Psychiatry Research 327 (2023) 115367

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

A new method for analyzing clinical trials in depression based on individual propensity to respond to placebo estimated using artificial intelligence

Roberto Gomeni [a,*], Françoise Bressolle-Gomeni [a], Maurizio Fava [b]

[a] Pharmacometrica, La Fouillade, France
[b] Department of Psychiatry, Massachusetts General Hospital, and Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Placebo response
Depression
Propensity analysis

ABSTRACT

One of the major reasons for trial failures in major depressive disorders (MDD) is the presence of unpredictable levels of placebo response as the individual baseline propensity to respond to placebo is not adequately controlled by the current randomization and statistical methodologies.

The individual propensity to respond to any treatment or intervention assessed at baseline was considered as a major non-specific prognostic and confounding effect. The objective of this paper was to apply the propensity score methodology to control for potential imbalance at baseline in the propensity to respond to placebo in clinical trials in MDD.

Individual propensity was estimated using artificial intelligence (AI) applied to observations collected in two pre-randomization occasions.

Cases study are presented using data from two randomized, placebo-controlled trials to evaluate the efficacy of paroxetine in MDD. AI models were used to estimate the individual propensity probability to show a treatment non-specific placebo effect.

The inverse of the estimated probability was used as weight in the mixed-effects analysis to assess treatment effect. The comparison of the results obtained with and without propensity weight indicated that the weighted analysis provided an estimate of treatment effect and effect size significantly larger than the conventional analysis.

1. Introduction

The randomized placebo-controlled clinical trial (RCT) design is considered as the "gold standard" design for investigating the efficacy of new treatments. However, accumulated evidence indicated that this design failed to assess the treatment effect (TE, defined as the baseline-corrected change from placebo) in a large number of trials conducted to investigate the efficacy of novel medications for CNS diseases. The main reason for study failure was identified as the high and unpredictable levels of placebo response (Benedetti et al., 2003).

There is an extremely large body of evidence showing that the level of placebo response has a critical prognostic relevance in the assessment of TE in RCTs conducted in major depressive disorders (MDD) (Khan et al., 2003; Li et al., 2019; Papakostas et al., 2009). Meta-analysis conducted on 81 RCTs in MDD submitted to the US Food and Drug Administration (FDA) between 1983 and 2008 showed that only 53% of the trials were successful in the last 25 years and that the placebo response rate was increasing over time (Khin et al., 2011; Colloca, 2019; Gopalakrishnan et al., 2020; Khan et al., 2017; Tuttle et al., 2015; Enck, 2016).

As a consequence, the level of placebo response can be considered as a relevant prognostic covariate that cannot be ignored in any inference even when randomization has been deployed (Senn, 2013). Therefore, new methodological approaches for designing, conducting, and analyzing RCTs are needed for controlling and mitigating the increasing confounding effect of placebo response. Several methods were proposed to address this issue, such as the identification and the exclusion of placebo responders during a placebo run-in period (Faries et al., 2001; Scott et al., 2022), and the two stages sequential parallel comparison designs (Fava et al., 2003; Chen et al., 2011). In addition, the band-pass methodology was proposed to improve signal detection in antidepressant clinical trials using an enrichment window approach that identifies sites with extremely low/high mean placebo responses and excludes data from those sites from the analysis (Merlo-Pich et al. 2008; Gomeni

* Corresponding author at: Pharmacometrica, Lieu dit Longcol, 12270, La Fouillade, France.
E-mail address: roberto.gomeni@pharmacometrica.com (R. Gomeni).

https://doi.org/10.1016/j.psychres.2023.115367
Received 18 June 2023; Received in revised form 28 June 2023; Accepted 23 July 2023
Available online 2 August 2023

et al., 2019). More recently, released a guidance for implementing enrichment strategies in clinical investigations to evaluate the effectiveness of new drugs in the attempt to identify and exclude patients who improve spontaneously or have large placebo responses (FDA guidance, 2019).

In the context of the present analysis, the following definitions were used: PR = placebo response associated with a clinical improvement in patients treated with placebo, and PE = placebo effect associated with a clinical improvement due to expectancies of positive outcomes of a treatment irrespectively of the assigned treatment (Colloca et al., 2019). PE is a common outcomes in RCTs conducted in many psychiatric diseases (Palpacuer et al., 2017), it is usually associated with the patient's interactions with the clinician (Kaptchuk et al., 2015), and it was identified as a major non-specific effect affecting the individual level of PE (Salanti et al., 2018).

TE can be considered as the resultant of treatment specific and non-specific responses and the individual propensity to respond to any treatment assessed using pre-randomization observations can be considered as a relevant prognostic factor. The larger is the propensity to respond to non-specific treatment, the lower will be the chance to detect any treatment-specific effect (Ioveno et al., 2012; Katz et al., 2008).

The computations of PE is essential in RCTs for separating the specific effects of treatments from unspecific effects associated with the therapeutic intervention. Thus, the identification of placebo responders is critical for testing the efficacy of new interventions and drugs (Aslaksen, 2021).

In RCTs, subjects are randomly assigned to treatment arms to insure comparability of the study outcomes by balancing the distribution of potential confounders over treatment arms. Despite the randomization, the groups to be compared may remain unbalanced, and thus incomparable as relevant baseline prognostic covariates (i.e., PE propensity) not accounted by the randomization have been ignored.

Propensity weighting is a novel statistical inference approach aimed to reduce and control baseline imbalances between treatment arms (Moons, 2020). This methodology was developed for mitigating the confounding bias in non-randomized comparative studies and to facilitate causal inference for TE estimate (Rosenbaum et al.,1983). The methodology was used mainly in epidemiological and social science studies, until it was adopted in a regulatory setting by the FDA, where it was used in observational studies to support marketing applications for medical devices (Yue, 2007; Campbell et al., 2016; Li and Yue, 2023; Levenson et al.,2013).

This methodology is based on the calculation of the individuals' probability of showing PE using pre-randomization response (Li et al., 2020). The use of propensity weighted approach for analyzing RCTs in MDD was recently proposed and the comparative analysis of data generated in one RCT was presented as a case study to compare the performances of conventional and propensity weighted approaches (Gomeni et al., 2023).

In the present paper, we are further elaborating the propensity weighted approach using data of two additional RCTs in MDD. The estimated individual propensity to PE will be used as weight of the individual observations in the mixed-effect model for repeated measures (MMRM) conducted to assess TE.

The higher the individual PE will be, the lower the contribution of this subject to the TE assessment will be. The expected effect of the weighted analysis will be to enhance signal detection and effect-size due to a better control of the inter-individual variability, as the contribution of subjects with high/low placebo responders will be minimized by the weighting procedure.

The individual propensity probability to respond to placebo was estimated using the change from screening to baseline of the individual 17-item of the Hamilton Depression Rating Scale (HAMD-17) (Hamilton, 1960) as potential predictors of the placebo response at end of study (EOS) using the multilayer artificial neural network (ANN) method (Yu et al., 2019).

The predictive power of the model to estimate the response at EOS was assessed using an artificial intelligence (AI) approach. The ANN model, developed using the placebo data, was applied to the individual HAMD-17 item changes from screening to baseline of each subject to estimate the individual probability of PE. The inverse of this value was used as an individual weight in the MMRM analysis conducted to assess the TE.

A comparative analysis was conducted to estimate TE and effect-size with and without propensity weight in the two selected RCTs in MDD. A sensitivity analysis was also conducted to evaluate the potential risk of inconsistent assessment of TE and study failure in new trials in presence of high or low level of PE by comparing the outcomes of a propensity weighted and traditional MMRM approach.

2. Methods

2.1. Data

The data of two antidepressant RCTs were used. The first trial (study 449) was a double-blind, placebo controlled trial evaluating the effects of immediate (IR) and controlled release (CR) paroxetine in MDD using a flexible dose design. Subjects ($N = 108$, 112, and 110) were randomized to either CR (25–62.5 mg/day), IR (20–50 mg/day), or placebo.

The second trial (study 874) was a randomized, double-blind, parallel-group, placebo-controlled fixed-dose study evaluating the effect of paroxetine CR in MDD elderly outpatients using a fixed-dose design. Subjects ($N = 168$, 177, and 180) were randomized to either paroxetine CR (12.5 mg), paroxetine CR (25 mg), or placebo. The primary efficacy endpoint for the two RCTs was the change from baseline to the week 8 in the HAMD-17 total score. Details on these two trials were previously reported (Merlo-Pich et al., 2010).

2.2. Model development

The data of the two trials were independently analyzed using a sequential approach:

1) ANN model development using screening and baseline observations and EOS data (i.e., visit at 8 weeks) in subjects randomized to placebo to estimate the probability to be placebo responder at EOS.
2) ANN model validation by comparing model-predicted probability and observed placebo response and by estimating the area under the Receiver Operator Characteristic (ROC) curve.
3) Prediction of the individual probability of PE using the pre-randomization data of each subject randomized in the study using the ANN model.
4) Longitudinal MMRM analysis using the inverse individual probability as weighting factor to estimate TE.

The propensity to respond to placebo was defined as the probability of a clinically relevant reduction from baseline of the HAMD-17 total score at EOS. The relevant improvement in HAMD-17 was estimated by linking the change of HAMD-17 to the clinical global impression-severity scale (CGI-I) using the equipercentile linking method (Guy et al., 1976; Kolen et al., 2014). This analysis indicated that a CGI-I score of 3 ('minimally improved') was associated to an average reduction from baseline in the total Montgomery-Åsberg depression rating scale (MADRS) score (Montgomery, 1979) of 24.5%, a CGI-I score of 2 ('much improved') was associated to an average reduction of 52.5%; and a CGI-I score of 1 ('very much improved') to an average reduction of 82% (Leucht et al., 2017). A robust improvement in the disease severity was estimated as a percent change from baseline in MADRS scale of 38%: the median value between minimally and much improved CGI-I. The equivalent clinically relevant reduction in the HAMD-17 scale was estimated using the equipercentile linking method developed to estimate equivalence between MADRS and HAMD-17 assessments. The

percent reduction in HAMD-17 of 41% was identified as the equivalent percent reduction of 38% in MADRS (Leucht et al., 2018). This value was used in ANN analysis for identifying placebo-responders.

A binary score (0 or 1) was associated to each subject for absence or presence of response at EOS (i.e., HAMD-17 $\geq$ 41%). The model development and validation process was based on a random split of the original data into three datasets:

1) Training set including 75% randomly selected data in the placebo arm for ANN model development.
2) Validation set including the remaining 25% data used for assessing model performance in the placebo arm by comparing the model predictions with observed data
3) Working dataset, including the data of all subjects randomized in the RCTs, used to provide individual estimates of the propensity probability applying the ANN model.

Many potential predictors of placebo response evaluated at baseline can be considered such as demographic data, habits and quality of life, or disease-related information, etc. in the attempt to improve the overall predictive performance of the model. For simplicity, we decided to limit our exploration to the 17 items of the HAMD scale as these items are assumed to capture specific and independent symptoms of depression. The performance of the changes in these 17 items to predict the placebo response at EOS was evaluated using ANN as this methodology was shown to provide one of the most performing predictive tools (Hulsen, 2022). The ANN model requires the definition of the number of hidden layers and the number of nodes in each hidden layer (Rosenblatt,1961; Rumelhart et al., 1985).

In Step 1 of the analysis, a grid search was conducted for identifying the optimal number of layers and the optimal number of nodes in the ANN models. In Step 2 of the analysis, the validation dataset was used to evaluate the predictive performance of the best performing model. The criterion for model validation was the area under the ROC curve, with the associated 95% confidence interval. The ANN analysis was conducted using the '*neuralnet*' library in R (R Core Team, 2023). In Step 3 of the analysis, the ANN models developed using only placebo data were used to predict the individual PE in each subject using the individual pre-randomization data.

2.3. Longitudinal analysis

The inverse of the individual estimated probability was used as weight in the MMRM model for the longitudinal analysis of the HAMD-17 total score change from baseline (PROC MIXED, Version 9.4, SAS Institute, Carry, NC, USA). The analysis was conducted on changes from baseline using a random effect on the change from baseline, using an unstructured covariance matrix, time as a classification variable, baseline measurement as a covariate, baseline x time interaction, and treatment x time interaction. A level of $\alpha = 0.05$ was used to establish the significance of the TE. The effect-size was estimated using the least square (LS) mean active-placebo difference divided by the pooled standard deviation obtained as the standard error of the LS mean difference divided by the square root of the sum of inverse treatment group sample sizes.

2.4. Sensitivity analysis

A sensitivity analysis was conducted to assess the impact of excessively high/low propensity to PE on the estimated TE in the analyses conducted with and without a propensity weight in the three scenarios:

1. Exclude subjects with high probability of PE (PE $>$ 0.8)
2. Exclude subjects with very low probability of PE (PE $<$ 0.1)
3. Include all subjects

3. Results

The descriptive statistics on demographic data and on HAMD-17 total score at screening and baseline are presented in Table 1 by RCT.

The grid search analysis indicated that the optimal number of layers was 3 and the optimal number of nodes per layer was 10, 2, and 5, and 10, 13, and 3 for the 449 and the 874 studies, respectively. The optimality criteria was based on the best predictive performance of the model.

The final neural network layouts of the ANN analysis with the relative importance of the changes from screening to baseline of each individual HAMD-17 item for the prediction of placebo response at EOS is presented in Fig. 1 by study. In the left panel plots, each column represents:

- column 1, the change from screening to baseline of the 17 HAMD individual items *(dHAMD_x, with x = 1 to 17)* evaluated as potential predictors of placebo response ('*resp*'),
- column 2, the combined items characterizing the first layer,
- column 3, the combined items defining the second layer,
- column 4, the combined items defining the final layer.

The black color indicates an increasing effect and the grey color a decreasing effect. The size of the lines determines the relative influence of information associated with the connected variables in the network.

The relative importance of each explanatory variable for the response, presented in the right panel of Fig. 1, was determined by

Table 1
Descriptive statistics on demographic data and on the HAMD-17 total score at screening and baseline for the 449, and 874 studies.

Study 449			Study 874		
Treatment	Variable	Mean (StdErr)	Treatment	Variable	Mean (StdErr)
CR (*N* = 108)	Age (year)	42.01 (1.05)	**12.5 mg (*N* = 168)**	Age (year)	67.13 (0.47)
	Weight (kg)	80.54 (1.99)		Weight (kg)	79.64 (1.28)
	BMI (kg/m^2)	28.35 (0.71)		BMI (kg/m^2)	28.81 (0.42)
	Day*	7.34 (0.1)		Day*	7.7 (0.19)
	Screening HAMD-17	23.69 (0.32)		Screening HAMD-17	23.26 (0.32)
	Baseline HAMD-17	24.46 (0.33)		Baseline HAMD-17	22.49 (0.28)
IR (*N* = 112)	Age (year)	40.64 (1.14)	**25 mg (*N* = 177)**	Age (year)	67.03 (0.49)
	Weight (kg)	79.24 (1.52)		Weight (kg)	81.46 (1.32)
	BMI (kg/m^2)	28.27 (0.54)		BMI (kg/m^2)	28.63 (0.43)
	Day*	7.58 (0.17)		Day*	7.93 (0.2)
	Screening HAMD-17	23.77 (0.29)		Screening HAMD-17	23.22 (0.31)
	Baseline HAMD-17	24.38 (0.32)		Baseline HAMD-17	23.08 (0.3)
Placebo (*N* = 110)	Age (year)	40.7 (1.1)	**Placebo (*N* = 180)**	Age (year)	67.97 (0.5)
	Weight (kg)	78.73 (1.78)		Weight (kg)	82.41 (1.55)
	BMI (kg/m^2)	27.35 (0.6)		BMI (kg/m^2)	29.47 (0.47)
	Day*	7.31 (0.14)		Day*	7.96 (0.16)
	Screening HAMD-17	23.54 (0.3)		Screening HAMD-17	23.08 (0.28)
	Baseline HAMD-17	24.33 (0.31)		Baseline HAMD-17	22.71 (0.3)

* Days between screening and baseline visits.

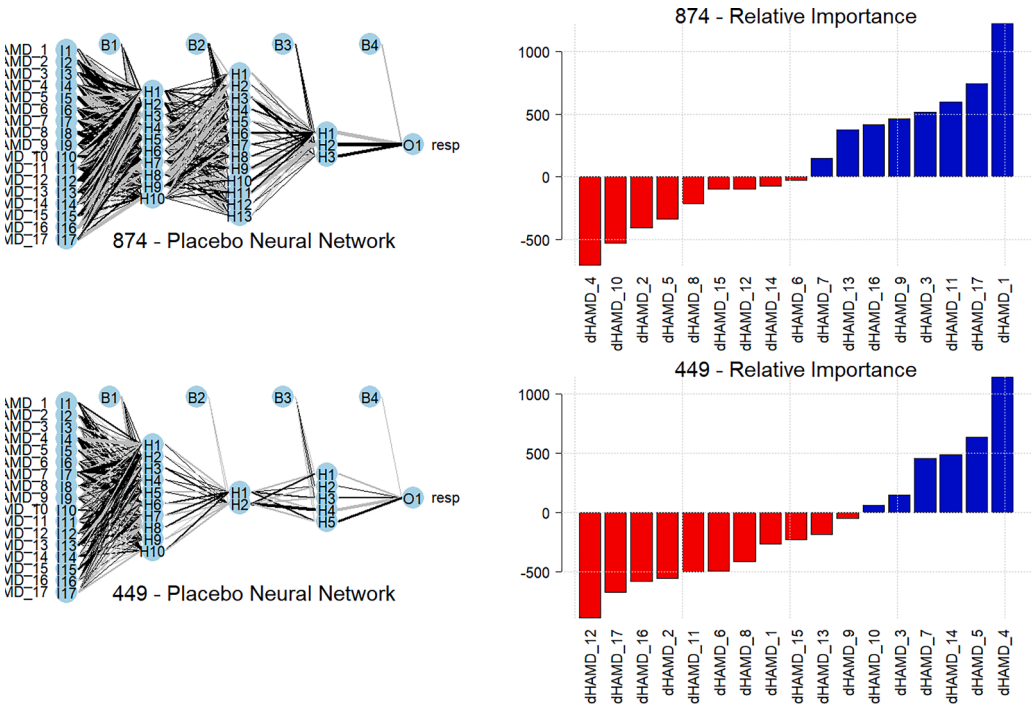

Fig. 1. Left panels: final neural network layouts for the analysis conducted using the changes from screening to baseline of the individual items of the HAMD-17 clinical scale (dHAMD_x) used as potential predictors of the response to placebo (resp). Right panels: relative importance of the changes from screening to baseline of each individual HAMD-17 item in the prediction of placebo response at EOS.

identifying all weighted connections between the nodes of interest (Olden et al., 2004). The connections were tallied for each input node and scaled relative to all other inputs. A single value was obtained for each explanatory variable that describes the relationship with response variable in the model. The estimated relative importance of each individual HAMD-17 item was presented as a bar plot where the size on the bar identifies the individual item weight in the prediction and the color identifies the positive (blue) or negative (red) contribution to the prediction.

The results of the analysis indicated that the predictive performance of the individual HAMD-17 items evaluated in the pre-randomization period varied study by study. As a consequence, the predictive performance of the data evaluated in one study cannot be translated to the data of another study as the predictive power is specific to the individual subjects enrolled in a study.

The predictive performance of the ANN models was assessed using the area under the ROC curve (AUC). The value of the AUC was 0.923 (95% confidence interval of 0.772–1.0) and 0.881 (95% confidence interval of 0.766–0.997) for the 449 and 874 studies, respectively. The ROC AUC values were statistically greater than the noninformative threshold of 0.5. As the ANN models were considered as appropriate for predicting the individual propensity probability using the pre-treatment data in the placebo arm, we assumed that the predictions for the individual propensity probability in the active treatment arms was also appropriate when the pre-treatment data were used.

The ANN models were used to estimate the individual propensity to respond to placebo for each subject included in the two RCTs. The percentage of subjects with an estimated PE to respond to non-specific treatment effects in the intervals <0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and >0.8 is presented in Fig. 2.

The distribution of PE indicated that a large majority of subjects in study 449 have a high probability (PE > 0.8) to negatively affect the estimated TE. Differently from the 449 study, the distribution of the propensity probability indicated that a large majority of the subjects in the study 874 have a high probability (PE < 0.2) to inflate the estimated TE. The results of the MMRM analyses are presented in Table 2.

The analysis with and without propensity weight indicated that the weighted analysis provided an estimate of TE and effect-size larger than

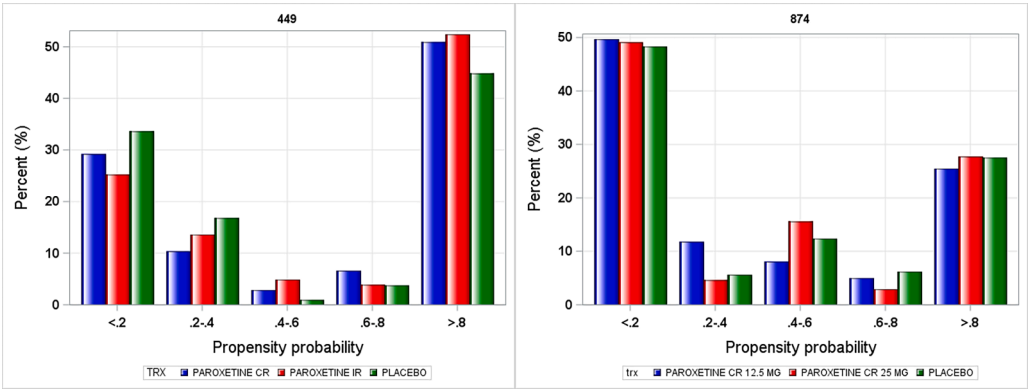

Fig. 2. Distribution of the propensity probability to a placebo effect by study and treatment.

Table 2
Sensitivity analysis results to evaluate the impact of the excessively high and excessively low propensity to a placebo effect on the estimated TE with and without a propensity weight in the MMRM analysis.

Study 449				
Analysis	Comparison	TE	P	Effect_size
Propensity Weight	CR_vs_Plac	−3.4801	0.0002	0.50955
	IR_vs_Plac	−5.9237	<0.0001	0.84602
No data with prob < 0.2	CR_vs_Plac	−0.5794	0.6489	0.06243
	IR_vs_Plac	−5.0775	<0.0001	0.58228
No data with prob > 0.8	CR_vs_Plac	−3.4704	0.009	0.36194
	IR_vs_Plac	−6.1762	<0.0001	0.62443
No Propensity Weight	CR_vs_Plac	−3.2792	0.0006	0.47321
	IR_vs_Plac	−2.7507	0.0044	0.39641
No data with prob < 0.2	CR_vs_Plac	−2.3931	0.0419	0.28069
	IR_vs_Plac	−1.2545	0.277	0.15049
No data with prob > 0.8	CR_vs_Plac	−3.2681	0.0157	0.33505
	IR_vs_Plac	−6.0379	<0.0001	0.60172
Study 874				
Analysis	Comparison	TE	P	Effect_size
Propensity Weight	12.5mg_vs_Plac	−4.0935	<0.0001	0.61279
	25mg_vs_Plac	−4.7777	<0.0001	0.7042
No data with prob < 0.2	12.5mg_vs_Plac	−1.3849	0.1871	0.14361
	25mg_vs_Plac	−0.3925	0.7061	0.04027
No data with prob > 0.8	12.5mg_vs_Plac	−4.3422	<0.0001	0.56168
	25mg_vs_Plac	−4.9945	<0.0001	0.63443
No Propensity Weight	12.5mg_vs_Plac	−1.1302	0.1549	0.15494
	25mg_vs_Plac	−1.9458	0.0129	0.26637
No data with prob < 0.2	12.5mg_vs_Plac	2.4892	0.026	0.2436
	25mg_vs_Plac	2.097	0.0531	0.20756
No data with prob > 0.8	12.5mg_vs_Plac	−4.118	<0.0001	0.5239
	25mg_vs_Plac	−4.3809	<0.0001	0.55087

the non-weighted analysis. As expected, the size of the TE was differently affected in the analyses with and without weight due to the different level of imbalance in the baseline PE, as shown in Fig. 2.

The plots of the longitudinal LS mean changes from baseline of the HAMD-17 total score by treatment and study resulting from non-weighted and weighted MMRM analyses are presented in Fig. 3.

A sensitivity analysis was conducted to evaluate how the estimated values of TE and effect-size were affected by the level of PE as expected from the meta-analysis conducted to evaluate the correlation between different levels of placebo response rate and clinical trial outcome in MDD (Ioveno et al., 2012).

Three analyses were conducted. In the first analysis, the subjects with high probability of PE (PE > 0.8) were excluded, in the 2nd analysis were excluded the subjects with low probability of PE (PE < 0.2) and in the final analysis all subjects were included. The same analyses were conducted with and without propensity weight (Table 2).

The results of the analyses, presented in Fig. 4, indicated that TE increases when the subjects with high probability of PE were removed, and TE decreases when the subjects with low probability of PE were removed. These finding are in agreement with the expected effect of low/high placebo response on the estimated TE (Ioveno et al., 2012).

The% absolute deviation from the TE value estimated in the total population and in the populations without subjects with high or low probability of PE was considered as measure of the potential risk of inconsistent assessment of TE and study failure in new trials in presence of high or low level of PE. The estimated risk was 0.503 and 0.255 for conventional and propensity weighted analyses in study 449 and 2.294 and 0.421 for study 874, respectively.

This large difference in the risk indicates that the propensity analysis is less sensitive to excessively low/high placebo responders due to the effect of the weight probability. On the contrary, the estimated TE in conventional MMRM analyses was significantly influenced by the baseline distribution of different level of PE.

4. Discussion

As previously reported (Fava, 2015), one may classify treated patients in a MDD trial based on each participant's propensity to respond to a given type of treatment. The propensity weighted methodology assumes that the TE in a MDD trial can be viewed as the resultant of treatment-specific and non-specific effects. While the specific effect can be associated with the active drug, the non-specific effect, defined by the individual probability to respond to any treatment or intervention, can be estimated using the ANN model applied to pre-randomization data.

The larger will be the imbalance in the individual baseline propensities of subjects allocated to the different treatment arms, the lower will be the chance to properly estimate TE and effect size. This because, the estimated TE and effect size derived using the current statistical methodologies will not represent the 'true' properties of the treatment but a working estimate of these values strongly correlated with the level of imbalance in the individual propensity distribution (Ioveno et al., 2012).

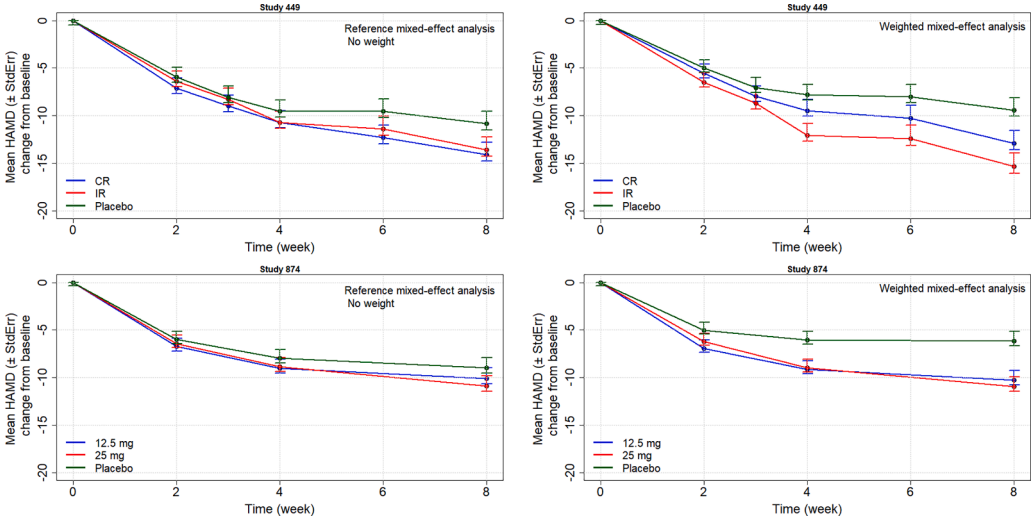

Fig. 3. Results of the non-weighted and weighted MMRM analyses with the estimation of the effect sizes. The LS mean (± standard error) of the longitudinal HAMD-17 total score changes from baseline are presented by treatment and study.

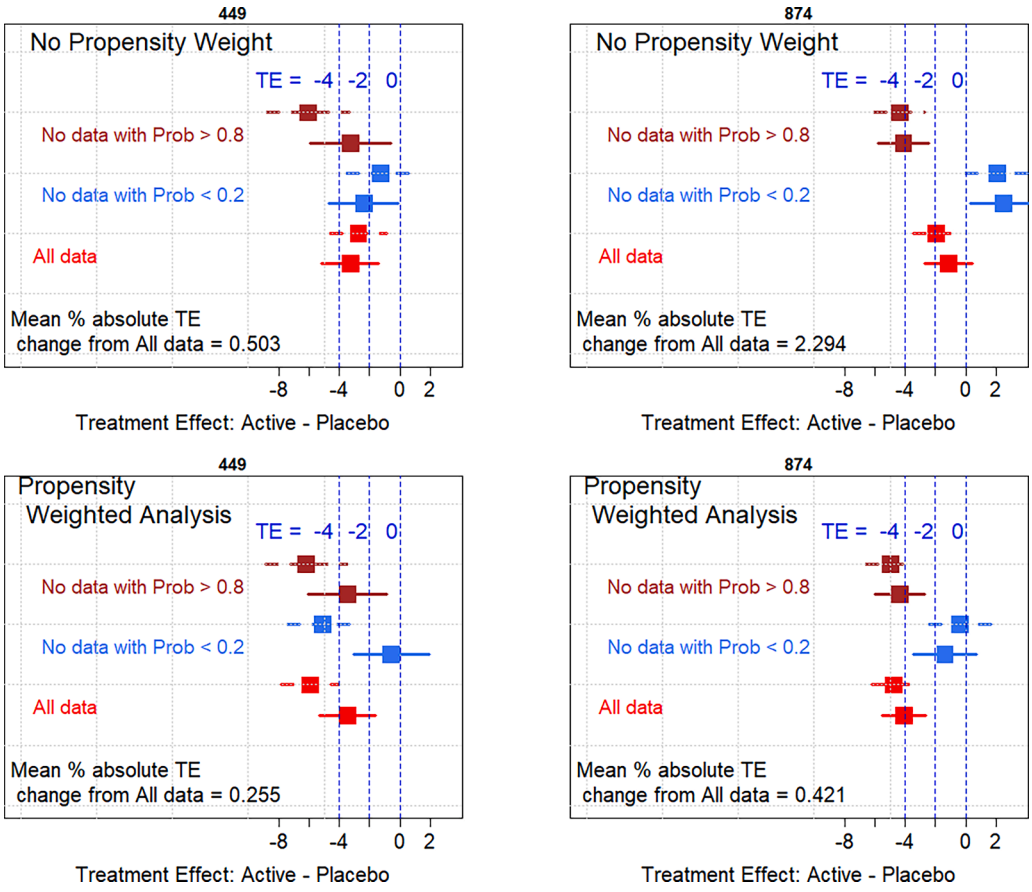

Fig. 4. Sensitivity analysis. Propensity weighted and non-weighted analyses: comparison of the estimated TE in the total population (All data) and in population without high (Prob > 0.8) and without low (Prob < 0.2) placebo response. The dots represent the TE value estimated in the MMRM analysis, the horizontal lines represent the 95% confidence intervals (the solid lines correspond to the 12.5 mg arm and the dotted lines corresponds to the 25 mg arm). The vertical blue dotted lines represent some reference TE values of −4, −2, and 0.

As shown in Fig. 2, the imbalance in the distribution of the individual propensity to PE varies study by study and remains an unaddressed issue for the comparability of treatments as this imbalance is not accounted by the randomization of the subjects included in the RCTs.

The proposed methodology assumes that the changes in the individual HAMD-17 items between screening and baseline contains relevant information on the time course of the disease, as also reported in a study in schizophrenia conducted using PANSS score (Hopkins et al., 2022). The response to placebo was defined as a clinically relevant change from baseline at EOS in the HAMD-17 total score. The relevant change was estimated by connecting HAMD-17 total score to the CGI-I scores using the equipercentile linking method and by selecting the percentage reduction associated with minimal and much improved CGI-I score.

An ANN model was initially developed to estimate the PE in the placebo treated subjects as a function of the HAMD-17 individual items evaluated in two pre-randomization occasions. This model was then validated by assessing the predictive performances of the individual items on data not used for model development. Finally, this model was applied to the pre-randomization data of each subject in the RCTs to estimate the individual propensity to respond to placebo. The inverse of the estimated propensity probability was included as weight in the MMRM model used to assess the TE in order to reduce baseline imbalances between arms (Zhang et al., 2023; Austin, 2011).

A case study was presented using data of two RCTs. The ANN models performed satisfactorily well in term of predictive performance estimated by the area under the ROC curve: 0.92 (95% confidence interval of 0.77–1.0) and 0.88 (95% confidence interval of 0.77–1.0) for the 449 and 874 studies, respectively.

The results of the analysis with and without the propensity weight indicated that the weighted analysis, corrected by the different and largely unbalanced distribution in baseline propensity probability, provided a larger estimate of both TE and effect-size.

The proposed methodology can be prospectively or retrospectively applied to any RCT when: (i) the study was designed to collect screening and pre-treatment baseline data, (ii) the criteria for assessing the clinical response to placebo were pre-specified in the analysis plan, (iii) the acceptable criteria for qualifying the predictive performance of the ANN model were defined in the analysis plan specifying that the acceptable ROC AUC cut-offs should be statistically greater than 0.5.

A relevant issue associated with the proposed analysis is related to the generalizability to a different population of the results. We are faced by two distinct issues: (a) the generalizability of the ANN model to predict the individual probability to PE and (b) the generalizability of the outcomes (i.e., the estimated TE and effect size) of the propensity weighted analysis. About point (a), the outcomes of the ANN model cannot be used for predicting the individual propensity probability as the subjects and the study designs are study specific as shown by the comparison of the 449 and 874 data. This because, the individual propensity to respond to placebo is associated with the individual expectations specific to each individual. However, the ANN model can be used with the pre-randomization data of different RCTs to estimate the individual propensity in different trials. About point (b), randomized trials remain the most accepted design for estimating the TE, but they do not necessarily answer a question of primary interest about the effectiveness and the generability of TE in a large scale target population. Recent literature indicates that a promising approach for assessing generability of TE size can be based on the use of propensity-score-based metrics using the TE adjusted and normalized by the study specific levels of confounding factors. Therefore, propensity weighting score offers a promising tool to developers, regulators or prescribers to best identify the performance of a new treatment in a target population by accounting for potential confounding effect of excessively low/high placebo response (Stuart et al., 2001, 2015; Loux and Huang, 2023).

The major difference and advantage of the propensity weighted approach with respect to the historical study designs and/or analysis procedures (Chen et al., 2011; Scott et al., 2022; Fava et al., 2003) is that all subjects randomized in the trial are included in the analysis consistently with the intention-to-treat (ITT) paradigm.

The propensity weighting method provides: (i) a model based strategy to associate to each subject a weight accounting for potential individual confounding factor of non-specific response, (ii) an estimate of the TE adjusted for the difference in the individual propensity to respond to placebo, and (iii) a better control of the impact of subjects with low/high PE. In absence of any propensity adjustment, the estimated TE will be conditioned by the proportion of subjects with excessively high/low PE.

A sensitivity analysis was conducted to evaluate potential risk of inconsistent assessment of TE and study failure in new trials in presence of high or low level of PE associated with the use (or not) of a propensity adjustment. This analysis indicated that the propensity methodology was associated with a reduced risk of inconsistent assessment of TE and study failure in new trials.

Among the benefit associated with the propensity score approach, recent papers advocate the uses of this approach to ensure balance between groups at the time of randomization, and to account for chance imbalances in observed randomization (Travis et al., 2023). While propensity scores were originally developed to address confounding in observational studies of causal effects, recent literature has shown that they are also helpful in randomized studies as well (Stuart et al., 2001). Propensity scores can be used to minimize this imbalance at the randomization stage, or to adjust for between-group differences in the analysis of outcomes. Both uses of propensity scores can improve the power of RCTs, especially in small samples or in investigating subgroup effects. Propensity scores, or propensity-based tools, can also be used to account for selection bias into randomized trials in hopes of generalizing or translating evidence from a randomized trial to a broader population (Freedman and Berk., 2008; Raad et al., 2020).

Several limitations of the current investigation should be noted. The HAMD-17 rating scale was the only clinical score evaluated. Other relevant clinical scores such as the MADRS scale have to be analyzed in trials conducted in MDD. In addition, as the unpredictable high placebo response rate is one of the major factor associated with the failure of randomized clinical trials in a large majority of psychiatric disorders such as bipolar disorders, schizophrenia, anxiety, etc., the propensity weighted approach would need to be also evaluated in trials conducted on these disorders. A further limitation of the current investigation is the restricted number of RCTs evaluated.

In conclusion, propensity score is an extensively used methodology in observational studies for improving treatment comparison by adjusting data for potentially confounding baseline factors. The results of the presented analysis indicate that this methodology can be profitably extended to deal with the control of the placebo effect in randomized placebo-controlled clinical trials.

Funding

No funding was received for this work.

CRediT authorship contribution statement

Roberto Gomeni: Conceptualization, Formal analysis, Methodology, Writing – original draft. **Françoise Bressolle-Gomeni:** Writing – review & editing. **Maurizio Fava:** Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors have no conflict of interest.

References

Aslaksen, P.M., 2021. Cutoff criteria for the placebo response: a cluster and machine learning analysis of placebo analgesia. Sci. Rep. 11 (1) https://doi.org/10.1038/s41598-021-98874-0.

Austin, P.C., 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav. Res. 46, 399–424. https://doi.org/10.1080/00273171.2011.568786.

Benedetti, F., Pollo, A., Lopiano, L., Lanotte, M., Vighetti, S., Rainero, I., 2003. Conscious expectation and unconscious conditioning in analgesic, motor, and hormonal placebo/nocebo responses. J. Neurosci. 23, 4315–4323. https://doi.org/10.1523/JNEUROSCI.23-10-04315.2003.

Campbell, G., Yue, L.Q., 2016. Statistical innovations in the medical device world sparked by the FDA. J. Biopharm. Stat. 26, 3–16. https://doi.org/10.1080/10543406.2015.1092037.

Chen, Y.F., Yang, Y., Hung, H.M., Wang, S.J., 2011. Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials. Contemp. Clin. Trials 32, 592–604. https://doi.org/10.1016/j.cct.2011.04.006.

Colloca, L., 2019. The Placebo Effect in Pain Therapies. Annu. Rev. Pharmacol. Toxicol. 59, 191–211. https://doi.org/10.1146/annurev-pharmtox-010818-021542.

Enck, P., 2016. Placebo response in depression: is it rising? Lancet Psychiatry 3, 1005–1006. https://doi.org/10.1016/s2215-0366(16)30308-x.

Faries, D.E., Heiligenstein, J.H., Tollefson, G.D., Potter, W.Z., 2001. The double-blind variable placebo lead-in period: results from two antidepressant clinical trials. J. Clin. Psychopharmacol. 21, 561–568. https://doi.org/10.1097/00004714-200112000-00004.

Fava, M., 2015. Implications of a biosignature study of the placebo response in major depressive disorder. JAMA Psychiatry 72, 1073–1074. https://doi.org/10.1001/jamapsychiatry.2015.1727.

Fava, M., Evins, A.E., Dorer, D.J., Schoenfeld, D.A., 2003. 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. https://doi.org/10.1159/000069738.

FDA Guidance, 2019. Enrichment Strategies for Clinical Trials to Support Determination of Effectiveness of Human Drugs and Biological Products. on. Available online at. https://www.fda.gov/media/121320/download. Accessed April 3, 2023.

Freedman, D.A., Berk, R.A., 2008. Weighting regressions by propensity scores. Evaluat. Rev. 12, 392–409. https://doi.org/10.1177/0193841x08317586.

Gomeni, R., Bressolle-Gomeni, F., Fava, M., 2023. Artificial intelligence approach for the analysis of placebo-controlled clinical trials in major depressive disorders accounting for individual propensity to respond to placebo. Transl. Psychiatry 13 (1), 141. https://doi.org/10.1038/s41398-023-02443-0.

Gomeni, R., Rabinowitz, J., Goyal, N., Bressolle-Gomeni, F.M.M., Fava, M., 2019. Model-informed approach to assess the treatment effect conditional to the level of placebo response. Clin. Pharmacol. Ther. 106, 1253–1260. https://doi.org/10.1002/cpt.1584.

Gopalakrishnan, M., Zhu, H., Farchione, T.R., Mathis, M., Mehta, M., Uppoor, R., et al., 2020. The trend of increasing placebo response and decreasing treatment effect in schizophrenia trials continues: an update from the US Food and Drug Administration. J. Clin. Psychiatry 81, 19r12960. https://doi.org/10.4088/jcp.19r12960.

Guy, W., 1976. Clinical Global Impressions. ECDEU Assessment Manual For Psychopharmacology. National Institute of Mental Health., Rockville, MD.

Hamilton, M., 1960. A rating scale for depression. J. Neurol. Neurosurg. Psychiatry 23, 56–62. https://doi.org/10.1136/jnnp.23.1.56.

Hopkins, S.C., Tomioka, S., Ogirala, A., Loebel, A., Koblan, K.S., Marder, S.R., 2022. Assessment of negative symptoms in clinical trials of acute schizophrenia: test of a novel enrichment strategy. Schizophr. Bull. Open 3 (1), 1–11.

Hulsen, T., 2022. Literature analysis of artificial intelligence in biomedicine. Ann. Transl. Med. 10, 1284–1298. https://doi.org/10.21037/atm-2022-50.

Iovieno, N., Papakostas, G.I., 2012. Correlation between different levels of placebo response rate and clinical trial outcome in major depressive disorder: a meta-analysis. J. Clin. Psychiatry 73, 1300–1306. https://doi.org/10.4088/jcp.11r07485.

Kaptchuk, T.J., Miller, F.G., 2015. Placebo effects in medicine. N. Engl. J. Med. 373, 8–9. https://doi.org/10.1056/nejmp1504023.

Katz, J., Finnerup, N.B., Dworkin, R.H., 2008. Clinical trial outcome in neuropathic pain: relationship to study characteristics. Neurology 70, 263–272. https://doi.org/10.1212/01.wnl.0000275528.01263.6c.

Khan, A., Detke, M., Khan, S.R., Mallinckrodt, C., 2003. Placebo response and antidepressant clinical trial outcome. J. Nerv. Ment. Dis. 191 (4), 211–218. https://doi.org/10.1097/01.nmd.0000061144.16176.38.

Khan, A., Fahl Mar, K., Brown, W.A., 2017. Does the increasing placebo response impact outcomes of adult and pediatric ADHD clinical trials? Data from the US Food and Drug Administration 2000-2009. J. Psychiatr. Res. 94, 202–207. https://doi.org/10.1016/j.jpsychires.2017.07.018.

Khin, N.A., Chen, Y.F., Yang, Y., Yang, P., Laughren, T.P., 2011. Exploratory analyses of efficacy data from major depressive disorder trials submitted to the US Food and Drug Administration in support of new drug applications. J. Clin. Psychiatry 72, 464–472. https://doi.org/10.4088/jcp.10m06191.

Kolen, M.J., Brennan, R.L., 2014. Observed score equating using the random groups design. Test Equating, Scaling, and Linking. Springer, New York, NY, pp. 29–63.

Leucht, S., Fennema, H., Engel, R.R., Kaspers-Janssen, M., Lepping, P., Szegedi, A., 2017. What does the MADRS mean? Equipercentile linking with the CGI using a company database of mirtazapine studies. J. Affect. Disord. 210, 287–293. https://doi.org/10.1016/j.jad.2016.12.041.

Leucht, S., Fennema, H., Engel, R.R., Kaspers-Janssen, M., Szegedi, A., 2018. Translating the HAM-D into the MADRS and vice versa with equipercentile linking. J. Affect. Disord. 226, 326–331. https://doi.org/10.1016/j.jad.2017.09.042.

Levenson, M.S., Yue, L.Q., 2013. Regulatory issues of propensity score methodology application to drug and device safety studies. J. Biopharm. Stat. 23, 110–121. https://doi.org/10.1080/10543406.2013.735778.

Li, H., Chen, W.C., Lu, N., Wang, C., Tiwari, R., Xu, Y., Yue, L.Q., 2020. Novel statistical approaches and applications in leveraging real-world data in regulatory clinical studies. Health Serv. Outcomes Res. Methodol. 20, 237–246.

Li, H., Yue, L.Q., 2023. Propensity score-based methods for causal inference and external data leveraging in regulatory settings: from basic ideas to implementation. Pharm Stat. 16 https://doi.org/10.1002/pst.2294.

Li, Y., Huang, J., He, Y., Yang, J., Lv, Y., Liu, H., Liang, L., Li, H., Zheng, Q., Li, L., 2019. The impact of placebo response rates on clinical trial outcome: a systematic review and meta-analysis of antidepressants in children and adolescents with major depressive disorder. J. Child Adolesc. Psychopharmacol. 29, 712–720. https://doi.org/10.1089/cap.2019.0022.

Merlo-Pich, E., Alexander, R.C., Fava, M., Gomeni, R., 2010. A new population-enrichment strategy to improve efficiency of placebo-controlled clinical trials of antidepressant drugs. Clin. Pharmacol. Ther. 88, 634–642. https://doi.org/10.1038/clpt.2010.159.

Merlo-Pich, E., Gomeni, R., 2008. 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. https://doi.org/10.1038/clpt.2008.70.

Montgomery, S.A., Asberg, M., 1979. A new depression scale designed to be sensitive to change. Br. J. Psychiatry 134, 382–389. https://doi.org/10.1192/bjp.134.4.382.

Moons, P., 2020. Propensity weighting: how to minimise comparative bias in non-randomised studies? Eur. J. Cardiovasc. Nurs. 19, 83–88. https://doi.org/10.1177/1474515119888972.

Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol. Model. 178, 389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013.

Palpacuer, C., Gallet, L., Drapier, D., Reymann, J.M., Falissard, B., Naudet, F., 2017. Specific and non-specific effects of psychotherapeutic interventions for depression: results from a meta-analysis of 84 studies. J. Psychiatr. Res. 87, 95–104. https://doi.org/10.1016/j.jpsychires.2016.12.015.

Papakostas, G.I., Fava, M., 2009. 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. https://doi.org/10.1016/j.euroneuro.2008.08.009.

R Core Team, 2023. R: 2022: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at. https://www.R-project.org/. Accessed April 3.

Raad, H., Cornelius, V., Chan, S., Williamson, E., Cro, S., 2020. An evaluation of inverse probability weighting using the propensity score for baseline covariate adjustment in smaller population randomised controlled trials with a continuous outcome. BMC Med. Res. Methodol. 20 (1), 70. https://doi.org/10.1186/s12874-020-00947-7.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55. https://doi.org/10.2307/2335942.

Rosenblatt, F., 1961. Principles of neurodynamics: perceptrons and the theory of brain mechanisms. Ryan T.A. (ed). (Spartan Books, Washington DC).

Rumelhart, D.E., Hinton, G.E., Williams, R.J., the PDP research group, 1985. Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. (Eds.), *Parallel Distributed processing: Explorations in the Microstructure of cognition*, Vol. 1: Foundations. Bradford Books/MIT Press, Cambridge MA.

Salanti, G., Chaimani, A., Furukawa, T.A., Higgins, J.P.T., Ogawa, Y., Cipriani, A., et al., 2018. Impact of placebo arms on outcomes in antidepressant trials: systematic review and meta-regression analysis. Int. J. Epidemiol. 47, 1454–1464. https://doi.org/10.1093/ije/dyy076.

Scott, A.J., Sharpe, L., Quinn, V., Colagiuri, B., 2022. Association of single-blind placebo run-in periods with the placebo response in randomized clinical trials of antidepressants: a systematic review and meta-analysis. JAMA Psychiatry 79, 42–49. https://doi.org/10.1001/jamapsychiatry.2021.3204.

Senn, S., 2013. Seven myths of randomisation in clinical trials. Stat. Med. 32, 1439–1450. https://doi.org/10.1002/sim.5713.

Stuart, E.A., Bradshaw, C.P., Leaf, P.J., 2015. Assessing the generalizability of randomized trial results to target populations. Prev. Sci. 16, 475–485. https://doi.org/10.1007/s11121-014-0513-z.

Stuart, E.A., Cole, S.R., Bradshaw, C.P., Leaf, P.J., 2001. The use of propensity scores to assess the generalizability of results from randomized trials. J. R. Stat. Soc. Ser. A Stat. Soc. 174, 369–386. https://doi.org/10.1111/j.1467-985x.2010.00673.x.

Loux, Travis, Huang, Yi, 2023. The uses of propensity scores in randomized controlled trials. Observational Stud. 9, 77–85. https://doi.org/10.1353/obs.2023.0007.

Tuttle, A.H., Tohyama, S., Ramsay, T., Kimmelman, J., Schweinhardt, P., Bennett, G.J., et al., 2015. Increasing placebo responses over time in U.S. clinical trials of neuropathic pain. Pain 156, 2616–2626. https://doi.org/10.1097/j.pain.0000000000000333.

Yu, H., Samuels, D.C., Zhao, Y.Y., Guo, Y., 2019. Architectures and accuracy of artificial neural network for disease classification from omics data. BMC Genomics 20, 167–178. https://doi.org/10.1186/s12864-019-5546-z.

Yue, L.Q., 2007. Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. J. Biopharm. Stat. 17, 1–13. https://doi.org/10.1080/10543400601044691.

Zhang, D., Li, H., Jia, W., 2023. Exploration of the prognostic value of the resection of adult brainstem high-grade glioma based on competing risk model, propensity score matching, and conditional survival rate. Neurol. Sci. https://doi.org/10.1007/s10072-022-06557-z.