Improving study design for antidepressant effectiveness assessment

Florian Naudet, Bruno Millet, Jean-Michel Reymann, Bruno Falissard

Improving Study Design for Antidepressant Effectiveness Assessment

Short title: Improving Antidepressant Effectiveness Assessment

Florian Naudet, Bruno Millet, Jean Michel Reymann, Bruno Falissard

Florian Naudet

- MD, MPH, Research fellow,

- INSERM U669, Paris, France

- Université de Rennes 1, EA-425 Behavior and Basal Ganglia unit, Rennes, France

- Centre d'Investigation Clinique CIC-P INSERM 0203, Hôpital de Pontchaillou, Centre Hospitalier Universitaire de Rennes & Université de Rennes 1, Rennes, France

Bruno Millet

- MD, PhD, Professor

- Université de Rennes 1, EA-425 Behavior and Basal Ganglia unit, Rennes, France

- Centre Hospitalier Guillaume Régnier, Service Hospitalo-Universitaire de Psychiatrie, Rennes, France

Jean Michel Reymann

- PhD

- Centre d'Investigation Clinique CIC-P INSERM 0203, Hôpital de Pontchaillou, Centre Hospitalier Universitaire de Rennes & Université de Rennes 1, Rennes, France

- Laboratoire de Pharmacologie Expérimentale et Clinique, Faculté de Médecine, CS34317, 2 avenue du Pr Léon Bernard, 35043 Rennes, France

Bruno Falissard

- MD, PhD, Professor

- INSERM U669, Paris, France

- Université Paris-Sud and Université Paris Descartes, UMR-S0669, Paris, France

- AP-HP, Hôpital Paul Brousse, Département de santé publique, Villejuif, France

Location of work : INSERM U669, Maison de Solenn, 97 Boulevard de Port Royal, 75679 Paris cedex 14, France

Address for reprints / Correspondence to :

Florian Naudet

INSERM U669, Maison de Solenn, 97 Boulevard de Port Royal, 75679 Paris cedex 14, France

tel: (+33) 1 58 41 28 50

fax: (+33) 1 58 41 28 43

email: floriannaudet@gmail.com

This paper was supported by the French Institut National de la Santé et de la Recherche Médicale (INSERM).

Abstract:

Antidepressants effectiveness in Major Depressive Disorder (MDD) is still questioned because the extrapolation of Randomized Controlled Trial (RCT) results to “real life” settings is problematic. The application of the RCT paradigm in a disorder of this type, where global care plays a central role, raises questions regarding the internal and external validity of this type of study. Outcome measurement, attrition rates, the ability of the double-blind design to control for expectations, placebo response, the representativeness of trial participants and publication bias are major methodological pitfalls. This review discusses these issues. It is illustrated using original data and proposes some alternatives for assessing antidepressant effectiveness via different approaches. Some are easy to implement, such as ecological measures, qualitative approaches, improvement of analytical strategy and improvement of blinding procedures. Some are sophisticated, involving temporary deception to deal with the counfounding effect of expectations, and they raise ethical issues. Others resort to external validity, this being the case in observational studies. But all are necessary to explore antidepressant effectiveness.

Keywords: antidepressants, clinical trials, depression, effectiveness, methodology

Introduction

The usefulness of Antidepressants in major depressive disorder is still questioned (Ioannidis, 2008) and the issue goes beyond the scientific debate, with a backdrop of conflicts of interest and some concerns about the medicalisation of modern society (Lacasse and Leo, 2005). The usefulness of a drug is usually reflected by its efficacy (under optimal circumstances), its effectiveness (in routine care), and its efficiency (does it maximize value for money?) (Bombardier and Maetzel, 1999). Among these concepts, effectiveness seems to be the most relevant question for clinicians. Randomised controlled trials (RCTs) are generally used to address this issue.

Since the first published RCT versus placebo to explore the efficacy of streptomycin in tuberculosis (1948), this design has become a gold standard. But tuberculosis is quite different from a mood disorder that fits a bio-psychosocial model (Garcia-Toro and Aguirre, 2007), where therapeutic benefits could result from the context in which the study is performed. Global care (including the ethical meaning of this term) (Beauchamp and Childress, 2008) plays a central role, and psychological factors such as expectations may influence the results. Outcome definition, analysis and extrapolation of the results are likewise somewhat specific in mood disorders.

The NICE guidelines on the treatment and management of depression in adults (NationalInstituteforClinicalExcellence, 2010) advise caution when considering the application of RCTs results in routine practice, and suggests that better ways of assessing effectiveness have yet to be developed.

The present paper has two aims: 1/ To present various methodological shortcomings in the evaluation of antidepressant effectiveness concerning internal (reliability of the results) and external (scope for generalisation) validity of trials, which are two fundamental inter-connected and sometimes contradictory guarantees. 2/ To present certain alternatives to

address these issues in order to achieve a balance between these two concepts. We have illustrated our reflexion with secondary analyses from a previous meta-analysis (Naudet et al., 2011).

1. Methodological issues

1.1. Outcome measurement and analysis is problematic

Outcome measurement

The efficacy of antidepressants is assessed using continuous outcomes (the mean change on a scale), or categorical outcomes (response rate and remission rates).

Concerning continuous outcomes, the HDRS (Hamilton, 1960) and the MADRS (Montgomery and Asberg, 1979) are recognized as gold standards (Duru and Fantino, 2008). Nevertheless they can show up differences that are statistically significant in formal terms even for differences that are not significant from a clinical point of view (Ioannidis, 2008): the identification of a minimum clinically relevant difference is not straightforward (Falissard et al., 2003). In addition, they contain items that are not specific to depression (sleeping difficulties, anxiety, agitation and somatic complaints) and may highlight non-mood-related benefits (Moncrieff, 2002). Moreover, these scales tend to be used by tradition rather than because of their perfect validity and reliability. For example, a review showed that the HDRS was not optimal psychometrically and was conceptually flawed (Bagby et al., 2004).

The Clinical Global Impression (Guy, 1976) (CGI) is used as a global assessment. It comprises a single item with high “face validity”, but it may be more prone to rater bias (Gaudiano and Herbert, 2005). Its validity is debated mainly because the response format used in the CGI is more likely to be ambiguous (what is the definition of a patient who is "Severely ill"?) and is prone to cultural misunderstanding (for example concerning the meaning of "moderate") (Kadouri et al., 2007).

The scales considered here are clinician-version evaluations, while self-administered questionnaires like the Beck Depression Inventory (Beck et al., 1961) (BDI) or ecological measures such as computerised assessments of depression (Greist et al., 2002; Mundt et al., 2006) are far from being systematically reported in antidepressant studies.

Clinicians, in their day-to-day practice, are used to dealing with binary outcomes such as response and remission, which have considerable prognostic value (Judd et al., 1998). Although these concepts appear intuitive, no real gold standard exists and categorical outcomes are generally calculated from continuous data, and are provided by the proportion of people who meet a predefined level of improvement (response) or fall below a predefined threshold score (remission) at a given time point. This does not take into account the longitudinal aspect of these concepts, and creates the impression of clear-cut patterns where the data does not suggest any. This phenomena is interpreted as a major bias (Kirsch and Moncrieff, 2007) or as proof of antidepressant effectiveness (Gibbons et al., 2012) depending on the authors' preconceived beliefs.

Attrition rates

Among patients enrolled in a RCT, typically 20 to 40% fail to complete the study. Whereas a loss to follow-up of 5% or lower is usually of little concern, a loss of 20% or more prevents good quality intention-to-treat (ITT) analysis, can cause biased estimates of the treatment effect (Dumville et al., 2006) and restricts the scope for generalising results (Leon et al., 2006). The two approaches to the analysis of incomplete data used in most of the studies by which efficacy of new-generation antidepressants is established (conducted between 1990 and 2010) are far from ideal: 1/ complete case analysis assumes that missing data are “Missing Completely At Random” (dropout is unrelated to the phenomenon studied or to patient characteristics) which is not likely to be valid. 2/ the last observation carried forward (LOCF) procedure, which is the most frequently used method, negatively impacts treatment arm

results since dropouts are assumed not to improve beyond their removal from the study. It ignores the natural history of MDD (Posternak et al., 2006). Differential dropout rates between groups may artificially inflate the superiority of one study condition over another. This method does not incorporate the uncertainty surrounding the imputed data in the analyses (Leon et al., 2006).

Regarding categorical outcomes, the maximum bias hypothesis (non-assessed patients are recorded as in remission if they belong to the placebo group and as having not responded if they belong to the antidepressant group) could be considered as the most complete and accurate measure of robustness for an analysis. It is rarely performed in MDD trials. In fact it leads to the reverse conclusion (placebo superiority) as shown in table 1.

Insert table 1 about here.

1.2. The response rate in the placebo group affects internal validity

Response rate in placebo group

Response in the placebo group is substantial in RCTs on antidepressants and has led to many negative trials (Enserink, 1999). Evidence of an increasing placebo response rate over the years has been documented (Walsh et al., 2002) and justifies the continued use of placebo-control trials even if there are a number of proven treatments for MDD (Benedetti et al., 2005). However, the cause of this increase is unclear; it could be hypothesised that increasing antidepressant availability, greater social acceptability (Olfson et al., 2002) and the changes in the methods by which patients are recruited into therapeutic trials (Walsh et al., 2002) could have resulted in changes in clinically important population characteristics (for example outpatients with less severe episodes) and could have contributed to changes in the placebo effect. Moreover, expectations about the therapeutic benefit of treatment may have changed and could affect the results of antidepressant RCTs (Krell et al., 2004; Noble et al., 2001; Rutherford et al., 2010; Sotsky et al., 1991) because they are linked with the placebo effect.

Meta-analyses (Fournier et al., 2010; Khan et al., 2002; Kirsch et al., 2008) suggest that the baseline severity of depressive symptoms is related to clinical trial outcome. The minimum baseline HDRS score needed to reach a clinically meaningful difference between antidepressant and placebo was found to be approximately 28 (very severely depressed patients) (Kirsch et al., 2008) or 25 (Fournier et al., 2010). Despite disagreements regarding whether the increasing superiority of antidepressants relative to placebo as severity increases is due to an increasing efficacy of antidepressants or a declining efficacy of placebo, the association between the drug-placebo difference and baseline severity is consistent and robust in the different meta-analyses.

Placebo response and internal validity

Randomisation (Vandenbroucke, 2004) is a cornerstone of internal validity: it enables unbiased allocation of treatment (Schulz and Grimes, 2002) and complies with statistical theory for random sampling. Blind allocation of treatment makes it possible to infer the specific treatment effect, thus addressing the problem of patient expectations (Fisher, 1971). However, regarding antidepressants, the ability of a double-blind design to preserve the benefit of randomisation is disputed (Perlis et al., 2010). The first reason is that the majority of patients and doctors correctly distinguish between placebo and active medication (Bystritsky and Waikar, 1994; Rabkin et al., 1986): this will be referred to as "unblinding". For instance the blinding could be compromised by the emergence of adverse effects (Perlis et al., 2010) known to be associated with a specific medication; informed consent forms, which list common adverse effects, may increase this risk (Brownell and Stunkard, 1982). Moreover, the possibility of belonging to a placebo group could lead to lower expectations of how much a patient is likely to improve during the trial. The likelihood of response and remission is significantly higher in comparator versus placebo-controlled trials (Naudet et al., 2011; Rutherford et al., 2010; Rutherford et al., 2009; Sinyor et al., 2010; Sneed et al., 2008). This

phenomenon could be differential between antidepressant arms and placebo arms: a greater probability of receiving placebo predicted a greater antidepressant efficacy versus placebo (Papakostas and Fava, 2009), without influencing attrition rates (Tedeschini et al., 2010). This has methodological implications since it could lead to an under-estimation of the placebo effect in placebo-controlled trials, and ethical consequences because patient improvement in such studies is poorer.

It has been hypothesised by certain authors that the apparent antidepressant effect is actually an active placebo effect (Kirsch and Sapirstein, 1998) (the different physiological experiences resulting from the ingestion of an active drug and an inert placebo may lead patients and assessors to suspect the nature of the medication and this will then introduce bias due to different expectations for treatment effect).

Understanding the placebo response

While a high response rate in a placebo group is a major methodological problem, there is considerable debate about the size, the nature and the mechanism of the placebo effect in depression. The placebo effect is quite difficult to define and has two main interpretations: the effect of the placebo intervention, and the effect of patient-provider interaction.

In a first approach, the effects of a placebo can be estimated as the difference between the placebo arm and a no-treatment arm. A meta-analysis utilizing a controversial approach found that a placebo effect accounted for about 50 % of the response, "natural history" for about 25% and antidepressant effect for 25 % (Kirsch and Sapirstein, 1998). In contrast, in another controversial and underpowered meta-analysis of RCTs of placebo versus no-intervention (Hrobjartsson and Gotzsche, 2010), there was no statistically significant effect of placebo interventions in depression.

Beyond these two contrasted approaches, scientific knowledge about the placebo effect in MDD is derived from RCTs, which are pragmatic and compare an antidepressant to a placebo

in order to prove the superiority of the antidepressant, regardless of the underlying mechanisms. The calculation of the antidepressant–placebo difference by comparing marginal response rates is thus based on the postulate that all placebo responders should be antidepressant responders (additive model, Figure 1) whereas theoretically, antidepressant response and placebo response could be independent or, at least, substantially overlapping phenomena (non-additive model, Figure 2) with four different types of patients : 1/ placebo-only responders 2/ treatment-only responders 3/ placebo and treatment responders and 4/ non-responders (Kirsch, 2000; Rihmer and Gonda, 2008). It is also noteworthy that RCTs endeavour to reduce placebo effect, typically by eliminating subjects who show a strong placebo response before randomization (Benedetti et al., 2005; Muthen and Brown, 2009). These different aspects limit our understanding of the placebo effect.

Moreover response to placebo is not strictly a placebo effect (the psychobiological reaction to the administration of an inert treatment based on expectation and conditioning or other learning processes) (Ernst and Resch, 1995; Finniss et al., 2010): the clinical improvement following administration of a placebo (placebo response) could result from many different factors, such as spontaneous improvement (Posternak and Zimmerman, 2000), statistical regression to the mean, co-interventions, biases as well as the placebo effect. Indeed, spontaneous improvement can result from environmental or biological (e.g. seasonal) factors which afford scope for scientific investigation. Spontaneous improvement may be common in clinical practice (Posternak et al., 2006; Posternak and Zimmerman, 2000) ; the number of follow-up assessments (Posternak and Zimmerman, 2007) is related to a significant therapeutic effect.

1.3. The extrapolation of RCT results is problematic

External validity of RCTs in Major Depressive Disorder

Whilst the vast majority of patients with clinical depression are catered for in primary care, most of the research findings upon which decisions are based have involved secondary care patients. In a Cochrane Review, the authors found only fourteen studies versus placebo in primary care with extractable data, of which ten studies examined tricyclic agents, two examined SSRIs and two included both classes (Arroll et al., 2009). This contrasts with a plethora of literature on antidepressants in secondary care outpatients. These patients differ from primary care patients (Araya, 1999; Suh and Gallo, 1997): they are less severely depressed, experience a milder course of illness, have a distinct symptom profile with more complaints of fatigue and somatic symptoms, and are more likely to have accompanying physical complaints (Linde et al., 2011).

Antidepressant RCTs use numerous exclusion criteria (comorbid medical condition, short duration of depressive episode, comorbid personality disorder, mild depression, treatment response during placebo lead-in period, comorbid anxiety disorder, long duration of depressive episode, comorbid substance use disorder, prior non-response to treatment, comorbid dysthymia, current suicidal ideation). Some of these criteria are arguable from a fundamental viewpoint. Efficacy trials are designed to answer specific questions and they are required to investigate the disorder independently from co-morbidities, which undoubtedly affect response, depending partly on the agent tested: for example, the fact that antidepressants have anxiolytic effects justifies the exclusion of comorbid anxious disorders so as to explore efficacy in depression on its own. Nevertheless, this greatly reduces scope for generalisation (Posternak et al., 2002) in a disorder where comorbidity is the rule and a conclusion of effectiveness should not derive solely from these studies.

Subjects treated in antidepressant trials represent a minority of patients treated for MDD in routine clinical practice (Zimmerman et al., 2002). One study among psychiatric outpatients suggests that patients that were excluded were a more chronically ill group with more

numerous previous episodes, greater psychosocial impairment, and more frequent personality disorders (Zimmerman et al., 2005). Furthermore, participants are generally recruited by newspaper advertisement, paid for their participation in the study and may not be representative of "real life" patients (Greist et al., 2002). Even the main inclusion criterion (i.e. suffering from MDD) could reduce the external validity of such studies since there could be deficits in knowledge and in the application of this criterion by clinicians (Zimmerman and Galione, 2010).

Some data suggest that antidepressants may not or not adequately assist recovery in a "real life" setting (Brugha et al., 1992; Ronalds et al., 1997). In a retrospective analysis of a cohort of inpatients (Seemuller et al., 2010), patients eligible (applying classic inclusion criteria) for a RCT and patients not eligible differed significantly for several baseline measures and for final Global Assessment of Functioning scores, but not for any other outcome measure, such as depression rating scores. However, this study only recruited inpatients (a more homogenous population) and the analysis was not adjusted on prognostic factors at baseline or on associated treatment.

In another similar analysis applied to an outpatient cohort (Wisniewski et al., 2009), patients eligible for a RCT had a better response to treatment, which persisted even after adjustments for baseline differences. The design of this study provides a better control for confounders. A meta-regression comparison (Naudet et al., 2011) showed that antidepressant response is lower in observational studies compared to RCTs. This result has recently been replicated (van der Lem et al., 2012).

Finally, RCTs typically last 6-8 weeks whereas it is recommended that an antidepressant treatment be continued for at least 6 months after remission of the episode of depression (NationalInstituteforClinicalExcellence, 2010).

Meta-analysis limitations

The limitations of a meta-analysis are linked to the limitations of the individual studies included (Egger et al., 2001) and all the above-mentioned methodological problems have to be considered. Moreover, most studies on the effects of drugs are sponsored by the pharmaceutical industry. These studies have been shown to be more likely to demonstrate positive effects for the sponsor's drug than independent studies (Lexchin et al., 2003). When meta-analyses are not based on registered trials (e.g. FDA-registered trials), a publication bias can occur (Turner et al., 2008). It has been shown that the publication bias can lead to considering reboxetine as a serious antidepressant agent, whereas it is probably an ineffective and potentially harmful antidepressant (Eyding et al., 2010). Since 2005, RCTs need to be registered prior to participant enrolment, but two points could be improved: unpublished but registered study results must be accessible and selective outcome reporting (Mathieu et al., 2009) must be avoided.

These considerations should lead to caution in the interpretation of efficacy meta-analyses, and also in interpretation of meta-analyses concerning the influence of methodological factors (Huf et al., 2011). These are precisely some of the studies on which some of the above remarks are based. It gives an idea of the uncertainty surrounding the discussion presented here.

Insert table 2 about here.

Table 2 illustrates all the points discussed above with a descriptive analysis of 26 randomized controlled trials on venlafaxine or fluoxetine.

2. Methodological alternatives to answer the question of antidepressant effectiveness

2.1. Improving outcome measurement and analysis

Outcome measurement

Determination of the effectiveness of antidepressants should not be based exclusively on mere interviewer ratings of outcome, which can be prone to statistical noise and/or bias. A more robust approach is needed, and outcomes should be assessed in multi-modal fashion (Gaudiano and Herbert, 2005).

Categorical outcomes like response and remission should not be exclusively calculated from continuous data such as the HDRS (Kirsch and Moncrieff, 2007).

Assessment of categorical self-report (remission and response) using valid instruments is needed for sensitivity analysis. It has been suggested that depressed patients consider symptom resolution as only one of the factors in determining the state of remission, and that the presence of positive features of mental health such as optimism, vigor, and self-confidence is a better indicator of remission than the absence of the symptoms of depression (Zimmerman et al., 2006).

Furthermore these two concepts should not be assessed at a single time point but should address the question of passing time, and whether there is stability over several weeks (Bandelow et al., 2006). Continuous (BDI), collateral information, behavioural ratings and physiological indices should be obtained to complete the information derived from clinician-rated scales and to examine convergence of these data (Petkova et al., 2000).

Finally, the use of qualitative approaches should be developed in RCTs (Lewin et al., 2009) and could be of interest in antidepressant trials to understand the effects of interventions and to focus on patients' experiences, as these processes are difficult to explore using quantitative methods alone. Mixed (qualitative-quantitative) methods could be of interest in this way

(Falissard et al., Submitted). The procedure is simple, and is at present under development: video-interviews based on the iCGI procedure (Kadouri et al., 2007) are performed and are randomly shown in a blind manner to different groups of raters (experts, clinicians or medical students...) who classify them according to whether the patient received a placebo or an antidepressant. The test is a permutation test. It enables the identification of differences between groups. A qualitative analysis of the videos will enable comparison of the experiences of patients under antidepressants and under placebo in a phenomenological perspective. This would also enable a broader measurement of adverse outcomes including unwanted psychological effects as an important aspect, which could contribute to a more fine-grained comparison of conditions. In addition it tackles the limitations of the CGI mentioned above.

Attrition rate management could be improved

Before dealing with missing data, it is important to prevent them. Nevertheless, "attrition-reduced studies" can present problems for generalisation to clinical practice where the attrition rate is high. We therefore recommend that for patients who are lost to follow up, an effort should be made to obtain the principal outcome without interfering with their adherence to treatment using more accessible assessments by telephone (Greist et al., 2002) or home visits: investigators should try to obtain complete follow-up data on all subjects, irrespective of their adherence to the treatment protocol (Lavori, 1992).

Secondly, the outcomes of subjects who withdrew should be described and compared to those of completers (Dumville et al., 2006). Concerning the handling of missing values, no universally applicable method can be recommended. Nevertheless, it should be well thought out, and pre-defined in the protocol. Three general approaches to the analysis of incomplete data can be used: 1) analysis of complete cases; 2) missing data imputation (LOCF or multiple imputation); and 3) analysis of incomplete data (survival analysis, mixed model, model of

missingness). ITT analysis should, as in all RCTs, be the rule for the main analysis. Here, mixed-effect models are useful because subjects who have missing data are not completely excluded from the analyses and the missing data are not imputed. Nevertheless, it is performed under the Missing At Random hypothesis (i.e. "missingness" is explained by observed outcomes or covariates, presumably pre-dropout, but not unobserved outcomes). This type of analysis is therefore likely to favour arms with attrition. Finally, collecting data that can help predict attrition, for instance by asking participants to rate the likelihood of attending the subsequent assessment session, can change the problem of dropout from Not Missing at Random (i.e. missingness is explained by unobserved outcomes) to Missing At Random, but this should be used cautiously in the analysis of data (Leon et al., 2006).

Multiple imputation (Little and Rubin, 1987) procedures assume that data are Missing At Random: all non-missing values of outcomes at all time points and baseline demographics are used in the models, which generate imputed estimates. Generally, 5 imputation data sets are generated and estimates are combined so that standard errors reflect the variability introduced by the imputation process.

Present-day studies tend to implement these two approaches (Lynch et al., 2011) whereas other analyses such as “completer-only analysis”, LOCF and analysis under the maximum bias hypothesis are used as sensitivity analyses to assess the robustness of the analytical strategy.

2.2. Controlling for the placebo response

Placebo response improvement should be sought in effectiveness studies

RCTs versus placebo aim to reduce the placebo effect, whereas in day-to-day clinical practice, everything is done to enhance placebo effect. Thus to asses antidepressant effectiveness it is reasonable to consider certain adjustments and explanatory designs potentiating the placebo

effect in depression, allowing comparison with conditions that mimic all the theoretically important elements of placebo response associated with pharmacotherapy (e.g., expectation of improvement, doctor involvement and contact, credible treatment rationale...) (Gaudiano and Herbert, 2005).

As the risk of unblinding is substantial an assessment of the integrity of double-blind procedures should be performed routinely (Antonuccio et al., 1999; Even et al., 2000) by asking clinicians and patients to report the study condition to which they think or guess they have been assigned. Concerning clinician rating scales, keeping raters blind to the study design and hypothesis can protect against bias from their expectations.

Multi-arm studies where different doses that may or may not be effective are used alongside a similar active comparator and placebo can address this question. Nevertheless, such studies are not valid when side effects are dose-dependent. The use of an "active" placebo, with side effects mimicking those of the active drug, has been proposed. This method was developed in the early days of antidepressant research, but is rarely used in modern psychotropic studies (Perlis et al., 2010). A meta-analysis of antidepressant trials using active placebos suggested smaller effect sizes than those observed in the presumably less blinded trials using inert substances (Moncrieff et al., 2004). However, the ability of a design of this sort to prevent unblinding is not established, as the raters were able to guess better than by chance what medication the patients were taking (Uhlenhuth and Park, 1964; Weintraub and Aronson, 1963).

Thus a four-arm "balanced placebo trial design" using antidepressants, active placebo controls and intentional deception of subjects (patients are given information in a way that produces false beliefs) in a latin square design has been proposed (Kirsch and Sapirstein, 1998) (Figure 3) and this could diminish the ability of subjects to discover the study condition to which they have been assigned. Subjects are randomised in four arms: 1/ a "deception" arm where

patients receive the real drug and they are told they are receiving a placebo 2/ a "deception" arm where the patients receive the active placebo and are told they are receiving the real drug 3/ a "non-deception" arm where patients receive the real drug and they are told they are receiving the real drug 4/ a "non-deception" arm where the patients receive the active placebo and are told they are receiving the placebo. This design makes it possible to distinguish between an additive model and a non-additive model. Nevertheless, using an active placebo deliberately induces risk of adverse effects (even if they are benign or even potentially therapeutic) and this is an ethical problem (Perlis et al., 2010). The "balanced placebo trial design" has not yet been used in clinical trials on antidepressant medication, because of the ethical issues involved with temporary deception (Dowrick et al., 2007; Waring, 2008).

Alternatives designs to control for expectations

However, as in standard trials unblinding can be highly problematic, temporary deception is a key point in controlling for expectations (because accurately informing subjects could bias response to treatment). Although its mechanisms are unclear, it is undeniable that deception is a key element in placebo potency (Lakoff, 2002). Two approaches have been suggested to minimize the ethical difficulties linked to temporarily deceiving subjects (Dowrick et al., 2007): 1/ pre-consent (subjects are informed that the study involves deception, and are asked to consent to its use, without being informed of the nature of the deception) and 2/ "minimised" deception. This can take the form of a three-arm randomised controlled trial in which the effects of placebo, active medication, and usual care are examined and where there is temporary deception concerning the placebo arm (Figure 4). Patients are told that they will be randomized to receive "usual care + nothing" or "usual care + antidepressant". Pre-consent (1/) (Wendler and Miller, 2004) ("You should be aware that the investigators have intentionally left out information about certain aspects of this study") respects the subject's

autonomy but could reduce the pragmatic effectiveness of the study because participants may guess the nature of the deception. “Minimised” deception (2/) is likewise possible because the information given about risk and benefit in the “usual care + nothing” group at the time when they provide consent is correct, but this nevertheless provides a placebo group. In both cases, the subjects are informed of the nature of the deception at the end of their participation.

This design is useful to preserve the methodological benefit of randomisation and to obtain an unbiased assessment of the benefit of the antidepressant against the placebo and the benefit of the placebo against nothing. Certain criteria may justify deceiving the patient: 1/ The use of deception is necessary and no equally effective, non-deceptive approach is feasible, 2/ the use of deception is justified by the study’s social value, 3/ subjects are not deceived about aspects of the study that would affect their willingness to participate, including potential risks and benefits, 4/ subjects are informed of the nature of the deception at the end of their participation and 5/ in case of pre-consent, subjects are informed prospectively of the use of deception and consent to its use (Wendler and Miller, 2004).

Nevertheless, another objection against studies involving deception is the risk of psychological harm to research participants (Bortolotti and Mameli, 2006). A number of studies performed among healthy volunteers participating in psychology experiments have found that being deceived does not upset most subjects (Wendler and Miller, 2004) but the impact of a design that involves deceiving subjects among depressive patients is not known. It could undermine patients’ trust in physicians in general, as has been suggested in a qualitative study (Dowrick et al., 2007). Thus if a trial uses deception techniques, investigators should obtain data on the impact of the deception on mood and the therapeutic alliance.

Even if they do not provide the same information, alternative trial designs can be considered. One option is to adopt a design in which all study participants are informed that they will start with a placebo and that an active drug may be substituted after a while and that they may (or

may not) be informed when this switch is made. This protocol could provide information for three of the four arms of the balanced placebo design without any deception being required - the exception being "told drug/no drug"- (Colloca et al., 2004; Dowrick et al., 2007). Nevertheless, it is prone to "unblinding" because subjects can guess when the switch is made, even if they are not told.

Another design to preserve the benefit of randomisation could be a non-inferiority study comparing a placebo (presented as a new therapeutic alternative with fewer side effects) to an active antidepressant. This design is well justified for patients with a baseline HDRS score of 25 which was identified as the score needed to reach a clinically meaningful difference (Fournier et al., 2010). Here there is no deception because in this case, the placebo is a real therapeutic alternative. Nevertheless, an inclusion criterion of this sort limits the scope for generalising the results.

In this respect, it has been recently argued that consent forms in RCTs versus placebo should generate positive expectations regarding the possible effect of a placebo (spontaneous improvement without the use of medication) to reduce patient fears of a negative outcome following study participation (Severus et al., 2012).

Another idea could be a double-blind trial comparing an antidepressant to homeopathy. In major depressive disorder, there is not enough evidence about the efficacy of homeopathy (Pilkington et al., 2005) but it elicits expectations in patients and could be considered as a good comparator to control for expectations if we postulate that the clinical effects of homoeopathy are placebo effects (Shang et al., 2005). A comparison of this sort could be performed in a double-blind design, but to enhance the effect of expectations about the treatment, it should be performed in open label, or better, in a four-arm design using antidepressants, homeopathy, blinding and open-label, in a latin square design (Figure 5). This design can evaluate both efficacy and effectiveness of antidepressant and homeopathy

(i.e. placebo). Nevertheless, it is prone to "unblinding" and the randomisation process does not take patient preferences into account between antidepressant and alternative medicine, and it can interfere with the treatment process. As an example, one study tried to compare homeopathy to fluoxetine and placebo in primary care, but failed because of recruitment difficulties, many of them linked to patient preferences (Katz et al., 2005). Indeed, this design can only meaningfully be applied in those depressed patients who feel that either anti-depressants or homeopathic anti-depressants could potentially work for their disorder. This results in a selection bias, with a restriction of the target population, and can in fact go against the concept of effectiveness. This is also the case for sophisticated designs ensuring internal validity such as the "balanced placebo trial design". Recommendations concerning external validity are thus necessary.

2.3. Enabling extrapolation of RCT results

The external validity of antidepressant studies should be improved

Recruitment difficulties arising from patient preferences can lead to a selection bias, yielding a non-representative sample of patients, and affect external validity. At the very least, patients who have been screened, patients who are eligible and patients who refuse to participate should be identified (Moher et al., 2001; Schulz et al., 2010). An interesting alternative is to perform a randomised trial with patient preference arms (for patients who agree to randomisation, treatment is allocated by randomisation, and for patients who refuse randomisation but agree to participate, a choice of treatment is offered). Treatment and follow-up are identical in the different groups (Brewin and Bradley, 1989; Chilvers et al., 2001; Howard and Thornicroft, 2006). This has been proposed for homeopathy (Figure 6) (Katz et al., 2005). This type of design directly synchronizes a randomized controlled trial and an observational study to generate alternative evidence for assessing antidepressant drug

treatment. The double-blind design makes it possible to control for the indication bias, and the two preference arms make it possible to partly reduce the selection bias introduced by the randomization process. A simple method of analysis is the use of a model with the principal outcome as the dependent variable and treatment, design, and treatment-design interaction as explanatory variables. Nevertheless a design of this type requires an even larger number of patients than a RCT and the analysis should be interpreted with caution because of the potential influence of unmeasured confounders (Gemmell and Dunn, 2011).

As the use of restrictive eligibility criteria limits the scope for generalising RCT results, populations in the next generation of (sophisticated) RCTs should differ from the target populations of "real-life" depressive patients as little as possible. Studies among primary care patients are needed. The only inclusion criterion should be "patient needing an antidepressant for depression". The only exclusion criterion should be "contraindication of the treatment". Using current suicidal ideation as an exclusion criterion could be argued for from an ethical point of view. But depressed patients who are assigned to a placebo in antidepressant clinical trials are not at greater risk for suicide than those assigned to active treatment (Khan et al., 2000) whereas patients assigned to antidepressant treatment could well be at greater risk (Fergusson et al., 2005). Moreover, these patients are treated with antidepressant in "real life" and antidepressants are not studied in these particular patients.

To assess whether the patients included are truly representative of patients treated in a real-life setting, we suggest comparing them with registries for their principal clinical and socio-demographical characteristics.

A study of effectiveness should last at least 6 months after patient remission to obtain more information on the longitudinal effect of antidepressants. Large observational studies comparing antidepressants to usual care or to alternative medicine are needed, because they have other characteristics that make them useful sources of evidence, in that they tend to last

longer and to enrol more patients than do randomized trials (Bluhm, 2009). Statistical modelling should enable adjustment on confounding factors (Concato and Horwitz, 2004; Lawlor et al., 2004) which should be prespecified in the protocol and assessed with as little measurement error as possible to avoid misclassification bias (Mertens, 1993).

Conclusion

Methodological alternatives to the orthodox RCT should be developed to interpret results accurately and ensure internal and external validity. Some are simple and could be implemented in RCT easily. Others are sophisticated and raise ethical issues because they involve temporary deception of the patient. Nevertheless, improvements in study design for antidepressant effectiveness assessment are needed to further knowledge, to improve patient care and to determine what costs health authorities should cover. It is a challenge to develop study designs addressing the inevitable tension between internal and external validity, which can often appear as contradictory. The methodological tools presented here can be useful. The concept of antidepressant effectiveness should be developed along different axes and based on a convergence of arguments from a range of different study designs.

References

Streptomycin Treatment of Pulmonary Tuberculosis: A Medical Research Council Investigation. Br Med J (Clin Res Ed) 1948; 2: 769-82.
Antonuccio DO, Danton WG, DeNelsky GY, Greenberg RP, Gordon JS. Raising questions about antidepressants. Psychother Psychosom 1999; 68: 3-14.
Araya R. The management of depression in primary health care. Current Opinion in Psychiatry 1999; 12: 103-7.
Arroll B, Elley CR, Fishman T, Goodyear-Smith FA, Kenealy T, Blashki G, Kerse N, Macgillivray S. Antidepressants versus placebo for depression in primary care. Cochrane Database Syst Rev 2009: CD007954.
Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? Am J Psychiatry 2004; 161: 2163-77.
Bandelow B, Baldwin DS, Dolberg OT, Andersen HF, Stein DJ. What is the threshold for symptomatic response and remission for major depressive disorder, panic disorder, social anxiety disorder, and generalized anxiety disorder? J Clin Psychiatry 2006; 67: 1428-34.
Beauchamp TL, Childress JF. Principles of Biomedical Ethics: Oxford University Press, USA, 2008.
Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961; 4: 561-71.
Benedetti F, Mayberg HS, Wager TD, Stohler CS, Zubieta JK. Neurobiological mechanisms of the placebo effect. J Neurosci 2005; 25: 10390-402.
Bluhm R. Some observations on observational research. Perspect Biol Med 2009; 52: 252-63.
Bombardier C, Maetzel A. Pharmacoeconomic evaluation of new treatments: efficacy versus effectiveness studies? Ann Rheum Dis 1999; 58 Suppl 1: I82-5.
Bortolotti L, Mameli M. Deception in psychology: moral costs and benefits of unsought self-knowledge. Account Res 2006; 13: 259-75.
Brewin CR, Bradley C. Patient preferences and randomised clinical trials. Bmj 1989; 299: 313-5.
Brownell KD, Stunkard AJ. The double-blind in danger: untoward consequences of informed consent. Am J Psychiatry 1982; 139: 1487-9.
Brugha TS, Bebbington PE, MacCarthy B, Sturt E, Wykes T. Antidepressants may not assist recovery in practice: a naturalistic prospective survey. Acta Psychiatr Scand 1992; 86: 5-11.
Bystritsky A, Waikar SV. Inert placebo versus active medication. Patient blindability in clinical pharmacological trials. J Nerv Ment Dis 1994; 182: 485-7.
Chilvers C, Dewey M, Fielding K, Gretton V, Miller P, Palmer B, Weller D, Churchill R, Williams I, Bedi N, Duggan C, Lee A, Harrison G. Antidepressant drugs and generic counselling for treatment of major depression in primary care: randomised trial with patient preference arms. Bmj 2001; 322: 772-5.
Colloca L, Lopiano L, Lanotte M, Benedetti F. Overt versus covert treatment for pain, anxiety, and Parkinson's disease. Lancet Neurol 2004; 3: 679-84.
Concato J, Horwitz RI. Beyond randomised versus observational studies. Lancet 2004; 363: 1660-1.
Dowrick CF, Hughes JG, Hiscock JJ, Wigglesworth M, Walley TJ. Considering the case for an antidepressant drug trial involving temporary deception: a qualitative enquiry of potential participants. BMC Health Serv Res 2007; 7: 64.
Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomised controlled trials. Bmj 2006; 332: 969-71.

Duru G, Fantino B. The clinical relevance of changes in the Montgomery-Asberg Depression Rating Scale using the minimum clinically important difference approach. Curr Med Res Opin 2008; 24: 1329-35.
Egger M, Smith GD, Sterne JA. Uses and abuses of meta-analysis. Clin Med 2001; 1: 478-84.
Enserink M. Can the placebo be the cure? Science 1999; 284: 238-40.
Ernst E, Resch KL. Concept of true and perceived placebo effects. Bmj 1995; 311: 551-3.
Even C, Siobud-Dorocant E, Dardennes RM. Critical approach to antidepressant trials. Blindness protection is necessary, feasible and measurable. Br J Psychiatry 2000; 177: 47-51.
Eyding D, Lelgemann M, Grouven U, Harter M, Kromp M, Kaiser T, Kerekes MF, Gerken M, Wieseler B. Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. BMJ 2010; 341: c4737.
Falissard B, Lukasiewicz M, Corruble E. The MDP75: a new approach in the determination of the minimal clinically meaningful difference in a scale or a questionnaire. J Clin Epidemiol 2003; 56: 618-21.
Falissard B, Milman D, Cohen D. Testing on randomized records (TR2): a generalization of the ‘‘lady tasting tea’’ procedure to test qualitative hypotheses in psychiatric research. Submitted.
Fergusson D, Doucette S, Glass KC, Shapiro S, Healy D, Hebert P, Hutton B. Association between suicide attempts and selective serotonin reuptake inhibitors: systematic review of randomised controlled trials. Bmj 2005; 330: 396.
Finniss DG, Kaptchuk TJ, Miller F, Benedetti F. Biological, clinical, and ethical advances of placebo effects. Lancet 2010; 375: 686-95.
Fisher RA. The design of experiments. 8th edn 1971.
Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC, Fawcett J. Antidepressant drug effects and depression severity: a patient-level meta-analysis. Jama 2010; 303: 47-53.
Garcia-Toro M, Aguirre I. Biopsychosocial model in Depression revisited. Med Hypotheses 2007; 68: 683-91.
Gaudiano BA, Herbert JD. Methodological issues in clinical trials of antidepressant medications: perspectives from psychotherapy outcome research. Psychother Psychosom 2005; 74: 17-25.
Gemmell I, Dunn G. The statistical pitfalls of the partially randomized preference design in non-blinded trials of psychological interventions. Int J Methods Psychiatr Res 2011; 20: 1-9.
Gibbons RD, Hur K, Brown CH, Davis JM, Mann JJ. Benefits From Antidepressants: Synthesis of 6-Week Patient-Level Outcomes From Double-blind Placebo-Controlled Randomized Trials of Fluoxetine and Venlafaxine. Arch Gen Psychiatry 2012.
Greist JH, Mundt JC, Kobak K. Factors contributing to failed trials of new agents: can technology prevent some problems? J Clin Psychiatry 2002; 63 Suppl 2: 8-13.
Guy W. ECDEU Assessment manual for psychopharmacology. In EaW U.S. Department of Health, Public Health Service, Alcohol, Drug Abuse and Mental Health administration ed. Rockville, 1976.
Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960; 23: 56-62.
Howard L, Thornicroft G. Patient preference randomised controlled trials in mental health research. Br J Psychiatry 2006; 188: 303-4.
Hrobjartsson A, Gotzsche PC. Placebo interventions for all clinical conditions. Cochrane Database Syst Rev 2010: CD003974.
Huf W, Kalcher K, Pail G, Friedrich ME, Filzmoser P, Kasper S. Meta-analysis: fact or fiction? How to interpret meta-analyses. World J Biol Psychiatry 2011; 12: 188-200.

Ioannidis JP. Effectiveness of antidepressants: an evidence myth constructed from a thousand randomized trials? Philos Ethics Humanit Med 2008; 3: 14.
Judd LL, Akiskal HS, Maser JD, Zeller PJ, Endicott J, Coryell W, Paulus MP, Kunovac JL, Leon AC, Mueller TI, Rice JA, Keller MB. Major depressive disorder: a prospective study of residual subthreshold depressive symptoms as predictor of rapid relapse. J Affect Disord 1998; 50: 97-108.
Kadouri A, Corruble E, Falissard B. The improved Clinical Global Impression Scale (iCGI): development and validation in depression. BMC Psychiatry 2007; 7: 7.
Katz T, Fisher P, Katz A, Davidson J, Feder G. The feasibility of a randomised, placebo-controlled clinical trial of homeopathic treatment of depression in general practice. Homeopathy 2005; 94: 145-52.
Khan A, Leventhal RM, Khan SR, Brown WA. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. J Clin Psychopharmacol 2002; 22: 40-5.
Khan A, Warner HA, Brown WA. Symptom reduction and suicide risk in patients treated with placebo in antidepressant clinical trials: an analysis of the Food and Drug Administration database. Arch Gen Psychiatry 2000; 57: 311-7.
Kirsch I. Are drug and placebo effects in depression additive? Biol Psychiatry 2000; 47: 733-5.
Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. PLoS Med 2008; 5: e45.
Kirsch I, Moncrieff J. Clinical trials and the response rate illusion. Contemp Clin Trials 2007; 28: 348-51.
Kirsch I, Sapirstein G. Listening to Prozac but hearing placebo: A meta-analysis of antidepressant medication. Prevention & Treatment 1998; 1.
Krell HV, Leuchter AF, Morgan M, Cook IA, Abrams M. Subject expectations of treatment effectiveness and outcome of treatment with an experimental antidepressant. J Clin Psychiatry 2004; 65: 1174-9.
Lacasse JR, Leo J. Serotonin and depression: a disconnect between the advertisements and the scientific literature. PLoS Med 2005; 2: e392.
Lakoff A. The mousetrap: managing the placebo effect in antidepressant trials. Mol Interv 2002; 2: 72-6.
Lavori PW. Clinical trials in psychiatry: should protocol deviation censor patient data? Neuropsychopharmacology 1992; 6: 39-48; discussion 9-63.
Lawlor DA, Davey Smith G, Bruckdorfer KR, Kundu D, Ebrahim S. Observational versus randomised trial evidence. Lancet 2004; 364: 755.
Leon AC, Mallinckrodt CH, Chuang-Stein C, Archibald DG, Archer GE, Chartier K. Attrition in randomized controlled clinical trials: methodological issues in psychopharmacology. Biol Psychiatry 2006; 59: 1001-5.
Lewin S, Glenton C, Oxman AD. Use of qualitative methods alongside randomised controlled trials of complex healthcare interventions: methodological study. Bmj 2009; 339: b3496.
Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. Bmj 2003; 326: 1167-70.
Linde K, Schumann I, Meissner K, Jamil S, Kriston L, Rucker G, Antes G, Schneider A. Treatment of depressive disorders in primary care--protocol of a multiple treatment systematic review of randomized controlled trials. BMC Fam Pract 2011; 12: 127.
Little RJA, Rubin DB eds. Statistical analysis with missing data Wiley & Sons, New York, 1987.

Lynch FL, Dickerson JF, Clarke G, Vitiello B, Porta G, Wagner KD, Emslie G, Asarnow JR, Jr., Keller MB, Birmaher B, Ryan ND, Kennard B, Mayes T, DeBar L, McCracken JT, Strober M, Suddath RL, Spirito A, Onorato M, Zelazny J, Iyengar S, Brent D. Incremental cost-effectiveness of combined therapy vs medication only for youth with selective serotonin reuptake inhibitor-resistant depression: treatment of SSRI-resistant depression in adolescents trial findings. Arch Gen Psychiatry 2011; 68: 253-62.
Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. Jama 2009; 302: 977-84.
Mertens TE. Estimating the effects of misclassification. Lancet 1993; 342: 418-21.
Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001; 357: 1191-4.
Moncrieff J. The antidepressant debate. Br J Psychiatry 2002; 180: 193-4.
Moncrieff J, Wessely S, Hardy R. Active placebos versus antidepressants for depression. Cochrane Database Syst Rev 2004: CD003012.
Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979; 134: 382-9.
Mundt JC, Katzelnick DJ, Kennedy SH, Eisfeld BS, Bouffard BB, Greist JH. Validation of an IVRS version of the MADRS. J Psychiatr Res 2006; 40: 243-6.
Muthen B, Brown HC. Estimating drug effects in the presence of placebo response: causal inference using growth mixture modeling. Stat Med 2009; 28: 3363-85.
NationalInstituteforClinicalExcellence ed. The treatment and management of depression in adults (updated edition) National Clinical Practice Guideline 90: The British Psychological Society and The Royal College of Psychiatrists, 2010.
Naudet F, Maria AS, Falissard B. Antidepressant Response in Major Depressive Disorder: A Meta-Regression Comparison of Randomized Controlled Trials and Observational Studies. PLoS One 2011; 6: e20811.
Noble LM, Douglas BC, Newman SP. What do patients expect of psychiatric services? A systematic and critical review of empirical studies. Soc Sci Med 2001; 52: 985-98.
Olfson M, Marcus SC, Druss B, Elinson L, Tanielian T, Pincus HA. National trends in the outpatient treatment of depression. Jama 2002; 287: 203-9.
Papakostas GI, Fava M. Does the probability of receiving placebo influence clinical trial outcome? A meta-regression of double-blind, randomized clinical trials in MDD. Eur Neuropsychopharmacol 2009; 19: 34-40.
Perlis RH, Ostacher M, Fava M, Nierenberg AA, Sachs GS, Rosenbaum JF. Assuring That Double-Blind Is Blind. Am J Psychiatry 2010; 167: 250-2.
Petkova E, Quitkin FM, McGrath PJ, Stewart JW, Klein DF. A method to quantify rater bias in antidepressant trials. Neuropsychopharmacology 2000; 22: 559-65.
Pilkington K, Kirkwood G, Rampes H, Fisher P, Richardson J. Homeopathy for depression: a systematic review of the research evidence. Homeopathy 2005; 94: 153-63.
Posternak MA, Solomon DA, Leon AC, Mueller TI, Shea MT, Endicott J, Keller MB. The naturalistic course of unipolar major depression in the absence of somatic therapy. J Nerv Ment Dis 2006; 194: 324-9.
Posternak MA, Zimmerman M. Short-term spontaneous improvement rates in depressed outpatients. J Nerv Ment Dis 2000; 188: 799-804.
Posternak MA, Zimmerman M. Therapeutic effect of follow-up assessments on antidepressant and placebo response rates in antidepressant efficacy trials: meta-analysis. Br J Psychiatry 2007; 190: 287-92.
Posternak MA, Zimmerman M, Keitner GI, Miller IW. A reevaluation of the exclusion criteria used in antidepressant efficacy trials. Am J Psychiatry 2002; 159: 191-200.

Rabkin JG, Markowitz JS, Stewart J, McGrath P, Harrison W, Quitkin FM, Klein DF. How blind is blind? Assessment of patient and doctor medication guesses in a placebo-controlled trial of imipramine and phenelzine. Psychiatry Res 1986; 19: 75-86.
Rihmer Z, Gonda X. Is drug-placebo difference in short-term antidepressant drug trials on unipolar major depression much greater than previously believed? J Affect Disord 2008; 108: 195-8.
Ronalds C, Creed F, Stone K, Webb S, Tomenson B. Outcome of anxiety and depressive disorders in primary care. Br J Psychiatry 1997; 171: 427-33.
Rutherford B, Sneed J, Devanand D, Eisenstadt R, Roose S. Antidepressant study design affects patient expectancy: a pilot study. Psychol Med 2010; 40: 781-8.
Rutherford BR, Sneed JR, Roose SP. Does study design influence outcome?. The effects of placebo control and treatment duration in antidepressant trials. Psychother Psychosom 2009; 78: 172-81.
Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. PLoS Med 2010; 7: e1000251.
Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet 2002; 359: 614-8.
Seemuller F, Moller HJ, Obermeier M, Adli M, Bauer M, Kronmuller K, Holsboer F, Brieger P, Laux G, Bender W, Heuser I, Zeiler J, Gaebel W, Schennach-Wolff R, Henkel V, Riedel M. Do efficacy and effectiveness samples differ in antidepressant treatment outcome? An analysis of eligibility criteria in randomized controlled trials. J Clin Psychiatry 2010; 71: 1425-33.
Severus E, Seemuller F, Berger M, Dittmann S, Obermeier M, Pfennig A, Riedel M, Frangou S, Moller HJ, Bauer M. Mirroring everyday clinical practice in clinical trial design: a new concept to improve the external validity of randomized double-blind placebo-controlled trials in the pharmacological treatment of major depression. BMC Med 2012; 10: 67.
Shang A, Huwiler-Muntener K, Nartey L, Juni P, Dorig S, Sterne JA, Pewsner D, Egger M. Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy. Lancet 2005; 366: 726-32.
Sinyor M, Levitt AJ, Cheung AH, Schaffer A, Kiss A, Dowlati Y, Lanctot KL. Does inclusion of a placebo arm influence response to active antidepressant treatment in randomized controlled trials? Results from pooled and meta-analyses. J Clin Psychiatry 2010; 71: 270-9.
Sneed JR, Rutherford BR, Rindskopf D, Lane DT, Sackeim HA, Roose SP. Design makes a difference: a meta-analysis of antidepressant response rates in placebo-controlled versus comparator trials in late-life depression. Am J Geriatr Psychiatry 2008; 16: 65-73.
Sotsky SM, Glass DR, Shea MT, Pilkonis PA, Collins JF, Elkin I, Watkins JT, Imber SD, Leber WR, Moyer J, et al. Patient predictors of response to psychotherapy and pharmacotherapy: findings in the NIMH Treatment of Depression Collaborative Research Program. Am J Psychiatry 1991; 148: 997-1008.
Suh T, Gallo JJ. Symptom profiles of depression among general medical service users compared with specialty mental health service users. Psychol Med 1997; 27: 1051-63.
Tedeschini E, Fava M, Goodness TM, Papakostas GI. Relationship between probability of receiving placebo and probability of prematurely discontinuing treatment in double-blind, randomized clinical trials for MDD: a meta-analysis. Eur Neuropsychopharmacol 2010; 20: 562-7.
Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med 2008; 358: 252-60.
Uhlenhuth EH, Park LC. THE INFLUENCE OF MEDICATION (IMIPRAMINE) AND DOCTOR IN RELIEVING DEPRESSED PSYCHONEUROTIC OUTPATIENTS. J Psychiatr Res 1964; 69: 101-22.

van der Lem R, van der Wee NJ, van Veen T, Zitman FG. Efficacy versus effectiveness: a direct comparison of the outcome of treatment for mild to moderate depression in randomized controlled trials and daily practice. Psychother Psychosom 2012; 81: 226-34.
Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004; 363: 1728-31.
Walsh BT, Seidman SN, Sysko R, Gould M. Placebo response in studies of major depression: variable, substantial, and growing. Jama 2002; 287: 1840-7.
Waring DR. The antidepressant debate and the balanced placebo trial design: an ethical analysis. Int J Law Psychiatry 2008; 31: 453-62.
Weintraub W, Aronson H. CLINICAL JUDGMENT IN PSYCHOPHARMACOLOGICAL RESEARCH. J Neuropsychiatr 1963; 4: 65-70.
Wendler D, Miller FG. Deception in the pursuit of science. Arch Intern Med 2004; 164: 597-600.
Wisniewski SR, Rush AJ, Nierenberg AA, Gaynes BN, Warden D, Luther JF, McGrath PJ, Lavori PW, Thase ME, Fava M, Trivedi MH. Can phase III trial results of antidepressant medications be generalized to clinical practice? A STAR*D report. Am J Psychiatry 2009; 166: 599-607.
Zimmerman M, Chelminski I, Posternak MA. Generalizability of antidepressant efficacy trials: differences between depressed psychiatric outpatients who would or would not qualify for an efficacy trial. Am J Psychiatry 2005; 162: 1370-2.
Zimmerman M, Galione J. Psychiatrists' and nonpsychiatrist physicians' reported use of the DSM-IV criteria for major depressive disorder. J Clin Psychiatry 2010; 71: 235-8.
Zimmerman M, Mattia JI, Posternak MA. Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? Am J Psychiatry 2002; 159: 469-73.
Zimmerman M, McGlinchey JB, Posternak MA, Friedman M, Attiullah N, Boerescu D. How should remission from depression be defined? The depressed patient's perspective. Am J Psychiatry 2006; 163: 148-50.

Authors contributions

Conceived and designed the experiments: N.F. F.B.

Performed the experiments: N.F.

Analyzed the data: N.F.

Contributed reagents/materials/analysis tools: N.F. F.B.

Wrote the paper: N.F.

Revised the paper critically for important intellectual content: M.B. R.J.M. F.B.

Final approval of the version to be published: N.F. M.B. R.J.M. F.B.

Disclosures

There are no conflicts of interest regarding this paper. All authors have completed the Unified Competing Interest form at http://www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare that (1) All authors have no support at all from any company for the submitted work; (2) N.F. has relationships (board membership or Travel/accommodations expenses covered/reimbursed) with Servier, BMS, Lundbeck and Janssen who might have an interest in the work submitted in the previous 3 years ; M.B has relationship (consultancy and Travel/accommodations expenses covered/reimbursed) with Janssen, BMS, Otsuka, Lundbeck, Lilly, Servier, Astra Zeneca, Medtronics, Syneïka and has received grants for research from Medtronic, Lilly and Astra Zeneca in the previous 3 years ; R.J.M. has no relationships with any company that might have an interest in the submitted work in the previous 3 years; F.B has relationship (board membership or consultancy or payment for manuscript preparation or Travel/accommodations expenses covered/reimbursed) with Sanofi-Aventis, Servier, Pierre-Fabre, MSD, Lilly, Janssen-Cilag, Otsuka, Lundbeck, Genzime, Roche, BMS who might have an interest in the work submitted in the previous 3

years (3) N.F. R.J.M. F.B. spouses, partners, or children have no financial relationships that may be relevant to the submitted work. M.B. spouse is an employee of Janssen; none of the authors have any non-financial interests that may be relevant to the submitted work.

Acknowledgements

This paper was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM). We thank Eric Bellissant (M.D., PhD) for his very interesting comments, Claudine Naudet and Angela Swaine Verdier for revising the English.

Studies	Optimistic bias	ITTLOCF	OC	Attrition = failure	Maximal bias
Sheehan	3.35 [2.32; 4.84]	1.34 [0.96; 1.87]	1.58 [1.09; 2.28]	1.52 [0.98; 2.37]	0.56 [0.41; 0.75]
Rudolph	4.12 [2.71; 6.26]	1.66 [1.19; 2.32]	1.71 [1.16; 2.52]	1.77 [1.13; 2.78]	0.51 [0.41; 0.65]
Mendels	2.02 [1.54; 2.66]	1.27 [0.99; 1.63]	1.31 [1.03; 1.65]	1.44 [1.08; 1.91]	0.82 [0.69; 0.98]
WXL101497	1.67 [1.39; 2.00]	1.38 [1.15; 1.67]	1.34 [1.12; 1.60]	1.38 [1.14; 1.69]	1.03 [0.87; 1.21]
AK130940	1.82 [1.53; 2.16]	1.33 [1.11; 1.59]	1.31 [1.11; 1.55]	1.30 [1.06; 1.58]	0.87 [0.74. 1.02]
Total	**2.31 [1.75; 3.06]**	**1.36 [1.23; 1.51]**	**1.36 [1.23; 1.50]**	**1.39 [1.24; 1.57]**	**0.74 [0.59; 0.94]**

Table 1: Meta-analysis of response rates using a random effect model of 5 venlafaxine versus placebo studies using different hypothesis about missing data. Data were extracted from our previous meta-analysis: out of 26 randomised double-blind trials, five studies on venlafaxine versus placebo had extractable data. Meta-analyses of response rates using a random effect model were performed under different hypotheses about missing data. Four situations were considered:

- Optimistic bias analysis: non-assessed patients are recorded as in remission if they belong to the antidepressant group and as having not responded if they belong to the placebo group;
- ITTLOCF: patient status is derived from the LOCF method on continuous outcomes;
- OC: observed case analysis;
- Attrition = failure: non-assessed patient are recorded as not having responded in both groups;
- Maximum bias: non-assessed patients are recorded as in remission if they belong to the placebo group and as not having responded if they belong to the antidepressant group.

Results are presented as relative risk. Positive relative risk favours venlafaxine and negative relative risk favours placebo.

This example illustrates the uncertainty that arises from missing data when assessing antidepressant effect, which can vary from a marked superiority of antidepressants over placebo to a superiority of placebo over antidepressants, depending on the imputation method used for missing data.

Outcome measurement	
Is a clinician-version evaluations used?	
Yes	26 (100%)
No	0 (0%)
Is a self-administered questionnaire used?	
Yes	16 (62 %)
No	10 (38 %)
Is an ecological measure used?	
Yes	0 (0%)
No	26 (100%)
Attrition and its management	
Percent of patients failing to complete the study	0.14, 0.25, 0.33, 0.37, 0.50 (NA = 3)
Is last observation carried forward method used?	
Yes	25 (96 %)
No	1 (4 %)
Is a mixed model used?	
Yes	1 (4 %)
No	25 (96 %)
Is complete case analysis used?	
Yes	7 (27 %)
No	19 (73 %)
Response rate in placebo group and internal validity	
Percentages of responders	0.26, 0.34, 0.41, 0.48, 0.63 (NA = 4)
Is the "unblinding" phenomena evaluated?	
Yes	0 (0%)
No	26(100%)
External validity of RCTs in Major Depressive Disorder	
What category of patients is studied?	(NA = 2)
Inpatients	3 (11.5 %)
Outpatients	18 (69%)
Outpatients in Primary Care	3 (11.5 %)
Is a severity score used as an inclusion criterion?	
Yes	26 (100 %)
No	0 (0 %)
Is a treatment response during placebo lead-in period a non-inclusion criterion?	(NA = 1)
Yes	22 (88 %)
No	3 (12 %)
Study duration (months)	4, 6, 8,12, 13 (NA = 1)
Meta-analysis limitations	
Is there an industry sponsorship in the study?	(NA = 2)
Yes	24 (100 %)
No	0 (0%)

Table 2: Descriptive analysis of the 26 randomized controlled trial on venlafaxine or fluoxetine considered in our previous meta-analysis. Results are presented as numbers (percent) for qualitative outcomes and as minimum, first quartile, median, third quartile, maximum for quantitative outcomes.

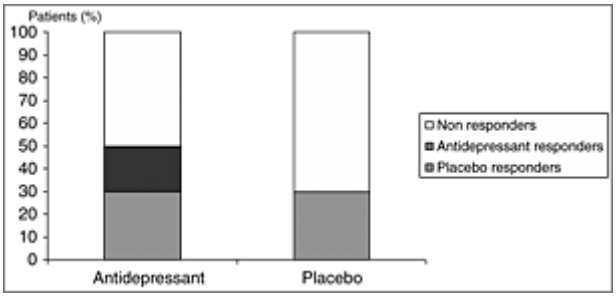

Figure 1. All placebo responders are antidepressant responders.

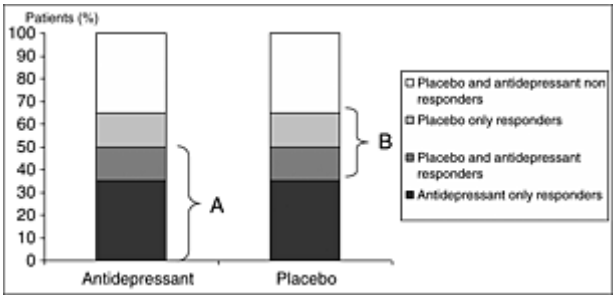

Figure 2. Placebo responders and antidepressant responders overlap each other. A, responders in treatment group; B, responders in placebo group.

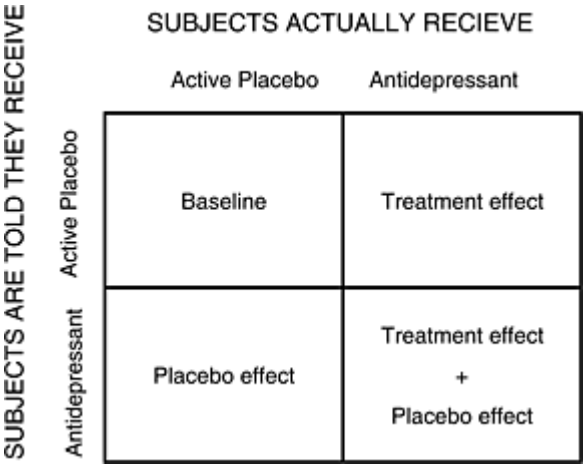

Figure 3. Balanced-placebo design. Four groups are formed following the combination of what the patients are told and what treatment they get.

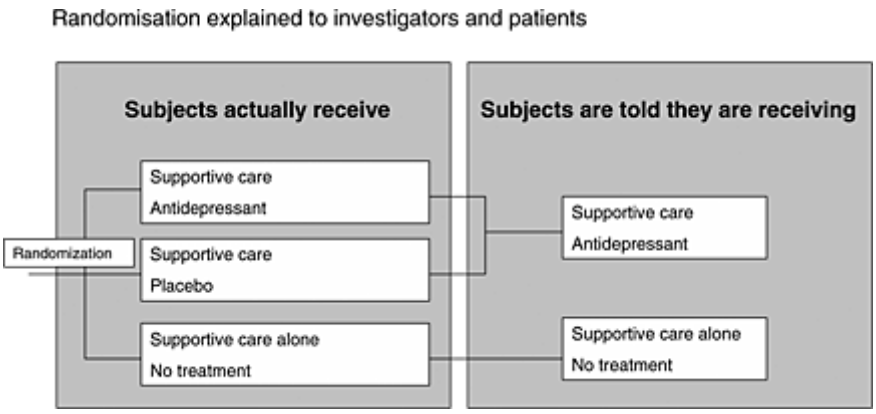

Figure 4. Three-arm RCT in which the patient is deceived.

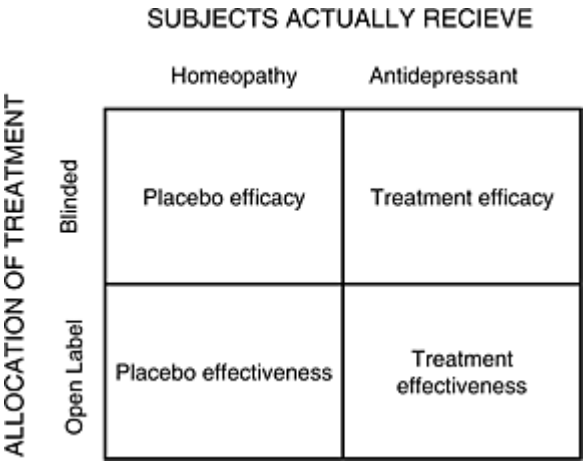

Figure 5. Four arm design. Four groups are formed according to the treatment they receive and their allocation. The effect of homeopathy is assumed to be a placebo effect.

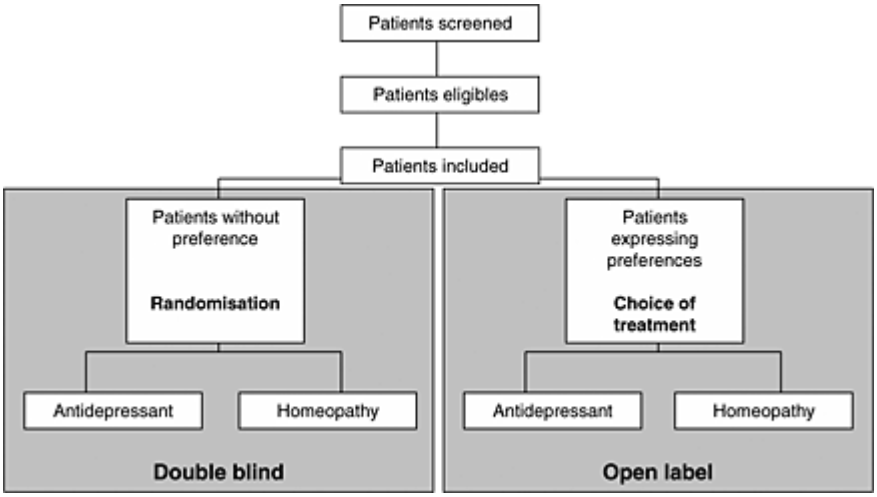

Figure 6. RCT, with patient preference arms. For patients who agreed to randomization, treatment is allocated with a randomization strategy. For patients who refused randomization but agreed to participate in the trial treatment is given according to their choice.