J Pharmacokinet Pharmacodyn (2009) 36:63–80
DOI 10.1007/s10928-009-9110-3

Evaluation of structural models to describe the effect of placebo upon the time course of major depressive disorder

Elizabeth Y. Shang · Megan A. Gibbs · Jaren W. Landen · Michael Krams · Tanya Russell · Nicholas G. Denman · Diane R. Mould

Received: 2 January 2008 / Accepted: 15 January 2009 / Published online: 10 February 2009

Abstract Major depressive disorder (MDD) is the leading cause of disability in many countries. Designing and evaluating clinical trials of antidepressants is difficult due to the pronounced and variable placebo response which is poorly defined and may be affected by trial design. Approximately half of recent clinical trials of commonly used antidepressants failed to show statistical superiority for the drug over placebo, which is partly attributable to a marked placebo response. These failures suggest the need for new tools to evaluate placebo response and drug effect in depression, as well as to help design more informative clinical trials. Disease progression modeling is a tool that has been employed for such evaluations and several models have been proposed to describe MDD. Placebo data from three clinical depression trials were used to evaluate three published models: the inverse Bateman (IBM), indirect response (IDR) and transit (TM) models. Each model was used to describe Hamilton Rating Scale for major depression (HAMD) data and results were evaluated. The IBM model had several deficiencies, making it unsuitable. The IDR and TM models performed well on most evaluations and

E. Y. Shang (✉) · M. A. Gibbs · J. W. Landen · M. Krams · T. Russell
Pfizer Global Research and Development, Pfizer Inc., 50 Pequot Ave, New London, CT 06320, USA
e-mail: eyshang@gmail.com

Present Address:
M. A. Gibbs
Pharmacokinetics and Pharmacometrics, Amgen, 1201 Amgen Ct. West, Seattle, WA 98119, USA

Present Address:
M. Krams
Wyeth Pharmaceuticals Inc., 500 Arcola Road, Collegeville, PA 19426, USA

N. G. Denman
School of Physical Sciences, University of Queensland, Brisbane, Australia

D. R. Mould
Projections Research Inc., 535 Springerview Lane, Phoenixville, PA 19460, USA

appear suitable. Comparing the IDR and TM models showed less clear distinctions, although overall the TM was found to be somewhat better than the IDR model. Model based evaluation can provide a useful tool for evaluating the time course of MDD and detecting drug effect. However, the models used should be robust, with well estimated parameters.

Keywords NONMEM · Major depressive disorder · HAMD · Model discrimination · Disease progression model

Introduction

Effective treatment of major depressive disorder (MDD) has been available for many years including antidepressant agents. However, designing clinical trials of antidepressants is challenging because of the pronounced placebo response that is not well defined and may be affected by trial design. More than half of recent clinical trials of marketed antidepressants failed to show statistical superiority for the drug over placebo [1]. This is not necessarily because of ineffectiveness of the antidepressant, but because of a large placebo response [1]. Existing data suggest that placebo response in antidepressant clinical trials is associated with a reduced likelihood of demonstrating statistical superiority of antidepressant treatment over placebo [2]. Results from reviewing 52 randomized, double-blind, placebo-controlled clinical trials obtained from the FDA [3] showed that trials with large placebo response, only 21% found active treatment superior. In trials with low placebo response, 74% showed superiority of active treatment.

Many attempts have been made to improve the design of MDD trials [4, 5], however, they were ineffective in decreasing placebo response, possibly partly due to the cyclic nature of MDD [6]. Another factor making the detection of drug effect difficult may be related to the statistical analysis used to interpret data [7]. Newer methods for analyzing data, both in terms of exploiting advances in statistical methods and quantitative analysis including population modeling could be considered [8].

The purpose of this work was to evaluate three published disease progression models to quantify the placebo response in MDD, enabling better detection of drug effect and ultimately guide antidepressant trial design. In addition, the utility of each model to predict placebo response in different trial designs was evaluated.

Methods

Study design

Three randomized, double-blinded, placebo-controlled clinical trials sponsored by Pfizer evaluating the efficacy of antidepressants in MDD were included. All studies were approved by internal review boards, and informed consents were collected from all subjects. All subjects met DSM-IV criteria for MDD.

All three trials had a run-in period where eligible subjects were observed without treatment and Hamilton Rating Scale for major depression (HAMD) scores were collected at Screening and the end of run-in period, which is a standard procedure to eliminate placebo responders [9]. Subsequently, subjects without improvement were randomized to active or placebo treatment. Trials 2 and 3 had 7-day run-in periods. Subjects randomized to placebo received daily oral placebo doses for 6 weeks. These two trials reflect the typical design for Phase 2 antidepressant trials [10] and each included over 150 subjects. Trial 3 was used as an external validation dataset. Trial 1 had ~50 subjects enrolled. Paroxetine was administered to all subjects and continued in the placebo arm. A 3-week run-in period was implemented in this trial to eliminate responders to paroxetine. This trial was considered relevant for use in the present evaluation because subjects were refractory to paroxetine therapy. Subjects received only single doses of placebo via intravenous infusion for 1.5 h and stayed in-house for Day 1, however, HAMD scores were measured at various time points after the cessation of placebo treatment, which was not available in the other two trials.

The version of HAMD used in these studies contains 17 items (HAMD-17) with each item containing 3 or 5 options described by a short sentence [11]. HAMD-17 questionnaires were administered to patients by qualified raters, who were asked to select one of several options for each item. Depending on the options selected, a score of 0 to 4 with increments of 1 per item contributed toward the total HAMD-17 score.

A brief listing of study designs and summary demography of each trial evaluated is presented in Table 1. A representative spaghetti plot describing individual time courses of HAMD-17 scores after multiple-dose placebo treatment (Trial 2) is shown in Fig. 1.

NONMEM database construction

NONMEM databases were constructed for all three trials using recorded dates and observed HAMD-17 data. Observation times were nominally set to 12:00 pm for all observations. For the inverse Bateman model (IBM) databases, run-in observations did not contribute information to the placebo response and would have had to have been described using a lag time, which introduces a discontinuity. Consequently these observations were removed from the databases used for evaluations of IBM. However, run-in data is informative and was retained in the database for the indirect response (IDR) and transit models (TM).

Structural models tested

Several structural models (IBM, IDR and TM) were tested to examine their ability to describe HAMD-17 scores over time. The pathophysiological basis for depression is not well understood and all functions are largely empirical in application. These functions were selected for evaluation because they had been used in published work to describe the time course of depression [12–15] and also

Table 1 Study designs and summary of baseline demographic information

Study design						Demographic		
Study	Number of subjects/observations	Route of administration	Treatment duration/study duration	Dose regimen	Schedule of HAMD-17 evaluations	Age (yr) mean (SD)/median (Range)	Sex	Baseline HAMD-17 (Units) mean (SD)/median (Range)
Trial 1	46/312	IV Infusion	1.5 h/15 Days	Single dose	Screening and days 1, 2, 5, 8, 12, and 15	39.5 (8.54)/39 [26 – 54]	Males = 41; Females = 5	21.5 (2.31)/21 [18–27]
Trial 2	177/848	Oral	6 weeks/ 6 weeks	Daily	Screening and weekly for 6 weeks	42.1 (12.9)/42 [19–78]	Males = 68; Females = 109	25.2 (2.38)/24 [22–32]
Trial 3 (External validation dataset)	201/1236	Oral	6 weeks/ 6 weeks	Daily	Screening and weekly for 6 weeks	39.3 (11.6)/40 [18–65]	Males = 68; Females = 133	23.2 (11.6)/23 [9–35]

NB—Trials 2 and 3 were pooled databases. Each database consisted of 2 studies of the same active agent. Study designs were comparable for all studies

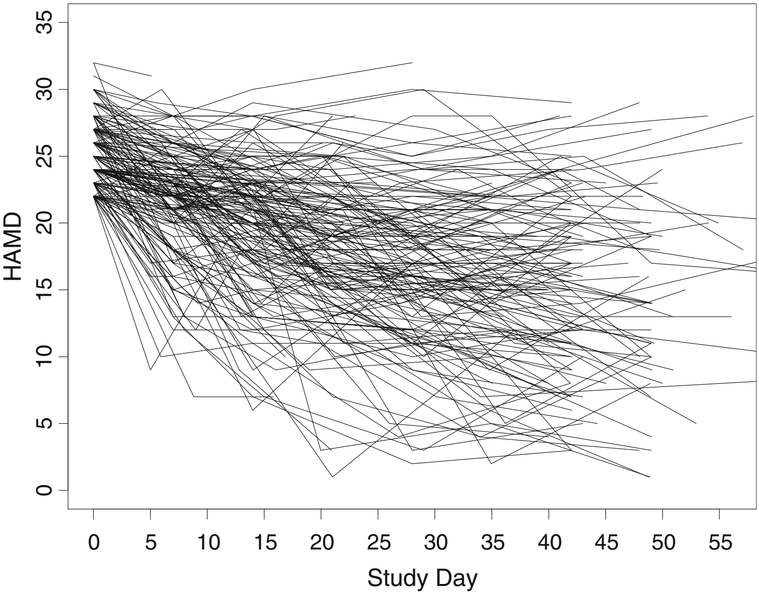

Fig. 1 Individual HAMD-17 scores vs. time profile from Trial 2 following daily placebo dosing. Solid lines are individual HAMD-17 data

provided a sufficiently flexible function that would allow a delay between the onset of treatment and observed response [16–18].

$$HAMD = S0 - DREC \cdot \left(\frac{\frac{Ln2}{THR}}{\frac{Ln2}{THR} - \frac{Ln2}{THO}} \right) \cdot \left(e^{\left(-\frac{Ln2}{THO} \cdot t\right)} - e^{\left(-\frac{Ln2}{THR} \cdot t\right)} \right) + \varepsilon_{Additive} \qquad (1)$$

For the IBM model (Eq. 1), S0 is the baseline HAMD-17 score. THR is the apparent half-life of spontaneous recovery from depression; the rate of spontaneous recovery is equal to the natural log of 2 divided by THR. THO is the half-life of spontaneous worsening of depression. The function describes a smooth decrease over time from a baseline with the maximum decrease from baseline being determined by DREC (a scaling factor). Between subject variability (BSV) was described assuming a log normal distribution with the exception of S0 which assumed a normal distribution. Residual error was described using a homoscedastic function. The HAMD-17 versus time curve described by this function is independent of treatment duration.

In the absence of treatment the IDR function assumes no change in HAMD-17 and consequently requires a forcing function to alter HAMD-17 scores. This was accomplished by assuming that the placebo effect is described by a “placebo concentration”. The placebo model was a one compartment linear model with zero order bolus input and first order elimination:

$$\frac{dPlacebo}{dt} = -ke_{placebo} \cdot Aplacebo \qquad (2)$$

The “dose” of placebo was assumed to be 1 unit and the dose regimen for placebo was set to mimic study treatment. The structural parameters for the placebo model were empirically selected to allow a single placebo dose to reach a uniform "concentration" of 1 and to be completely cleared by 24 h. For both trials, the infusion rate was 1 "placebo unit/hour". In Trial 1, two placebo doses were administered together to account for different components of the placebo effects: 1.5 units (7 subjects got 8 units) for placebo drug infusion and 24 units for one-day in-house stay which may contribute to the placebo responses [19, 20]. In Trial 2, the placebo was given for 6 weeks at a daily dose of 24 "placebo units". The clearance (CLplacebo) was 3 L/day and the volume of distribution (Vplacebo) was 1 L and $ke_{placebo} = CLplacebo/Vplacebo$. There were no terms for BSV or residual error in this model. The effect of placebo concentration (Cplacebo) was described as a linear function applied as a stimulatory effect on kout using a scale factor (Slope) with a BSV term to adjust for different degrees of individual response. Other variants of this model were evaluated but were rejected due to model performance.

$$\frac{dHAMD}{dt} = kin - kout \cdot HAMD \cdot (1 + Slope \cdot Cplacebo) \tag{3a}$$

$$HAMD = HAMAD(t) + \varepsilon_{Additive} \tag{3b}$$

For the IDR model (Eqs. 3a and 3b), kin is the rate of spontaneous worsening of depression and kout is the rate of spontaneous improvement. BSV was described assuming log normal distributions. Residual error was described with a homoscedastic function. The system was initialized to the baseline HAMD score.

$$\frac{dPREC}{dt} = k \cdot S0 \cdot (1 - Slope \cdot Cplacebo) - k \cdot PREC \tag{4a}$$

$$\frac{dT1}{dt} = k \cdot PREC - k \cdot T1 \tag{4b}$$

$$\frac{dT2}{dt} = k \cdot T1 - k \cdot T2 \tag{4c}$$

$$\frac{dHAMD}{dt} = k \cdot T2 - k \cdot HAMD \tag{4d}$$

$$HAMD = HAMAD(t) + \varepsilon_{Additive} \tag{4e}$$

For the TM model (Eqs. 4a–4e), S0 is the baseline HAMD-17 score, PREC is the amount in a precursor pool compartment, T1 and T2 are transit compartments and HAMD-17 is the observation compartment. Rate constants of transfer between compartments (k) were set to be the same. The value for k was determined based on the mean transit time (MTT) between compartments such that k = MTT/3. BSV was described assuming a log normal distribution. Residual error was described using a homoscedastic function. The system was initialized to the baseline HAMD-17 score.

Similar to IDR, the TM assumes that without intervention, the HAMD-17 values will not change. Consequently the same placebo forcing function and placebo pharmacokinetic model parameters described earlier was applied to the TM.

Model based evaluation

Model building and parameter estimation were performed using NONMEM Version V Level 1.1 [21–23]. NONMEM was compiled using Compaq Digital Visual Fortran 6.6.3C. The analyses were performed on an Intel Pentium 4 3.2 GHz processor running Windows XP.

Model building followed standard approaches. Various omega matrix structures were tested. All models used the First Order Conditional Estimation method. The IBM was run using $PRED with 3 significant digits requested. The IDR and TM models were run using ADVAN6 TRANS1 [24] with the tolerance set to 4 and 3 significant digits requested. All the parameters in the structure models were estimated. Placebo model parameters were fixed.

Model performance was evaluated by standard diagnostic plots (e.g., observed, typical predicted and individual predicted HAMD-17 versus time, observed versus typical predicted HAMD-17 values, and conditional weighted residuals versus time). 95% confidence intervals (CI) for all models were also generated using a nonparametric bootstrap approach [25]. One thousand bootstrap replicates were generated and fitted using the final models for each trial. CIs were determined from the results from all replicates.

A numerical predictive check (NPC) was conducted comparing summary statistics (mean and standard deviation) for the observed versus simulated HAMD-17 scores at the study day assumed to have the nadir HAMD-17 score. Simulations were conducted using the model parameters derived for both trials and compared against the observed nadir values. Days 5 and 42 after treatment initiation (after run-in period completed) were used for the evaluation of Trials 1 and 2, respectively.

$$Shrinkage = 1 - \frac{SD_{\eta parameter}}{\Omega_{parameter}} \tag{5}$$

Bayesian shrinkage [26, 27] (Eq. 5) was calculated where $\mathrm{SD}_{\eta \mathrm{parameter}}$ is the standard deviation of the individual estimates of η for each parameter and Ω is the square root of the estimated population variance.

Parameter reproducibility across different study designs was evaluated and the estimatability of the parameters for the different models was assessed by evaluating the determinant of the population Fisher information matrix [28–31]. The software package WinPOPT was used for construction and evaluation of the information matrix [32].

Model validation was performed by visual predictive check (VPC) [27, 33, 34] using the parameter values for each model evaluated for Trial 2 overlaid with the external validation database (Trial 3). Each simulation contained 1,000 replicates. For each model, all simulated profiles were pooled together and the 2.5 and 97.5% intervals were calculated and visually compared with the observed data. In addition, VPC was performed using published IBM parameters [35] and the 2.5 and 97.5% interval for each time point were calculated and compared with observed data from Trial 3.

Results

Parameter estimates, associated asymptotic standard errors, nonparametric bootstrapped 95% CIs, and the expected standard error for the IBM, IDR, and TM are presented in Tables 2, 3, and 4, respectively. The results of the NPC for all models are presented in Table 5.

Inverse Bateman model

The parameters generally appear to be reasonably well estimated, with low standard errors for most parameters, although THO generally had the poorest standard errors. However, the bootstrap CIs are generally wider than the parameter standard errors. Furthermore, parameter estimates for both trials exhibited a trend of the value for THO equaling or exceeding the length of the trial duration. Also the value for DREC was larger for the longer trial (Trial 2) than for Trial 1.

Trial 1, which collected HAMD-17 scores at various time points after the cessation of treatment, had acceptable shrinkage estimates for most parameters except THR. However, the shrinkage estimates are generally high (>0.2) for Trial 2, particularly on parameter DREC, suggesting that individual parameter estimates are poor.

The expected standard errors for THR and THO are large for both trial designs, particularly THO for Trial 2, suggesting that there is insufficient information in the present study designs to evaluate this parameter. The expected standard errors are considerably larger than the asymptotic standard errors obtained from the model based evaluations, but are consistent with the bootstrap CIs and the shrinkage assessments. Trial 1 had lower expected standard errors than Trial 2, which is likely due to the fact that this study had HAMD-17 data taken after treatment ceased.

Although not presented here, the structural parameter estimates were highly correlated in these models. Furthermore, covariance terms were often near unity when evaluated. Both findings suggest that the IBM was over-parameterized for both study designs evaluated.

The results of NPC indicated that the standard deviation (SD) of the observed data is greater than the SD of the simulated data for all evaluations. For the first evaluation which used the model and data from Trial 1, the summary statistics are in good agreement. However, for the second and third comparisons, which evaluated the model from one trial using the data from the other trial, the mean of the simulated data is approximately 3 units higher than the observed data. In the last comparison which used the model and data from Trial 2, the mean of the simulated data was approximately 2 points lower than observed. The deviations seen for the second and third comparisons are as expected given the differences in study designs and model parameters.

The result of the VPC of the IBM (Fig. 2 Panel a) shows a slight under-prediction of observed data, with a greater proportion of points above the 97.5% interval. The result of the VPC using published IBM parameters (Fig. 2 Panel b) shows similar under predicted the observed data. Furthermore, the lower bound of 95% interval became negative after 1 week, and therefore was fixed to 0 for presentation purpose.

Table 2 Parameter estimates (%CV), associated shrinkage estimates and expected standard errors of parameter estimates for the inverse Bateman model

Parameter	Model estimates		Shrinkage	Expected standard error (%CV)	
	Population value (%CV) [95% CI]	Between subject variability (%CV) [95% CI]		Population value	Between subject variability
Trial 1					
S0 (Units)	21.6 (2.10) [20.0–22.2]	NE	–	12.1	244
THR (Days)	0.962 (34.7) [0.329–6.87]	76.6 (56.4) [10.9–181]	0.34	66.0	74.4
THO (Days)	18.4 (54.9) [1.10–154]	178 (23.1) [114–432]	0.12	86.6	203
DREC (HAMD-17)	7.65 (15.2) [4.55–25.6]	51.1 (59.8) [10.9–897]	0.093	23.8	26.3
Residual Error (Units)		2.98 (8.3) [2.61–3.42]		5.33	
Bootstrap results[a] (success/terminated due to rounding error/abnormal) 488/510/2					
Trial 2[b]					
S0 (Units)	25.1 [24.8–25.6]	6.16 [4.31–7.54]	0.28	0.700	25.0
THR (Days)	50.2 [18.9–98.2]	59.0 [22.9–81.3]	0.11	126	50.7
THO (Days)	84.9 [17.6–500]	125 [70.1–141]	0.76	652	16.5
DREC (HAMD-17)	26.6 [12.8–70.3]	56.8 [2.27–189]	0.28	112	632
Residual error (Units)		2.69 [2.39–2.91]		2.95	
Bootstrap results[a] (success/terminated due to rounding error/abnormal) 530/468/2					

NE not estimated

[a] 1000 Nonparametric bootstrap runs

[b] Standard error not estimatable

Table 3 Parameter estimates (%CV) and associated shrinkage estimates and expected standard errors of parameter estimates for the indirect response model

Parameter	Model estimates		Shrinkage	Expected standard error (%CV)	
	Population value (%CV) [95% CI]	Between subject variability (%CV) [95% CI]		Population value	Between subject variability
Trial 1					
Kin (1/Day)	0.110 (297) [0.0791–15.0]	7.50 (52.5) [5.17–20.8]	0.326	231	68.4
Kout (HAMD-17/Day)	0.00485 (297) [0.0035–0.714]	NE	–	231	–
Slope	0.942 (275) [0.0110–1.50]	223 (76.6) [0–359]	0.313	209	28.1
Residual error (HAMD-17)		3.41 (0.300) [3.18–5.37]		4.17	
Bootstrap results[a] (success/terminated due to rounding error/abnormal)		620/380/0			
Trial 2					
Kin (1/Day)	0.360 (12.7)[0.195–0.623]	6.76 (24.7) [5.22–7.87]	0.294	25.2	17.9
Kout (HAMD-17/Day)	0.0148 (12.2) [0.00810–0.0259]	NE	–	25.2	–
Slope	0.125 (12.4) [0.081–0.209]	103 (15.8)[87.4–119]	0.173	18.6	20.1
Residual error (HAMD-17)		2.69 (4.10) [2.48–2.91]		2.31	
Bootstrap results[a] (success/terminated due to rounding error/abnormal)		996/4/0			

NE not estimated

[a] 1000 Nonparametric bootstrap runs

Table 4 Parameter estimates (%CV) and associated shrinkage estimates and expected standard errors of parameter estimates for the transit model

Parameter	Model estimates		Shrinkage	Expected standard error (%CV)	
	Population value (%CV) [95% CI]	Between subject variability (%CV) [95% CI]		Population value	Between subject variability
Trial 1					
Baseline (HAMD-17)	21.8 (3.00) [21.6–22.5]	3.34 (461) [0.100–5.62]	0.293	2.29	33.9
MTT (Hours)	241 (17.8) [42.6–816]	NE	–	20.3	–
Slope	0.0604 (59.1) [0.00100–0.119]	181 (104) [155–666]	0.291	37.6	58.4
Residual error (HAMD-17)		3.35 (9.30) [3.21–3.83]		4.17	
Bootstrap results[a] (success/terminated due to rounding error/abnormal) 340/660/0					
Trial 2					
Baseline (HAMD-17)	24.0 (1.00) [23.6–24.3]	6.43 (27.1) [3.08–4.16]	0.250	0.780	24.8
MTT (Hours)	203 (2.70) [166–251]	NE	–	4.58	–
Slope	0.0362 (7.00) [0.0310–0.0420]	67.0 (21.3) [56.7–80.3]	0.133	5.64	12.6
Residual error (HAMD-17)		3.02 (4.40) [2.75–3.27]		2.55	
Bootstrap results[a] (success/terminated due to rounding error/abnormal) 996/4/0					

NE not estimated

[a] 1000 Nonparametric bootstrap runs

Table 5 Numerical predictive check for inverse Bateman, indirect response and transit models of simulated and observed Nadir HAMD-17 using models and data from Trial 1 and Trial 2

	Model from Trial 1/Data from Trial 1		Model from Trial 1/Data from Trial 2		Model from Trial 2/Data from Trial 1		Model from Trial 2/Data from Trial 2	
	Simulated	Observed	Simulated	Observed	Simulated	Observed	Simulated	Observed
Inverse Bateman model								
Mean	15.4	15.8	19.5	16.5	19.0	15.8	14.4	16.5
SD	4.44	6.35	4.31	8.70	5.11	6.35	5.68	8.70
Indirect response model								
Mean	17.4	15.8	17.5	16.5	16.2	15.8	13.6	16.5
SD	4.16	6.35	7.09	8.70	5.39	6.35	6.37	8.70
Transit model								
Mean	16.2	15.8	15.9	16.5	15.3	15.8	15.0	16.5
SD	5.36	6.35	6.94	8.70	6.01	6.35	6.44	8.70

Resampling of the residual error term for negative values was not conducted due to the bias induced by this procedure.

Indirect response model

For Trial 1, the standard errors of the parameter estimates are high for all parameters. The bootstrapped 95% CIs are wide and consistent with estimated standard errors. For Trial 2, the standard errors are low, suggesting the parameters are well estimated. The bootstrapped 95% CIs are narrow and consistent with asymptotic standard errors.

Unlike the results for the IBM, there were no trends in parameter estimates with different study durations. The structural parameter estimates were not correlated and when evaluated, covariance terms were generally reasonable with correlations less than 0.5 (data not presented).

The estimates for shrinkage were generally high (>0.2) for most of the parameters estimated in each evaluation, although the shrinkage of the SLOPE parameter is good for Trial 2.

The expected standard errors for Trials 1 and 2 are presented in Table 3. For Trial 1, the expected standard errors are high and generally consistent with the observed standard errors, likely due to fewer subjects. For Trial 2, the expected standard errors are reasonable and consistent with observed values.

The results of the NPC indicated that the SD of the observed data is still greater than the SD of the simulated data for all evaluations, although the differences are smaller than were seen with the IBM. For the first three evaluations, the summary statistics are in good agreement with the mean of the simulated HAMD-17 scores. However for the last evaluation, the mean of the simulated data is approximately 3 units lower than the mean of the observed data, suggesting that the typical predicted placebo effect in this model may be biased.

The results of the VPC (Fig. 2 Panel c) show overall good agreement between simulated and observed data with a slight over-prediction on HAMD-17 scores in first 2 weeks.

Transit model

For both trials the standard errors for all structural parameters are low, suggesting that the parameters were well estimated. This is consistent with the narrow bootstrapped 95% CIs for all structural parameters. Standard errors for variance terms are larger for Trial 1 than for Trial 2 but this is likely due to the fact that Trial 1 had fewer subjects.

As was seen with the IDR, there were no trends in parameter estimates with different study durations for the TM. The structural parameter estimates were not correlated, and covariance terms were generally reasonable (data not presented).

The estimates for shrinkage were generally poor (>0.2) for many parameters, except the SLOPE parameter for Trial 2. Overall, shrinkage for TM parameters appears to be the best relative to the other two models evaluated.

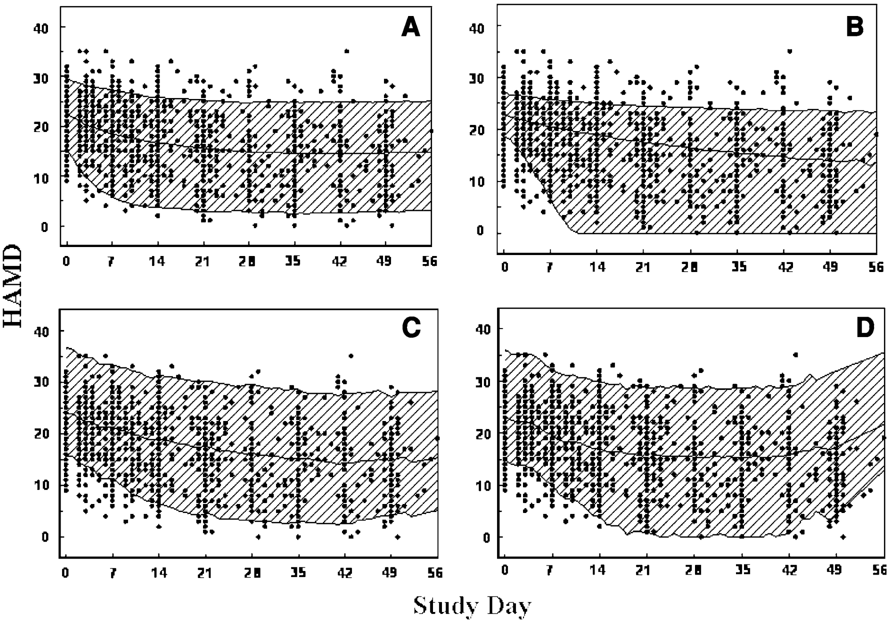

Fig. 2 External validation by visual predictive checks for inverse Bateman model, indirect response model and transit model. Filled circles are observed HAMD-17 data, the shaded area is the 95% prediction interval and the solid line is the median of the simulated data. Panel **a** Inverse Bateman model from Trial 2/data from Trial 3; Panel **b** Inverse Bateman model from published work [35]/data from Trial 3; Panel **c** Indirect response model from Trial 2/data from Trial 3; Panel **d** Transit model from Trial 2/data from Trial 3

The expected standard errors are presented in Table 4. They are generally low and consistent with the observed asymptotic standard errors for both models evaluated. The expected standard errors for the TM are consistently better than for the other models evaluated in the present assessment.

The results of the NPC indicated that the SD of the observed data is still greater than that of the simulated data for all evaluations, however these values are closer to the SDs of the observed data than for the other two models evaluated. For the TM, all comparisons made for the NPC are in good agreement.

The results of the VPC (Fig. 2 Panel d) were similar to IDR. There is good agreement between simulated and observed data with a slight over-prediction on HAMD-17 scores in first 2 weeks.

Discussion

Several models proposed for describing the time course of placebo response in MDD have been evaluated. The first model was the IBM [12] which has distinct advantages including using an integrated function with shorter run times. In addition, the general structure of this model is reasonably familiar to the users and the parameters can be associated with observable clinical events such as recovery and relapse of depression.

However the IBM has a tendency to estimate a half-life of recovery that is on the order of the duration of the study, and a half-life of relapse that is generally longer than the observation period. This trend has been noted in other presentations of this model [35] and is possibly due to the fact that MDD trials usually do not have observations from post-treatment follow-up periods, and HAMD-17 scores tend to continue decreasing over the course of the trial (Fig. 1). As the IBM is independent of treatment duration, parameter estimates for relapse are dependent on study duration. Only Trial 1, which had a follow-up period of observation, had recovery and onset parameter estimates that did not exceed the study observation period. The lack of data describing relapse may have caused the high shrinkage estimates for Trial 2, which was a large study. Given model complexity, and the lack of data on relapse, the observed correlations between the structural parameter estimates seen with this model for both study designs were expected. Although the estimated standard errors from this model are small, parameter values appear to be unsupported by the data resulting in high correlations between the structural parameters and the nearly perfect correlation between the variance terms in the model. Such results suggest that the model is over-parameterized, which would likely inflate Type I and Type II errors when evaluating the effect of new antidepressants.

The IBM was found to be unreliable for simulation of studies of different durations. Some of the issues with model performance for Trial 1 may be due to the fact that it is a small study, however, the same problems occurred in the larger trial (Trial 2).

The second model evaluated was the IDR. There have been several modifications of this model used to describe depression data. Gruwez et al. [36] proposed a slightly modified version of the IDR which included a transduction component to account for homeostatic control mechanisms. In this model, antidepressants exert their effect by either increasing the transduction set-point or the rate of feedback mechanisms. A second variant of the IDR was the K-PD model [37, 38] which uses a dose based forcing function (e.g., a virtual dose driving rate) instead of measured concentrations to affect a change in the observed depression scores. The model was successfully applied to placebo data [15]. Although it was not applied for physiological reasons, IDR permits a slow onset of effect and a relapse to baseline after treatment ceases. It requires the use of differential equations leading to longer run times, and the parameters are less translatable to clinical events than the IBM parameters.

There were no trends in the parameter estimates for the IDR and structural parameter estimates were uncorrelated suggesting that the model was not over-parameterized. Shrinkage estimates for the IDR were generally lower than for the IBM. For Trial 1, there was still substantial shrinkage, possibly due to the small number of subjects in the study. The shrinkage estimates for the larger study (Trial 2) were generally better. Other model diagnostics were comparable with the IBM, suggesting that both functions could adequately describe the data. The application of the IDR to simulate various study designs did not show the same liabilities as the IBM as the parameter estimates are more robust to study design because this model requires “dosing input” to drive response making the model sensitive to aspects of study design. Consequently, relapse would not be expected to occur during treatment which is consistent with the trends in the observed data in the present analysis.

The last model evaluated in the present work is the TM. We recently proposed this model to describe placebo response in antidepressant trials [14]. This model has the same strengths and weaknesses as the IDR and the model diagnostics and performance were generally consistent with the IDR. As with all the models evaluated, the model performance for Trial 1 suffered somewhat due to the limited number of subjects. There was no evidence of over parameterization and the TM was capable of simulating different study designs.

Overall, the IDR and TM showed reasonably robust parameter estimates that did not alter with different study durations, as the IBM did. However, the smaller database (Trial 1) had larger associated standard errors. The parameter estimates did not show a high correlation, and the covariance terms, when estimated, were reasonably small. The expected standard errors and the bootstrapped CIs were also small, suggesting that parameters for both models were better estimated than the IBM. The predictive capacity was good, with the numerical and visual predictive checks performing well on both models.

While the IBM was clearly the inferior model, comparison of the IDR and TM was less clear. The standard errors and expected standard errors for the TM were somewhat lower than for the IDR. The predictive check capacity was slightly better. Overall, the TM appeared to be the best choice for describing the time course of HAMD-17 data although the IDR was quite similar.

Interpreting the results of clinical studies of depression is difficult. Aside from issues of compliance and dropout, the endpoint is highly variable, and the placebo effect can be substantial. Model based evaluation can provide a useful tool as well as the greater statistical power [39] for estimating the time course of depression and detecting drug effect. As illustrated by Mould et al. [40], model based analysis using TM was superior to the empirical analysis in detecting antidepressant effect as small as a 2-unit change of HAMD-17 scores. In addition, a robust structural model provides a solid base for D-optimal design to optimize the clinical trial. Based upon the TM, Deman et al. [41] showed that the most informative antidepressant clinical trial design would be including HAMD-17 observations for a short period time after the cessation of placebo/active drug treatment.

Conclusion

We evaluated the performance of three empirical models with data that covered a wide range of study durations and treatment regimens (single vs. multiple doses; with and without follow-up observations). Although IBM was easy to estimate and parameters readily translated to clinical events, the parameters were poorly estimated, highly correlated, and dependant on study duration, making it unsuitable for clinical trial analysis and simulation. Both IDR and TM are more complex with non-intuitive parameters. However, parameter estimates are more robust, uncorrelated, and independent of study duration. Both IDR and TM appeared useful for analysis, simulation, and quantifying antidepressant effect, although the TM was the slightly better model.

Acknowledgment We are grateful to Dr. Lisa J. Benincosa and Thomas G. Tensfeldt for their support.

References

1. Khan A, Schwartz K (2005) Study designs and outcomes in antidepressant clinical trials. Essent Psychopharmacol 6:221–226
2. Dworkin RH, Katz J, Gitlin MJ (2005) Placebo response in clinical trials of depression and its implications for research on chronic neuropathic pain. Neurology 65:S7–S19. doi:10.1212/01.wnl.0000177925.92714.a9
3. Khan A, Detke M, Khan SR, Mallinckrodt C (2003) Placebo response and antidepressant clinical trial outcome. J Nerv Ment Dis 191:211–218. doi:10.1097/00005053-200304000-00001
4. Klein DF, Thase ME, Endicott J, Adler L, Glick I, Kalali A, Leventer S, Mattes J, Ross P, Bystritsky A (2002) Improving clinical trials: American Society of Clinical Psychopharmacology recommendations. Arch Gen Psychiatry 59:272–278. doi:10.1001/archpsyc.59.3.272
5. Emslie GJ, Ryan ND, Wagner KD (2005) Major depressive disorder in children and adolescents: clinical trial design and antidepressant efficacy. J Clin Psychiatry 66(Suppl 7):14–20
6. Storosum JG, Elferink AJ, van Zwieten BJ, van den Brink W, Huyser J (2004) Natural course and placebo response in short-term, placebo-controlled studies in major depression: a meta-analysis of published and non-published studies. Pharmacopsychiatry 37:32–36. doi:10.1055/s-2004-815472
7. Klein DF (1999) A checklist for developing and reviewing comparative treatment evaluations in the areas of psychotherapy and pharmacotherapy. In: Janowsky DS (ed) Psychotherapy indications outcomes. American Psychiatric Press Inc., Washington DC, pp 367–381
8. Greist JH, Mundt JC, Kobak K (2002) Factors contributing to failed trials of new agents: can technology prevent some problems? J Clin Psychiatry 63(Suppl 2):8–13
9. Trivedi MH, Rush H (1994) Does a placebo run-in or a placebo treatment cell affect the efficacy of antidepressant medications? Neuropsychopharmacology 11:33–43
10. FDA (1997) Guidance for industry: guideline for the clinical evaluation of antidepressant drugs. February 1997. U.S. Department of Health, Education, and Welfare, Food and Drug Administration, Center for Drug Evaluation and Research (CDER). http://www.fda.gov/cder/guidance/old050fn.pdf. Accessed 15 Oct 2008
11. Hamilton M (1967) Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 6:278–296
12. Holford N, Li J, Benincosa L, Birath M (2002) Population disease progress models for the time course of HAMD score in depressed patients receiving placebo in anti-depressant clinical trials. In: Abstracts of the XI annual meeting of the population approach group in Europe. Abstr. 311 www.page-meeting.org/?abstract=311. Accessed 15 Oct 2008
13. Mould DR (2007) Developing models of disease progression. In: Ette EI, Williams PJ (eds) Pharmacometrics: the science of quantitative pharmacology. Wiley, Hoboken, pp 547–581
14. Shang E, Gibbs MA, Landen J, Denman NG, Krams M, Russell T, Mould DR (2006) Model based analysis of placebo response in major depression. In: Abstract of the American College of Clinical Pharmacology annual meeting, Cambridge, MA
15. Cosson V, Gomeni R (2005) Modelling placebo response in depression using a mechanistic longitudinal model approach. In: Abstracts of the XIV annual meeting of the population approach group in Europe. Abstr. 818 www.page-meeting.org/?abstract=818. Accessed 15 Oct 2008
16. Mager DE, Wyska E, Jusko WJ (2003) Diversity of mechanism-based pharmacodynamic models. Drug Metab Dispos 31:510–518. doi:10.1124/dmd.31.5.510
17. Dayneka NL, Garg V, Jusko WJ (1993) Comparison of four basic models of indirect pharmacodynamic responses. J Pharmacokinet Biopharm 21:457–478. doi:10.1007/BF01061691
18. Karlsson MO, Molnar V, Bergh J, Freijs A, Larsson R (1998) A general model for time-dissociated pharmacokinetic–pharmacodynamic relationship exemplified by paclitaxel myelosuppression. Clin Pharmacol Ther 63:11–25. doi:10.1016/S0009-9236(98)90117-5
19. Preskorn SH (1996) A dangerous idea. J Pract Psychiatry Behav Health 2:231–234

20. Anonymous (1997) The appearance of knowledge. J Pract Psychiatry Behav Health 3:233–238
21. Sheiner LB, Rosenberg B, Marathe VV (1977) Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. J Pharmacokinet Biopharm 5:445–479. doi:10.1007/BF01061728
22. Sheiner LB, Beal S, Rosenberg B, Marathe VV (1979) Forecasting individual pharmacokinetics. Clin Pharmacol Ther 26:294–305
23. Beal SL, Sheiner LB (1980) The NONMEM system. Am Stat 34:118–119. doi:10.2307/2684123
24. Beal SL, Boeckman AJ, Sheiner LB (1998) NONMEM: user's guide part I-VIII. University of California at San Francisco, San Francisco
25. Yafune A, Ishiguro M (1999) Bootstrap approach for constructing confidence intervals for population pharmacokinetic parameters. I: a use of bootstrap standard error. Stat Med 18:581–599. doi:10.1002/(SICI)1097-0258(19990315)18:5<581::AID-SIM47>3.0.CO;2-1
26. Karlsson MO (2005) Model-building diagnostics. In: DIA meeting, Philadelphia, PA
27. Karlsson MO, Savic RM (2007) Diagnosing model diagnostics. Clin Pharmacol Ther 82:17–20. doi:10.1038/sj.clpt.6100241
28. Mentre F, Mallet A, Baccar D (1997) Optimal design in random-effects regression models. Biometrika 84:429–442. doi:10.1093/biomet/84.2.429
29. Retout S, Duffull S, Mentre F (2001) Development and implementation of the population Fisher information matrix for the evaluation of population pharmacokinetic designs. Comput Methods Programs Biomed 65:141–151. doi:10.1016/S0169-2607(00)00117-6
30. Hooker A, Dodds MG, Vicini P (2004) Evaluating the predictive power of the Fisher information matrix in population optimal experimental design. In: Abstracts of the XIII annual meeting of the population approach group in Europe. Abstr. 526 www.page-meeting.org/?abstract=526. Accessed 15 Oct 2008
31. Retout S, Mentre F, Bruno R (2002) Fisher information matrix for non-linear mixed-effects models: evaluation and application for optimal design of enoxaparin population pharmacokinetics. Stat Med 21:2623–2639. doi:10.1002/sim.1041
32. Duffull S, Denman NG, Eccleston JA, Kimko H (2006) WinPOPT user guide. University of Otago, Dunedin, New Zealand
33. Yano Y, Beal SL, Sheiner LB (2001) Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. J Pharmacokinet Pharmacodyn 28:171–192. doi:10.1023/A:1011555016423
34. Holford N (2005) The visual predictive check—superiority to standard diagnostic (Rorschach) plots. In: Abstracts of the XIV annual meeting of the population approach group in Europe. Abstr. 738 www.page-meeting.org/?abstract=738. Accessed 15 Oct 2008
35. Holford N (2005) The time course of placebo response in clinical trials—do antidepressants really take two weeks to work? In: AAPS annual meeting and exposition, Nashville, TN. http://www.aapspharmaceutica.com/inside/focus_groups/ModelSim/imagespdfs/Holford05.pdf. Accessed 15 Oct 2008
36. Gruwez B, Dauphin A, Tod M (2005) A mathematical model for paroxetine antidepressant effect time course and its interaction with pindolol. J Pharmacokinet Pharmacodyn 32:663–683. doi:10.1007/s10928-005-0006-6
37. Jacqmin P, Gieschke R, Jordan P, Steimer JL, Goggin T, Pillai G, Snoeck E, Girard P (2001) Modeling drug induced changes in biomarkers without using drug concentrations: introducing the K-PD model. In: Abstracts of the X annual meeting of the population approach group in Europe. Abstr. 232 www.page-meeting.org/?abstract=232. Accessed 15 Oct 2008
38. Pillai G, Gieschke R, Goggin T, Jacqmin P, Schimmer RC, Steimer JL (2004) A semimechanistic and mechanistic population PK-PD model for biomarker response to ibandronate, a new bisphosphonate for the treatment of osteoporosis. Br J Clin Pharmacol 58:618–631. doi:10.1111/j.1365-2125.2004.02224.x
39. Jonsson EN, Sheiner LB (2002) More efficient clinical trials through use of scientific model-based statistical tests. Clin Pharmacol Ther 72:603–614. doi:10.1067/mcp.2002.129307
40. Mould DR, Denman NG, Duffull S (2007) Using disease progression models as a tool to detect drug effect. Clin Pharmacol Ther 82:81–86. doi:10.1038/sj.clpt.6100228
41. Denman NG, Mould DR, Duffull S (2006) Study designs to assess the placebo response (PBO-R) for Hamilton Depression (HAMD) Scores for depression. AAPS Journal 8 (S2) Abstr T3371