Review

The placebo response in clinical trials: more questions than answers

Paul Enck¹,*, Sibylle Klosterhalfen¹,², Katja Weimer¹, Björn Horing¹ and Stephan Zipfel¹

¹Department of Internal Medicine VI: Psychosomatic Medicine and Psychotherapy, University Hospital, Tübingen, Germany
²Institute for Clinical Neuroscience and Medical Psychology, University of Düsseldorf, Düsseldorf, Germany

Meta-analyses and re-analyses of trial data have not been able to answer some of the essential questions that would allow prediction of placebo responses in clinical trials. We will confront these questions with current empirical evidence. The most important question asks whether the placebo response rates in the drug arm and in the placebo arm are equal. This ‘additive model’ is a general assumption in almost all placebo-controlled drug trials but has rarely been tested. Secondly, we would like to address whether the placebo response is a function of the likelihood of receiving drug/placebo. Evidence suggests that the number of study arms in a trial may determine the size of the placebo and the drug response. Thirdly, we ask what the size of the placebo response is in ‘comparator’ studies with a direct comparison of a (novel) drug against another drug. Meta-analytic and experimental evidence suggests that comparator studies may produce higher placebo response rates when compared with placebo-controlled trials. Finally, we address the placebo response rate outside the laboratory and outside of trials in clinical routine. This question poses a serious challenge whether the drug response in trials can be taken as evidence of drug effects in clinical routine.

Keywords: placebo; clinical trials; study designs; ethics

1. INTRODUCTION

The studies investigating neurobiological and psychological mechanisms underlying the placebo response have increased substantially over the last decade (figure 1). Despite this increase in scientific knowledge on the placebo response [1], application of this knowledge in clinical practice—be it either medical care or testing of novel compounds—has remained poor [2]. This may be due to a number of reasons. For one, placebo research is mostly experimental, and mostly dealing with placebo analgesia. It has been shown, however, that experimental placebo analgesia yields a sixfold higher efficacy than can be found in clinical studies with pain medication that include a placebo arm [3]. Secondly, meta-analyses of published trial data that have focused on the placebo response cannot overcome the fact that the reported variables usually contain only study design factors but do not include individual patient and physician characteristics that may mostly contribute to the placebo response [4]. Access to individualized patient data of such trials, on the other hand, is only feasible with re-analyses of trial raw data that are often restricted owing to secrecy policy of the drug companies. Finally, experimentally manipulating the placebo response in clinical practice—despite the wide use of placebos under routine condition [5]—is under severe ethical surveillance as soon as it implies incomplete information for the patient or even deception [6].

Placebo research, therefore, has not been able to answer some of the essential questions that started the current placebo debate and research some 30 years ago. While placebo control groups in drug trials had been in use since the middle of the nineteenth century [7], it was not before the middle of the twentieth century that physicians would accept the randomization and double-blinding of group assignment [8] that would prevent biases during efficacy assessment of drugs. Randomized double-blinded and placebo-controlled (RDBPC) studies have made it evident that placebo response rates can yield high levels as compared with drug efficacy, and that they may include not only individual patient response characteristics but also spontaneous symptom variation, and methodological biases (regression to the mean). To separate these and to understand the mechanisms underlying the patient responses have been the targets of placebo research.

In the following, we will reformulate these questions in the light of current knowledge on the placebo response, and will attempt to draw conclusions for
how the placebo research may shape the way clinical trials should be conducted.

2. ARE THE PLACEBO RESPONSE RATES IN THE DRUG ARM AND IN THE PLACEBO ARM EQUAL?

It is nowadays regarded as good scientific practice to study the efficacy of drugs in randomized, placebo-controlled and double-blinded trials by subtracting the placebo response (in the placebo arm) from the response in the drug arm of such studies for estimation of the ‘true’ drug efficacy. The implicit assumption of this ‘additive model’ [9,10] is that in both the drug and the placebo arm, the drug-unspecific responses (which include the placebo response) are equal. This model reflects a general assumption in almost all placebo-controlled drug trials that have been performed since its dawn in the 1940s [7,8,11] (figure 2). Interestingly, the underlying hypothesis that the placebo response is equal in size irrespective of whether an active drug or a placebo was given has never been tested thoroughly. Some novel findings even argue against it.

This is not merely an academic question: on the one hand, if the true drug response cannot be accurately estimated from placebo-controlled trials by subtracting the observed placebo response from the observed drug response, then these trials fail to inform the clinician regarding the size of the response, and novel drugs are either overestimated or underestimated in their efficacy. On the other hand, if the major purpose of placebo-controlled trials is to determine whether drugs have ‘specific efficacy’ beyond the placebo response, then failure of the additive model is less relevant. Unfortunately, both points of view are frequently interchanged, both in clinical practice as well as in pharmacology.

The first critical arguments against the additive model were generated with meta-analyses of placebo-controlled trials. Under conditions of high placebo response rates (such as in depression, functional disorders, pain), meta-analyses and re-analyses of trial data have identified some factors that contribute to the placebo response, e.g. lower symptom severity at study onset [12–15] or improvement of symptoms during drug-free run-in [13,14]. These factors did not even marginally contribute to the drug response in the same trials, as was shown for depression in children [15] and in adults [16], for menopausal symptoms [13], and in functional dyspepsia [14,17]. If, however, the same factors are of no relevance for the drug efficacy, the placebo response in the drug arm of the trials must be driven by other mechanisms—which hence argues against the additive model.

An entirely different possibility to test the additive model was generated by a mathematical approach. In a statistical modelling approach, Muthén & Brown [18], rather than using the conventional separation in drug responders/drug non-responders and placebo responders and non-responders, differentiate between four virtually exclusive types of participants in placebo-controlled trials: patients that would respond only with drug, patients that would respond only with placebo, patients that would respond both with drug and placebo (‘always responder’) and patients that would not respond to drug and placebo (‘never responder’). Based on real data from depression and
schizophrenia trials, they model (with a ‘growth mixture model’) the distribution of individual patients to these four classes of responses using continuous symptom monitoring data rather than fixed endpoints. For the drug arm of the depression trial, the overall drug responder rate was 68 per cent while ‘drug-only’ responder rate was only 26 per cent. For the schizophrenia trial, the drug responder rate was 70 per cent, while the drug-only responder counted only 35 per cent. The same discrepancy occurs between overall placebo response and ‘placebo responder only’. While this does not preclude that the number of always responders (= placebo responders) is similar in both study arms, it would provide a way of testing the hypothesis of the additive model using real data from a randomized placebo-controlled trial. If at least one example will be found where the assumption is falsified, the model itself is at question.

Finally, doubts about the validity of the additive model derive from neurobiological evidence. In a recently published placebo analgesia investigation using functional magnetic resonance imaging, Petrovic et al. [19] demonstrated that separate mechanisms have to account for the placebo response in an (open) drug trial (with an opioid agonist) and following application of placebo in an expectancy trial. While the drug caused greater activation than placebo in the rostral anterior cortex, placebo caused a greater increase in the lateral orbitofrontal and the ventrolateral prefrontal cortex; both however, were effective in reducing experimental pain. This can be taken as evidence that the placebo effect during open drug treatment—which is inherent to all drug applications—must be different from the placebo response during the placebo analgesia experiment, and that the placebo response during drug administration ‘underestimates’ the placebo response possible with expectation-induced placebo analgesia—hence argues against the additive model [20].

In summary, therefore, the additive model is at question, and alternative models of drug testing are needed that account for the possibilities of non-additive placebo effects. We will discuss some potential alternatives, at least for use in experimental settings, later in this chapter.

3. IS THE PLACEBO RESPONSE A FUNCTION OF THE LIKELIHOOD OF RECEIVING DRUG/PLACEBO?

Evidence from animal research suggests that tonic dopamine release in the striatum, the ‘reward centre’ of the brain, is maximal with a 50:50 chance of receiving reward in a money game [21], and dopamine release has been shown to be a major component of the placebo response in humans as well [22]. It still needs to be shown whether altering the chances of reward is associated with higher or lower activation of respective reward mediators in the human brain. First evidence derives from a recent paper by Lidstone et al. [23] on placebo-initiated dopamine responses in Parkinson’s disease: the clinical response to varying likelihood of active treatment mimics the animal data with maximal response for 50 and 75 per cent chances of receiving active treatment compared with 25 and 100 per cent, while for the neurobiological response, only the 75 per cent chance of active treatment led to significant dopamine release in the caudate, putamen and ventral striatum.

Some clinical data also suggest that the number of study arms in a trial, e.g. with various dosages of the drug against placebo, codetermines the size of the placebo and the drug response. In two meta-analyses of depression trials [24,25], it was shown that the lower the likelihood of receiving active treatment (when compared with placebo), the lower the response to placebo and to drug. Similar findings were made for migraine earlier [26] and for schizophrenia treatment recently [27]: with trials designs that randomized 50 per cent of patients to either drug or placebo (called 1:1 ratio trials here), the placebo response would be minimal compared with trials with two or more drug arms and higher numbers of patients assigned to active treatment compared with placebo (called 2:1 or ≥2:1 ratio trials).

These findings underline a speculation by Colagiuri [28] that one of the three ways by which patient expectations limit the validity of RDBP trials refers to the lower expectancy of receiving active treatment in these trials when compared with clinical routine (see below). Halpern et al. [29] found that fewer patients were willing to participate in a hypertension trial as the percentage who would receive placebo increased (from 10 to 50%), but randomly assigning half of patients to placebo still yielded maximal recruitment efficiency.

Interestingly, most meta-analyses would ignore this and would instead process various drugs arms against the same placebo arm without adjusting for the likelihood of receiving active treatment. This is further supported by data from other areas: among 100 trials with various drugs in functional bowel disorders (irritable bowel syndrome, IBS) [30], 17 were identified that used a drug : placebo ratio greater than 1:1, and these studies yielded a significantly higher placebo response rate than 1:1 studies [31].

RDBPC trials in depression, in IBS, in migraine headache and in schizophrenia showed that maximal differences between drug and placebo are achieved with a 1:1 ratio. This generates an important dilemma: if exposing patients to placebo carries an ethical burden that requires the minimal number of patients to be assigned to placebo treatment [32], more active treatment arms would be in favour. On the other hand, 1:1 trials would require fewer patients to be tested to prove efficacy of the drug over placebo, and thus would claim the same ethical argument to be in favour of 1:1 trials. This dilemma becomes even more virulent with comparator trials.

4. WHAT IS THE SIZE OF THE PLACEBO RESPONSE IN ‘COMPARATOR’ STUDIES?

At first glance, it appears paradoxical to ask for the placebo response in trials where no placebo is given but a novel drug is tested against drugs already available (comparator). Comparator trials have been favoured for ethical reasons, especially under conditions where providing no treatment would be regarded as highly unethical and where providing ‘the best available therapy’ is required, i.e. in the case of severe diseases, when effective treatment is available, or in Third-World
countries with poor medical infrastructure [33]. Testing new drugs against comparators also generates specific statistical problems (proving non-inferiority rather than superiority of the new compound) [34]. Comparator trials provide 100 per cent security to receive active treatment, hence they resemble the ultimate extreme to 1 : 1 trials, and from the above-discussed changes in placebo response rates depending on the likelihood of receiving active treatment, one may expect a further increase in the placebo response, as in the depression and the IBS trials. Unfortunately, no comparator trials are available for IBS because this area is lacking highly effective treatment options [31].

In depression treatment, effective drugs are available and have been used as comparators. Rutherford et al. [35] compared the efficacy of various antidepressants in 48 placebo-controlled studies with 9515 patients treated with the efficacy of the same drugs in 42 comparator studies with 7030 patients. They found on average a 15 per cent higher response rate of the drugs in the comparator trials, which they attributed to expectancy responses (patients knowing that they would receive active treatment anyway). Since the average placebo response in the placebo-controlled trials was 35 per cent, they calculate a total of 50 per cent placebo response in comparator trials (table 1).

This again leads us back to the argument that comparator trials are of higher ethical value because they provide treatment to all patients. Even if so, they simultaneous also increase the burden for the development of new compounds as they require more patients to be included into a trial to prove non-inferiority of the new drug, which in turn casts doubts on their ethical ‘superiority’.

Table 1. Estimation of the placebo response in ‘comparator trials’ when no placebo is given but a novel drug is tested against drugs already on the market. The meta-analysis is based on trials with antidepressants; see Rutherford et al. [35] for details. SSRI, serotonin reuptake inhibitors; SNRI, serotonin noradrenaline reuptake inhibitors; TCA, tricyclic antidepressants; MAOI, monoamine oxidase inhibitors.

<table>
<thead>
<tr>
<th>comparator trial</th>
<th>placebo-controlled trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of studies</td>
<td>42</td>
</tr>
<tr>
<td>number of active treatment groups</td>
<td>84</td>
</tr>
<tr>
<td>number of patients included</td>
<td>7300</td>
</tr>
<tr>
<td>drugs tested</td>
<td>SSRI, SNRI, TCA, MAOI, atypical</td>
</tr>
<tr>
<td>duration of trials</td>
<td>6–12 weeks</td>
</tr>
<tr>
<td>average placebo response</td>
<td>—</td>
</tr>
<tr>
<td>average drug response</td>
<td>65%</td>
</tr>
<tr>
<td>true placebo response</td>
<td>34 + 13 = 47%</td>
</tr>
</tbody>
</table>

*This difference in drug efficacy can be attributed to the higher expectations of effective treatment in comparator trials.

response in clinical routine, it is not ‘routine’ any longer), it nevertheless poses a serious challenge: whether the treatment response in trials can be taken as evidence of treatment effects in clinical routine.

Hegerl & Mergl [36] argue that features of daily practice that are not applicable in randomized trials (such as adjustment of dosing, switching of drug and also its pricing [37]) alter the efficacy of drugs in daily routine, as does the level of health insurance a patient has, the status as inpatient or outpatient and the relapse and recurrence history of an individual patient. Furthermore, the clinical equivalent of placebo treatment is what they call ‘watchful waiting’, which is frequently used by physicians before prescribing a drug, and that does generate less hope (and less of a placebo response) than providing a 50 per cent chance of receiving a drug in a randomized placebo-controlled trial. According to their model assumptions, drug efficacy in daily routine might be substantially higher than in controlled trials (figure 3) owing to significantly higher non-specific effects of treatment (placebo response).

Similarly, Colagiuri [28] argues that the knowledge that a participant will be allocated active treatment or placebo in double-blind placebo-controlled RCTs is likely to lead to weaker treatment responses than would be expected in standard clinical practice, in which patients are unlikely to doubt that they have been given an active treatment.

It is well known that the frequency of visits and the average time spent with a patient significantly contribute to patient satisfaction, and that both are usually much higher in clinical trials, where they drive the placebo response [38–41], than in clinical routine. However, this argument does not exclude the possibility that under certain circumstances the placebo response in clinical routine may even be higher than in clinical trials. Why otherwise would on average 50 per cent of all questioned physicians in many
countries acknowledge that they have used inappropriate (non-effective = placebo) medication at least once during the course of a year in their patients [5]. While this is highly questionable from an ethics point of view, it indicates that physicians are well aware of the power of the placebo response and make use of it, at least in individual cases.

6. NOVEL STUDY DESIGNS FOR DRUG TESTING

Kirsch & Weixel [9] had proposed to use different drug-testing models such as the ‘balanced placebo design’ (BPD) (figure 4a) for separating the true drug effect and the true placebo effect from compound effects in clinical trials, but a BPD has two major disadvantages: it provides (false or correct) information that raises substantial ethical concerns [42], and it provides this information prior to drug testing, which raises suspicion among subjects as to why this information is provided.

We [43] have recently presented a number of alternative trial designs that have been developed, among them the ‘balanced cross-over design’ (BCD) (figure 4b): in this case, subjects are divided into four groups, and all are told they participate in a conventional randomized double-blinded and placebo-controlled cross-over trial, in which they will receive both the drug and the placebo at two different occasions in a randomized and double-blinded order. However, only groups 2 and 3 will be exposed to drug and placebo in a balanced way, that is half the subjects will receive the drug first and the placebo at the second occasion, while the other half will receive first placebo and then the drug. Group 1 will receive the drug twice, and group 4 will receive placebo twice instead.

In this case, groups 2 and 3 represent the conventional drug trial assuming the additive model for drug and placebo effects. In group 1, the difference between test 1 and 2 should be attributable to placebo, and the minimal value of both should represent the true drug effect. In group 4, the maximum value should represent the true placebo effect. The additive model can be tested by comparing the difference between drug and placebo in groups 2 and 3 with the true placebo response in group 4, and the ‘assumed’ drug effect in groups 2 and 3 with the true drug effect in group 1.

Such a design would overcome some of the limitations of the BPD, such as the irritation that may occur in subjects when they are informed prior to drug testing on what they are receiving. With increased basic knowledge of subjects about placebo-controlled trials, this may cause questioning of the surreptitiously given study rationale and may induce second thoughts about its true purpose that would interfere with the study goals.

A downside of the BCD—similar to the BPD—is the deception of subjects, which would prohibit its use in patients. A way out of this dilemma of placebo research—false information for subjects to answer the research question—has been shown by Miller et al. [42] and is called ‘authorized deception’: subjects are informed that they will receive incomplete information of the purpose of the study to not corroborate its goals, but will be briefed completely after its termination. It has recently been shown in an experimental placebo analgesia trial that authorized deception produces similar results to a deceptive trial design [44].

Another model called the delayed response design would allow a proof of the assumptions of the additive model (figure 4c). With this design, all subjects receive the same information that they will receive either a drug or a placebo in a double-blinded fashion, and no information is given about the timing of drug action; instead, a rationale (‘cover story’) is provided for prolonged drug action monitoring, e.g. for 24 h. Group 1 will receive the drug with immediate action, group 2 the respective placebo. Group 3 receives the delayed response medication, e.g. with drug release after 12 h. To confirm the non-additive model, the following equations need to be fulfilled: P1 ≠ P2, M1 + P1 ≠ M2 + P3 with P3 = P2 and M1 = M2 (see figure 4 for details).

Placebo-study designs may be grouped into different classes that either manipulate patient/subject expectancy by providing false information (which restricts their use in the patient) or by manipulating the timing of drug application. The latter has been shown to dramatically alter the efficacy of drugs: when drugs such as benzodiazepines for anxiety control or cholecystokinin antagonists for pain control were given in a hidden fashion, i.e. without notice as to the time when the drug was provided through an infusion line, they lost most of their clinical efficacy in comparison to open and visible application [43].
Finally, a unique but never tested model would not randomize patients between drug and placebo but would allow them ‘free choice’ between two pills, one being the active drug and one the placebo. Provided that they are restricted (by technical means) to take both (‘not to take chances’), assessment of drug-placebo efficacy would not rely on symptom improvement reports (which may contain reporting biases) but rather on choice behaviour. Other restrictions may be short-acting effects of the drug, no need for steady drug levels and effects on symptoms rather than biochemical disease indicators, hence symptomatic endpoints rather than disease biomarkers. It would also be adherent to ethical restrictions, as it does not provide no or non-adequate treatment to patients.

In the case that one or more of the above-discussed alternative study designs prove that the additive model is not valid, many drug and other studies performed in the past will lose their credibility—but as long as we do not have alternatives for clinical use, insisting on statistically significant differences between drug and placebo arms of trials may still be a reasonable and valid option.

7. SUMMARY AND CONCLUSION

While our neurobiological and psychobiological knowledge of what drives the placebo response has increased substantially, it has at the same time made it evident that some of the assumptions about placebos that we use both in research and routinely are at least questionable if not false.

Among them is the central proposition of all of currently conducted placebo-controlled trials, i.e. that the placebo response may be similar in all arms of the trials and, therefore, true drug efficacy can be assessed by subtracting unspecified (placebo) effects from the response in the drug arm of studies. If this does not hold true any longer, novel study designs are needed that assess the true effect size of drugs and placebos in a different fashion. Some potential future drug designs are discussed here, but they raise ethical issues that have to be kept in mind when studying patients.

We also identified an ethical paradox that requires a solution: while higher randomization rates of patients to drug arms of trials (either by enrichment of the drug arm or by testing a novel compound against a comparator drug) seem ethically needed to provide effective treatment to as many patients as possible and to minimize the number of patients randomized to placebo, increasing the a priori likelihood of effective treatment increases the placebo response rate itself, thus requiring more patients to be included into the study to demonstrate efficacy of drugs.

Finally, an entirely unanswered question addresses the problem of placebo response rates in medical routine beyond clinical trials and laboratory testing. While we know that placebos are frequently used in everyday medicine, their efficacy still needs to be determined. First evidence points towards the high contribution of physician behaviours as a major modulating factor. All these problems may be the true reason why many of the new insights have not yet found their way into clinical research and clinical practice.

Supported by a grant from Volkswagen Foundation, I/83805.

REFERENCES


