and 95th percentiles of the observation were mostly included in the prediction interval of these percentiles in simulation, representing a satisfactory ability of the model to describe the data. Un-stratified VPC (not shown) indicate the model was able to predict the drop-out rates. However, stratified VPC in Fig. 4 shows under- and over- predictions of the drop-out rates notably in the TO+P (top row) and the P+T (bottom row) groups with up to 0.1 points of difference.

Model application

Impact of dropout on the computation of the difference from placebo

We simulated 100 clinical trials using the final model to evaluate the expected difference between treated and placebo arm. The associated difference between active and placebo treatment, shown in Fig. 5 (left), reflects the beneficial effect of agomelatine. The maximum beneficial effect of agomelatine reaches 5.2 points (90% predicted interval: [4.5, 5.9]) after 12.7 weeks with an associated dropout-corrected SMD of 0.77 ([0.68,0.86]). The median difference at 6 weeks is 4.4 points on the HAMD-17 total score and the dropout-corrected SMD equals to 0.70 ([0.61, 0.80]).

Accepted Manuscript

However, because of the dropout, good responders in placebo and active groups are over-represented over time due to the MNAR process, reducing this difference. To mimic this process, we simulated the time to dropout for each subject and used the same longitudinal profiles up to the individual time to dropout. The median difference between agomelatine and placebo treatment is shown in Fig. 5 (right). The beneficial effect of agomelatine reaches a maximum at 6 weeks with a median difference of 3.1 points on the HAMD-17 scale (90% predicted interval: [2.7, 3.7]) and a SMD equals to 0.58 ([0.49,0.68]) but decreases afterward.

Sensitive subset of items

We used equations 3 and 25 (in supplementary materials) to convert total HAMD-17 scale into the scale of the latent variable. The depression levels measured using cutoffs on the HAMD-17 score translate on average to the following categories on the scale of the latent variable: no depression ($D < -4.9$), mild depression $D \in [-4.9, -2.5[$, moderate depression $D \in [-2.5, -0.95[$ and severe depression ($D \geq -0.95$).

The informativity content of the most representative items by disease severity is shown in Fig. 6. The more informative an item is, the more sensitive it is to detect a small change in the level of depression. Overall, 9 items accounted for at least 75% of the information, meaning that they were the most sensitive to detect a change in the depression level. The symptoms "Depressed mood" combined with "work and activities" represented around 40% of the information, suggesting a high link with the severity of the depression. Informativeness can change over the severity of the disease; for example "suicide" was more informative in case of severe depression, whereas "general somatic symptoms" was more informative for patients in remission (no depression).

Discussion

Depression is a common and complex disease with a strong impact on quality of life. Major depression (MDD) is characterised clinically by a persistence of severe symptoms over several weeks, and tends to evolve in a recurrent condition with cyclic patterns of improvement and relapse. To assess the depressed status of patients, clinicians use the HAMD-17 scale, a composite score reflecting the responses to a series of questions covering mood and feelings and covering anxiety, melancholic and psychotic features of the disease. The clinical outcome in trials on antidepressants typically focuses on a comparison between HAMD scores after 6 weeks of treatment, but characterising the time-course of depression may prove more informative than considering the outcome at a single time point[56]. In this work, we analysed data from 5 clinical trials on agomelatine using item-response theory (IRT) to account for each item of the HAMD-17, while simultaneously modelling changes in underlying depression status and the probability to drop out. Our goal was first to describe the time-course of depression in a population of patients with MDD, and second to evaluate the informativeness of the items of the HAMD-17 scale.

The IRT framework has been used mainly to calibrate and combine different aspects of a disease into a single score, in order to define the severity of a multifaceted disease. Ueckert et al. pioneered its use in pharmacometrics when they applied IRT to analyse the ADAS-cog score which measures cognitive decline in patients suffering from Alzheimer's disease [21]. By using a latent variable reflecting underlying cognitive decline and influencing the individual assessments, they showed that disease evolution is better characterised, allowing a greater power to detect the effect of a treatment [22, 24]. In the present study, an additional challenge was to take into account the specificities of clinical trials, such as the dropout mechanism associated with patient relapse, the possibility to have a dose adjustment if there is no improvement at the 2nd week and an open treatment period. The present study represents the first application of IRT in depression, and allowed to characterise the time course of the disease in placebo/active treatment in the context of a randomised clinical trial, through a latent variable representing the depression status.

Accepted Manuscript

The model describing the evolution of the latent variable symbolising depression was driven by the available data. The pattern of placebo response was compatible with the evolution of patients in real life: in this work, we assumed a flexible model with an initial decrease in the level of depression followed by a relapse with a much longer time-frame. Indeed, most MDD patients recover from their depressive episode within the first year. Some of them remain in a state of remission while others alternate between episodes of depression and remission [6]. We showed that remission is faster and relapse is slower if patients are on active therapy, reflecting the known response to agomelatine. Indeed, it has been shown in the literature that agomelatine has a dual effect on melatonin and serotonergic receptors patients [7] resulting in a faster remission. The exact mechanism of how agomelatine decreases relapse has not been elucidated; however it was observed in a rat model of chronic mild stress that the antidepressant effects of agomelatine were still observed on week after cessation of agomelatine treatment [57] suggesting that downstream agomelatine activation mechanisms (i.e. neuroplasticity, increase on trophic factors including brain-derived neurotrophic factor) could have a main contribution to this effect. Several models have been proposed in the literature [51, 52, 53, 54] to characterise the evolution of depression. Russu et al. presented a way to integrate the flexible dosing designs in the model, but all of these models aimed to describing the HAMD-17 score time course. Our structural model of the placebo response is based on the model of Holford [52] and integrates the model used by Gomeni for the remission. For the drug effect model, Mould described the drug effect evolution as a symptomatic effect, i.e. a temporary evolution from placebo until the end of treatment, which corresponds to the second model we tested. It has been shown that the time spent in a remission period and the history of recurrent depressive episodes were predictive of the risk of additional episodes [2]. This suggests an alternative model where relapse could be modelled using a time-to-relapse model. However, the study would require a very long observation period as cycles in depression may occur over periods of years, which would be infeasible for clinical trials. To allow for some patients not relapsing, we attempted to use a mixture model on the parameter controlling relapse rate, but we could not identify a subpopulation with near to zero relapse.

In the first stage of model building, we tested different combinations of longitudinal model, dropout model and link function to select the one which best explained the available data. Because of the relationship between dropout and evolution of depression status, a sequential analysis considering only the evolution of depression without taking into account different dropout rates would not have been appropriate. For the IRT model, we assumed that the latent variable affected all the symptoms of the depression captured by the HAMD-17 scale, while the difficulty of the different items reflected their association with the levels of depression. Applications of this methodology in other therapeutic areas such as Parkinson's disease [30] suggested that several latent variables may be introduced in multifaceted pathologies, each of them being associated with different symptom groups. We evaluated this possibility here using the method proposed by Gottipati et al. [30] which looks at correlations between items on the basis of standardised residuals. We found that the correlations were small (Fig. S6). Despite the multidimensional nature of the HAMD scale, we therefore chose the most parsimonious model with a single latent variable.

The model selected by combining the longitudinal and dropout components provided a first description of the data, but evaluation graphs showed discrepancies when stratified according to treatment group. In particular, we noticed a systematic difference between non-responders, who were prescribed an increase in dose, and responders who continued on the same regimen. This assessment was performed early during the clinical trial, as it took place after 2 weeks while the entire study duration was 6 months or 1 year. This was included, as in the work of Russu et al., in the model as an a priori covariate, here delaying and slowing the remission phase. To take into account this information in their model, Russu et al. assumed that parameters associated with remission and relapse take distinct values before and after the time of dose change, but this model was not identifiable with our data. Another discrepancy between model predictions and observations was apparent for study [Goodwin2009], which started with an open treatment period of 8 to 10 weeks. Accounting for the fact that subjects had a faster remission when they knew they were being

Accepted Manuscript

given an active treatment improved the model. We estimated that this effect, which can be considered as an expectancy effect due to the open label design, is about 166% of the drug effect at that time (for patients who don't have a dose escalation). This result must however be interpreted cautiously as the expectancy effect we observed in a controlled clinical trial may not necessarily be extrapolated to a general population. The quantification of those effects which are much higher than agomelatine effect shows their importance in depressed patients and the interest of lifestyle or psychotherapy interventions.

One of the challenges of this work was to account for dropout. As the dropout is mainly related to the level of the depression, values are missing not at random. Several models for dropout were tested, and the model selected involved MNAR depending on the improvement of depressed status relative to the baseline value, which was jointly modelled with the longitudinal evolution of the depression. Note that the depression level is modelled as an unobserved latent variable inferred by the changes in the items of the HAMD-17 score. One issue when parametrically modelling the dropout is that it can be sensitive to the assumptions made, for instance concerning the shape of the baseline hazard model. To evaluate these assumptions, we looked at visual predictive checks. Given the complexity of the data, with several treatment sequences and studies with specific design characteristics, we considered that the adequacy between the prediction and the Kaplan-Meier estimation of the dropout was acceptable, despite a trend toward a small overestimation of the dropout for patients in a treatment period. We tried to fix this issue by adding an impact of the treatment sequence on the dropout without success. In our model, the dropout process was essentially driven by the latent variable, and dropout due to adverse event or non-medical reasons would not be well captured without adding competing dropout mechanisms. Another limitation is that we did not take into account the dropout instigated by the clinicians based on the CGI-I score which was implemented in some studies at weeks 6, 10 and 12. Although this is clearly a limit and a potential source of bias in our modelling, the impact on the drop-out seems to be minimal (up to 0.1 points of difference) which would not raise serious doubts about the predictions of the disease's evolution.

Data-splitting allowed predictive performance to be evaluated more realistically, and could be performed here given the large number of observations. The predictive performance of the model in the evaluation dataset was good. Describing the placebo and active treatment response jointly with the risk to dropout yielded more accurate predictions from the model.

We estimated the magnitude of the effect of agomelatine versus placebo by simulating clinical trials using the model developed and assessing the difference in HAMD score at week 6 between the two arms. A difference in HAMD-17 at week 6 (42 days) is considered a common statistical endpoint in depression clinical trials. The effect estimated here (SMD: 0.58) was however higher than was previously reported for agomelatine in the various meta-analyses performed on the drug (SMD: 0.26 [58], 0.18 [59], 0.24 [10], 0.26 [9]). This was mostly due to the limited number of studies included here compared to the other meta-analysis and also driven by the study [Kennedy2014]. The five trials combined in the present analysis are a subset of the clinical trials conducted on agomelatine and presented in [10]. Trials were selected based on a similar clinical management of the placebo arms, as well as the availability of individual HAMD scores to allow longitudinal modelling of the evolution of depression with IRT. However, the difference between placebo arm and treated arm in [Kennedy2014] was larger than in the other studies with agomelatine, and is consistent with the SMD of 0.7 reported in [10] for that particular study and in the meta-analysis performed by [9]. From the time-course of HAMD-17 score in the different studies, this appears to be due to a slower remission in the placebo group in [Kennedy2014], whereas the impact of the drug in the treated group was similar when compared to the other studies (see Fig. 1). A more conservative estimate of 0.26 for the effect size has been reported in the meta-analysis by [9] and reflects the significant but modest effect of agomelatine."

We performed additional simulations using the model to assess the benefit of agomelatine when taking into account differential dropouts between treatment arms. Indeed, dropout is a censoring mechanism and the remaining patients are those who have responded most effectively to placebo or active treatment. Thus

Accepted Manuscript

the difference from placebo is biased since it does not reflect the actual one. Characterising the dropout and jointly estimated the IRT sub-model and the latent variable allows to have an unbiased estimation of the parameters, and thus to offer the possibility to simulate individual uncensored profiles. These simulations showed a difference of 4.3 points on the HAMD-17 scale at the 6th week compared to 3.1 points taking into account patient dropout. As illustrated here in the context of depression, taking dropout into account in the modeling allows to access to the actual value in these clinical trials of the drug effect. This shows the benefice of this framework for the design of future clinical trial with more powerful statistical tools.

One of the advantages of modelling each item using IRT is that we can evaluate which items are sensitive to a change in the depression level, and thus are sensitive to differentiate between active and placebo treatment. Since the 60s, the 17-item version of the HAMD has become the standard for clinical trials with a widespread use. Limitations have been largely discussed and mainly concern a lack in the identification of a change in the severity of the disease [17], the fact that the items measure different constructs, the presence of unequal weights attributed to different symptoms domains and the failure to include all symptom domains such as reverse neurodegenerative symptoms [60]. Thus, several scales have been developed including Bech [16], Maier[19], Gibbons[17], Toronto[20] subscales. The top 9 items we found to be most informative are all part of these subscales. The strength here is that the association between symptoms and diseases is quantitatively assessed. Several publications showed that this methodology increases the chances of detecting a difference between treatments compared to an analysis based on the total score. In their example, Buatois et al. [22] found that IRT methodology in the context of NLMEM is at least as powerful as using a subscale. As presented by Ueckert [50], this is an ideal framework to include other outcomes as the association is quantified and would allow a better characterisation of the level of severity of the disease. It would be very interesting for future work to integrate different measures such as neurodegenerative symptoms or the CGI scale. Also given the complexity of such models, a simplified subscale with a single factor structure could be envisaged.

Conclusion

In conclusion, the evolution of the depression level in five phase III clinical trials was analysed by combining IRT and pharmacometric modeling. The joint framework with integration of the dropout considers the symptoms of the HAMD-17 scale, the disease progression and the drug effect which were more accurately assessed by integrating and modelling the dropout. This allowed us to have access to the actual therapeutic effect in clinical trial setting. The obtained sensitive set of items was in accordance with the literature and through the IRT methodology, the link between symptom and disease was quantitatively assessed.

Acknowledgements

Marc Cerou received funding from Institut de Recherches Internationales Servier, as part of a PhD research fellowship program. The authors thank Valérie Olivier, Pierre-François Penelaud and Cécilia Gabriel Gracia for their clinical insight and challenging discussions. We would like to thank also Donato Teutonico and Karl Brendel for their valuable contribution to this work as well as Hervé Le Nagard and Lionel de la Tribouille for the use of the computer cluster services hosted on the "Centre de Biomodélisation UMR1137". This work is also indebted to the investigators in the [Goodwin2009], [Olie2007], [Kennedy2006], [Kennedy2014] and [Heun2013] studies.

Accepted Manuscript

References

[1] WHO. World Health Organisation; 2018. Available from: https://www.who.int/en/news-room/fact-sheets/detail/depression.

[2] Solomon DA, Keller MB, Leon AC, Mueller TI, Lavori PW, Shea MT, et al. Multiple recurrences of major depressive disorder. Am J Psychiatry. 2000;157(2):229–233.

[3] Kessler RC, Bromet EJ. The epidemiology of depression across cultures. Annu Rev Public Health. 2013;34:119–138.

[4] Weissman MM, Bland RC, Canino GJ, Faravelli C, Greenwald S, Hwu HG, et al. Cross-national epidemiology of major depression and bipolar disorder. JAMA. 1996 Jul;276(4):293–299.

[5] Van de Velde S, Bracke P, Levecque K. Gender differences in depression in 23 European countries. Cross-national variation in the gender gap in depression. Soc Sci Med. 2010 Jul;71:305–313.

[6] Bentley SM, Pagalilauan GL, Simpson SA. Major depression. Med Clin North Am. 2014;98(5):981–1005.

[7] De Bodinat C, Guardiola-Lemaitre B, Mocaër E, Renard P, Muñoz C, Millan MJ. Agomelatine, the first melatonergic antidepressant: discovery, characterization and development. Nat Rev Drug Discov. 2010;9(8):628.

[8] Racagni G, Riva MA, Molteni R, Musazzi L, Calabrese F, Popoli M, et al. Mode of action of agomelatine: synergy between melatonergic and 5-HT2C receptors. World J Biol Psychiatry. 2011;12(8):574–587.

[9] Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet. 2018;391(10128):1357–1366.

[10] Taylor D, Sparshatt A, Varma S, Olofinjana O. Antidepressant efficacy of agomelatine: meta-analysis of published and unpublished studies. Br Med J. 2014;348:g1888.

[11] Khan A, Detke M, Khan SR, Mallinckrodt C. Placebo response and antidepressant clinical trial outcome. J Nerv Ment Dis. 2003;191(4):211–218.

[12] Iovieno N, Papakostas GI. Correlation between different levels of placebo response rate and clinical trial outcome in major depressive disorder: a meta-analysis. J Clin Psychiatry. 2012;73(10):1300–6.

[13] Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23(1):56–62.

[14] Faries D, Herrera J, Rayamajhi J, DeBrota D, Demitrack M, Potter WZ. The responsiveness of the Hamilton depression rating scale. J Psychiatr Res. 2000;34(1):3–10.

[15] Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? Am J Psychiatry. 2004;161(12):2163–2177.

[16] Bech P, Rafaelsen O. The use of rating scales exemplified by a comparison of the Hamilton and the Bech-Rafaelsen Melancholia Scale. Acta Psychiatr Scand. 1980;62(S285):128–132.

[17] Gibbons RD, Clark DC, Kupfer DJ. Exactly what does the Hamilton depression rating scale measure? J Psychiatr Res. 1993;27(3):259–273.

[18] Montgomery SA, Åsberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979;134(4):382–389.

Accepted Manuscript

[19] Maier W, Philipp M. Improving the assessment of severity of depressive states: a reduction of the Hamilton Depression Scale. Pharmacopsychiatry. 1985;18(01):114–115.

[20] McIntyre R, Kennedy S, Bagby RM, Bakish D. Assessing full remission. J Psychiatry Neurosci. 2002;27(4):235.

[21] Ueckert S, Plan EL, Ito K, Karlsson MO, Corrigan B, Hooker AC, et al. Improved utilization of ADAS-cog assessment data through item response theory based pharmacometric modeling. Pharm Res. 2014;31(8):2152–2165.

[22] Buatois S, Retout S, Frey N, Ueckert S. Item response theory as an efficient tool to describe a heterogeneous clinical rating scale in de novo idiopathic Parkinson's disease patients. Pharm Res. 2017;34(10):2109–2118.

[23] Bock RD. A brief history of item theory response. J Educ Meas. 1997;16(4):21–33.

[24] Verma N, Beretvas SN, Pascual B, Masdeu JC, Markey MK. New scoring methodology improves the sensitivity of the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) in clinical trials. Alzheimer's Res Ther. 2015;7(1):64.

[25] Schneider LS, Kennedy RE, Wang G, Cutter GR. Differences in Alzheimer disease clinical trial outcomes based on age of the participants. Neurology. 2015;84(11):1121–1127.

[26] Dowling NM, Bolt DM, Deng S. An approach for estimating item sensitivity to within-person change over time: An illustration using the Alzheimer's Disease Assessment Scale–Cognitive subscale (ADAS-Cog). Psychol Assess. 2016;28(12):1576.

[27] Novakovic AM, Krekels EH, Munafo A, Ueckert S, Karlsson MO. Application of item response theory to modeling of expanded disability status scale in multiple sclerosis. AAPS J. 2017;19(1):172–179.

[28] Vandemeulebroecke M, Bornkamp B, Krahnke T, Mielke J, Monsch A, Quarg P. A longitudinal item response theory model to characterize cognition over time in elderly subjects. CPT Pharmacometrics Syst Pharmacol. 2017;6(9):635–641.

[29] Välitalo P, Krekels E, van Dijk M, Simons S, Tibboel D, Knibbe C. Morphine Pharmacodynamics in Mechanically Ventilated Preterm Neonates Undergoing Endotracheal Suctioning. CPT Pharmacometrics Syst Pharmacol. 2017;6(4):239–248.

[30] Gottipati G, Karlsson MO, Plan EL. Modeling a Composite Score in Parkinson's Disease Using Item Response Theory. AAPS J. 2017;19(3):837–845.

[31] Krekels E, Novakovic AM, Vermeulen A, Friberg LE, Karlsson MO. Item response theory to quantify longitudinal placebo and paliperidone effects on PANSS scores in schizophrenia. CPT Pharmacometrics Syst Pharmacol. 2017;6(8):543–551.

[32] Guk J, Chae D, Son H, Yoo J, Heo JH, Park K. Model-based assessment of the benefits and risks of recombinant tissue plasminogen activator treatment in acute ischaemic stroke. Br J Clin Pharmacol. 2018;84(11):2586–2599.

[33] Chae D, Park K. An item response theory based integrated model of headache, nausea, photophobia, and phonophobia in migraine patients. J Pharmacokinet Pharmacodyn. 2018;45(5):721–731.

[34] Baker FB. The basics of item response theory. ERIC; 2001.

[35] Gomeni R, Lavergne A, Merlo-Pich E. Modelling placebo response in depression trials using a longitudinal model with informative dropout. Eur J Pharm Sci. 2009;36(1):4–10.

Accepted Manuscript

[36] Goodwin GM, Emsley R, Rembry S, Rouillon F. Agomelatine prevents relapse in patients with major depressive disorder without evidence of a discontinuation syndrome: a 24-week randomized, double-blind, placebo-controlled trial. J Clin Psychiatry. 2009;70(8):1128–1137.

[37] Claghorn JL, Feighner JP. A double-blind comparison of paroxetine with imipramine in the long-term treatment of depression. J Clin Psychopharmacol. 1993;.

[38] Mitsikostas DD, Mantonakis L, Chalarakis N. Nocebo in clinical trials for depression: a meta-analysis. Psychiatry Res. 2014;215(1):82–86.

[39] Mazumdar S, Tang G, Houck PR, Dew MA, Begley AE, Scott J, et al. Statistical Analysis of Longitudinal Psychiatric Data with Dropouts. J Psychiatr Res. 2007;41(12):1032–1041.

[40] Little RJ. Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc. 1995;90(431):1112–1121.

[41] Hu C, Sale ME. A joint model for nonlinear longitudinal data with informative dropout. J Pharmacokinet Pharmacodyn. 2003;30(1):83–103.

[42] Björnsson MA, Friberg LE, Simonsson US. Performance of nonlinear mixed effects models in the presence of informative dropout. AAPS J. 2015;17(1):245–255.

[43] Pierre Olié J, Kasper S. Efficacy of agomelatine, a MT1/MT2 receptor agonist with 5-HT2C antagonistic properties, in major depressive disorder. Int J Neuropsychopharmacol. 2007;10(5):661–673.

[44] Kennedy S, Emsley R. Placebo-controlled trial of agomelatine in the treatment of major depressive disorder. Eur Neuropsychopharmacol. 2006;16(2):93–100.

[45] Kennedy SH, Avedisova A, Giménez-Montesinos N, Belaïdi C, agomelatine study group, et al. A placebo-controlled study of three agomelatine dose regimens (10 mg, 25 mg, 25–50 mg) in patients with major depressive disorder. Eur Neuropsychopharmacol. 2014;24(4):553–563.

[46] Heun R, Ahokas A, Boyer P, Giménez-Montesinos N, Pontes-Soares F, Olivier V. The efficacy of agomelatine in elderly patients with recurrent major depressive disorder: a placebo-controlled study. J Clin Psychiatry. 2013;74(6):587–94.

[47] Jacqmin P, Snoeck E, Van Schaick E, Gieschke R, Pillai P, Steimer JL, et al. Modelling response time profiles in the absence of drug concentrations: definition and performance evaluation of the K–PD model. J Pharmacokinet Pharmacodyn. 2007;34(1):57–85.

[48] Cohen J. Statistical power analysis for the behavioral sciences 2nd edn. Erlbaum Associates, Hillsdale; 1988.

[49] Kjellsson MC, Zingmark PH, Jonsson EN, Karlsson MO. Comparison of proportional and differential odds models for mixed-effects analysis of categorical data. J Pharmacokinet Pharmacodyn. 2008;35(5):483.

[50] Ueckert S. Modeling composite assessment data using item response theory. CPT Pharmacometrics Syst Pharmacol. 2018;7(4):205–218.

[51] Gomeni R, Merlo-Pich E. Bayesian modelling and ROC analysis to predict placebo responders using clinical score measured in the initial weeks of treatment in depression trials. Br J Clin Pharmacol. 2007;63(5):595–613.

Accepted Manuscript

[52] Holford N, Li J, Benincosa L, Birath M. Population disease progress models for the time course of HAMD score in depressed patients receiving placebo in anti-depressant clinical trials. Abstracts of the XI annual meeting of the population approach group in Europe. 2002;Abstr. 311. Available from: www.page-meeting.org/?abstract=311.

[53] Mould DR. Developing models of disease progression. In: Ette EI, Williams PJ (eds) Pharmacometrics: the science of quantitative pharmacology. 2007;p. 547–581.

[54] Shang EY, Gibbs MA, Landen JW, Krams M, Russell T, Denman NG, et al. Evaluation of structural models to describe the effect of placebo upon the time course of major depressive disorder. J Pharmacokinet Pharmacodyn. 2009;36(1):63–80.

[55] Bauer JR, ICON S Development. NONMEM users guide: Introduction to NONMEM 7.3.0. Maryland: 2013;.

[56] Russu A, Marostica E, De Nicolao G, Hooker AC, Poggesi I, Gomeni R, et al. Joint Modeling of Efficacy, Dropout, and Tolerability in Flexible-Dose Trials: A Case Study in Depression. Clin Pharmacol Ther. 2012;91(5):863–871.

[57] Papp M, Gruca P, Boyer PA, Mocaër E. Effect of Agomelatine in the Chronic Mild Stress Model of Depression in the Rat. Neuropsychopharmacology. 2003;28(4):694.

[58] Singh SP, Singh V, Kar N. Efficacy of agomelatine in major depressive disorder: meta-analysis and appraisal. Int J Neuropsychopharmacol. 2012;15(3):417–428.

[59] Koesters M, Guaiana G, Cipriani A, Becker T, Barbui C. Agomelatine efficacy and acceptability revisited: systematic review and meta-analysis of published and unpublished randomised trials. Br J Psychiatry. 2013;203(3):179–187.

[60] Cusin C, Yang H, Yeung A, Fava M. Rating scales for depression. In: Handbook of clinical rating scales and assessment in psychiatry and mental health. Springer; 2009. p. 7–35.

[61] Holford N, Peace KE. Methodologic aspects of a population pharmacodynamic model for cognitive effects in Alzheimer patients treated with tacrine. Proc Natl Acad Sci U S A. 1992;89(23):11466–11470.

[62] Zimmerman M, Martinez JH, Young D, Chelminski I, Dalrymple K. Severity classification on the Hamilton depression rating scale. J Affect Disord. 2013;150(2):384–388.

[63] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2018.

[64] Lavielle M. mlxR: Simulation of Longitudinal Data; 2018. R package version 3.3.0.

Accepted Manuscript

Tables and Legend to Figures

Table I: Study design

Study	Treatment sequence (compulsory/optional period)	Number of patients	Period length (week)	Study duration (week)	Median number of visits (min-max)
[Goodwin2009]	TO+P/P	174	8*+24/20	52	13 (6-15)
	TO+T/T	164			14 (6-15)
[Olie2007]	P/T	119	6/46	52	12 (2-12)
	T/T	116			12 (2-12)
[Kennedy2006]	P/T	105	6/46	52	12 (2-12)
	T/T	106			12 (2-12)
[Kennedy2014]	P/P	141	6/18	24	8 (2-8)
	T/T	406			8 (2-8)
[Heun2013]	P/P	70	8/16	24	9 (2-9)
	T/T	148			9 (2-9)

*for some patients, the open treatment period was extended to 10 weeks. TO: open treatment period, P: Placebo period, T: Active treatment period

Table II: Characteristics of the population.

Characteristic	Median (min-max)
Weight (kg)	70.5 (36.3-210)
Body mass index (kg/m^2)	25.4 (14.2-58.2)
Body surface area (m^2)	1.79 (1.25-3.13)
Creatinine (μmol/l)	73 (19-179)
Creatinine clearance (mL/min)	96.45 (28.36-427.81)
Age (year)	48 (18-87)
Height (cm)	166 (138-195)
Smoking habits (% of no smoker, has stopped, smoker)	64-10-27
Gender (% Female)	70.5

Table III: Parameter estimates wih their relative standard errors (RSE%)

Parameter	Value (RSE%)	Between-subject variability (RSE%)
D_0	0 (-)	0.45 (4.3%)
$Drem$	8.39 (2%)	0.37 (3.0%)
$Trem$ (year)	0.16 (4%)	0.82 (1.7%)
$Trel$ (year)	1.66 (7%)	1.27 (3.4%)
γ (-)	0.88 (2%)	0.32 (3.7%)
α_{Trem}	3e-04 (11%)	1.41 (3.8%)
α_{Trel}	1e-04 (34%)	-
Teq (year)	0.06 (6%)	-
k (-)	3.29 (5%)	-
λ (year^{-1})	5.99 (11%)	-
β	12.72 (6%)	-
β_{open}	1.52 (12%)	-
β_{adj}	0.27 (12%)	-
T_{open} (year)	0.11 (13%)	-

Accepted Manuscript

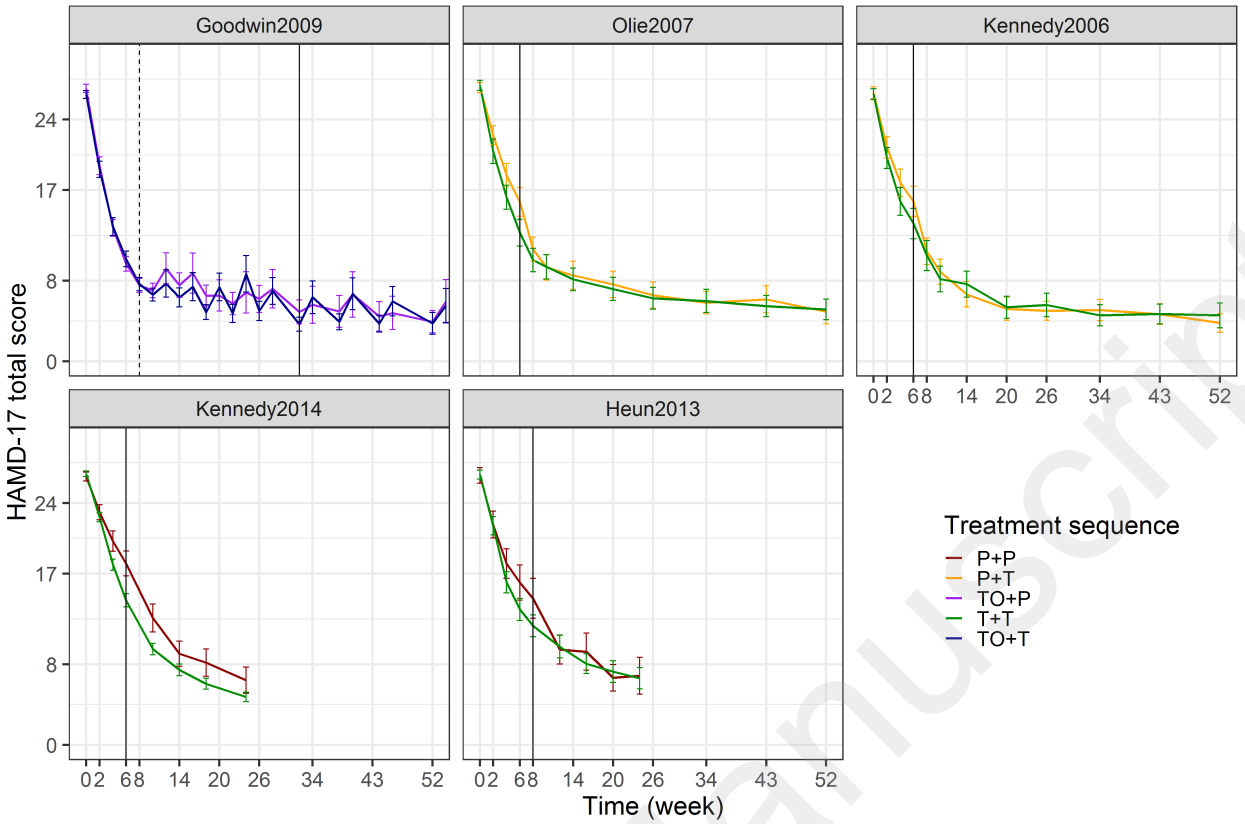

Figure 1: Mean value with the 95% confidence interval of the total HAMD-17 score, stratified by study. Each colour represents the treatment sequence. The vertical line shows the end of the compulsory period,except for study [Goodwin2009] where two bars denote respectively the end of the open treatment period (dashed line) and the end of the compulsory period (continuous line).

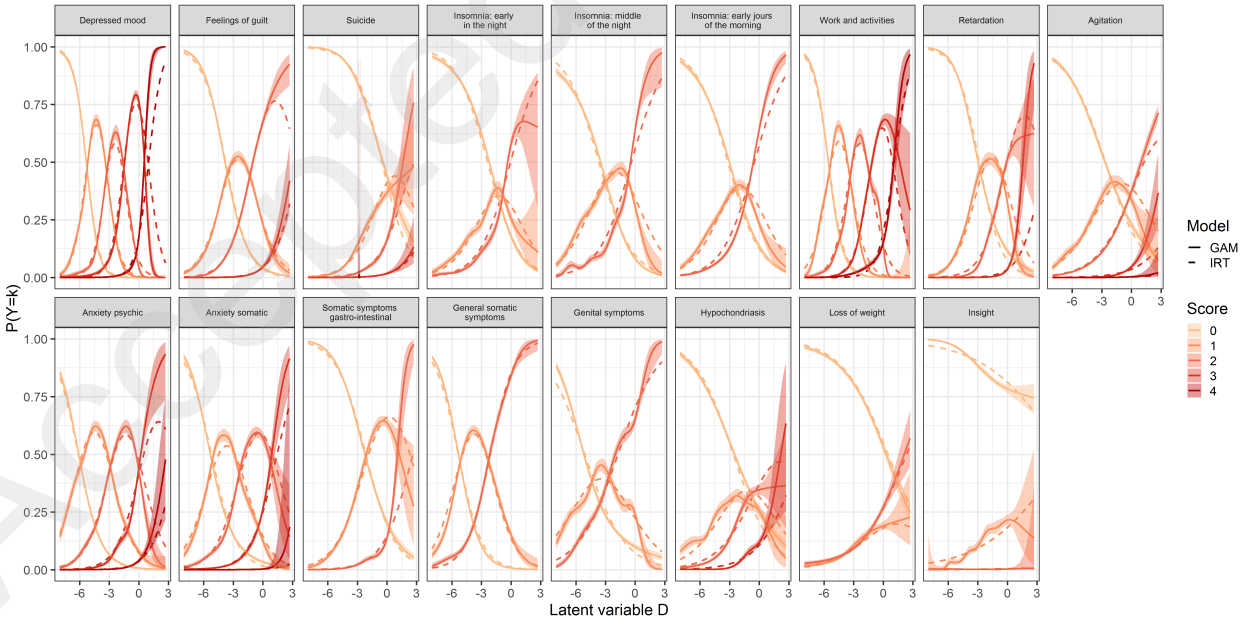

Figure 2: Probability for each score for all the items of the HAMD scale predicted by the IRT model (based on item specific parameters - dashed line) compared to the fit of a generalized additive model (GAM) with cross-validated cubic spline as a smoothing function (continuous line with 95% confidence interval)

Accepted Manuscript

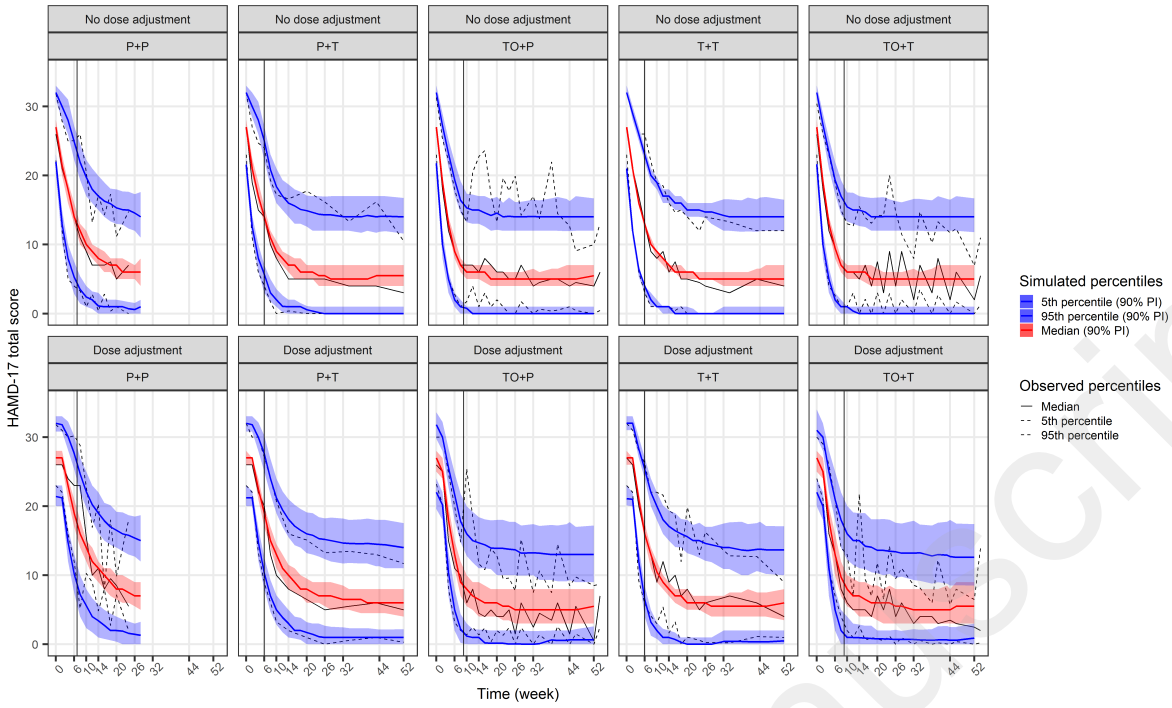

Figure 3: Internal evaluation: visual predictive check of the HAMD-17 total score stratified by dose adjustment (by row) and treatment sequence (by column). Median, 5th and 95th percentiles in the data are compared to the model simulated median, 5th and 95th percentiles (with the 90% prediction interval in shaded area). The vertical black line represents the end of the first period.

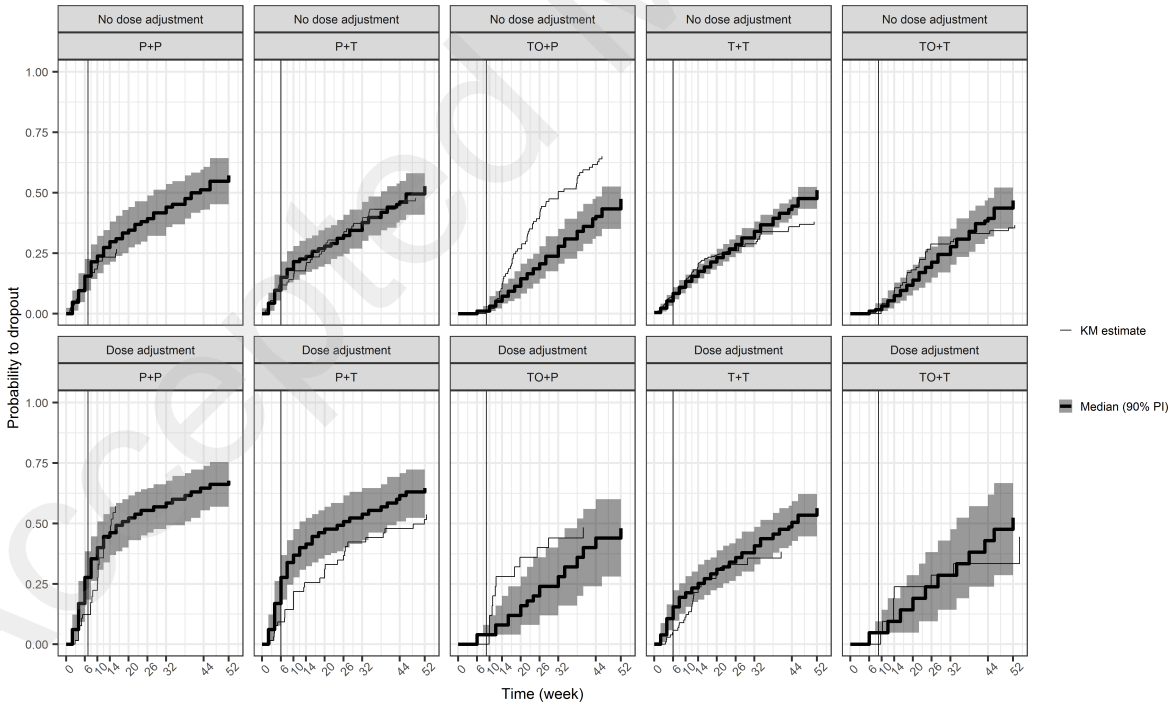

Figure 4: Internal evaluation: visual predictive check of the dropout stratified by dose adjustment (by row) and treatment sequence (by column). The Kaplan-Meier estimates based on the data (mean and 90% confidence interval) is compared to the model simulated median and the 90% prediction interval (shaded area). The vertical black line represents the end of the first period.

Accepted Manuscript

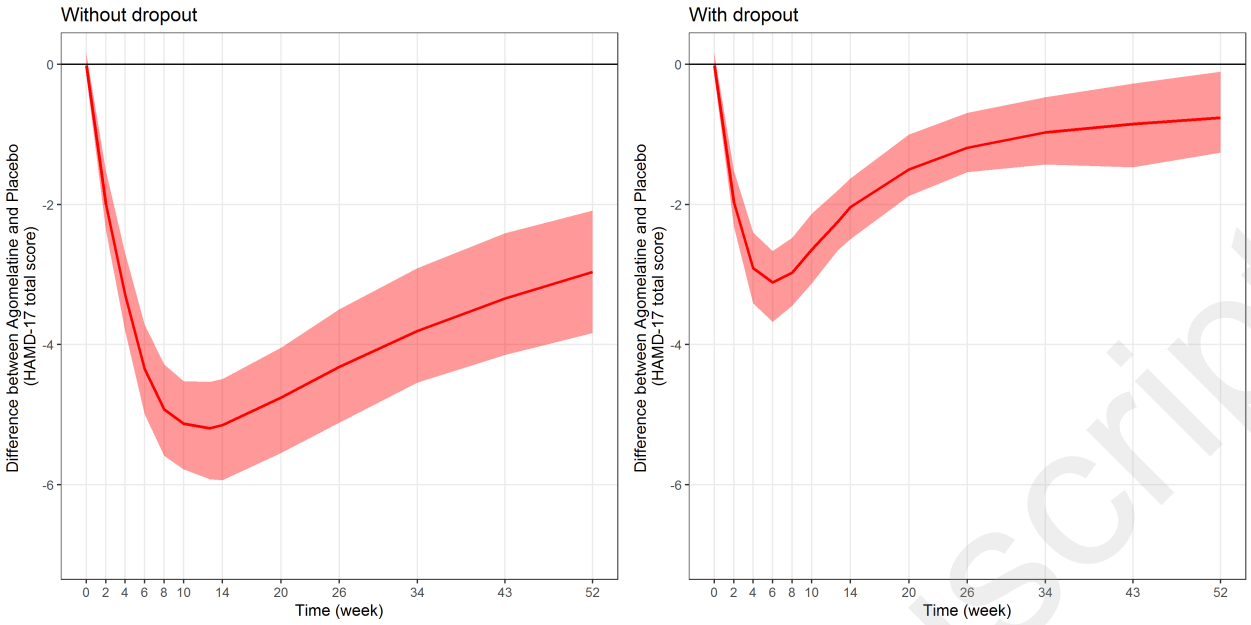

Figure 5: Predicted agomelatine improvement in HAMD in the context of a clinical trial with dropout (Left) or assuming no dropout (Right). The black continuous line represents the threshold between a beneficial (negative values: agomelatine is better) and a non-beneficial effect (positive values: placebo is better). The red continuous line represents the median difference from placebo with its 95% confidence interval (red shaded area)

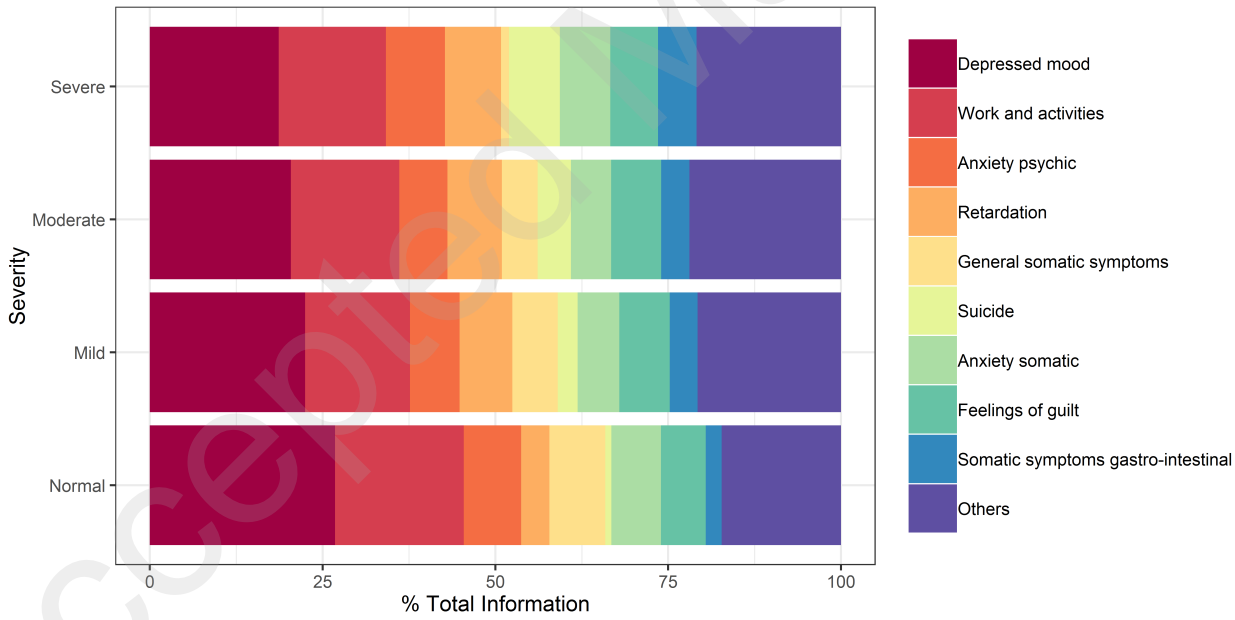

Figure 6: Contribution of each item based on the Fisher information and stratified by severity. The greater the contribution for a given item, the more differences in the corresponding score are informative on underlying depression. Only items where the fisher information represented over 5% of the total information at least for one degree of severity are included.

Accepted Manuscript

Supplementary material

Methods

Data

The data for this paper was collected in five phase III ([36, 43, 44, 45, 46]), multi-centre, placebo-controlled, double-blind studies with similar inclusion criteria. One clinical trial ([Goodwin2009]) was designed to evaluate the efficacy of agomelatine in the prevention of depressive relapse. In the other four trials, the objectives were to assess the efficacy of the drug compared to placebo after a 6-week treatment, and to provide additional safety data on agomelatine. A summary of the main characteristics of each trial can be found in table I. All studies included men and women aged between 18 and 65 years, except for study [Heun2013] where patients were over 65 years old. The main inclusion criterion was a diagnosis of moderate to severe major depressive episode according to the DSM-IV-TR criteria and requiring an antidepressant treatment. Written informed consent was obtained prior to inclusion.

In all studies, the main clinical endpoint was the value of the Hamilton Depression Rating Scale, a questionnaire composed of 17 items, at the end of the treatment period. With 3 to 5 possible answers to each item, the total HAMD-17 score can range from 0 (not depressed) to 52, with severe depression characterised as a score of more than 24. The questions are shown in Fig. S7. They evaluate different components of depression acting through sleep, anxiety, mood, somatic and psychotic disruptions. Depression status was also monitored throughout the study by recording the HAMD-17 scale at each visit. The schedule of planned visits varied across studies, but overall patients were followed every 2 weeks during the mandatory period, then visits were scheduled every 4 to 9 weeks. At each visit, CGI scale was also assessed. It was developed to briefly give the clinician's point of view by summarising all the clinician's knowledge of the patient in two items: CGI-S; severity of psychopathology from 1 to 7 and CGI-I; change from the initiation of treatment on a similar seven-point scale. A higher scores on CGI-I represents more severe depression.

All trials included a mandatory double-blind treatment period and an optional period, based on investigators and/or patients agreement. During the optional phase, patients continued to receive the treatment they had been randomised into in a double-blind manner except for patients under placebo in studies [Olie2007, Kennedy2006] who were switched to a 25 mg dose of agomelatine in the extension phase. During the double-blind periods (both mandatory and extended), patients were instructed to take two tablets orally once a day, during dinner, with a glass of water. When patients were randomised in the treatment group (non fixed 25 mg), they were assigned during the first 2 weeks to the dose 25mg. If the improvement of the patient's depressive condition was deemed insufficient at the end of the second week for all studies (even in the open phase), the dosage in agomelatine was increased from 25mg to 50mg for the rest of the study, by substituting a placebo tablet in double-blind conditions. As a result, the treatment consisted in 2 placebo tablets, agomelatine 25 mg (25 mg agomelatine tablet + 1 placebo tablet) or agomelatine 50 mg (2 x 25 mg agomelatine tablets).

In [Goodwin2009], the goal was to assess the long term efficacy through a prevention to relapse study. Patients were considered to have relapsed when HAMD-17 is greater than 16 during the relapse phase (where the severity of the disease increase). This study was the only one to begin with an open treatment period (TO) before the randomisation into the Placebo (TO+P) or Treatment group (TO+T) at week 8. If the patient was not eligible for randomisation (according to independent expert) at week 8, he/she continued the open treatment period until week 10 where a second assessment took place. Patients were withdrawn from the study if they did not fulfil the randomisation criteria. To summarise, only the good responders at week 10 are randomised. If they dropped out before randomisation, their data was not included in the analysis.

In the other four studies, the statistical analysis assessed short term efficacy by comparing the HAMD-17 score at week 6 or week 8 between active and placebo treatment. In studies [Olie2007] and [Kennedy2006],

Accepted Manuscript

patients were randomised at initiation in the placebo or treatment group; at week 6, patients in the treatment arm continued under the same dosage and all patients under placebo received 25mg of agomelatine (P+T or T+T respectively). In [Kennedy2006], only patients with CGI-I score ≤ 3 at week 10 were eligible to continue the study after this visit. In [Heun2013], patients were randomised at initiation in the placebo or treatment group, but patients continued in the same arm after 8 weeks during the extension period (P+P or T+T respectively), and the total duration of follow-up was 6 months instead of one year. However during the extension period at W12 visit, only the patients having CGI-I ≤ 2 were allowed to continue.

Study [Kennedy2014] also had, as study [Heun2013], a follow-up duration of 6 months, but patients were randomised in 4 parallel groups at treatment initiation: placebo, dose 10mg, fixed dose 25mg and dose 25mg increased up to 50mg if no improvement after 2 weeks. For [Kennedy2014], all the patients having CGI-I ≤ 3 at W6 visit were entered in the extension period with the same treatment subject to patient's agreement. Patients having CGI-I > 3 at W6 visit did not continue into the double-blind extension treatment period. Then all the patients having CGI-I ≤ 2 at W10 visit could continue in the extension period. The patients having CGI-I > 2 at W10 visit were withdrawn from the study.

Height (HT), age (AGE), smoking habits (SMOK) and gender (SEX) were recorded at inclusion. Weight (WT), body mass index (BMI), body surface area (BSA), creatinine (CREAT) and creatinine clearance (CRCL) were recorded at each visit. A summary of the characteristics at inclusion of the populations in each clinical trial can be found in table II in the main text.

Modelling strategy

The IRT model, longitudinal model and the main components of the final drug effect and dropout models are described in Results. In this appendix, we give additional details on the models tested during model building and on the model building strategy.

Drug effect

As stated in the main text, in the absence of individual pharmacokinetic measurements, agomelatine concentrations in the effect compartement, $CE(t)$, were predicted using a K-PD model based on the individual dose regimens recorded for all studies.

Patients treated for depression usually manifest an improvement in their condition, apparent in a decreased HAMD score. When the treatment is discontinued, the depression status tends to return to a similar evolution as would be observed under placebo. This has been termed a symptomatic effect, as the drug is seen to act on the symptoms of the disease and not modify its underlying course [61]. Considering the time course of depression, three specific mechanisms of action were investigated for agomelatine. The first mechanism assumes that the drug speeds up the decrease (i.e. improvement) of the depressed status D(t) represented by the remission scale $Krem(t)$, through a linear or an Emax model:

$$\begin{aligned} E(t) &= \alpha \times CE(t) \\ E(t) &= \frac{Emax \times CE(t)}{EC_{50} + CE(t)} \end{aligned} \tag{10}$$

where E was modelled as a multiplicative effect on the remission scale:

$$Krem(t) = Krem(t)(1 + E(t)) \tag{11}$$

A second mechanism assumes an immediate improvement in depression when the patient receives treatment, and is modelled as an additive effect on the placebo response through a linear or Emax model:

$$D(t) = D(t) - E(t) \tag{12}$$

Accepted Manuscript

where $D(t)$ is from equation (3) (main text).

A third mechanism is through preventing relapse, and we modelled this as a drug effect decreasing the relapse rate. Again, both linear and Imax model for the dose-effect relationship were tested.

$$\begin{aligned} I(t) &= \alpha \times CE(t) \\ I(t) &= \frac{Imax \times CE(t)}{IC_{50} + CE(t)} \end{aligned} \tag{13}$$

where I is the fraction of decrease of the relapse rate ($Krel(t)$).

$$Krel(t) = Krel(t)\,(1 - I(t)) \tag{14}$$

The parameters associated with the drug effect were assumed to follow a log-normal distribution, excepted for Imax which was assumed to follow a logit-normal distribution.

Dropout Model

During the course of the five trials 35% of patients dropped out, for lack of effect (58%), non-medical reasons (20%), adverse events (13%), remission (5%), protocol deviation (3%) or loss to follow-up (1%). In the context of depression and accounting for dose escalation, Russu et al. [56] showed that the dropout must be considered as MNAR since the main cause is a problem of efficacy and therefore a high level of disease severity. They modelled the dropout mechanism through a parametric time-to-event model to ensure unbiased estimators. Patients treated for depression tend to stop taking their medication or to drop out when they do not feel any improvement in their status, therefore the risk of dropout was assumed to be influenced by the overall depression score captured by the latent variable. The study duration acts as a censoring mechanism.

Let the random variable T denotes the time to dropout. In standard survival analysis, the survival function $S(t)$ is a meaningful measure giving at each time $t > 0$ the probability to have survived up to time t event-free. Assuming that this probability equals 1 at $t = 0$, we have:

$$S(t) = Pr(T \geq t) \tag{15}$$

The hazard function, denoted by $h(t)$, describes the instantaneous risk of having an event at time t for an individual who survived up to that time. It can be expressed as:

$$h(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt | T \geq t)}{dt} \tag{16}$$

Survival can be directly expressed as a function of the hazard (17):

$$S(t) = \exp\left(-\int_0^t h(x)dx\right) \tag{17}$$

Modelling the hazard is often more relevant than modelling survival as the hazard function can adjust throughout the study to reflect changes in disease status and treatment. Here we adopted the usual decomposition of hazard in a parametric baseline risk function $h_0(t)$ and an exponential function of risk factors:

$$h(t|\theta_i) = h_0(t) \times \exp\left(\beta \cdot f\left(t|\theta_i\right)\right) \tag{18}$$

where β is the strength of the link between the risk and a function f, depending on the individual parameters θ_i.

Accepted Manuscript

In this study, we tested different models for the hazard. First, we assumed a base model where hazard does not depend on depression status (no link, $\beta = 0$). Then, we tested various models for the link function, estimating the strength of the link through the parameter β:

- dropout at time t may be related to the depression level at that time, $f\left(t|\theta_i\right) = D\left(t|\theta_i\right)$
- dropout may be related to the cumulative burden of depression up to time t, $f\left(t|\theta_i\right) = \int_0^t D\left(x|\theta_i\right) dx$
- dropout at time t may be related to a difference between the depression level at baseline and at that time, weighted by the maximal amplitude of the placebo effect, $f\left(t|\theta_i\right) = 1 - \frac{D(t|\theta_i) - D_0}{Drem}$

To represent the baseline risk h_0, standard risk functions were used, like the Weibull ($h_0(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1}$), and Gompertz ($h_0(t) = \lambda e^{k \cdot t}$) models.

Model building strategy and evaluation

The data was split into a building and an evaluation dataset, comprising respectively 70% and 30% of the data. Data-splitting was stratified on treatment sequence and study. The model building process was performed on the learning dataset (steps 1 to 6 below), and the final model was evaluated both on the learning (internal evaluation) and then on the evaluation dataset (external evaluation). Model selection was based on the Bayesian Information Criterion (BIC) and parameter precision (relative standard error RSE%). For covariate selection, statistical relevance was evaluated using a Wald test evaluating the significance of the slope of a linear model for continuous covariates. For categorical covariates, we performed a Wilcoxon test (binary covariate), or a Kruskal-Wallis test (multi-level covariate). Model building proceeded in the following series of steps.

Step 1: initialisation

The first step was to obtain initial estimates for the item-specific parameters in the IRT model. We estimated them from the population of patients who did not receive the treatment, corresponding to data at baseline for all subjects, and data from the placebo arms. This dataset provided information on the evolution of D up to 24 weeks, which was the maximal duration of follow-up for placebo patients in [Kennedy2014, Heun2013]. The value of the latent variable at baseline was supposed to be normally distributed, and for maximum flexibility an unstructured latent variable model was used to take into account disease progression over time, estimating the mean values of D at weeks $\{2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 24\}$:

$$\begin{aligned} &D_0 \sim N(0, 1) \\ &D = D_0 - \beta_j \quad \text{with } j \text{ the } j^{th} \text{ occurrence} \end{aligned} \tag{19}$$

To evaluate the adequacy of ICC and identify misfit, we used non-parametric smoothing splines to fit the relation between the probability for each score depending on the level of the predicted value of the latent variable. We then compared the non-parametric smoothing splines with the predicted ICC from the estimated 69 item specific parameters. Generalised additive models using binomial distribution were used with cross-validated cubic splines.

Step 2: structural and variability model

The 69 item specific parameters estimated above were then fixed during the model building process (steps 2 to 4). The goal was to determine a structural KPD and dropout model describing the data. D_0 was fixed to 0 but its inter-individual variability was estimated. Since there is no consensus in the model building strategy for joint models, combinations of possible models of placebo effect, drug effect and dropout were tested. The best model amongst the 48 models tested was selected as the one with the lowest BIC.

Accepted Manuscript

The parameters for all these models were estimated assuming a diagonal variance-covariance matrix to describe the variability of the random effects. Then for the best model, a full variance-covariance matrix was estimated. Then, block matrix was built based on the full matrix but with non-diagonal elements associated with moderate correlation and adequate estimation accuracy (RSE $< 50\%$). The model with the lowest BIC was selected.

Step 3: refining the model

Goodness of fit plots, including VPC, were produced to evaluate the intermediate models. The unstratified VPC from the final model in Step 2 showed a strong underestimation of the effect, both in the remission and relapse phases. The plots were then stratified by treatment sequence and dose adjustment (at week 2). The group who received treatment in an open administration period showed consistent overprediction of the disease. Also, a delay was apparent in the HAMD decrease for subjects who required a dose adjustment after two weeks, suggesting the treatment acts more slowly in these patients or that its effect only appears after the dose increases. These factors were included in the model as follows.

First, systematic differences between open and blinded treatment periods were accounted for using an additional model. A simple way to do this is to consider open treatment as a binary covariate that has an impact on the remission scale, as in the following equations:

$$Krem(t) = Krem(t)(1 + \beta_{open}) \tag{20}$$

Alternatively, we assumed that open treatment acts like a symptomatic effect with a temporary improvement:

$$\begin{array}{ll} D(t) = D(t) - \beta_{open}\left(1 - \exp\left(-\frac{log(2)}{T_{open}} \cdot t\right)\right) & \text{if } t \leq t_{rand}, \\ D(t) = D(t) - \beta_{open}\left(1 - \exp\left(-\frac{log(2)}{T_{open}} \cdot t_{rand}\right)\right) \times \exp\left(-\frac{log(2)}{T_{open}} \cdot (t - t_{rand})\right) & \text{else} \end{array} \tag{21}$$

where t_{rand} corresponds to the randomisation end of the open treatment period and is defined by the design (8 or 10 weeks). Two parameters were estimated in this model: T_{open}, representing the half-life of improvement, and β_{open}, reflecting the maximal improvement from $D(t)$. Those two models were tested and the one with the lower BIC and best parameter precision. Moreover, the primary outcome in [Goodwin2009] was the relapse so patients who dropped out in the open treatment period were not included in the dataset. Thus only for this study, the risk to dropout was null during this period.

Second, dose adjustment after 2 weeks was handled by defining a binary covariate with value 1 for the subjects who required a dose change. As these patients did not experience an improvement in the level of depression by the end of the second week, we modelled this by adding a fixed lag-time of two weeks, and introducing a covariate effect on the remission scale, as follows:

$$Krem(t) = Krem(t)(1 - \beta_{adj}) \tag{22}$$

Third, we tested combinations of drug effect on $Trem$ and $Trel$.

Step 4: joint estimation of all parameters

During steps 2 and 3, the item specific parameters remained fixed to the estimates obtained using an unstructured model of the latent variable in step 1. Once the structural model and the impact of design elements had been defined, a simultaneous estimation of all of the parameters for the IRT/K-PD/dropout model was performed to obtain unbiased estimates of the item specific parameters, taking into account the model for the latent variable. Individual estimates of all the parameters in the model (Empirical Bayes Estimates, or EBE) were obtained and goodness of fit plots were checked for model adequacy.

Accepted Manuscript

Step 5: covariate selection

The value at inclusion of continuous covariates (WT, HT, AGE, BMI, BSA, CREAT, CRCL) and categorical covariates (SMOK, SEX) were recorded. They were tested in a Cox model for time to dropout using a stepwise procedure based on likelihood-ratio test with a threshold on the p-value of 0.05 in the forward selection and 0.001 in the backward elimination. Then they were integrated in the exponential term of the hazard function (proportional hazard). For their impact on longitudinal process, covariates were tested for their potential correlation with the EBE, assuming a linear model with a Wald test on the slope for continuous covariates, a Wilcoxon test for gender covariate, and a Kruskal-Wallis test for smoking habit covariate. To take into account multiple tests, a relationship was considered to be statistically significant only if the p-value was lower than 0.001. If the relation was both statistically significant and clinically relevant, they were integrated in the model as follows. For continuous covariates the model was:

$$\begin{array}{rl} \theta_i = \mu + \eta_i + \beta_\Lambda \times (\Lambda - median(\Lambda)) & \text{for parameters normally distributed} \\ log(\theta_i) = log(\mu) + \eta_i + \beta_\Lambda \times (\Lambda - median(\Lambda)) & \text{for parameters log-normally distributed,} \end{array} \tag{23}$$

and for categorical covariates we assumed:

$$\begin{array}{rl} \theta_i = \mu + \eta_i + \beta_{\Lambda,n} \times \mathbb{1}_{\Lambda=n} & \text{for parameters normally distributed} \\ log(\theta_i) = log(\mu) + \eta_i + \beta_{\Lambda,n} \times \mathbb{1}_{\Lambda=n} & \text{for parameters log-normally distributed} \end{array} \tag{24}$$

with $n \in \{2, \cdots, p\}$ and p the number of categories. β_Λ represents the impact of the covariate on parameters and Λ the considered covariate (e.g. Age, Sex). For categorical covariates, μ is the value for the reference and $\beta_{\Lambda,n}$ is the fractional change for other categories.

The covariates retained both in the longitudinal and dropout sub-models were then directly integrated in the full model.

Step 6: evaluation

To evaluate the best model, we performed visual predictive checks (VPCs) of the total score (HAMD-17) by comparing the observed 5th, median and 95th percentiles to the predicted ones (with the 90% prediction interval for each boundary) both in the learning and the evaluation dataset. Predictions of the total score were based on simulations of 1000 datasets on item level using the best model. For each dataset and patient, time-to-dropout was simulated and longitudinal observations (score for each of the 17-th items) according to the study design were simulated up to the time-to-dropout. The total HAMD-17 score was derived from the item-level scores in the simulations and used to plot a VPC for the clinical outcome.

Model application

Impact of dropout on the computation of the difference from placebo

Based on the model and parameter estimates, the median difference between active and placebo treatment was predicted over time by simulating individual trajectories over 1 year. 100 trials were simulated according to table SII. The sample size in each group were based on the data. In clinical trials, patients who relapse drop-out from the study and their clinical data are not recorded after this time. Thus the best responders are over-represented over time decreasing the difference. This is illustrated by using the same simulations but by taking into account only observations occurring before the dropout time, which was simulated by our model. Then the median difference between both arm was computed over time. Because of the computational burden, we only did 100 simulations, so the 90% prediction interval of this difference was computed instead of the 95% interval.

Accepted Manuscript

We obtained the predicted HAMD-17 score as a sum of the probabilities of the score for each item:

$$HAMD_{17} = \sum_{s=1}^{17} \left(\sum_{K=0}^{\{2,4\}} K \times P(Y = K) \right) \tag{25}$$

where the probability of each score over time was defined in the IRT model according to equation (2), s is the item index, and K the possible score for each item which can be in $\{0, 2\}$ or $\{0, 4\}$.

Sensitive subset of items

Another application of the model was to determine items which were the most sensitive to a change in the level of the depression. We used the Fisher information matrix and optimal design theory to determine the most informative items depending on the severity of the disease [21]. If each item is equally informative, then they each must represent 5% of the information. The information is expressed depending on the latent variable, i.e. the severity of the disease. The most influent items are defined as the items which represent more than 5% of the information at least once over the range of the latent variable.

The average proportion of informativeness of the most informative items was calculated according to different categorised levels of disease severity. To do so, we first categorised the level of severity on the latent variable scale, then we computed the average proportion in a categorised level of severity. Zimmerman et al. [62] recommended the following severity ranges for the HAMD score: no depression (0-7); mild depression (8-16); moderate depression (17-23); and severe depression ($\geq$24). The corresponding value on the latent variable scale can then be derived using eq. 25. The AUC of the information within each category was computed and the ratio between the AUC of each of the most influent items and the sum of the AUC of all the items was determined to compute the proportion of informativeness for the most influent items.

Software

Parameter estimation was performed in the software NONMEM version 7.3 [55] using the second-order conditional estimation algorithm with Laplacian approximation. NONMEM was also used to perform simulations for the computation of the VPC for the intermediate models, and R-3.5.1 [63] was used for diagnostic graphs and statistical analyses. For the final model, simulations of clinical trial was performed using the function simulx of the package mlxR [64] in R. Due to the size of the data, we used our own R script to create all the VPCs.

Results

Model building

IRT model and item-specific parameters

First, the 69 item-specific parameters in the IRT model were successfully estimated using only placebo data. At this stage, the relative standard error of these parameters could not be obtained since the covariance estimate step did not complete and bootstrap could not be performed due to the very long runtime. The comparison between the fit of a GAM or the IRT (Fig. S1) was however satisfying, supporting the use of these parameters for the next stage of model building, and these parameters were fixed for model selection.

The second step aimed to determining a structural and variability model which could describe the data sufficiently well. The parameters and objective function were estimated for the 48 models combining different features. Parameters T_{eq}, EC_{50} and IC_{50} were estimated as fixed effects without variability. At this stage, almost all of the tested models ran into computational errors and terminated due to rounding errors. All of

Accepted Manuscript

the runs were then re-estimated with updated initial parameter values. The best structural model assumed a linear model for the dose-effect relationship, with the compound acting on the remission scale. The baseline hazard model was modelled by a Weibull function, and the link between the latent variable and the hazard was best described using a weighted difference between the current and baseline value of the latent variable. The full variance-covariance matrix was tested and all correlations between the random effects were assessed. We kept the model with correlations not overly poor and maintaining adequate estimation accuracy. The final model included a block correlation between η_{Drem}, η_{Trem}, η_{Trel} and η_{α}, which best described the variability structure. At the end of this phase, the VPC for the total HAMD score showed a reasonably well estimated remission but an underestimation of the remission scale in the open treatment period. Moreover, there were an overestimation of the median and the 95th percentile for late time points. For patients who needed a dose adjustment, the HAMD-17 scores in the data did not decraese while the model, on average, predicted an instantaneous improvement. Incorporating design elements highly improved the fit. Adding a fixed lag-time of 2 weeks in the longitudinal evolution and a decreased remission scale for patients who required a dose adjustment at the second week improved the BIC by 441 points. Adding a fixed lag-time (set to the end of the open treatment period) in the hazard model and a symptomatic open treatment period effect according to (eq. 8) further improved the BIC by 173 points.

Finally, we added a second effect of the drug, assuming it also acted to slow the relapse rate, and this improved the fit again (60 points). The 95th confidence interval on this parameter, from $3.4E-05$ to $1.7E-04$, computed based on the RSE excluded 0.

The 82 parameters (i.e. 69 for IRT, 10 for the KPD submodel and 3 for the time-to-event model submodel) and the 12 components of the variance-covariance matrix were then jointly estimated. The run successfully converged, and we could obtain estimates of the standard errors for all parameters. The BIC was much lower (a drop of 2194 points) after re-estimation, with a final objective function of 242311.

The link between the covariates and the EBE was investigated using a linear relationship (expressed as a fraction of increase from reference for categorical covariates). Out of the 54 tests, only an association between age and the value at baseline of the latent variable was found to be significant ($p = 2.8e-05$). However, the magnitude of the effect was not clinically relevant with 0.4 points of difference on the total HAMD score for a patient 10 years older compared to the median age, so the age was not retained in the final model.

Parameter estimates

Except for the IRT submodel, parameters of the model and their relative standard errors are reported in table III in main text. The parameter estimates for each item of the IRT submodel can be found in supplementary Table SI.

All parameters were precisely estimated with very small estimation errors reported by NONMEM. On the other hand, the variability for some parameters, $Trel$ and α_{Trem}, was found to be quite high. The high variability for the half-life of relapse $Trel$ can be due to the fact that some patients will relapse (short half-life), and some will stay stable at low values of depression status (very high half-life). As a consequence, the estimated time to relapse for a typical individual is very high, greater than the maximal duration of clinical trials of one year. This is confirmed by the individual distribution of the parameter Trel conditional on having relapsed or not (dropout) which indicates a significant difference in the distribution mode. Indeed, the relapse half-life mode of patients who dropout from the study is 0.25 years while that of patients who complet the study is 4.25 years. We attempted to use a mixture model to account for this but this was not successful. D_0, $Drem$, α_{Trem}, α_{Trel}, and β_{open} are variables on the logit scale, i.e. the scale of the latent variable. The typical individual starts at $D_0 = 0$ and tends to the estimated $Drem$ value (-8.39 points) if there is no relapse. For patients under placebo, the "half-life" of the remission is equal to 0.16 year (slightly less than 2 months) and the half-life of the relapse is estimated to be about 1.66 year. The drug acts both by increasing the scale of the remission, implying a faster improvement (24% of increase for the 25mg dose),

Accepted Manuscript

and by reducing the slope of relapse (8% of decrease for the 25mg dose), so that it takes longer to relapse.

To better illustrate the effect of drug treatment and dose adjustment according to the model and equation 25, we predicted in Fig. S2 the evolution of the total HAMD-17 score for a typical subject under placebo/active treatment, experiencing or not a dose adjustment at the second week; we also considered whether the treatment was blind or open. After 6 weeks, the mean HAMD-17 score for patients who don't have a dose adjustment at week 2 was 13.2 versus 18.5. This is due both to the delayed improvement due to the lag-time of 2 weeks included in the model, and to the remission scale, reduced by a factor of 27% for patients who have a dose increase. The impact of the open treatment period was modelled through a temporary improvement (symptomatic effect) which increases over time. For a typical patient without dose adjustment, the difference with the reference (placebo/treatment) at baseline was null. The additional predicted open effect at the 8th week (randomisation week) was equal to 3.15 points which corresponded to 166% of the drug effect to that time (difference active/placebo = 1.9 points). At week 8, patients were randomised into placebo or active arm. The figure shows the predicted residual effect of the open label period and how the curve tends to reference. The half-life of this decrease equals 5.72 weeks, indicating that this residual effect disappears after 30 weeks and is negligible (less than 1 point difference) 16 weeks after the end of the open label period.

The dropout was modelled with a Weibull function for the baseline hazard with an increasing risk (shape parameter is over 1). If the latent variable is assumed not to have an impact on the hazard ($\beta = 0$), the probability to drop out at one year roughly equals 0, suggesting that the actual dropout is mainly driven by a change from baseline of the latent variable. We estimate the hazard ratio for a decrease of 5 point on the total HAMD-17 score to be equal to 8.35. This leads to a 20% decrease in dropout at 1 year.

Model performance

The ability of the model to represent the data and describe different individual profiles is shown in Fig. S3. The time course of the individual observed and predicted HAMD-17 scores from equation 25 shows the good adequacy of the model which can describe for example a non-relapser patient as well as a relapser. In each panel, two individuals were randomly sampled.

The predictive performances of the longitudinal model on the evaluation dataset (30% of the data) are shown in Fig. S4. Only three patients were in the group with a dose adjustment and treatment sequence $TO + T$ so the graphs from this panel are omitted. In general, the evaluation shows a reasonably adequacy between the predictions and the observations. However, for the some strata, the model tends to overestimate the level of depression for time after 6 months (eg 'No dose adjustment' & 'TO+P') which was not the case in the internal evaluation. This can be due to the low number of patients in this group (40 at the beginning) and the high rate of dropout which was over 50% after 6 months.

Fig. S5 illustrates the performance of the dropout model in the evaluation dataset. Results are in accordance to the VPC for the interval evaluation: the model tends to overpredict (Dose adjustment & P+T / TO+P / T+T) or underpredict (No dose adjustment & TO+P) the dropout in some groups.

Accepted Manuscript

Complementary Tables and Figures

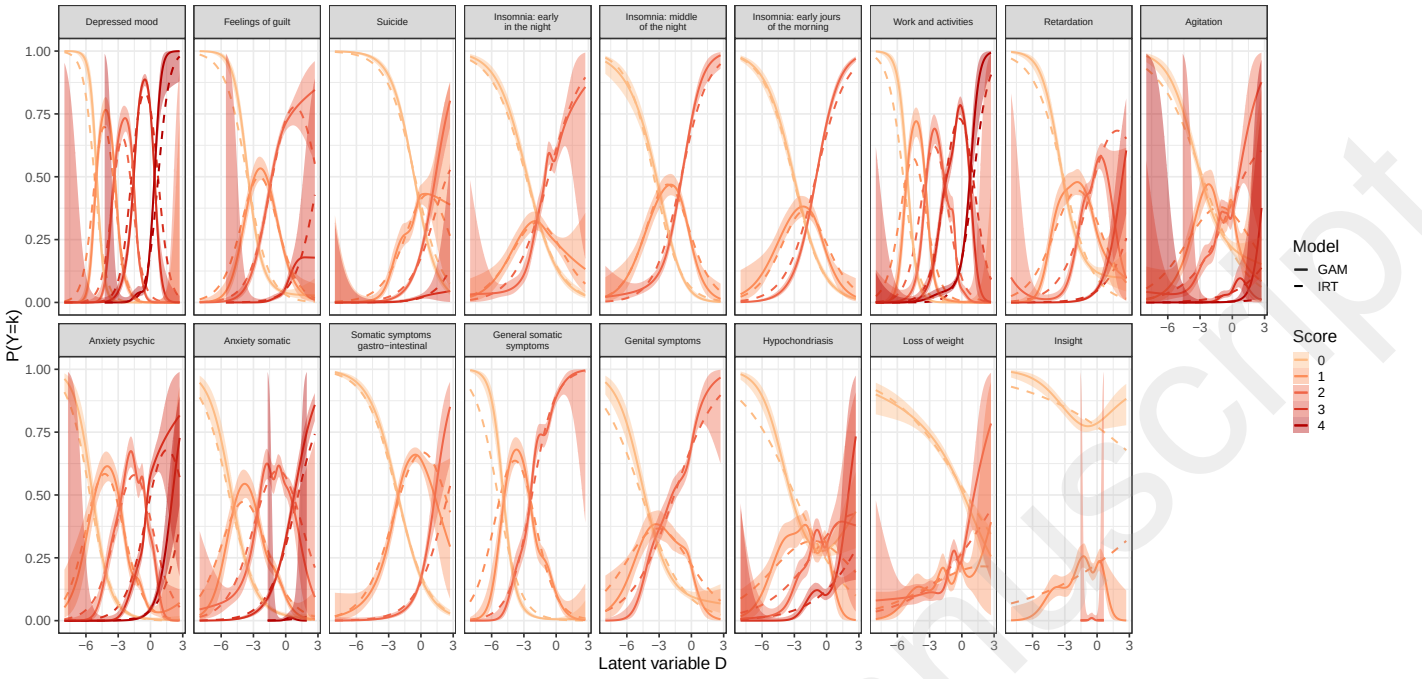

Figure S1: Probability for each score for all items of the HAMD scale, predicted by the IRT model using only the placebo data (based on item specific parameters - dashed line). This was compared to the fit of a generalized additive model (GAM) with cross-validated cubic spline as a smoothing function (continuous line with 95% confidence interval)

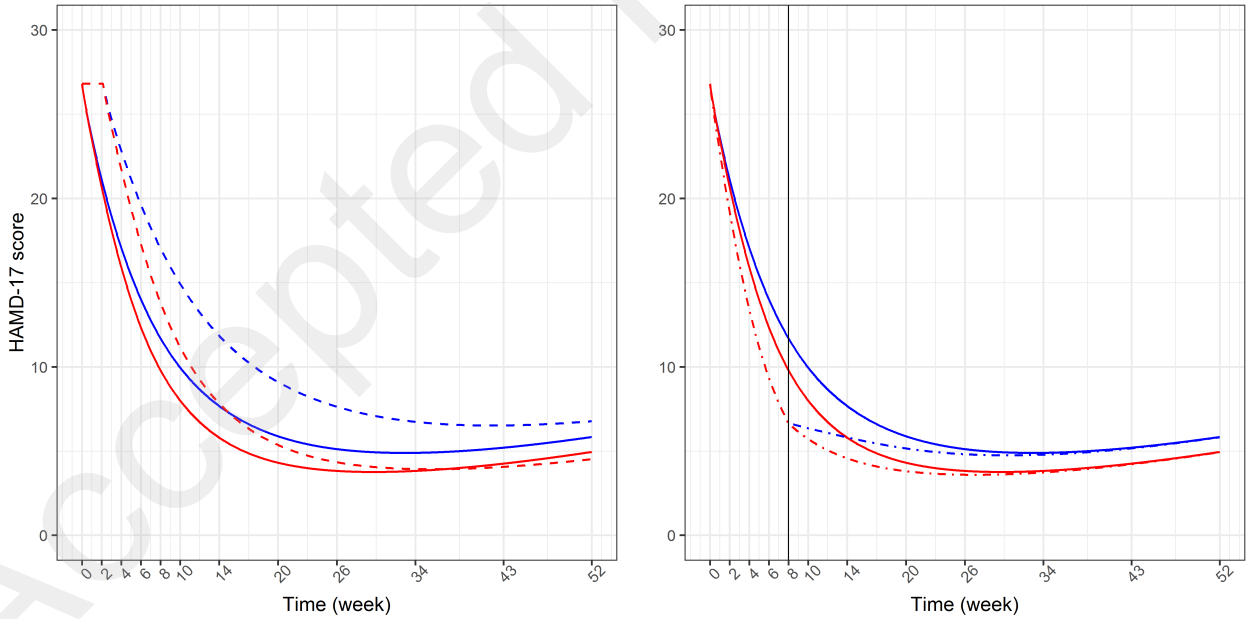

Figure S2: Predicted evolution for typical individuals of the total HAMD-17 score depending on the dose level (randomised into placebo: blue, randomised into active with 25mg: red) and design elements: dose adjustment at the second week (left) and open treatment period effect (right).

Accepted Manuscript

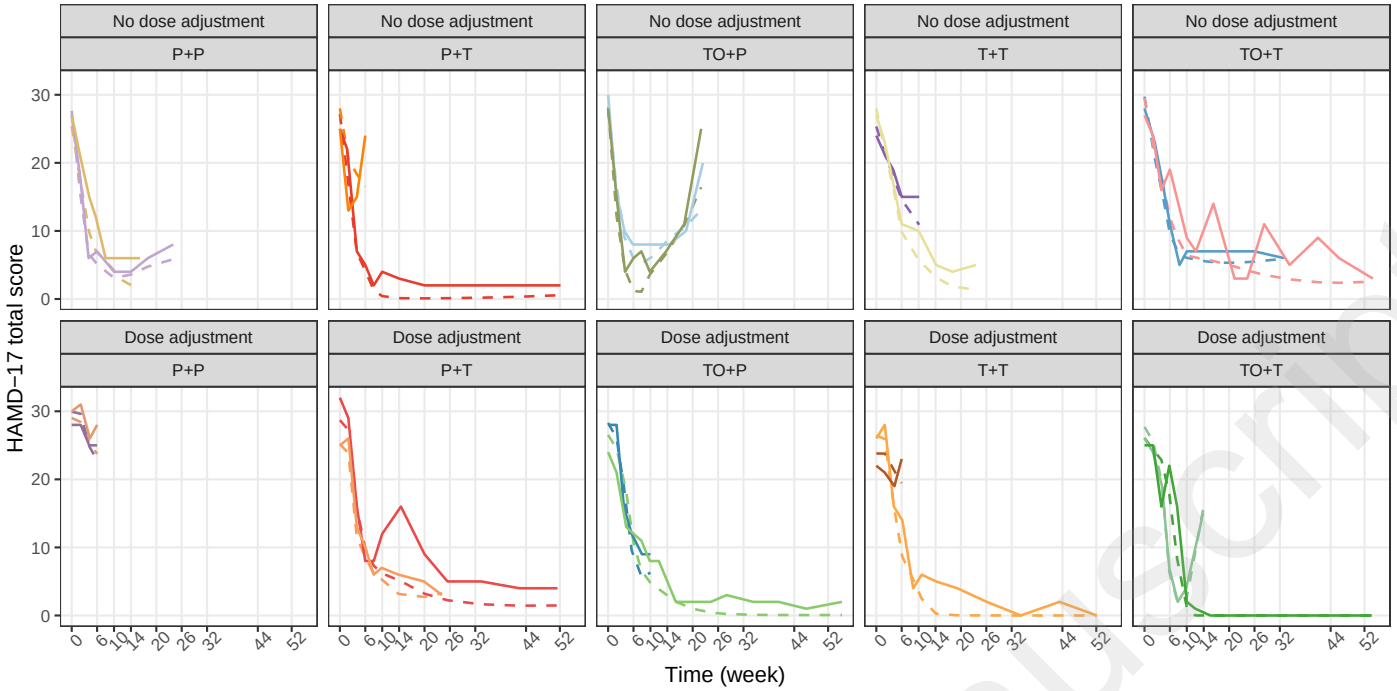

Figure S3: Individual trajectories observed (continuous lines) compared to the predicted ones (dashed lines) from the fit. Each colour represents one sampled individual.

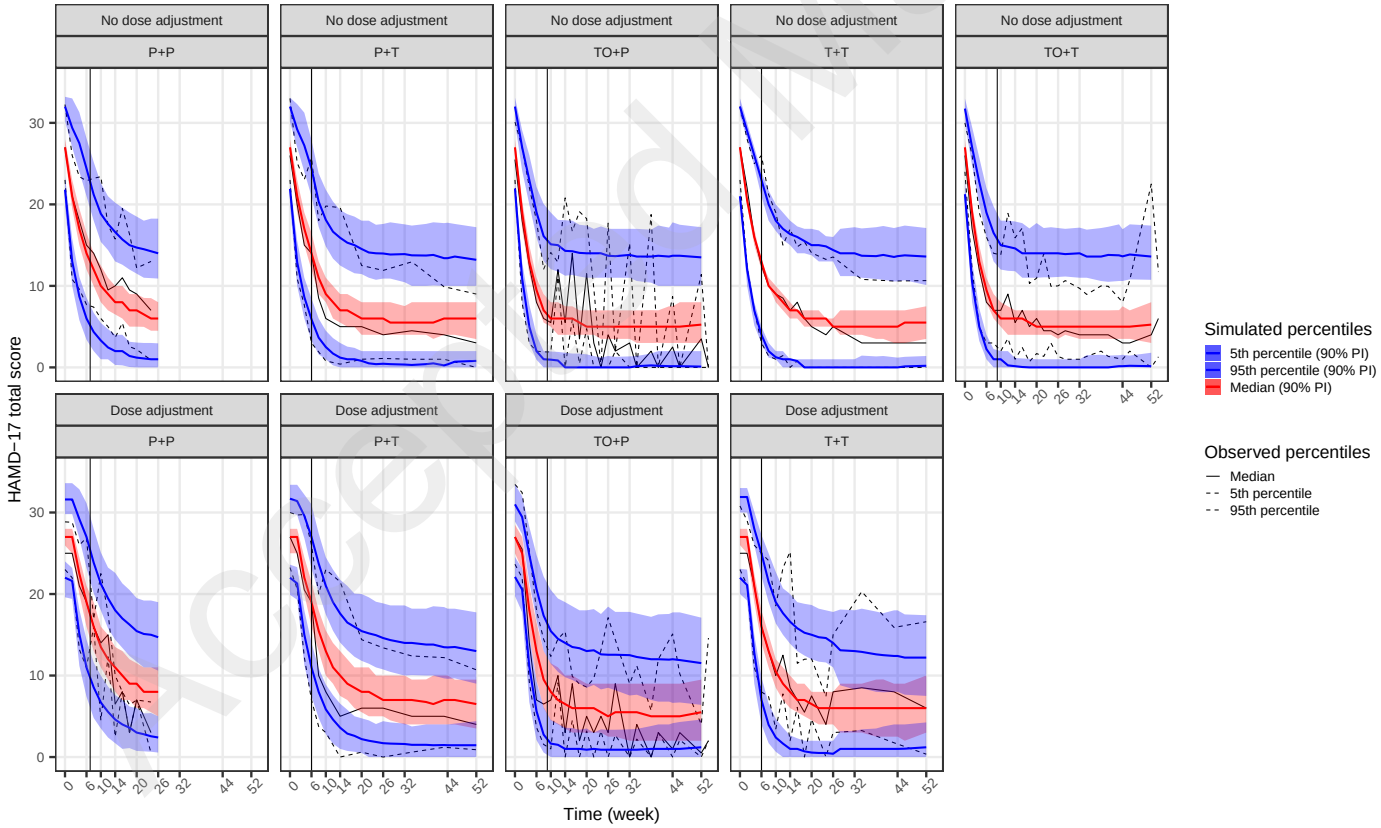

Figure S4: External validation: visual predictive check of the HAMD-17 total score stratified by dose adjustment (by row) and treatment sequence (by column). Median, 5th and 95th percentiles in the data are compared to the model simulated median, 5th and 95th percentiles (with the 90% prediction interval in shaded area). The vertical black line represents the end of the first period. Due to the very low number of patient (3) in the botttom right box, the results are not shown.

Accepted Manuscript

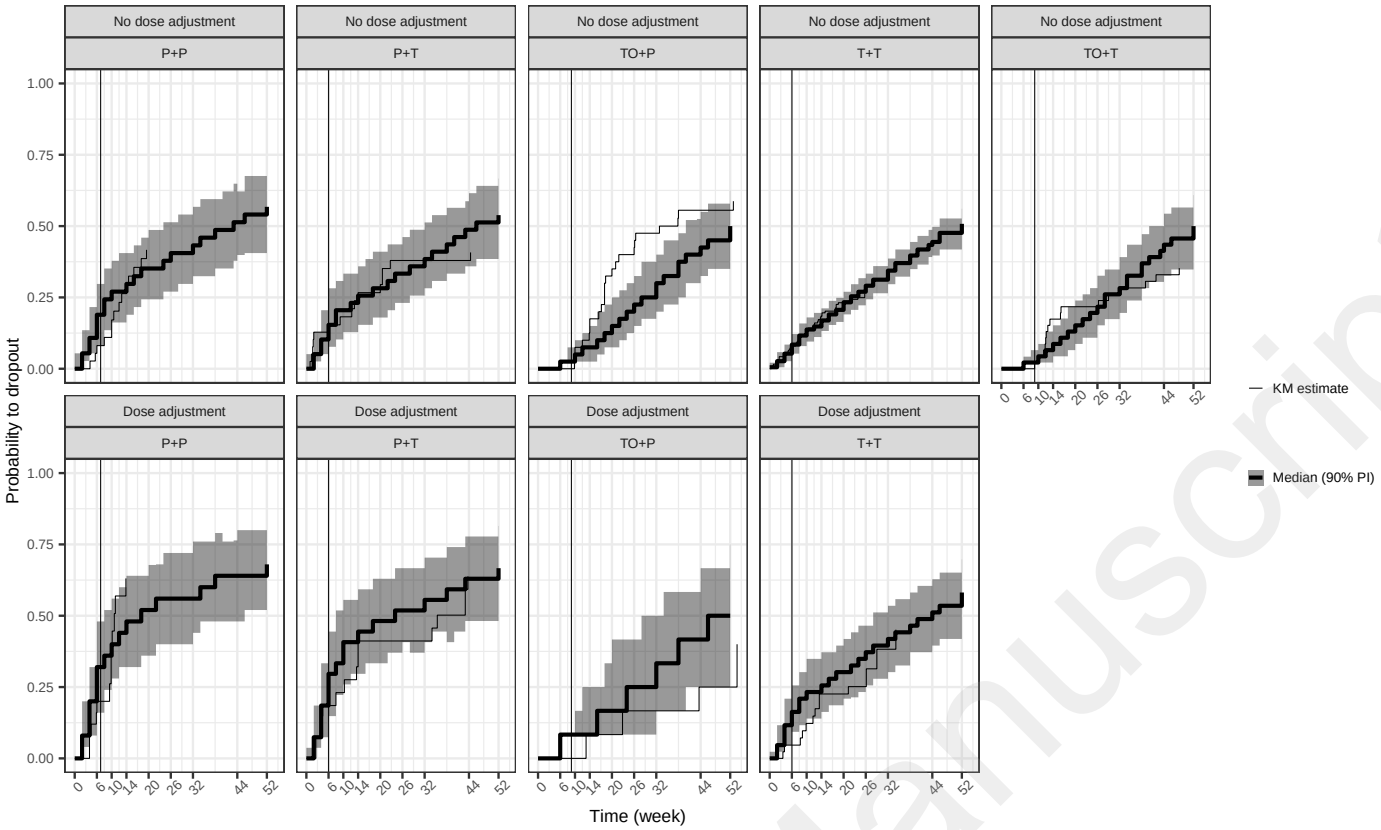

Figure S5: External validation: visual predictive check of the dropout stratified by dose adjustment (by row) and treatment sequence (by column). The Kaplan-Meier estimated based on the data (mean and 90% confidence interval) is compared to the model simulated median and the 90% prediction interval (shaded area). The vertical black line represents the end of the first period. Due to the very low number of patients (3) in the TO+T group with dose adjustment, the bottom right box is not shown.

Accepted Manuscript

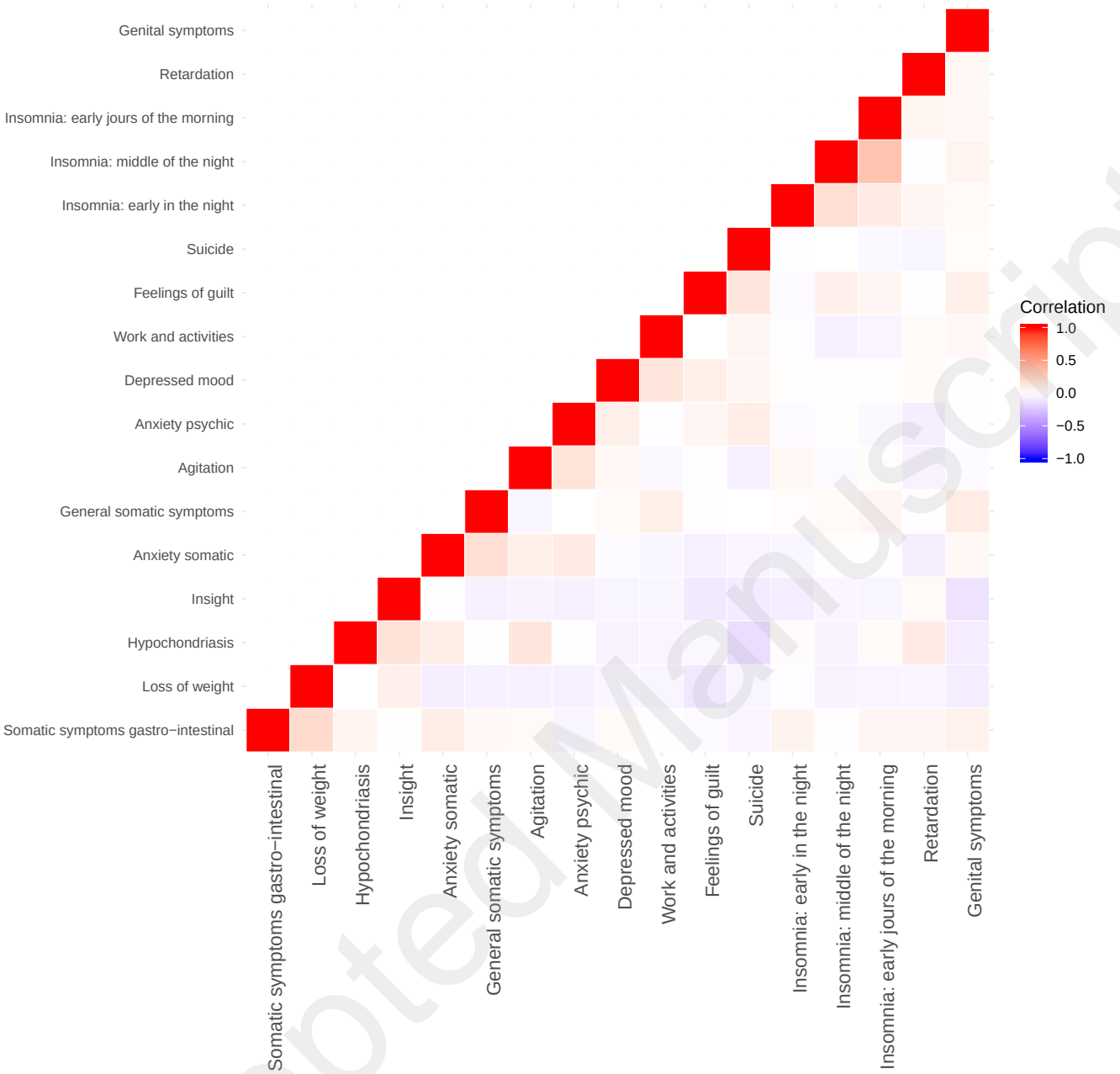

Figure S6: Correlation between the residuals obtained using a single latent variable IRT model

Accepted Manuscript

Hamilton Depression Rating Scale (HDRS)

Reference: **Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960; 23:56–62**

Rating Clinician-rated

Administration time 20–30 minutes

Main purpose To assess severity of, and change in, depressive symptoms

Population Adults

Commentary

The HDRS (also known as the Ham-D) is the most widely used clinician-administered depression assessment scale. The original version contains 17 items ($HDRS_{17}$) pertaining to symptoms of depression experienced over the past week. Although the scale was designed for completion after an unstructured clinical interview, there are now semi-structured interview guides available. The HDRS was originally developed for hospital inpatients, thus the emphasis on melancholic and physical symptoms of depression. A later 21-item version ($HDRS_{21}$) included 4 items intended to subtype the depression, but which are sometimes, incorrectly, used to rate severity. A limitation of the HDRS is that atypical symptoms of depression (e.g., hypersomnia, hyperphagia) are not assessed (see SIGH-SAD, page 55).

Scoring

Method for scoring varies by version. For the $HDRS_{17}$, a score of 0–7 is generally accepted to be within the normal range (or in clinical remission), while a score of 20 or higher (indicating at least moderate severity) is usually required for entry into a clinical trial.

Versions

The scale has been translated into a number of languages including French, German, Italian, Thai, and Turkish. As well, there is an Interactive Voice Response version (IVR), a Seasonal Affective Disorder version (SIGH-SAD, see page 55), and a Structured Interview Version (HDS-SIV). Numerous versions with varying lengths include the HDRS17, HDRS21, HDRS29, HDRS8, HDRS6, HDRS24, and HDRS7 (see page 30).

Additional references

Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 1967; 6(4):278–96.

Williams JB. A structured interview guide for the Hamilton Depression Rating Scale. Arch Gen Psychiatry 1988; 45(8):742–7.

Address for correspondence

The HDRS is in the public domain.

Hamilton Depression Rating Scale (HDRS)

PLEASE COMPLETE THE SCALE BASED ON A STRUCTURED INTERVIEW

Instructions: for each item select the one "cue" which best characterizes the patient. Be sure to record the answers in the appropriate spaces (positions 0 through 4).

1 DEPRESSED MOOD *(sadness, hopeless, helpless, worthless)*

0 |__| Absent.
1 |__| These feeling states indicated only on questioning.
2 |__| These feeling states spontaneously reported verbally.
3 |__| Communicates feeling states non-verbally, i.e. through facial expression, posture, voice and tendency to weep.
4 |__| Patient reports virtually only these feeling states in his/her spontaneous verbal and non-verbal communication.

2 FEELINGS OF GUILT

0 |__| Absent.
1 |__| Self reproach, feels he/she has let people down.
2 |__| Ideas of guilt or rumination over past errors or sinful deeds.
3 |__| Present illness is a punishment. Delusions of guilt.
4 |__| Hears accusatory or denunciatory voices and/or experiences threatening visual hallucinations.

Accepted Manuscript

3 SUICIDE
0 |__| Absent.
1 |__| Feels life is not worth living.
2 |__| Wishes he/she were dead or any thoughts of possible death to self.
3 |__| Ideas or gestures of suicide.
4 |__| Attempts at suicide (any serious attempt rate 4).

4 INSOMNIA: EARLY IN THE NIGHT
0 |__| No difficulty falling asleep.
1 |__| Complains of occasional difficulty falling asleep, i.e. more than ½ hour.
2 |__| Complains of nightly difficulty falling asleep.

5 INSOMNIA: MIDDLE OF THE NIGHT
0 |__| No difficulty.
1 |__| Patient complains of being restless and disturbed during the night.
2 |__| Waking during the night – any getting out of bed rates 2 (except for purposes of voiding).

6 INSOMNIA: EARLY HOURS OF THE MORNING
0 |__| No difficulty.
1 |__| Waking in early hours of the morning but goes back to sleep.
2 |__| Unable to fall asleep again if he/she gets out of bed.

7 WORK AND ACTIVITIES
0 |__| No difficulty.
1 |__| Thoughts and feelings of incapacity, fatigue or weakness related to activities, work or hobbies.
2 |__| Loss of interest in activity, hobbies or work – either directly reported by the patient or indirect in listlessness, indecision and vacillation (feels he/she has to push self to work or activities).
3 |__| Decrease in actual time spent in activities or decrease in productivity. Rate 3 if the patient does not spend at least three hours a day in activities (job or hobbies) excluding routine chores.
4 |__| Stopped working because of present illness. Rate 4 if patient engages in no activities except routine chores, or if patient fails to perform routine chores unassisted.

8 RETARDATION (slowness of thought and speech, impaired ability to concentrate, decreased motor activity)
0 |__| Normal speech and thought.
1 |__| Slight retardation during the interview.
2 |__| Obvious retardation during the interview.
3 |__| Interview difficult.
4 |__| Complete stupor.

9 AGITATION
0 |__| None.
1 |__| Fidgetiness.
2 |__| Playing with hands, hair, etc.
3 |__| Moving about, can't sit still.
4 |__| Hand wringing, nail biting, hair-pulling, biting of lips.

10 ANXIETY PSYCHIC
0 |__| No difficulty.
1 |__| Subjective tension and irritability.
2 |__| Worrying about minor matters.
3 |__| Apprehensive attitude apparent in face or speech.
4 |__| Fears expressed without questioning.

11 ANXIETY SOMATIC (physiological concomitants of anxiety) such as:
gastro-intestinal – dry mouth, wind, indigestion, diarrhea, cramps, belching
cardio-vascular – palpitations, headaches
respiratory – hyperventilation, sighing
urinary frequency
sweating
0 |__| Absent.
1 |__| Mild.
2 |__| Moderate.
3 |__| Severe.
4 |__| Incapacitating.

12 SOMATIC SYMPTOMS GASTRO-INTESTINAL
0 |__| None.
1 |__| Loss of appetite but eating without staff encouragement. Heavy feelings in abdomen.
2 |__| Difficulty eating without staff urging. Requests or requires laxatives or medication for bowels or medication for gastro-intestinal symptoms.

13 GENERAL SOMATIC SYMPTOMS
0 |__| None.
1 |__| Heaviness in limbs, back or head. Backaches, headaches, muscle aches. Loss of energy and fatigability.
2 |__| Any clear-cut symptom rates 2.

14 GENITAL SYMPTOMS (symptoms such as loss of libido, menstrual disturbances)
0 |__| Absent.
1 |__| Mild.
2 |__| Severe.

15 HYPOCHONDRIASIS
0 |__| Not present.
1 |__| Self-absorption (bodily).
2 |__| Preoccupation with health.
3 |__| Frequent complaints, requests for help, etc.
4 |__| Hypochondriacal delusions.

16 LOSS OF WEIGHT ***(RATE EITHER a OR b)***

a) According to the patient:	**b) According to weekly measurements:**
0 |__| No weight loss.	0 |__| Less than 1 lb weight loss in week.
1 |__| Probable weight loss associated with present illness.	1 |__| Greater than 1 lb weight loss in week.
2 |__| Definite (according to patient) weight loss.	2 |__| Greater than 2 lb weight loss in week.
3 |__| Not assessed.	3 |__| Not assessed.

17 INSIGHT
0 |__| Acknowledges being depressed and ill.
1 |__| Acknowledges illness but attributes cause to bad food, climate, overwork, virus, need for rest, etc.
2 |__| Denies being ill at all.

Total score: |__|__|

This scale is in the public domain.

29

Figure S7: HAMD-17 scale

Accepted Manuscript

Table SI: Parameter estimates with their relative standard errors (%). a is the discrimination parameter and b the difficulty parameter for item s and score K.

	Parameter	Value (RSE%)	Parameter	Value (RSE%)
1	OBJ	242310.79	$b_{s=11,K=3}$	3.51 (1.64%)
2	AIC	242500.8	$b_{s=11,K=4}$	6.06 (7.64%)
3	BIC	243975	$a_{s=12}$	0.67 (1.78%)
4	$a_{s=1}$	1.55 (1.37%)	$b_{s=12,K=1}$	-2.14 (1.48%)
5	$b_{s=1,K=1}$	-5.32 (0.62%)	$b_{s=12,K=2}$	4.76 (1.89%)
6	$b_{s=1,K=2}$	2.05 (1.51%)	$a_{s=13}$	0.83 (1.36%)
7	$b_{s=1,K=3}$	1.72 (1.78%)	$b_{s=13,K=1}$	-5.35 (0.74%)
8	$b_{s=1,K=4}$	2.54 (1.71%)	$b_{s=13,K=2}$	3.24 (1.35%)
9	$a_{s=2}$	0.86 (1.26%)	$a_{s=14}$	0.5 (1.42%)
10	$b_{s=2,K=1}$	-3.78 (0.72%)	$b_{s=14,K=1}$	-5.1 (0.73%)
11	$b_{s=2,K=2}$	2.67 (1.35%)	$b_{s=14,K=2}$	3.36 (1.56%)
12	$b_{s=2,K=3}$	4.7 (2.33%)	$a_{s=15}$	0.51 (1.49%)
13	$a_{s=3}$	0.72 (2.21%)	$b_{s=15,K=1}$	-2.62 (1.22%)
14	$b_{s=3,K=1}$	-0.5 (8.38%)	$b_{s=15,K=2}$	2.8 (1.68%)
15	$b_{s=3,K=2}$	2.61 (2.76%)	$b_{s=15,K=3}$	3.98 (2.28%)
16	$b_{s=3,K=3}$	3.43 (5.96%)	$a_{s=16}$	0.43 (2.18%)
17	$b_{s=3,K=4}$	5.19 (28.82%)	$b_{s=16,K=1}$	0.47 (17.03%)
18	$a_{s=4}$	0.64 (1.55%)	$b_{s=16,K=2}$	1.89 (3.06%)
19	$b_{s=4,K=1}$	-2.38 (1.33%)	$a_{s=17}$	0.26 (5.73%)
20	$b_{s=4,K=2}$	2.32 (2.08%)	$b_{s=17,K=1}$	5.69 (8.26%)
21	$a_{s=5}$	0.6 (1.57%)	$b_{s=17,K=2}$	15.61 (8.14%)
22	$b_{s=5,K=1}$	-3.68 (0.88%)	D0	0
23	$b_{s=5,K=2}$	3.3 (1.69%)	Drem	8.39 (1.92%)
24	$a_{s=6}$	0.61 (1.66%)	Trem (year)	0.16 (3.51%)
25	$b_{s=6,K=1}$	-3.18 (0.93%)	Trel (year)	1.66 (6.59%)
26	$b_{s=6,K=2}$	2.7 (1.75%)	γ (-)	0.88 (1.38%)
27	$a_{s=7}$	1.27 (1.21%)	α_{Trem}	3e-04 (10.98%)
28	$b_{s=7,K=1}$	-5.55 (0.47%)	Teq (year)	0.06 (5.43%)
29	$b_{s=7,K=2}$	2.14 (1.4%)	k (-)	3.29 (4.51%)
30	$b_{s=7,K=3}$	2.09 (1.41%)	λ (year $^{-1}$)	5.99 (10.82%)
31	$b_{s=7,K=4}$	2.43 (1.7%)	β	12.72 (5.72%)
32	$a_{s=8}$	0.88 (1.44%)	β_{Open}	1.52 (12.04%)
33	$b_{s=8,K=1}$	-2.85 (0.92%)	Tlag (year)	0.04
34	$b_{s=8,K=2}$	2.64 (1.57%)	β_{adj}	0.27 (11.56%)
35	$b_{s=8,K=3}$	3.94 (3.13%)	α_{Trel}	1e-04 (33.68%)
36	$a_{s=9}$	0.52 (1.64%)	T_{open} (year)	0.11 (12.61%)
37	$b_{s=9,K=1}$	-2.63 (1.23%)	ω_{D0}	0.45 (4.3%)
38	$b_{s=9,K=2}$	3.34 (1.87%)	ω_{Drem}	0.37 (3.0%)
39	$b_{s=9,K=3}$	5.45 (3.07%)	corr_{Trem_Drem}	0.7 (1.7%)
40	$b_{s=9,K=4}$	5.18 (9.73%)	ω_{Trem}	0.82 (1.7%)
41	$a_{s=10}$	0.85 (1.31%)	corr_{Trel_Drem}	-0.43 (8%)
42	$b_{s=10,K=1}$	-6.13 (0.68%)	corr_{Trel_Trem}	-0.15 (26.2%)
43	$b_{s=10,K=2}$	3.22 (1.3%)	ω_{Trel}	1.27 (3.4%)
44	$b_{s=10,K=3}$	3.14 (1.42%)	$\text{corr}_{\alpha_{Trem}\_Drem}$	0.84 (1.7%)
45	$b_{s=10,K=4}$	3.56 (3.12%)	$\text{corr}_{\alpha_{Trem}\_Trem}$	0.58 (5.3%)
46	$a_{s=11}$	0.78 (1.2%)	$\text{corr}_{\alpha_{Trem}\_Trel}$	-0.84 (2.6%)
47	$b_{s=11,K=1}$	-5.21 (0.49%)	$\omega_{\alpha_{Trem}}$	1.41 (3.8%)
48	$b_{s=11,K=2}$	3.09 (1.2%)	ω_{γ}	0.32 (3.7%)

Accepted Manuscript

Table SII: Settings of the clinical trial simulation

Dose Adjustment at week 2	Arm	Sample size
No	Placebo	253
Yes	Placebo	182
No	Active	630
Yes	Active	146

Accepted Manuscript