See COMMENTARY page 592

A New Population-Enrichment Strategy to Improve Efficiency of Placebo-Controlled Clinical Trials of Antidepressant Drugs

E Merlo-Pich[1,2], RC Alexander[1], M Fava[3] and R Gomeni[2]

The rate-limiting factor in the discovery of novel antidepressants is the inefficient methodology of traditional multicenter randomized clinical trials (RCTs). We applied a model-based approach to a large clinical database (five RCTs in major depressive disorder (MDD), involving 1,837 patients from 124 recruitment centers) with two objectives: (i) to learn about the role of center-specific placebo response in RCT failure and (ii) to apply what is learned to improve the efficiency of RCTs by enhancing the detection of treatment effect (TE). Sensitivity analysis indicated that center-specific placebo response was the most relevant predictor of RCT failure. To reduce the statistical "noise" generated by centers with nonplausible, excessively high/low placebo responses, we developed an enrichment-window strategy. Clinical trial simulation was used to assess the enrichment strategy applied before the standard statistical analysis, resulting in an overall reduction in failure of RCTs from ~50 to ~10%.

Major depressive disorder (MDD), a chronic psychiatric illness with a lifetime prevalence of 16.2% in the United States, is characterized by recurrent episodes of low mood, loss of interest, and impaired psychosocial functioning.[1–3]

Antidepressant drugs such as selective serotonin reuptake inhibitors have long been used as first-line treatment for MDD, despite the high incidence of failure in trials.[4,5]

A recent meta-analysis, conducted on the US Food and Drug Administration database and including 12 approved antidepressant drugs, 74 randomized clinical trials (RCTs), and 12,564 patients, indicated that 49% of the clinical trials failed.[6] Given the high health burden of MDD as predicted by the World Health Organization[7,8] and the limitations of current treatments (i.e., modest efficacy as well as tolerability issues such as sexual dysfunction and weight gain), the development of novel drugs for MDD remains a critical medical goal.[9] An exploration in May 2009 of ClinicalTrials.gov, a public database in which the majority of clinical trials in progress are collected, indicated that 72 trials for novel or approved antidepressants were in progress in 2009, involving more than 15,000 subjects with a diagnosis of MDD. According to the statistics cited above, ~50% of these subjects will potentially be exposed to some degree of safety risk. There will be no probability of a contribution to knowledge about the efficacy of these novel agents because these RCTs will fail. From the point of view not only of science but also of ethics and health economics, there is therefore an urgent need to improve the efficiency and accuracy of RCTs in MDD so as to reduce the frequency of failed and uninformative trials. Despite the current high standard of RCT design, conduct, and management, RCTs still fail. These failures may be attributable to (i) the low sensitivity of the assays used to measure the clinical improvement;[10] (ii) modest efficacy of the antidepressant drugs; (iii) the heterogeneity of the MDD population, often including misdiagnosed or noncompliant subjects; and (iv) high level of placebo response.[11–14]

Converging findings indicate placebo response as the most relevant among these factors in contributing to the failure of RCTs. In fact, high response in the placebo arm makes it very difficult to detect drug effects.[12,15] Of note, the magnitude of placebo response has tended to increase in MDD studies over the years, with greater placebo response observed in more recent RCTs relative to RCTs performed in the 1980s (ref. 16).

Recently, we proposed to move the focus on placebo-response analysis from the entire RCT to individual recruitment centers.[17] In some centers, several factors, including the interactions between investigators and patients, could have favored the occurrence of extremely high or low placebo responses.

[1]Neurosciences Center of Excellence for Drug Discovery, GlaxoSmithKline R&D, Verona, Italy; [2]Pharmacometrics, GlaxoSmithKline R&D, Upper Merion, Pennsylvania, USA; [3]Psychiatry Department, Massachusetts General Hospital, Boston, Massachusetts, USA. Correspondence: E Merlo-Pich (emilio.v.merlo-pich@gsk.com)

Received 2 March 2010; accepted 4 June 2010; advance online publication 22 September 2010. doi:10.1038/clpt.2010.159

Therefore, the level of placebo response in these centers could have prevented the detection of a drug treatment signal, thereby leading to failure of the RCT.

A signal-detection approach was proposed to characterize each center's performance by computing the posterior probability of detecting a clinically relevant separation of active treatment from placebo. This approach appeared to be a useful methodology to rank the performances of recruitment centers and to classify each center on the basis of information yielded for detecting clinically relevant signals of efficacy. In a typical trial, only 60% of the centers were classified as informative.[17]

Our proposal was to identify the informative centers by applying an enrichment window defined by two boundaries (cutoffs) placed on the high and low ends of the placebo response distribution. The application of the enrichment window will filter out the data from centers in which the mean values of the placebo response fell outside these boundaries. As is true for all enrichment strategies, its implementation in this context should result in an increased efficiency of RCTs in detecting signals of clinically relevant treatment effects (TEs).

In this article, we describe the implementation of a model-based approach recently proposed by the US Food and Drug Administration to enhance the productivity of drug discovery.[18] This goal was achieved first by learning the role of center-specific placebo response in the detection of TE (using a database of multicenter RCTs) and then by applying what was learned to implement the enrichment-window strategy. In particular, we consider this methodology appropriate for optimizing a proof-of-concept (PoC) RCT, the first evidence of antidepressant effects of a novel compound.

RESULTS

Development and validation of the models

We developed our models using data from five RCTs that tested the efficacy of paroxetine in a total of 1,837 patients with MDD, from 124 recruitment centers. The TE was measured by the

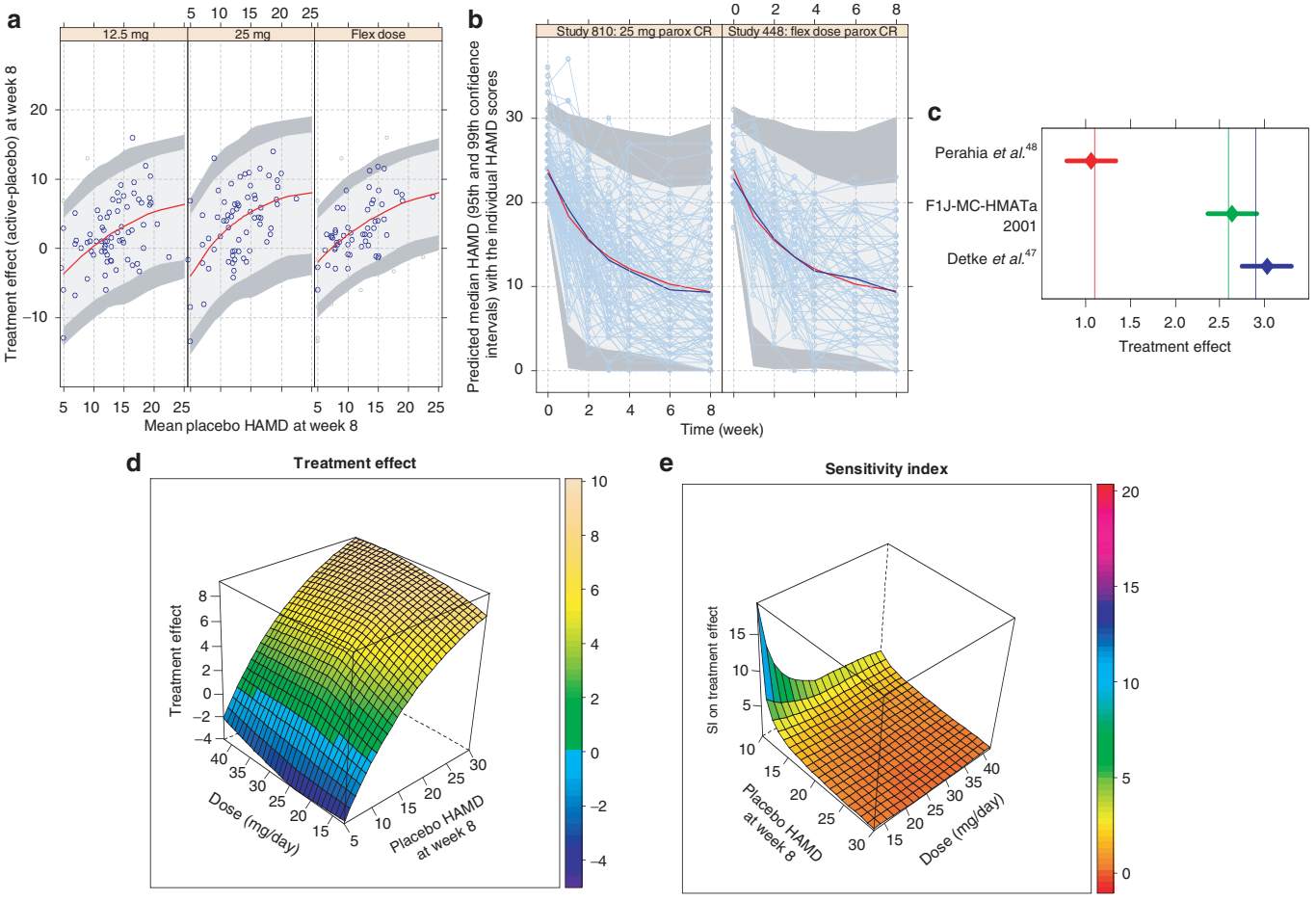

Figure 1 Quantitative characterization of disease-drug-trial model. (**a**) Mean model prediction (red curve) with 95% (light gray area) and 99% (gray area) confidence intervals of the relationship between mean TE vs. placebo HAMD clinical scores at week 8 in each recruitment center (blue empty circles). (**b**) Internal model validation: mean observed HAMD trajectories (blue lines), observed individual HAMD trajectories (light gray lines), and mean model-predicted HAMD trajectories (red lines) with 95 and 99% confidence intervals (light and dark gray areas, respectively). (**c**) External model validation: the mean model-predicted TEs of three RCTs that were not used for model development are shown as vertical lines. The observed TEs are represented as dots for the mean values and as horizontal bars for the 95% confidence interval. (**d**) Integrated surface response model of TE as a function of paroxetine daily doses (mg) and of center-specific placebo responses (HAMD at week 8). Color-coded green and yellow areas represent the typical outcomes of positive RCTs in MDD. (**e**) Normalized sensitivity index (SI) of treatment response surface. This index (*z* axis) represents the percentage change of model output (TE) produced by changes in dose strength (*x* axis) and center-specific placebo response (*y* axis). CR, modified release; HAMD, Hamilton Rating Scale for Depression; MDD, major depressive disorder; RCT, randomized clinical trial; TE, treatment effect.

difference in Hamilton Rating Scale for Depression (HAMD) score between paroxetine response and placebo response at end point, that is, after 8 weeks of treatment. The clinical response model was obtained by independently fitting the HAMD scores in each study, and in each trial arm, to a mixed Weibull linear equation (Equation 1).[19] The distribution of the maximum *a posteriori* Bayesian individual parameter estimates was used to characterize the center-specific response, as previously shown.[17]

The TE model was developed using the results of the clinical response model, by describing the functional relationship between center-specific paroxetine TE and center-specific placebo response.[2] Three independent analyses were conducted on pooled data from the five RCTs, grouped by daily drug dosage regimen (12.5 mg, 25 mg, and flex-dose regimen ≤62.5 mg, respectively). **Table 1** summarizes the parameter estimates, and **Figure 1a** shows the model-predicted mean TE (with the 95 and 99% confidence intervals) by dose. Both internal and external validations confirmed the reliability and predictive performances of the model (**Figure 1b,c**).

Table 1 Treatment effect model parameters (SE) associated with paroxetine CR 12.5 mg, paroxetine CR 25 mg, and paroxetine CR flex dose (42 mg)

	Hbas	Slope	δ	Res. error
Parox 12.5 mg	−9.96 (2.76)	0.08 (0.01)	9.31 (2.94)	4.10 (0.09)
Parox 25 mg	−15.1 (5.98)	0.12 (0.03)	9.33 (1.70)	4.51 (0.12)
Parox flex dose	−7.79 (4.58)	0.07 (0.05)	11.4 (4.36)	3.17 (0.05)

CR, modified release; HAMD, Hamilton Rating Scale for Depression; Hbas, treatment effect when the placebo HAMD score approaches zero; Res. error, residual error.

Surface response and sensitivity analysis

An integrated surface-response representation was then derived to arrive at a consolidated description of the changes in TE of paroxetine as a function of the dose (*d*) and of the center-specific level of placebo response (Hend) (**Figure 1d**). The analysis of the surface response indicated the predominant role of center-specific placebo response as compared with the dose strength in determining the TE of paroxetine.

Although there is no accepted standard, most experts recognize a TE as clinically relevant if the mean HAMD difference relative to placebo is >3 (TE > 3), a criterion generally supported by the US Food and Drug Administration and the European Medicines Agency. When this criterion was applied, the surface response model showed that recruitment centers reporting high placebo responses—defined as a mean HAMD score <10 at end point—were not expected to deliver relevant TE at any paroxetine dose tested.

We also used the normalized sensitivity index (SI) to estimate the relative weightages of paroxetine dose and center-specific placebo response (*d* and Hend) as factors that are influential in determining the drug TE. This methodology is widely used in mathematical modeling to identify critical determinants of model response, to support dimension reduction, and to help the design of informative experiments.[20] TE was only marginally sensitive to paroxetine at doses <20 mg/day when the center-specific placebo response with HAMD at end point was >15 (1 > SI > 2). Higher placebo response per center (HAMD <10 at end point) was always associated with SI >2, indicating the relevant impact on TE, independent of the paroxetine dose (**Figure 1e**). In this region of the surface response (SI >2), the

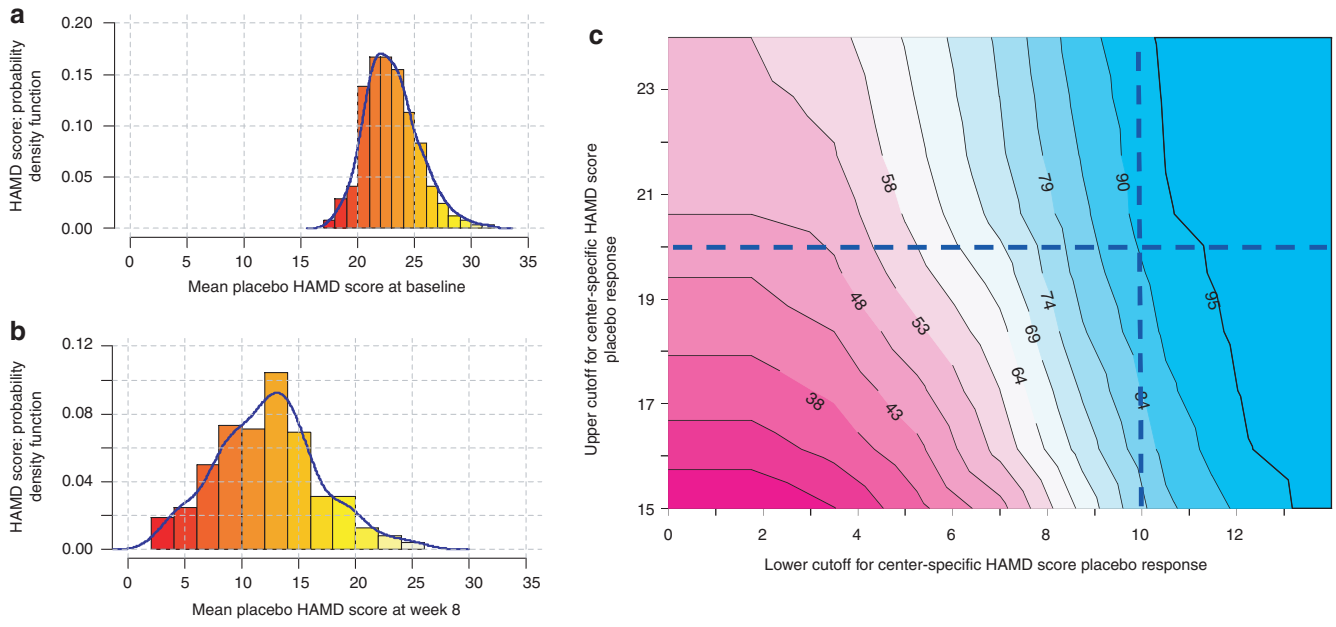

Figure 2 Clinical trial simulation. (**a**, **b**) Distribution of the mean placebo HAMD scores at (**a**) week 0 and (**b**) week 8, measured in 239 recruitment centers of nine RCTs with the associated density function (blue solid lines). (**c**) Contour plot representing the probability of detecting clinically relevant effects of paroxetine treatment (TE >3) when the informative RCT population is defined using an enrichment-window approach. The *x* axis represents the lower boundary, and the *y* axis represents the high boundary of the enrichment window, both expressed as mean center-specific HAMD scores at week 8. The contour plot represents the outcome of 100 simulated RCTs based on the model described in **Figure 1**, empirically tested for the various cutoff combinations. HAMD, Hamilton Rating Scale for Depression; RCT, randomized clinical trial; TE, treatment effect.

TE variability will increase, with a potential negative influence on study power.

Clinical trial simulation

The results of the clinical trial simulation conducted to explore the relationship between the probability of success of an RCT and specific enrichment-window boundaries are summarized in the contour plot shown in **Figure 2c**. The equiprobability curves displayed here represent the probability of detecting clinically relevant TEs (TE >3) when different cutoff values were used to define the upper and lower boundaries of the enrichment window.

The contour plot can be used to assess the impact on signal detection when the cutoffs are selected according to clinically relevant criteria. For example, the low boundary could be set to HAMD = 10 at the end of the study, corresponding to a condition of partial remission as defined by Rush *et al.*[21] Similarly, the high boundary could be set at HAMD = 20 at the end of the study, corresponding to a reduction, relative to baseline, of <10% (given the inclusion criterion of HAMD score ≥ 23), as indicated in the meta-analysis performed by Khan and collaborators.[12] Their analysis showed that the percentage of mean reduction in HAMD from baseline at the end of the study in the placebo arms of 52 trials was 31.4%, with values ranging between 10.5 and 49.2%. These data indicate that reduction of placebo response trajectories to <10% at the end of the study are highly unlikely.

In the contour plot, the quadrant identified by values lower than HAMD = 20 and higher than HAMD = 10 represents the probability of detecting the paroxetine clinical effect (TE >3). This probability ranged from 84 to 100%.

A summary of the main steps of the implementation of the enrichment-window strategy is shown in **Figure 3**. Note that the distribution of center-specific placebo response is significantly affected by the enrichment, whereas less effect was observed

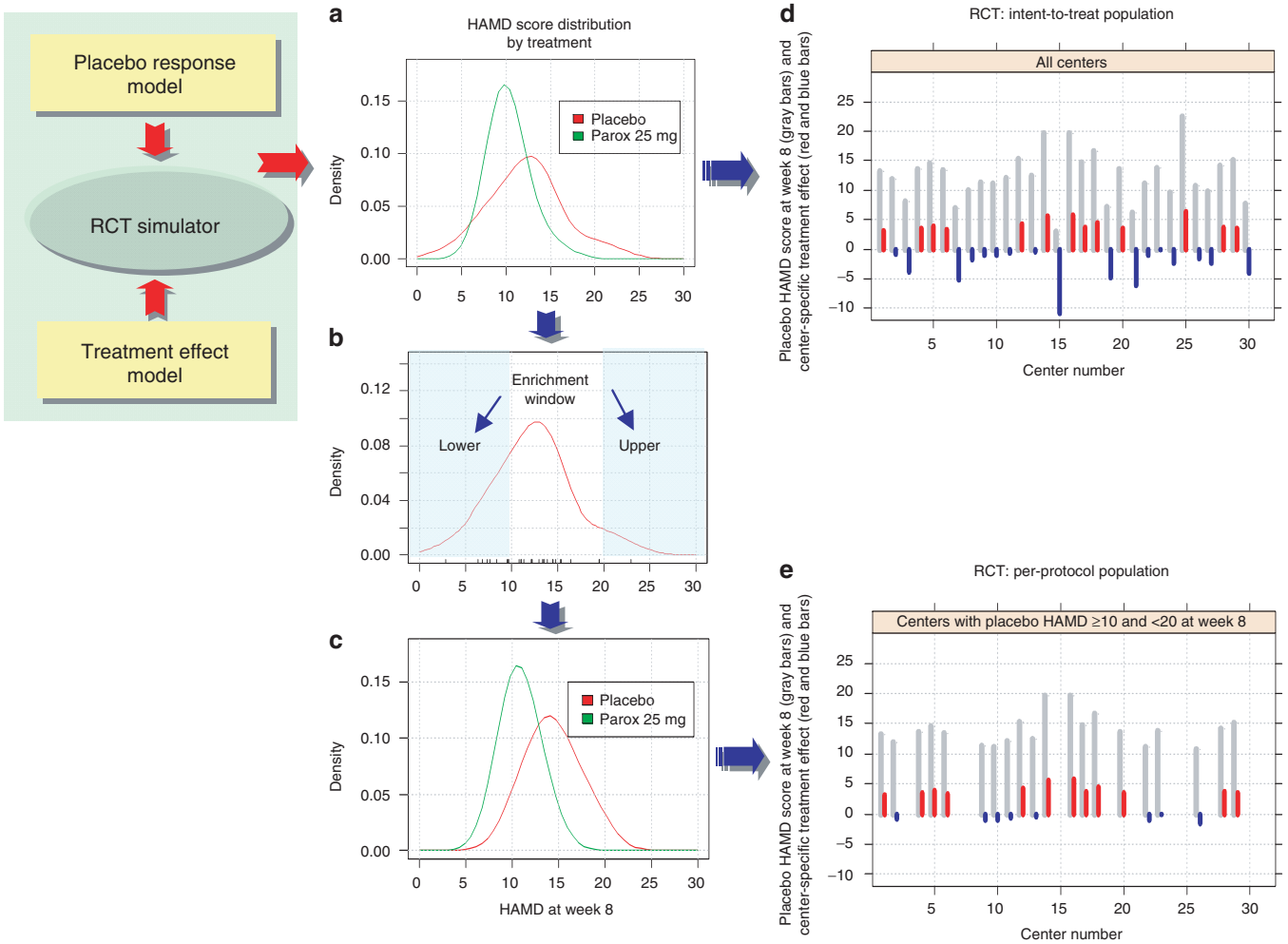

Figure 3 Enrichment-window strategy. (**a**) Representation of all the mean center-specific HAMD score distributions at week 8, for the placebo arm (red curve) and the paroxetine 25 mg/day arm (blue curve) from one of the 100 simulated multicenter RCTs ($n = 30$ centers in each RCT). (**b**) Enrichment window (band-pass filter) operating on placebo scores only, with cutoffs (HAMD <10 and >20 at end point: light blue bands). (**c**) Result of the enrichment-window application expressed as mean center-specific HAMD score distributions at end point for the placebo and the paroxetine arms. (**d**, **e**) Representations of a single simulated center-specific RCT outcome before and after enrichment-window application. Gray bars represent the mean placebo HAMD score at end point specific for each center (*y* axis, left). For each center, the paroxetine-induced changes from placebo (*y* axis, right) are shown in the lower part of the figure as red bars if numerically positive and as blue bars if numerically negative. HAMD, Hamilton Rating Scale for Depression; RCT, randomized clinical trial; TE, treatment effect.

in the distribution in the center-specific paroxetine response (mean HAMD ± SD, before enrichment: 12.3 ± 4.2 (placebo), 10.3 ± 1.7 (paroxetine, after enrichment: 14.4 ± 2.6 (placebo), 10.9 ± 1.4 (paroxetine)) (**Figures 3a–c**)). A possible explanation of the differential impact of enrichment in the placebo and paroxetine arms is that paroxetine treatments tend to normalize the response, reducing the skewness of the distribution. Informative centers selected with this procedure in a typical RTC (**Figure 3d**) are shown in **Figure 3e**.

DISCUSSION

In this study, we developed a model-based approach to capture key information from a database of multicenter RCTs on the role of center-specific placebo response in the detection of antidepressant TE. On the basis of this information, we propose an enrichment strategy to enhance signal detection in early drug development RCTs.

We developed our model from five RCTs selected from the paroxetine database representing typical PoC studies and consisting of three-arm, placebo-controlled, parallel, randomized trials. The five trials were selected on the basis of similarity in the levels of depression severity in the patients at baseline (HAMD ≥23) and the year of publication (2002–2004). These factors were recognized as independent predictors of antidepressant clinical trial outcomes.[14] Controlling for the above variables allowed the analyses to focus on the role of the "drug"-specific characteristics (e.g., dosing regimen, dose strength, and duration of exposure) as well as on "disease" and "trial" mixed factors (e.g., placebo response per recruitment center).

Each individual recruitment center's efficiency in measuring actual clinical response is critically important for the overall success of a multicenter RCT. However, even in centers with proven logistical capacity and adherence to protocol, difficulties in detecting TE are commonly observed.

High levels of placebo response were observed in a substantial percentage of the centers within multicenter RCTs.[17] Several factors have been implicated in determining the level of placebo response, including management of patient expectation by the investigator, investigator bias about the efficacy of the new treatment, misdiagnosis, and regression to the mean.[5,7]

These factors are difficult to assess, and no methodologies are currently available to define a center's performance on the basis of the compounded contribution of each of the factors. The typical way to address this issue is to (i) select centers on the basis of the track record and (ii) provide awareness sessions and good clinical practice training to the selected investigators at the beginning of every multicenter RCT. Despite these efforts, heterogeneity among the performance levels of the centers remains relevant, calling for practical solutions.[22,23] In fact, converging findings indicate that a center's performance is inconsistent over time (possibly due to staff turnover and change in logistics) and that awareness sessions do not have as much impact as expected.[24,25]

In our study, the relevance of center-specific placebo response in determining TE was confirmed by surface response and sensitivity analyses, being much more important to outcome than the more modest contribution of drug dosage. The lack of dose response is consistently reported in trials of antidepressant drugs.[26] We showed that the presence of recruitment centers that report either very high or a very low placebo response in a multicenter trial limit the ability to detect drug TE.

One way to increase the ability to detect TE in RCTs is to apply an enrichment-window approach conceptually derived from band-pass filter. The main function of a band-pass filter is to maximize the signal to be detected in the system by filtering out noise-related signals that fall outside the low and high cutoff limits of the filter. When applied to clinical trials, the enrichment window will help to identify the nonplausible placebo response trajectories generated in a given center that fall below or above the low and high enrichment-window boundaries. These boundaries need to be anchored to clinically relevant criteria. If centers report data that fall outside the enrichment window, such data will be excluded from the per-protocol (PP) RCT statistical analysis.

The choice of the boundaries is critical for determining the outcome of a PP analysis. For this reason, we used clinical trial simulation to explore the relationship between the probability of detecting a clinically relevant signal of efficacy and different enrichment-window criteria in RCTs for an antidepressant drug known to be clinically effective (paroxetine 25 mg/day).

The simulation showed that, when no enrichment was applied, the rate of RCT failure ranged from 48 to 53%, which is in agreement with a recent meta-analysis.[6]

When the enrichment window was defined by the lower boundary equal to a score compatible with partial remission from the MDD episode[27,28] and the higher boundary to a score corresponding to a reduction vs. baseline of <10% (ref. 12), the simulation indicated that the rate of failure of RCTs would be ~10%. Therefore, the implementation of the enrichment window increased the probability of successful RCTs from ~50% to as high as ~90%.

Application of band-pass filter is a common practice in economics, high-energy physics, engineering, and physiology as an enrichment strategy when optimization of signal detection is sought.[29–32] This has not been used in clinical trial design because data enrichment is often perceived as a way of improperly introducing a source of bias. Prospectively, the enrichment-window methodology can be applied to (i) conventional RCTs, (ii) run-in phase RCTs, and (iii) adaptive RCTs. In conventional RCTs, in order to overcome the bias risk, the enrichment strategy should be accounted for and pre-planned in the study protocol. Investigators at each recruitment center must be informed about the implications of the enrichment approach. For example, the protocol could state that, if the placebo response at end point at a given center is beyond the enrichment boundaries, then all data produced by that center, including those of drug treatments, would not be considered for the PP statistical analysis. In this way, only data from subjects tested in centers that are able to provide acceptable placebo responses would contribute to the final RCT signal detection, whereas the data from all the subjects would contribute to the safety, tolerability, and intention-to-treat analyses. In the run-in RCT, the enrichment window would be applied at the end

of the run-in period to identify the centers that should be included in the randomized treatment phase. Finally, in the adaptive RCT, the enrichment-window approach could be applied to scrutinize the quality of the centers on an ongoing basis.[33,34] In this latter application of the window approach, a shift of new randomized subjects toward the most efficient recruitment centers could be agreed according to methodologies such as "playing the winner"[35] or sequential parallel-comparison design,[36,37] in which only the informative centers are used in the trial as the RCT progresses.

Our approach is not expected to affect the type I error because the centers excluded under the proposed enrichment strategy would also be under the same randomization rule as the rest of the trial. Therefore, the balance between treatment allocations is preserved even when the data from noninformative centers are left out of the analyses.

The enrichment-window methodology is expected to improve signal detection leading to a likely increase of the study power. In planning a new clinical trial, this increased study power could theoretically be used to estimate the sample size, but caution should be exercised, given the difficulty of predicting the proportion of subjects belonging to centers with abnormal placebo responses. An ideal solution could be the use of an adaptive powering strategy in which the study power is periodically reassessed on the basis of accrued data.

In this study, we considered the enrichment-window strategy mainly as a means to prevent the discontinuation of promising compounds in early clinical development (i.e., PoC). PoC trials are inherently exploratory in their conception, aimed at signal detection, open to innovation, and based on the method-effectiveness as defined by Sheiner and Rubin.[38] Method-effectiveness is a PP analysis focused on the effect of the actually administered therapy. This differs from the intention-to-treat analysis, which describes the effect of assignment to therapy in the RCTs (so-called use-effectiveness).[38] Our proposal is to provide a specific PP methodology as a complement for, and not as a replacement of, the intention-to-treat analysis.

In PoC RCTs, decisions should be based of method-effectiveness because the application of intention-to-treat analysis (based on all randomized subjects) would risk diluting the TE effect in the presence of excessive heterogeneous center-specific placebo response data. The proposed enrichment methodology identifies the study population that can better discriminate between placebo and active drug treatment, resulting in a more informed assessment of drug efficacy.

In principle, this approach could be also adapted to obtain better information about efficacy in late clinical development RCTs as well, in which the design and analyses are highly regulated. However, the implementation of this approach would require a more thorough assessment of the statistical implications by the clinical and regulatory scientific communities.

Despite the relatively large size of the clinical database considered, the main limitation of this study is the restricted level of generalization of the proposed enrichment-window methodology. The results need to be replicated in other collections of clinical trials with different study designs, inclusion and exclusion criteria, placebo treatment durations, arm numbers, and dates of execution. In addition, we simulated PoC RCTs aimed at testing unknown compounds, but the database also included larger phase III trials. Paroxetine was the only antidepressant considered, with the underlying assumption that similar score distributions of drug effects could be applied to other novel effective antidepressants. Finally, the recruitment center performances were characterized only by the HAMD clinical score, without considering other parameters such as dropout rate and number of queries.

In conclusion, within the relevant limitations, we found that (i) placebo-response data from recruitment centers is the main contributing factor to the failure of multicenter RCTs testing antidepressant drugs, (ii) learning from historical databases can help shape the problem, (iii) the rate of occurrence of PoC RCT failure can be significantly reduced by including an enrichment strategy to the placebo response for each center before starting standard inferential statistical analysis, and (iv) simulations using an enrichment window with boundaries anchored to clinically relevant prior information suggest the possibility of achieving this goal. We recommend that the enrichment-window strategy be implemented to improve efficiency in the following types of multicenter RCTs: (i) conventional, (ii) run-in phase, and (iii) adaptive.

The advantages of the proposed approach could be a reduction in unnecessary exposure of subjects to novel antidepressant compounds, a reduction in drug development costs, and a reduction in the time required to bring novel therapeutic agents to patients who need them.

METHODS

Data. Data were derived from GSK clinical databases (GSK clinical trial register (http://ctr.gsk.co.uk/medicinelist.asp)). For model development and internal validation, we considered five randomized, double-blind, placebo-controlled, three-arm, parallel-group studies for the treatment of MDDs. The five trials were selected on the basis of the similarities in their key design factors, i.e., depression severity at baseline (HAMD ≥23), number of treatment arms ($n = 3$), and year of publication (2002–2004). Selection of studies and data extraction were performed using the SAS software.[39] Models were developed using data up to week 8, a treatment duration common to all the selected trials. Studies 448, 449, and 487 used the flexible dose scheme to evaluate clinical efficacy of paroxetine immediate release and modified release (CR) formulation.[40,41] Studies 810 and 874 were fixed-dose studies to evaluate the clinical efficacy of paroxetine CR at the doses of 12.5 and 25 mg/day.[42]

Clinical response model. The clinical response (to either placebo or drug) was defined by the time-varying HAMD scores, considered to be the "standard" end point in MDD RCTs.[43] The trajectory of this curve usually shows a nonlinear decrement from a high initial score (~23) to lower values (~10) associated with clinical remission, within 6–8 weeks of treatment, the typical time lag for reliably detecting clinical effects in MDD.[27]

In each of the five studies, the HAMD time courses in the three treatment arms were independently analyzed using a mixed Weibull linear equation:

$$f(t) = Ae^{-(t/t_d)^b} + h_{rec}t \quad (1)$$

where A, b, t_d, and h_{rec} were the fixed-effect parameters. A represents the baseline HAMD score, t_d is the time corresponding to 63.2% of the maximal change from baseline, b is the shape or sigmoidicity factor, and h_{rec} is the remission rate. This model was successfully applied to describe the placebo response in a large subject population with MDD.[19] The model

parameters were estimated using NONMEM (GloboMax, Hanover, MD).[44] The random effects were assumed to be normally distributed for A and log-normally distributed for t_d, b, and h_{rec}, with a mean of zero and variance Ω, with a proportional residual error model. The mean placebo responses of each recruitment center were estimated by averaging the Bayesian *post hoc* individual center-specific parameter estimates. An alternative methodology for computing the mean center response could have been to consider the centers as fixed-effect parameters. The major drawback of the latter option is the very high number of parameters to be included in the model (the number of centers varied between 19 and 38 in the trials). In addition to the potential model instability associated with the high number of parameters, the heterogeneity in the sample size for the different recruitment centers would have constituted an additional hurdle to achieving stable and precise parameter estimates. As an example, the population parameter estimates for the three treatment arms of the study 810 are shown in **Table 2**.

Treatment effect model. Assuming that the placebo response specific to each recruitment center is a major determinant of the probability to detect signals of drug efficacy, the typical antidepressant response in each recruitment center was determined as a function of the dose of the drug tested and of the center-specific level of placebo response. Pooled data from clinical trials that had similar patient populations, the same study design, the same efficacy end point, and the same drug and dosage regimen were used to derive the TE model. The shape of the curve describing placebo response and TE was expected to be curvilinear with an asymptotic upper limit (the true TE size), attainable with an indefinite decrease of the placebo HAMD response (flat placebo response curve). Given that all the patients included in the RCTs whose data were used for model development had similar disease severity, the TE was assumed to dependent only on the placebo response at the end of the study (Hend). Alternative models were explored, and the following exponential equation was retained as the best-fitting model:

$$TE = Hbas + (\delta - Hbas) \cdot (1 - e^{-slope \times Hend}) \quad (2)$$

where Hbas is the TE when the placebo HAMD score approaches zero, δ is the maximal TE when the placebo HAMD score at study-end approaches very high values, and slope is the slope factor of the curve.

The assumptions in the model were: (i) the TE trajectory is curvilinear with an asymptotic upper limit representing the true TE size, (ii) TE is assumed to be dependent on the placebo response at end point (Hend), given that all the patients included in the analysis had similar HAMD scores at baseline (Hbas), and (iii) the drug treatment response is proportional to the dosage regimen used. Three analyses were independently performed to estimate the typical TE dose response curves: (Parox 12.5 mg) combining data in the arms of the studies 810 and 874, which used paroxetine 12.5 mg CR; (Parox 25 mg) combining data in the arms of the studies 810 and 874, which used paroxetine 25 mg CR; and (Parox flex dose) combining data in the arms of the studies 448, 449, and 487, which used flex doses. In each analysis, the mean placebo response and TE were fitted to Equation 2.

Internal validation of the model. Internal validation was performed by comparing the model-predicted distribution of the individual HAMD time courses with the observed ones in (i) the 25 mg CR paroxetine arm of study 810 and (ii) the paroxetine CR flex-dose arm of study 448 using the visual predictive check approach.[45,46] Using the identified model, several replications based on the structure of the original data were simulated using a Monte Carlo approach, and the median with the 5th and 95th percentiles of these replicates were compared with actual observations. The good consistency between observation and predictions qualified the model. Individual HAMD trajectories were simulated using Equation 1, with the following assumptions on model parameters: (i) values of A (baseline HAMD values) are the same in the reference placebo arm, (ii) t_d (onset time of the response) is 50% lower with respect to the reference placebo (as expected in an active treatment, according to the data shown in **Table 3**), (iii) the shape factor, b, is the same in the reference placebo arm, and (iv) h_{rec} is adjusted according to Equation 2 in order to provide an end-of-study separation from placebo similar to the one estimated in the "25-mg" and "flex-dose" TE model fittings. One hundred replicates of the simulations were generated using the reference placebo distribution, and the 95 and 99% confidence regions were computed.

External validation of the model. External validation was performed by comparing predicted paroxetine TE with the response observed in three clinical trials that were not used for model development. For this purpose, we used the HAMD scores in the paroxetine 20 mg/day arms

Table 2 Study 810: final population longitudinal model parameter estimates (SE) in the three treatment arms

	Placebo		12.5 mg		25 mg	
	Fixed effect	Random effect	Fixed effect	Random effect	Fixed effect	Random effect
A	23.8 (0.32)	5.41 (1.43)	23.3 (0.27)	3.28 (1.15)	23.4 (0.31)	5.46 (1.43)
t_d	5.74 (0.69)	0.62 (0.13)	4.22 (0.36)	0.27 (0.07)	4.32 (0.35)	0.31 (0.08)
b	0.68 (0.12)	0.21 (0.06)	0.79 (0.09)	0.19 (0.08)	0.60 (0.08)	0.16 (0.10)
h_{rec}	0.90 (0.07)	0.24 (0.07)	1.01 (0.07)	0.19 (0.05)	1.06 (0.06)	0.16 (0.04)
Res. error	3.08 (0.09)		2.88 (0.09)		2.83 (0.09)	

A, baseline HAMD score; *b*, shape or sigmoidicity factor; HAMD, Hamilton Rating Scale for Depression; h_{rec}, remission rate; Res. error, residual error; t_d, time point corresponding to 63.2% of the maximal change from baseline.

Table 3 Details of the five clinical trials included in the meta-analysis

			Arm 1		Arm 2	
Study	Number of centers	Number of patients	Treatment	Dose	Treatment	Dose
448	19	299	Parox IR	20–50 mg flex	Parox CR	25–62.5 mg flex
449	20	333	Parox IR	20–50 mg flex	Parox CR	25–62.5 mg flex
487	26	319	Parox IR	10–40 mg flex	Parox CR	12.5–50 mg flex
810	38	489	Parox CR	12.5 mg fixed	Parox CR	25 mg fixed
874	21	397	Parox CR	12.5 mg fixed	Parox CR	25 mg fixed

CR, modified release; IR, immediate release.

Table 4 Details of the three clinical trials of duloxetine used in the external validation of the model

Study	Number of centers	Number of patients in the placebo arm	Treatment effect (Parox 20 mg–placebo)	Mean placebo HAMD baseline	Mean placebo HAMD at week 8
F1J MC-HMAT	22	89	2.62	17.79	13.01
Perahia *et al.*[48]	22	90	1.1	20.6	9.8
Detke *et al.*[47]	21	93	2.9	19.9	11.1

HAMD, Hamilton Rating Scale for Depression.

of three placebo-controlled clinical trials conducted to compare different doses of duloxetine in an 8-week acute treatment of patients with MDD.[47–49] Table 4 shows the details of the placebo HAMD response and paroxetine TE reported in the three clinical trials of duloxetine used for external validation. The model-predicted response for paroxetine 20 mg/day was estimated using parameters shown in Table 1 and assuming a dose-proportional behavior between the response at 12.5 mg/day and that at 25 mg/day, with a similar baseline value of disease severity.

Surface response and sensitivity analysis. A surface response model ($STE = f(Hend, d)$) was derived to describe the TE (STE) as a joined function of center-specific placebo response (Hend) and paroxetine dose (d). The surface response model was derived using local linear interpolation of model parameters associated with the dosages of 12.5, 25, and 42 mg/day (flex-dose response), reported in Table 1.

HAMD scores for the 12.5 and 25 mg/day fixed-dose groups were obtained from studies 810 and 874, and information for the HAMD response at higher doses was derived from flex-dose studies 448, 449, and 487, in which dosages ranging from 12.5 to 62.5 mg/day were used. In these studies, the distribution of the doses of paroxetine used at week 8 indicated that 8.2, 18.8, 24.5, 24.9, and 23.7% of the subjects were treated with 12.5, 25, 37.5, 50, and 62.5 mg/day, respectively. Because the large majority of the subjects received paroxetine doses >37.5 mg/day, we considered that the dosage of 42 mg/day (the average dosage at week 8) was a reasonable approximation of the typical dose characterizing the response in the flex-dose studies.

Sensitivity analysis was performed to determine the relative contributions of the factors placebo response and dose to TE. The normalized SI was computed as the partial derivative of STE with respect to each predictor variable ($P = d$, Hend) and by normalizing the derivative with respect to the placebo and dose nominal values:[50]

$$SI = \frac{\partial STE}{\partial P} \cdot \frac{P}{STE} \qquad (3)$$

Clinical trial simulation. Clinical trial simulation was used to explore the relationship between probability of success of an RCT and the use of an enrichment window (cutoff values). The simulations were performed using the distributions of center-specific placebo scores at the beginning (Figure 2a) and at the end (Figure 2b) of the study from data collected in the GSK Clinical Data Register (http://ctr.gsk.co.uk/medicinelist.asp).[17] Two-arm (placebo and paroxetine 25 mg/day) multicenter (30 centers) RCTs were simulated with different levels of placebo response randomly selected using a two-stage approach. Initially, the SAS PROC SURVEYSELECT procedure was used to resample the distribution of the 239 HAMD scores at baseline and at week 8 and to generate 1,000 replicates of these scores in the 30 centers.[33] The mean and SD values of the 1,000 HAMD replicates at baseline and at week 8 in each of the 30 centers were used to characterize the placebo response in the simulated clinical trials.

Next, the HAMD scores at baseline and at week 8, associated with 100 clinical trials, were randomly generated from this distribution and inserted into Equation 2 (with the model parameters reported in Table 1) to obtain the distribution of TE delivered by each center in a trial comparing placebo to paroxetine 20 mg/day. The probability of detecting clinically relevant effects was estimated as the proportion of clinical trials delivering TE >3 with respect to the 100 simulated trials.

CONFLICT OF INTEREST

The authors declared no conflict of interest.

1. Kessler, R.C. *et al.*; National Comorbidity Survey Replication. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA* **289**, 3095–3105 (2003).
2. American Psychiatric Association. *Diagnostic and Statistical Manual of Manual Disorders* 4th edn (Washington, DC, 1994).
3. Fava, M. & Kendler, K.S. Major depressive disorder. *Neuron* **28**, 335–341 (2000).
4. Yang, H., Cusin, C. & Fava, M. Is there a placebo problem in antidepressant trials? *Curr. Top. Med. Chem.* **5**, 1077–1086 (2005).
5. Kirsch, I., Deacon, B.J., Huedo-Medina, T.B., Scoboria, A., Moore, T.J. & Johnson, B.T. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med.* **5**, e45 (2008).
6. Turner, E.H., Matthews, A.M., Linardatos, E., Tell, R.A. & Rosenthal, R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N. Engl. J. Med.* **358**, 252–260 (2008).
7. *The Global Burden of Disease: 2004 Update* (WHO Press, World Health Organization, Geneva, Switzerland, 2008).
8. Nelson, J.C., Pikalov, A. & Berman, R.M. Augmentation treatment in major depressive disorder: focus on aripiprazole. *Neuropsychiatr. Dis. Treat.* **4**, 937–948 (2008).
9. Gelenberg, A.J. *et al.* The history and current state of antidepressant clinical trial design: a call to action for proof-of-concept studies. *J. Clin. Psychiatry* **69**, 1513–1528 (2008) .
10. Rush, A.J., Gullion, C.M., Basco, M.R., Jarrett, R.B. & Trivedi, M.H. The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychol. Med.* **26**, 477–486 (1996).
11. Thase, M.E. Comparing the methods to use to compare antidepressants. *Psychopharmacol Bull.* **36**, 1 (2002).
12. Khan, A., Detke, M., Khan, S.R. & Mallinckrodt, C. Placebo response and antidepressant clinical trial outcome. *J. Nerv. Ment. Dis.* **191**, 211–218 (2003).
13. Kaptchuk, T.J. *et al.* Do "placebo responders" exist? *Contemp. Clin. Trials* **29**, 587–595 (2008).
14. Papakostas, G.I. & Fava, M. Does the probability of receiving placebo influence clinical trial outcome? A meta-regression of double-blind randomized clinical trials in MDD. *Eur. Neuropsychopharmacol.* **19**, 34–40 (2009).
15. Lakoff, A. The Right Patients for the Drug: Managing the Placebo Effect in Antidepressant Trials. *BioSocieties* **2**, 57–71 (2007).
16. Walsh, B.T., Seidman, S.N., Sysko, R. & Gould, M. Placebo response in studies of major depression: variable, substantial, and growing. *JAMA* **287**, 1840–1847 (2002).
17. Merlo-Pich, E. & Gomeni, R. Model-based approach and signal detection theory to evaluate the performance of recruitment centers in clinical trials with antidepressant drugs. *Clin. Pharmacol. Ther.* **84**, 378–384 (2008).
18. Gobburu, J.V. & Lesko, L.J. Quantitative disease, drug, and trial models. *Annu. Rev. Pharmacol. Toxicol.* **49**, 291–301 (2009).
19. 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.* **63**, 595–613 (2007).
20. Varkonyi, P., Bruckner, J.V. & Gallo, J.M. Effect of parameter variability on physiologically-based pharmacokinetic model predicted drug concentrations. *J. Pharm. Sci.* **84**, 381–384 (1995).
21. Rush, A.J. *et al.* STAR*D: revising conventional wisdom. *CNS Drugs* **23**, 627–647 (2009).
22. Schatzberg, A.F. & Kraemer, H.C. Use of placebo control groups in evaluating efficacy of treatment of unipolar major depression. *Biol. Psychiatry* **47**, 736–744 (2000).

23. Kobak, K.A., Feiger, A.D. & Lipsitz, J.D. Interview quality and signal detection in clinical trials. *Am. J. Psychiatry* **162**, 628 (2005).
24. Rickels, K. & Robinson, D.S. Why do clinical trials fail? *J. Clin. Psychopharmacol.* **27**, 420–421; author reply 421 (2007).
25. Kobak, K.A. *et al*. Sources of unreliability in depression ratings. *J. Clin. Psychopharmacol.* **29**, 82–85 (2009).
26. Baker, C.B., Tweedie, R., Duval, S. & Woods, S.W. Evidence that the SSRI dose response in treating major depression should be reassessed: a meta-analysis. *Depress. Anxiety* **17**, 1–9 (2003).
27. Nierenberg, A.A. & Wright, E.C. Evolution of remission as the new standard in the treatment of depression. *J. Clin. Psychiatry* **60** (suppl. 22), 7–11 (1999).
28. Rush, A.J. *et al*.; ACNP Task Force. Report by the ACNP Task Force on response and remission in major depressive disorder. *Neuropsychopharmacology* **31**, 1841–1853 (2006).
29. Christiano, L.J. & Terry, J.F. The band pass filter. *Int. Econo. Rev.* **44**, 435–465 (2003).
30. Yamada, S. & Murase, K. Effectiveness of flexible noise control image processing for digital portal images using computed radiography. *Br. J. Radiol.* **78**, 519–527 (2005).
31. Uchida, Y. *et al*. Detection of vulnerable coronary plaques by color fluorescent angioscopy. *JACC. Cardiovasc. Imaging* **3**, 398–408 (2010).
32. Rodts, S., Bytchenkoff, D. & Fen-Chong, T. Cardinal series to filter oversampled truncated magnetic resonance signals. *J. Magn. Reson.* **204**, 64–75 (2010).
33. Merlo-Pich, E., Bettica, P. & Gomeni, R. Bayesian monitoring and bootstrap trial simulation: A new paradigm to implement adaptive trial design for testing antidepressant drugs. *Open Psychiatr. J.* **3**, 20–32 (2009).
34. Müller, P., Berry, D.A., Grieve, A.P. & Krams, M. A Bayesian decision-theoretic dose-finding trial. *Deci. Anal.* **3**, 197–207 (2006).
35. Rosenberger, W.F. & Huc, F. Maximizing power and minimizing treatment failures in clinical trials. *Clin. Trials* **1**, 141–147 (2004).
36. Fava, M., Evins, A.E., Dorer, D.J. & Schoenfeld, D.A. The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychother. Psychosom.* **72**, 115–127 (2003).
37. Noble, R.E., Gelfand, L.A. & DeRubeis, R.J. Reducing exposure of clinical research subjects to placebo treatments. *J. Clin. Psychol.* **61**, 881–892 (2005).
38. Sheiner, L.B. & Rubin, D.B. Intention-to-treat analysis and the goals of clinical trials. *Clin. Pharmacol. Ther.* **57**, 6–15 (1995).
39. SAS Institute Inc. *SAS/STAT® User's Guide, Version 9.2* (SAS Institute Inc., Cary, NC, 2008).
40. Golden, R.N., Nemeroff, C.B., McSorley, P., Pitts, C.D. & Dubé, E.M. Efficacy and tolerability of controlled-release and immediate-release paroxetine in the treatment of depression. *J. Clin. Psychiatry* **63**, 577–584 (2002).
41. Rapaport, M.H., Schneider, L.S., Dunner, D.L., Davies, J.T. & Pitts, C.D. Efficacy of controlled-release paroxetine in the treatment of late-life depression. *J. Clin. Psychiatry* **64**, 1065–1074 (2003).
42. Trivedi, M.H., Pigotti, T.A., Perera, P., Dillingham, K.E., Carfagno, M.L. & Pitts, C.D. Effectiveness of low doses of paroxetine controlled release in the treatment of major depressive disorder. *J. Clin. Psychiatry* **65**, 1356–1364 (2004).
43. Hedlund, J. & Vieweg, B. The Hamilton rating scale for depression. *J. Operat. Psychiatry* **10**, 149–165 (1979).
44. Beal, S. & Sheiner, L.B. *NONMEM User Guides: Parts I–VIII*. (Hanover, MD, GloboMax, 1989–2008).
45. Yano, Y., Beal, S.L. & Sheiner, L.B. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. *J. Pharmacokinet. Pharmacodyn.* **28**, 171–192 (2001).
46. Post, T.M., Freijer, J.I., Ploeger, B.A. & Danhof, M. Extensions to the visual predictive check to facilitate model performance evaluation. *J. Pharmacokinet. Pharmacodyn.* **35**, 185–202 (2008).
47. Detke, M.J., Wiltse, C.G., Mallinckrodt, C.H., McNamara, R.K., Demitrack, M.A. & Bitter, I. Duloxetine in the acute and long-term treatment of major depressive disorder: a placebo- and paroxetine-controlled trial. *Eur. Neuropsychopharmacol.* **14**, 457–470 (2004).
48. Perahia, D.G., Wang, F., Mallinckrodt, C.H., Walker, D.J. & Detke, M.J. Duloxetine in the treatment of major depressive disorder: a placebo- and paroxetine-controlled trial. *Eur. Psychiatry* **21**, 367–378 (2006).
49. Eli Lilly and Company. Clinical Study Summary: Study F1J MC-HMAT Study Group A [online] <http://www.lillytrials.com/results_files/cymbalta/cymbalta_summary_4091a.pdf>.
50. Abraham, A.K., Krzyzanski, W. & Mager, D.E. Partial derivative-based sensitivity analysis of models describing target-mediated drug disposition. *AAPS J.* **9**, E181–E189 (2007).