

Deep Learning versus Linear Mixed Models for Estimating Treatment Effects in Longitudinal Repeated-Measures Studies

Roberto Gomeni

`roberto.gomeni@pharmacometrica.com`

Pharmacometrica

Françoise Bressolle-Gomeni

Pharmacometrica

Research Article

Keywords: Longitudinal data

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11318468/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations:

No competing interests reported.

Tables are available in the Supplementary Files section.

Deep Learning versus Linear Mixed Models for Estimating Treatment Effects in Longitudinal Repeated-Measures Studies

R. Gomeni¹, F. Bressolle-Gomeni¹

¹Pharmacometrica, Longcol, La Fouillade, France

Corresponding author:

Roberto Gomeni
Pharmacometrica, Lieu-dit Longcol, 12270, La Fouillade, France
E-mail: roberto.gomeni@pharmacometrica.com
Phone number: +33 760451976

ORCID references:

Roberto Gomeni: 0000-0003-3916-9578
Françoise Bressolle-Gomeni: 0009-0004-0183-8830

Keywords:

Longitudinal data
Repeated measures
Deep learning
Mixed-effects models
Synthetic data

Abstract

Background: Mixed models for repeated measures (MMRM) are widely used to analyze longitudinal clinical outcomes. Deep-learning (DL) methods can represent nonlinear relationships, but their performance in estimating treatment effects requires direct evaluation.

Objective: To compare DL and MMRM estimates of week-8 treatment effects in simulated major depressive disorder trials.

Methods: We simulated three-arm trials (placebo, 20 mg/day, and 40 mg/day; 80 participants per arm) using three HAM-D-17 trajectory profiles and two assessment schedules. Across 20 replicates per scenario, treatment effects from DL and MMRM analyses were compared with a reference analysis of simulated week-8 change from baseline. DL fit was assessed using RMSE, MAE, and R^2 .

Results: Mean treatment-effect estimates were broadly comparable across methods and preserved the dose ordering. However, DL estimates were closer to the reference estimates: the reported mean absolute difference was approximately 0.15 points for DL and 3.5 points for MMRM. Regression comparisons also showed fewer departures from the identity relationship for DL. DL fit metrics showed moderate variability in RMSE and MAE and low variability in R^2 across replicates.

Conclusions: Under the simulated conditions, the DL approach exhibited robust predictive performance and reliability in the estimation of treatment effect with respect to the established MMRM approaches while requiring substantially fewer a priori assumptions regarding the underlying system. These findings do not establish general superiority: conclusions are limited by the specific simulation models, sample size, analysis procedures, and number of replicates.

Introduction

Longitudinal clinical data require statistical methods that account for the correlation among repeated observations obtained from the same participant over time. Mixed-effects models are among the most established approaches for analyzing such data because they can incorporate fixed effects, such as time, treatment, and relevant covariates, while accounting for participant-level heterogeneity through random effects [1]. These models can estimate the overall trajectory of an outcome and evaluate interactions between time and key explanatory variables to determine whether trajectories differ between groups. They can also accommodate unequal numbers of observations across participants under appropriate assumptions regarding missingness. Model assumptions can be assessed using standard diagnostic procedures, with effect estimates reported alongside corresponding 95% confidence intervals and p-values.

Repeated measurements provide substantially more information than a single assessment because they allow researchers to characterize individual trajectories and examine how symptoms evolve during treatment and follow-up. However, observations obtained from the same individual are generally correlated and therefore cannot be treated as statistically independent. Ignoring this within-participant correlation may result in inappropriate estimates of uncertainty and potentially misleading statistical conclusions. Mixed-effects models address this dependence while providing parameters that are relatively straightforward to interpret in clinical research.

Despite these advantages, conventional mixed-effects models require researchers to specify an appropriate functional form for the relationships between predictors and outcomes. For example, a linear specification assumes a particular pattern of change in depressive symptoms over time unless additional nonlinear terms are introduced. Modeling complex interactions, nonlinear trajectories, or high-dimensional predictors may therefore require increasingly elaborate model specifications.

Deep-learning methods provide an alternative framework for representing complex relationships without requiring all functional forms to be specified in advance [2]. Neural networks can learn nonlinear associations and interactions directly from data, while recurrent neural networks, long short-term memory networks, temporal convolutional networks, and transformer-based architectures can be used to model sequential information. These methods have consequently attracted increasing attention in biomedical and clinical prediction.

However, greater model flexibility does not necessarily translate into superior performance. Deep-learning methods typically require larger datasets than conventional statistical models and may be more difficult to interpret. This consideration is particularly important in longitudinal clinical studies, where sample sizes may be modest and the number of repeated observations per participant limited. A direct comparison is therefore warranted rather than assuming that deep learning is inherently superior to traditional statistical approaches. Such a comparison may help identify the data characteristics and modeling conditions under which each framework performs most effectively.

The primary objective of this study was to compare the performances of a deep-learning approach (DL) and a mixed-effects model for the longitudinal analysis (MMRM) of

repeated measurements for the estimation of treatment effect using simulated major depressive disorder (MDD) data.

The simulations are expected to provide criteria for the comparison of the two approaches with respect to the predictive performance in the estimation of treatment effects and to the robustness of the performances in presence of missing observations.

By evaluating these dimensions under controlled simulation conditions, this study aims to investigate the relative strengths and limitations of MMRM and DL approaches for longitudinal clinical data..

Methods

Clinical trial simulation

Deep learning (DL) and mixed models for repeated measures (MMRM) were compared on their ability to estimate treatment effects in simulated randomized clinical trials.

The trials were designed to reflect a typical antidepressant efficacy study, with three parallel groups: placebo, low-dose active treatment (20 mg/day), and high-dose active treatment (40 mg/day). Each group included 80 participants, for a total sample size of 240. Efficacy was assessed using longitudinal measurements of the 17-item Hamilton Depression Rating Scale (HAM-D-17) total score, a widely used clinician-rated measure of depressive symptom severity [6]. The primary efficacy endpoint, TE, was the baseline-corrected change in HAM-D-17 total score at week 8 relative to placebo.

HAM-D-17 scores were simulated across multiple study visits using a previously developed model of HAM-D-17 trajectories in patients with major depressive disorder [3,4,5]. A sample size of 80 subjects per treatment arm was selected to provide 90% power to detect a 4-point difference in the treatment effect on the HAM-D-17 total score between the 40 mg/day treatment arm and placebo. We evaluated multiple scenarios that varied study-design and data-generating conditions. For each simulated trial, DL and MMRM independently estimated TE, enabling a direct comparison of their ability to recover the prespecified treatment effect.

The simulation scenarios assessed the effects of:

1. Longitudinal HAM-D-17 trajectory shape: Three distinct profiles (i.e., Model 1, 2, and 3) were used to represent differing patterns of symptom change during treatment.
2. Timing and frequency of HAM-D-17 assessments: Two assessment sampling scenarios were considered:
 - Sampling 1: All participants were assessed at Weeks 0 (baseline), 1, 2, 4, 6, and 8 (endpoint).
 - Sampling 2: Within each treatment arm, 50% of participants were assessed at weeks 0, 2, 4, and 8, while the remaining 50% were assessed at weeks 0, 6, and 8.

Individual HAMDD-17 trajectories were simulated using a Monte Carlo approach. Interindividual variability and residual unexplained variability were resampled under each modeling and sampling scenarios. For each scenario, 20 independent RCT replicates were simulated. Within each replicate, the treatment effect (TE) was estimated using the reference model, MMRM, and DL methodologies. For each simulated RCT, the difference between the TE estimated by the reference model and by each alternative method was calculated.

Model based simulation

The individual HAMDD-17 total scores for subject (i) at time (t) were simulated using a longitudinal model defined by a Weibull function:

$$P_i(t) = A_i e^{-\left(\frac{t}{td}\right)^b} + hrec_i + eps \quad (1)$$

where $P(t)$ is the HAMDD-17 total clinical score at time t, A is the baseline total score, td is the time corresponding to a 63.2% reduction of HAMDD-17 from baseline, b is the rate of change in response, $hrec$ is the recovery rate and eps is the residual error. This model was previously applied to describe the HAMDD-17 longitudinal response in different RCTs conducted in MDD populations [3,4,5].

Assuming that the primary endpoint of the analysis was the change from baseline of the HAMDD-17 score at study-end (week 8, $t = 8$), the model was re-parameterized in order to include a parameter defining the HAMDD-17 total score at week 8 ($F8$) and the change from baseline of the HAMDD-17 total score (TR) at week 8:

$$F8 = A e^{-(8/td)^b} + hrec \cdot 8 \quad (2)$$

where

$$hrec = \frac{F8 - A e^{-(8/td)^b}}{8} \quad (3)$$

The change from baseline of the HAMDD-17 total score (TR) at week 8 was estimated as:

$$TR = F8 - A \quad (4)$$

The inter-individual variability was modeled using a log-normal distribution, therefore, the values of the individual parameters (P_j) were expressed as:

$$P_j = P_{TV} e^{\eta p} \quad (5)$$

Where P_j is the value of parameter for the j^{th} individual, P_{TV} is the typical value of P for the population, and ηP denotes the difference between P_j and P_{TV} .

The residual error model (eps_j) was assumed to be normally distributed with mean 0 and variance σ^2 .

Dose and age were the identified as covariates impacting the $F8$ parameter value according to the following model.

$$F8 = (F8_base + Age * F8_age) * \exp(-Dose * F8_dose)$$

Where $F8_base$ is the value of the $F8$ parameter associated with an age = 0 and a dose = 0. The effect of dose on $F8$ was evaluated assuming a dose-related exponential reduction (defined by the $F8_dose$ parameter) of $F8$ with the increase of the dose.

The effect of age on $F8$ was evaluated assuming a linear age related increase (defined by the $F8_age$ parameter) of $F8$. A dose = 0 was associated to the data in the placebo arm.

The non-linear mixed-effect model included 7 fixed effect parameters (A , tb , $F8_base$, b , $F8_age$, $F8_dose$, and Age) and 5 random effect parameters (variances of A , tb , $F8$, b , and Age). The values of these parameters for the 3 models evaluated are presented in Table 1.

Reference	Parameter	Model 1	Model 2	Model 3
Fixed Effect	A (baseline)	25.00	25.00	25.00
	tb (week)	5.00	5.00	2.00
	F8_base	7.97	5.97	7.97
	b	1.07	4.07	4.07
	F8_age	0.08	0.09	0.09
	F8_dose	1.04	1.04	1.04
	Age	60.00	60.00	60.00
Random Effect*	A	0.02	0.02	0.02
	tb	0.20	0.20	0.20
	F8	0.20	0.20	0.20
	b	0.20	0.20	0.20
	Age	0.02	0.02	0.02
Residual Error	sigma**	2.32	2.32	2.32
* log-normally distributed variance				
** normally distributed variance				

Table 1. Fixed and random effect parameter values for the 3 models used in the synthetic data generation.

Simulations were conducted using the NONMEM software (version 7.5.1, ICON Development Solutions).

Deep Learning Methodology

The deep learning (DL) model was implemented using an artificial neural network (ANN) architecture comprising interconnected neurons organized into input, hidden, and output layers. The network learns optimal connection weights to model complex nonlinear relationships between input and the target variables (i.e., the longitudinal HAM-D-17 scores). Model training involved sequential feed-forward propagation and back-propagation. During feed-forward propagation, information was transmitted from the input layer through the hidden layer(s) to the output layer. The predicted outputs were then compared with the observed values, and the resulting errors were propagated backward to iteratively update the connection weights and minimize prediction error. A supervised learning framework with the resilient back-propagation algorithm was employed to

improve training stability and convergence efficiency [7]. A detailed implementation of the DL model was previously described [2]. The ANN analysis was conducted using the '*neuralnet*' library available in the statistical software R (Version 4.4.1). This analysis used a supervised learning methodology, along with the resilient back-propagation algorithm, to enhance convergence efficiency and stability during model training [7].

Statistical analysis

Three statistical analyses were performed to compare TEs across the simulation scenarios:

1. Analysis 1: A reference ANOVA analysis comparing TEs estimated from the simulated change from baseline in HAMD-17 total score at week 8.
2. Analysis 2: An ANOVA comparing TEs estimated from the week-8 change from baseline in DL-adjusted longitudinal HAMD-17 scores.
3. Analysis 3: A MMRM model was fitted to the simulated longitudinal HAMD-17 change-from-baseline trajectories and the model outcomes were used to estimate the TE at week 8.

In Analysis 1, the association between dose and TE at week 8 was assessed using a linear regression model, with dose treated as a categorical variable and age included as a continuous covariate. The overall effect of dose was evaluated using a type III analysis of variance (ANOVA). Adjusted least-squares means (LS-means) for change from baseline were estimated for each dose level from the fitted model.

Pairwise comparisons between dose groups were performed using the Tukey method to account for multiple comparisons. In addition, prespecified pairwise contrasts between dose levels were estimated from the LS-means, with corresponding estimated mean differences, standard errors, t-statistics, and p-values reported. The primary comparisons were between doses placebo and 20 mg, placebo and 40 mg. All statistical tests were two-sided, and statistical significance was assessed at the 5% level.

The second analysis was conducted using the DL-adjusted individual HAMD-17 trajectories and the same methodology as Analysis 1. Because this analysis used DL-adjusted longitudinal HAMD-17 scores, three complementary metrics were used to assess the DL model's fit: root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). RMSE is the square root of the mean squared prediction error, while MAE is the average absolute difference between predicted and observed values. R^2 measures the proportion of variability in the observed data explained by the model and provides a scale-independent measure of goodness of fit.

Furthermore, to assess the DL model's robustness and reproducibility, the distribution of each performance metric across the 20 replicated RCTs was summarized using the mean, coefficient of variation (CV), and 95% confidence interval (CI). The following plots were also generated: (1) individual DL-predicted values versus individual simulated values, by dose; (2) residuals versus time, by dose; and (3) individual simulated and model-predicted values versus time.

In the MMRM analysis the change from baseline were analyzed using a linear mixed-effects model based on restricted maximum likelihood. The analysis included post-baseline observations only.

The dependent variable was change from baseline at each post-baseline visit, calculated as the post-baseline observed value minus the corresponding baseline value. Treatment was represented by three groups: placebo and two active-dose groups. The model included treatment, time, age, baseline, treatment-by-time interaction, and baseline baseline-by-time interaction as fixed effects. Repeated measurements over time within subject were modeled using an unstructured covariance matrix, allowing the variances and correlations between measurements at different time points to be estimated without imposing a specific covariance structure. Degrees of freedom for the fixed effects were estimated using the Kenward–Roger method.

Least-squares mean estimates, together with 95% confidence intervals, were obtained for each treatment group at each time point and overall by treatment group. Pairwise treatment differences were estimated for placebo versus each active-dose group and between the two active-dose groups, with corresponding 95% confidence intervals. Treatment-by-time interaction terms were included to allow treatment effects to vary over time. The treatment comparisons at the prespecified assessment time point (week 8) were used to characterize the treatment effect at the primary assessment time point. All statistical tests were based on the model-derived least-squares mean differences and used a significance assessed at the 5% level.

The MMRM analysis was conducted using the *PROC MIXED* procedure available in the SAS software (Version 9.4). The ANOVA was conducted using the *lm* procedure available in the statistical software R (Version 4.4.1).

Comparing the performances of DL and MMRM in the estimation of TE

The performance of the DL and MMRM approaches in estimating the TE was assessed by comparing the TE estimates from Analyses 2 and 3 with those from the reference analysis (Analysis 1). Comparisons were performed using paired TE estimates obtained from the same simulated longitudinal datasets across the 20 RCT replicates.

Two complementary assessments were performed by dose group:

1. **Concordance of TE estimates:** The correlation between the TE estimates from the reference analysis and those from the DL and MMRM analyses was assessed across the 20 RCT replicates. Concordance was evaluated based on the correlation and the estimated regression intercept and slope, with an intercept of zero and a slope of one indicating complete agreement.
2. **Difference from the reference analysis:** The absolute mean difference in TE between the reference analysis and each of the DL and MMRM analyses was calculated to quantify deviations from the reference estimates.

Results

Fig.1 shows the mean longitudinal HAMD-17 scores simulated using the three selected models. As an illustrative example of the different shapes of the longitudinal observation, data from the 20th RCT replicate are shown, with the corresponding mean trajectories ($\pm$ Standard Error) for each of the three models used in the analysis.

As an example of the RCT simulations, Fig.2 shows of the distribution of the individual HAMD-17 longitudinal scores simulated using Model 1 and Sampling 1 generated in the 20th RCT replicate.

Assessment of DL model performances

An initial analysis was conducted using data from the first simulated RCT, considering Models 1, 2, and 3 and sampling times 1 and 2. The objective was to determine the optimal number of hidden-layer nodes for the artificial neural network (ANN) model used to implement the deep learning (DL) approach. A systematic grid search was performed, with the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) used as criteria for model selection [8,9]. Based on the criterion of minimizing both AIC and BIC, the one-hidden-layer ANN model with 11 nodes was selected as the best-performing model. This architecture was therefore used in all subsequent DL analyses.

A fixed random seed value was used in all runs to ensure reproducibility. The following diagnostic plots were generated to assess the ability of the DL model to fit the longitudinal HAMD-17 scores: (1) individual DL-predicted values versus individual simulated values, stratified by dose; (2) residuals versus time, stratified by dose; and (3) individual simulated and model-predicted values versus time. For simplicity, only the plots derived from RCT simulation 1, conducted using model 1 and sampling time 1, are presented in the supplemental material (Figures S1 and S2). Inspection of these plots confirmed that the DL model provided an accurate fit to the longitudinal HAMD-17 scores.

Furthermore, to assess the robustness and reproducibility of the DL model outcomes, the distributions of the RMSE, MAE, and R^2 metrics across the 20 replicated RCTs were summarized in Table S1 of the supplemental material using the mean, coefficient of variation (CV%), and 95% confidence interval (CI). The average CV% values for RMSE and MAE remained below 23% across the 20 replicated RCTs, indicating moderate variability and good reproducibility of the DL model outcomes. This finding was further supported by the R^2 results, with mean R^2 values showing consistently low variability, with a CV% remaining below 0.5%.

Assessment of treatment effect

The overall descriptive statistics on the TE values estimated with the reference, the DL and the MMRM methodologies in the 20 RCTs are presented in Table 2 by dose. The mean TE values estimated by models (1, 2, and 3), by sampling times (sampling 1 and

2), and by dose (20mg and 40mg) are presented in Table S2 of the supplemental material.

On average the TE values estimated with the reference, MMRM, and DL were broadly comparable with TE value ~1.7 larger at the dose of 40mg with respect to the dose of 20mg.

The overall mean TE differences by dose between reference and DL and between reference and MMRM methodologies in the 20 RCTs replicates is presented in Table 3. A table with the mean TE values by model, sampling times and dose is presented in the supplemental Table S3.

Concordance between the TE values estimated from the reference analysis and the values derived from the DL and MMRM analyses was assessed across the 20 RCT replicates using the regression intercept and slope. Fig.3 shows the correlation between reference TE estimates and TE values estimated from the DL and MMRM analyses assessed across the 20 RCT replicates using the data of Model 2 and Sampling 1. The blue solid lines represents the regression line and the light blue dotted curve represent the identity reference lines.

An intercept of zero and a slope of one indicate complete agreement between the TE values. However, deviation from the zero intercept and the unitary slope were considered. Figure S1 presents the results for model 2, sampling times 1 and doses of 20 mg and 40 mg. Supplementary Table S2 reports the detailed correlation analyses of TE estimates obtained using the reference, DL, and MMRM methods. The analyses covered 12 comparisons: three models, two sampling times, and two doses. The MMRM regression slopes were significantly below one in 11 of 12 comparisons (92%), compared with one of 12 (8%) for DL. The MMRM regression intercepts differed significantly from zero in nine of 12 comparisons, whereas none of the DL intercepts differed significantly from zero. These results indicate the presence of a potential systematic bias of the MMRM analysis in the estimation of TE.

The descriptive statistics on the mean absolute difference of TE between the values estimated in the reference analysis and the values estimated in the DL and MMRM analyses is presented in Table 4.

The estimated TE values obtained using either the DL or MMRM methodology showed a mean absolute difference from the reference value that was greater than zero. This finding indicates that both methodologies produced TE estimates that modestly deviated from the expected values. However, comparison of the average mean absolute difference in TE between the estimated and reference values indicates that the MMRM methodology showed a substantially greater deviation from the reference value, with an average difference of approximately 3.5 points, compared with only 0.15 points for the DL methodology. These results suggest that the DL methodology provided TE estimates that were considerably closer to the reference values than those obtained using the MMRM methodology.

Discussion

In this simulation study, we compared a deep-learning approach with a mixed model for repeated measures (MMRM) for estimating week-8 treatment effects from longitudinal HAMD-17 data. Across the scenarios examined, the methods produced broadly comparable average treatment-effect estimates, and both recovered the expected ordering of the dose groups. However, their agreement with the reference analysis differed: the reported mean absolute difference was approximately 0.15 points for DL and 3.5 points for MMRM. The regression comparisons likewise showed fewer departures from a zero intercept and unit slope for DL. Taken together, these results suggest that, under the data-generating and analysis conditions studied, DL estimates were closer to the reference estimates than MMRM estimates.

The broadly similar average treatment effects and the differences in agreement with the reference are not contradictory. Average estimates can be close while replicate-level estimates differ in their variability or systematic deviation. The paired comparisons across simulated datasets are useful in this respect because each method was evaluated on the same generated trial replicates. Still, the reference analysis is an operational benchmark in this simulation, rather than an independent empirical truth: its estimates are calculated from the simulated week-8 outcomes, and the relationship between the benchmark and each fitted method depends on the simulation and analysis specifications. The observed differences should therefore be interpreted as relative performance in these scenarios, not as evidence that one method is generally unbiased or superior in clinical trials.

The DL model also fitted the simulated longitudinal scores well according to the reported diagnostic assessment, with mean RMSE and MAE variability below 23% and low variability in R^2 across replicates. This supports the model's ability to reproduce the generated trajectories in the present setting. It does not, by itself, establish accurate treatment-effect estimation in new populations or under different data-generating mechanisms. In particular, flexible neural networks can represent nonlinear associations, but their fitted relationships and resulting estimates may depend on the training data, architecture, tuning procedure, and preprocessing. The selected one-hidden-layer network with 11 nodes was chosen using AIC and BIC in an initial replicate; further evaluation of tuning stability and out-of-sample performance would help determine how robust that choice is.

The results are relevant to longitudinal clinical studies because observation schedules can vary across participants. The two schedules considered here changed the timing and frequency of assessments, and the three trajectory profiles represented different patterns of symptom change. These scenarios provide a first assessment of how the methods behave when observation timing and trajectory shape vary.

As reported in Table S2, comparable mean TE estimated were found when observation timing (Sampling time 1 and 2) and trajectory shape vary (Model 1, 2, and 3).

MMRM is a model-based approach whose interpretation depends on several assumptions. The most consequential is that the specified mean model and within-participant covariance models adequately represent the data. A typical mean model includes treatment, visit, treatment-by-visit, baseline outcome, and, where appropriate, baseline-by-visit and randomization stratification factors. An unstructured covariance is often used because it does not impose a specific correlation pattern across visits.

However, it can be difficult to estimate when there are many visits or limited data; therefore, its suitability and the robustness of results to alternative covariance structures should be considered [10,11].

Differently from the MMRM, the deep learning methodology provides a data-driven approach capable of learning relationships between variables directly from data without requiring predefined model structures or mechanistic assumptions. In this work, the DL methodology was implemented using an artificial neural network, supported by the universal approximation theorem, which states that a feed-forward neural network with a sufficient number of neurons can approximate any continuous function to arbitrary accuracy [12]

Several limitations should guide interpretation. First, each scenario was evaluated in 20 simulated trial replicates. This provides limited precision for summaries of bias, variability, and regression-based concordance, particularly when inference is based on only 20 paired estimates. Second, the simulations used a specific HAMD-17 trajectory model and a sample size of 80 participants per arm. The findings may change with other sample sizes, treatment effects, outcome distributions, covariance structures, degrees of nonlinearity, or departures from the assumed generator. Third, the reference, DL, and MMRM analyses do not appear to target precisely the same estimand by construction: the reference analysis uses simulated week-8 change from baseline, DL-adjusted outcomes are analyzed with ANOVA, and MMRM estimates the week-8 contrast from all post-baseline observations. Differences in estimand, model specification, or implementation could contribute to the observed discrepancies and should be checked before attributing them to the modeling approach alone.

Additional work could strengthen the comparison by increasing the number of replicates and reporting Monte Carlo uncertainty for bias, root mean squared error, confidence-interval coverage, and type I error or power. The methods could be evaluated under prespecified dropout and intermittent-missingness mechanisms, including sensitivity analyses for departures from missing-at-random assumptions. Using aligned estimands and documenting the DL inputs, preprocessing, validation strategy, tuning procedure, and uncertainty estimation would improve interpretability and reproducibility. Performance could also be tested under data-generating models that differ from the fitted models, as well as across larger and smaller samples. These extensions would clarify whether the apparent advantage in closeness to the reference persists beyond the particular scenarios considered here.

Conclusion

Under the simulated conditions, the DL approach exhibited robust predictive performance and reliability in the estimation of treatment effect with respect to the established MMRM approaches while requiring substantially fewer a priori assumptions regarding the underlying system. These findings do not establish general superiority: conclusions are limited by the specific simulation models, sample size, analysis procedures, and number of replicates.

In the simulated MDD trials examined, DL and MMRM yielded broadly comparable mean week-8 treatment-effect estimates, while DL estimates were closer to the reference estimates by the reported mean absolute difference and regression-concordance summaries. This finding supports further evaluation of DL as a reliable option for estimating treatment effects from longitudinal outcomes, but it does not establish general superiority over MMRM. The conclusions are limited to the specified simulation models, sample size, assessment schedules, and analysis procedures. Broader simulation studies that align estimands and increase the number of replications are needed to confirm the performance of the DL approach as an effective tool for assessing TE in clinical trial practice.

Author contributions: RG – study design and conceptualization; data analysis; writing and reviewing the manuscript. FBG – analysis plan, data interpretation, and review of the draft and final manuscript

Conflicts of Interest: The authors declare no conflict of interest. R.G. and F.B.G. are consultant of PharmacoMetrica.

Funding: The authors declare that they did not receive any financial support for this study

Data Availability: The simulated datasets used in this analysis are available from the corresponding author upon reasonable request.

Acknowledgements: None

Declarations Ethics approval Ethics approval was not required

References

1. Mallinckrodt, C.H., Lane, P.W., Schnell, D. *et al* (2008). Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials. *Ther Innov Regul Sci.* **42**, 303–319. <https://doi.org/10.1177/009286150804200402>
2. Gomeni R, Bressolle-Gomeni F. (2026) Deep-Learning: An Emerging Tool to Support Model-Informed Drug Development. *Clin Transl Sci.* 19(8):e70691. doi: 10.1111/cts.70691.
3. Gomeni R, Bressolle-Gomeni F. (2024) Model-informed approach to estimate treatment effect in placebo-controlled clinical trials using an artificial intelligence-based propensity weighting methodology to account for non-specific responses to treatment. *J Pharmacokinet Pharmacodyn.* 52(1):5. doi: 10.1007/s10928-024-09950-7.
4. Gomeni R, Goyal N, Bressolle F, Fava M (2015) A Novel Methodology to Estimate the Treatment Effect in Presence of Highly Variable Placebo Response. *Neuropsychopharmacology* 40(11):2588-95. <https://doi.org/10.1038/npp.2015.105>
5. Gomeni R, Merlo-Pich E (2007) 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(5):595-613 <https://doi.org/10.1111/j.1365-2125.2006.02815.x>
6. Hamilton M. (1960) A rating scale for depression. *J. Neurol. Neurosurg. Psychiatr.* 23, 56–62.
7. Günther F, Fritsch S. (2010) neuralnet: Training of Neural Networks, *The R Journal*, (2):30-38. doi:10.32614/RJ-2010.
8. Bozdogan H. (1987) Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika.* 52:345–370. Doi:10.1007/BF02294361.
9. Schwarz G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics.* 6:461-464. Doi: 10.1214/aos/1176344136.
10. Mallinckrodt CH, Kaiser CJ, Watkin JG, Molenberghs G, Carroll RJ. (2004) The effect of correlation structure on treatment contrasts estimated from incomplete clinical trial data with likelihood-based repeated measures compared with last observation carried forward ANOVA. *Clin Trials.* 1(6):477-89. doi: 10.1191/1740774504cn049oa. PMID: 16279288.
11. FDA Guidance on E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials, May 2021, <https://www.fda.gov/media/148473/download>
12. Hornik K, Stinchcombe M, White H. (1989) Multilayer feedforward networks are universal approximators, *Neural Netw.* (2) 359-366. Doi: 10.1016/0893-6080(89)90020-8.

List of Tables

Table 1. Fixed and random effect parameter values for the 3 models used in the synthetic data generation.

Table 2. Overall descriptive statistics on the mean TE values by dose estimated with the reference, the DL and the MMRM methodologies in the 20 RCTs replicates.

Table 3. Overall mean TE difference by dose between the reference and the DL and the reference and the MMRM methodologies in the 20 RCTs replicates using the three models and the two sampling measurements

List of Figures

Fig1. Mean ($\pm$ Standard Error) longitudinal HAMD-17 scores simulated using the three selected models from the 20th RCT .

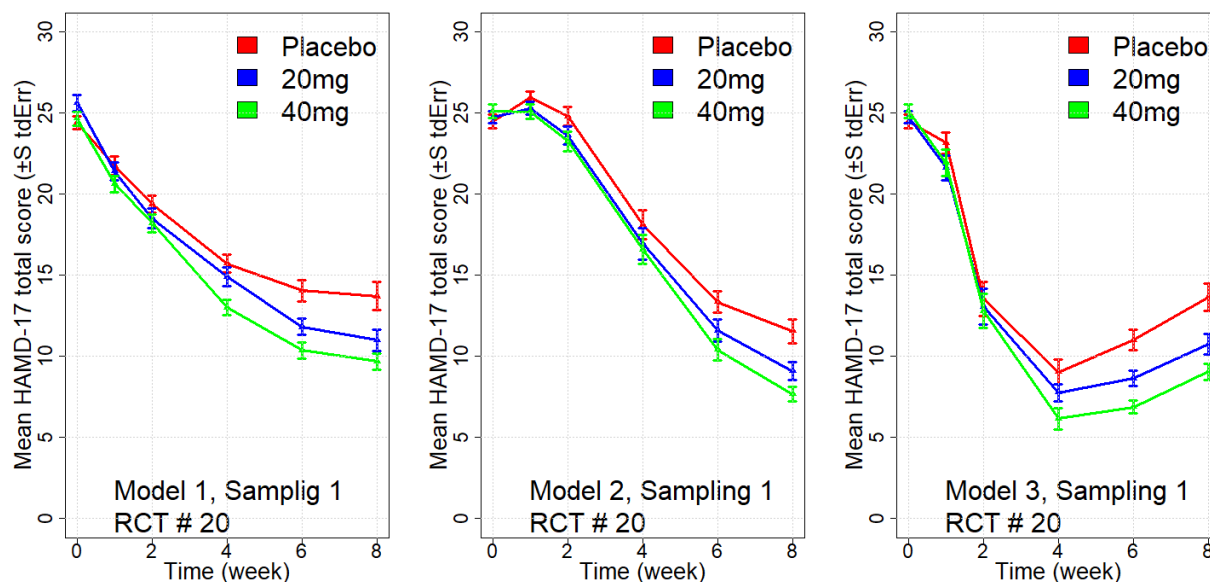


Fig2. Individual model simulated longitudinal HAMD-17 total score (dots) with the median predicted curves (solid lines) and the 95% prediction intervals (shaded area) for the simulations of RCT number 20 using the Model 1 and Sampling 1 definition. The gray horizontal dotted lines represent the HAMD-17 total score equal to 10 as a reference.

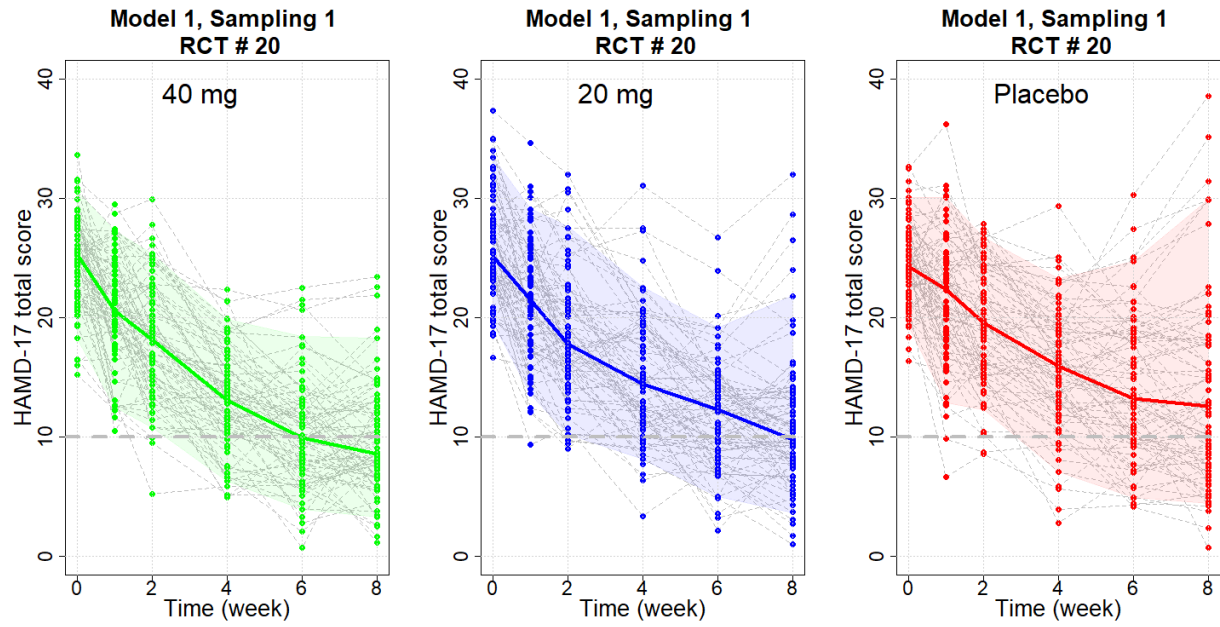
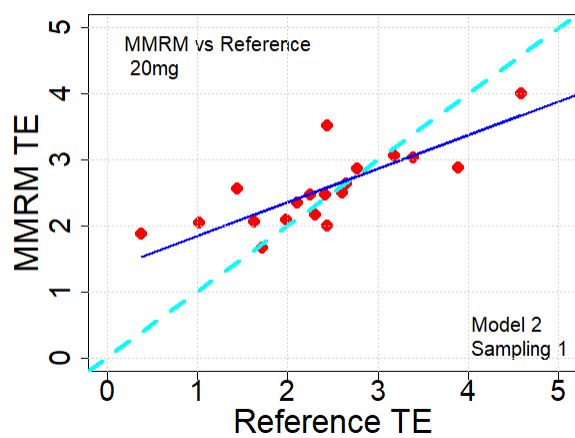
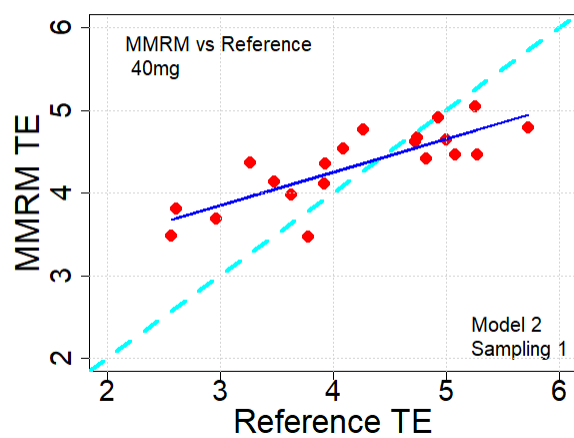
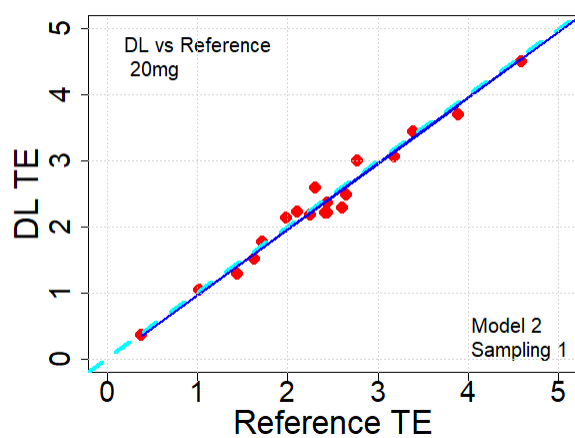
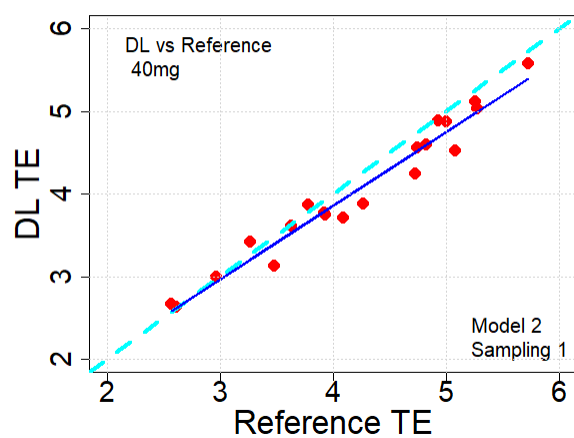


Fig3. Correlation between reference TE estimates and TE values estimated from the DL and MMRM analyses (red points) assessed across the 20 RCT replicates. The blue solid lines represents the regression line and the light blue dotted curve represent the identity reference lines.



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table1.docx](#)
- [Table3.docx](#)
- [Table2.docx](#)
- [Supplementalmaterial.docx](#)