4

Flaws in Antidepressant Research

Before I scrutinise the design, conduct, and reporting of antidepressant trials it is worthwhile to briefly outline under which medico-scientific framework healthcare services are assessed and provided nowadays. Contemporary healthcare is devoted to evidence-based medicine, a new approach to clinical decision making that developed in the early 1990s [641]. According to the founders of this paradigm, "Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research" [642]. To determine what the best clinical evidence is, the new approach established a hierarchy of scientific evidence. High-quality evidence is provided by double-blind randomised controlled trials, low-quality by observational studies (i.e. case-control and cohort studies), and very low-quality by any other evidence (i.e. case reports, animal research, in-vitro research, and expert opinion). However, the quality of evidence from observational studies can be upgraded when effect estimates are very large, when a dose–response relationship can be demonstrated, or when all relevant confounders (e.g. treatment selection) can be excluded. By

M. P. Hengartner, *Evidence-biased Antidepressant Prescription*,
https://doi.org/10.1007/978-3-030-82587-4_4

contrast, the quality of evidence from both randomised controlled trials and observational studies needs to be downgraded when the studies have serious limitations (e.g. unblinding of participants and/or clinical investigators), when the effect estimates are inconsistent across studies, when the effect estimate is indirect (e.g. due to unrepresentative samples or surrogate outcomes), when there is imprecision in effect estimates, or when reporting or publication bias is likely [641, 643]. By this means a systematic assessment of observational studies can obtain a high evidence grade, whereas a synthesis of randomised controlled trials can yield a low or very low evidence grade.

A large and well-controlled double-blind randomised clinical trial certainly provides the strongest evidence to evaluate efficacy and safety/tolerability of medical interventions, that is, the balance between benefits and harms in a large group of patients. Randomised means that patients are randomly assigned to treatment conditions (e.g. new drug vs. active comparator or placebo) to avoid that treatments are differently assigned to patients based on specific characteristics, which could produce incomparable treatment groups and thus biased efficacy estimates (e.g. when those with less severe illness are assigned to the new drug and those with more severe illness to the comparator drug). Double-blind means that neither the patients nor the clinical investigators ought to know which treatment a patient receives, as this could also bias the outcome due to treatment expectations. When a particular drug has been tested in several trials, as is mostly the case, a systematic review and meta-analysis of all available studies then provides the best evidence of efficacy and safety/tolerability, for results of individual studies could differ due to sampling variability. Depression, for example, is a very heterogeneous condition, and depending on treatment setting (e.g. urban psychiatric hospital vs. rural primary care practice), samples of patients with depression may differ substantially from study to study with respect to age, sex, ethnicity, socio-economic background, illness severity, symptomatology, medical comorbidity, and functional impairment. Differences in sample composition could thus affect treatment effects of antidepressants (i.e. efficacy and safety estimates).

In sum, evidence-based medicine attempts to provide the best healthcare according to the most reliable scientific evidence. To that end, it

incorporates not only the quality of individual studies but also considers the strength of evidence based on assessments of all available studies to determine whether a specific treatment is safe and effective in a specific patient population. These systematic reviews of clinical trials are the foundation of treatment guidelines and inform clinical decision making in modern healthcare. Put differently, evidence-based medicine replaced so-called eminence-based medicine, that is, treatment decisions based on unsystematic, uncontrolled observations and physiological reasoning. According to Drs Djulbegovic and Guyatt, "The basis for the first EBM [evidence-based medicine] epistemological principle is that not all evidence is created equal, and that the practice of medicine should be based on the best available evidence. The second principle endorses the philosophical view that the pursuit of truth is best accomplished by evaluating the totality of the evidence, and not selecting evidence that favours a particular claim" [641]. Arguably, the most important and best resource for the practice of evidence-based medicine is Cochrane, an international non-profit organisation that produces systematic reviews and meta-analyses of medical interventions to inform clinical decision making. According to the Cochrane website, "Our vision is a world of improved health where decisions about health and health care are informed by high-quality, relevant and up-to-date synthesized research evidence. Our mission is to promote evidence-informed health decision-making by producing high-quality, relevant, accessible systematic reviews and other synthesized research evidence" [644].

The Corruption of Evidence-Based Medicine

It comes without saying that evidence-based medicine was a great achievement that improved healthcare in various medical conditions and therapeutic domains. However, in some sense, the movement became a victim of its own success. Various doctors seem to ignore (or simply are unaware) that medical research is fallible and subject to falsification and correction. Research findings in support of a medical intervention remain valid temporarily and never provide a clear confirmation that will necessarily stand the test of time. Medicine is a probabilistic, not an exact science. We

cannot be certain about best medical practice, and quite often the scientific evidence allows for nothing more than a wild guess, that is, a treatment decision with huge uncertainty [383]. The history of medicine is replete with examples of new (breakthrough) interventions that quickly became established best medical practice (or standard of care) and later turned out to be in error [645]. Again and again medicine had to fundamentally change its "best" practice due to new results from methodologically superior clinical trials showing that the standard of care was ineffective or that its harms exceeded its benefits, a phenomenon now commonly termed "medical reversal" [646, 647]. Contending that a medical intervention clearly works for most patients because it became standard medical practice (an argument often made to defend the widespread prescription of psychiatric drugs) is thus utterly naïve and misinformed. Best medical practice may eventually turn out to be bad medical practice. I give an example.

During the 1980s, it was accepted best medical practice to treat patients who had suffered myocardial infarction with antiarrhythmic drugs in the conviction that this intervention would reduce mortality. But in 1989, the first placebo-controlled trial that examined mortality as treatment outcome showed that antiarrhythmic drugs accounted for an excess of deaths from major adverse cardiac events and also for higher all-cause mortality [648]. That is, antiarrhythmic drugs were not beneficial in this patient population, they were harmful. Instead of reducing mortality, they actually increased mortality! According to Dr. Jeremy Howick, an expert in evidence-based medicine from the University of Oxford, "Given the widespread use of the drugs, it has been estimated that tens of thousands of people were killed by the drugs each year" [649].

But how come antiarrhythmic drugs were even approved for the treatment of patients with myocardial infarction? Put simply, the pharmaceutical companies seeking marketing approval for these drugs made sure that they did not need to establish effectiveness based on mortality (even though, ironically, the aim of antiarrhythmic drugs was precisely to reduce mortality). Instead, the companies managed to claim effectiveness based on a surrogate outcome, namely ventricular extra beats, a common type of cardiac arrhythmia [649]. This example illustrates nicely how surrogate outcomes can mislead and give a false impression of treatment

benefits when in fact a drug is harmful. It also demonstrates that drug regulators may erroneously approve drugs as safe and effective when they don't require the pharmaceutical industry to study the right treatment outcomes.

So there can be no doubt that medical research is fallible. And because medical research is fallible and evidence-based medicine became so influential, it was soon corrupted by the biomedical industry [29, 376, 377, 650]. The pharmaceutical and medical device industry had quickly realised that by managing (or dominating) the scientific literature with study results that support their commercial interests, they could exert a significant influence over healthcare policy and clinical practice [649, 651, 652]. This led to a myriad of serious flaws that threaten (or undermine) the validity of evidence-based medicine. More to the point, given the inherent hierarchy of scientific evidence, with the synthesis of clinical trials on top, the industry was able to co-opt evidence-based medicine and turn this movement into an efficient marketing tool to boost its profits. The companies did so by sponsoring thousands of clinical trials, often systematically biased and selectively reported, that were then eligible for the systematic reviews and meta-analyses that guide clinical practice [377, 653]. How is this possible? Let me briefly explain.

Given that original research is expensive, most biomedical research, especially drug trials, are sponsored by the pharmaceutical industry [654–656]. The typical drug trial is thus sponsored by the pharmaceutical company that seeks regulatory marketing approval for a treatment indication (premarketing/preapproval trials) or increased market share in approved indications (postmarketing/postapproval trials). These trials are designed by the sponsoring company with input from industry advisors and conducted by contract-research organisations. These firms organise a network of study sites, which may number in dozens across the globe, designate the clinical investigators and implement the trial protocol at those sites. During the trial, the contract-research organisations monitor the study sites and send report forms to the sponsoring company. Once the trial is completed, the sponsoring company conducts the data analysis and evaluates and interprets the results. If the results are not too unfavourable to the sponsor's drug, the company will publish the results (or selected parts thereof). For it the company hires a medical

communication firm that produces several manuscript drafts based on instructions from the marketing department of the sponsoring company. When the company is satisfied with the manuscript, the marketing department selects key opinion leaders, often senior researchers from prestigious academic departments who serve on the company's advisory board and/or speaker's bureau, to be listed as "authors" on the publication. At this final stage, the academic researchers comment on the manuscript, make some edits, and lend the study the badge of scientific excellence and academic independence. However, the sponsoring company almost always has the final say on the manuscript to be submitted for publication and it owns the data. That is, the trial data are property of the sponsor and the eminent researchers from the prestigious universities listed as "authors" on the publication hardly ever have full access to the raw data. They only know the data and results the company was willing to show them, and often they only give intellectual inputs but don't write a single sentence, let alone an entire paragraph of the manuscript. Most articles are thus not written by the academic "authors" listed on the publication; they are largely ghostwritten, that is, drafted by industry employees and medical communication firms that are not declared as authors on the paper's by-line [29, 428, 459, 657, 658].

As detailed by Matheson, "Through a patchwork of diminutions, aggrandizements, omissions, euphemisms, fudges, and misnomers, academics are positioned as masters, and proprietors as their worthy aides. The company is placed in the shop window—but nobody is told it owns the shop … The language of corporate 'sponsorship' and academic 'investigators' and superficial arrangements of trial committees suggest that companies merely provide finance and that independent academic institutions are in true command, while the actual role of commerce in instigation, analysis, framing, writing, and data ownership is politely shepherded into the margins by diverse attributional tricks—and that is how medicine likes it." [659]. So the pharmaceutical company gets its commercially tailored research article, senior academics and medical institutions get reputation and credits, and the journals get impact points and revenues. According to Matheson, "each party benefits in its own way. Companies get the elixir of endorsement on which advocacy marketing depends; academics reap the rewards of authorial status and

generally feel that they deserve top billing; journals sell reprints; and culturally, I believe, academic medicine and its journals crave the sense that the research scene remains in their hands" [659]. Put differently, evidence-based medicine has been corrupted by the pharmaceutical industry, and academic medicine eagerly cooperated to advance its own agenda [377]. The interests least served by this commercial research enterprise are often those of both patients and the public [29, 459, 650, 660].

There is of course an inherent financial conflict of interest in industry-sponsored drug trials, for being critical towards the efficacy of its own drug and fully transparent (or honest) about adverse effects and safety issues undermines the company's commercial interest. As a result, industry-sponsored trials often have systematic methodological biases so that the sponsor's drug appears more effective, better tolerated and safer than it really is [29, 62, 661]. Another pervasive bias is the selective reporting of treatment outcomes [61, 662]. Trials with unfavourable results are either not published or the prespecified primary outcomes are not fully reported in the published article when they are negative [89, 663, 664]. Of course, selective reporting affects not only efficacy outcomes but also safety and tolerability data [87, 664, 665]. Likewise, statistical analyses often deviate from the model prespecified in the study protocol (the analysis intended before the data was inspected), for instance, by using a different statistical model or by focusing on a different analysis population (i.e. including only a subset of participants in the analysis) [666]. A last form of scientific misconduct briefly mentioned here is spin, that is, the deliberate misrepresentation and misinterpretation of negative trial results [667]. A typical example of spin is reporting and interpreting a trial with non-significant primary outcome as if the intervention was unequivocally effective [668–670].

Together these issues systematically bias the benefit–harm ratio of medical interventions reported in the scientific literature, resulting in the overestimation of efficacy and underestimation of harms [29, 85, 89, 428, 459]. Perhaps you'll doubt that the situation is that bad and you are right to challenge these conclusions, given that they call into question the whole of modern medicine that we need (and want) to rely on when we seek healthcare. Perhaps you assume that most medical interventions are supported by strong evidence and thus argue that twisting and bending

of the scientific evidence is only an issue for a small minority of interventions. If you think so, then you're mistaken. As a matter of fact, the scientific evidence supporting the effectiveness of contemporary medical treatments is generally poor. A substantial portion (presumably the majority) of the scientific literature on medical interventions is inconclusive and unreliable. But don't just take my word for it. Instead, let's have a quick look at two pertinent studies.

First, a recent systematic review showed that only 4% of contemporary medical interventions were supported by high-quality evidence. The quality of evidence was low or insufficient in 74% of surgical interventions, 82% of pharmacological interventions, and 86% of psychosocial interventions [671]. Second, according to a recent analysis of Cochrane reviews of medical interventions (mostly drug treatments), only 10% provided high-quality evidence for the effectiveness of treatments; 37% provided moderate-quality evidence, 31% low-quality evidence, and an alarming 22% very low-quality evidence [672]. I reiterate: about half of medical interventions (53%) are "supported" by low or very low quality evidence. That's not reassuring…

Another common view is that with the accumulating number of trials for a specific treatment, the scientific evidence on its benefits and harms will improve. However, this is not true either. On average, Cochrane reviews updated with new trial results did not provide improved quality of evidence. By tendency, it was rather the other way round. After inclusion of new trial results, the quality of evidence was downgraded in 58% of reviews and upgraded in 42% of reviews, but of the latter only a very small minority achieved a high-quality rating [672]. The evidence base for the effectiveness of antidepressants in depression is no exception to the rule. According to the most recent systematic reviews, the quality of evidence is in general low to very low, and this applies to both adult clinical trials [13, 141] and paediatric clinical trials [292, 294].

What does the low quality and unreliability (i.e. poor credibility) of the scientific evidence for medical interventions imply? A brilliant study by Heres and colleagues impressively demonstrated what the consequences are. They examined the comparative efficacy of popular antipsychotic drugs in the treatment of schizophrenia or schizoaffective disorder [673]. For the sake of simplicity, let's call them drug A, B, and C. Logic

dictates that if in head-to-head trials A beats B, and B beats C, then A must also beat C. By consequence, if the evidence is reliable, then A would be the most effective drug, B the second best, and C the least effective. However, the "reality" looks different. The scientific literature shows that when the manufacturer of A sponsors the trial, then A beats both B and C. If, however, the manufacturer of B sponsors the trial, then B beats both A and C, and, you certainly sense what's coming, if the manufacturer of C sponsors the trial, then C beats both A and B. So ultimately the scientific evidence provides no clue as to which drug is best in the treatment of schizophrenia or schizoaffective disorder. The confusing and conflicting evidence on the comparative efficacy of antipsychotics is thus basically meaningless.

Based on this study it is obvious that pharmaceutical companies can quite easily get the results they want (i.e. the results that present their own product in the most favourable light relative to competitors). And given that there is little reason to assume that the trial results from one company are more (or less) credible than the findings from the other companies, it follows that the industry-sponsored studies on the comparative efficacy of antipsychotics, and by extension all other drugs, are neither trustworthy nor reliable. I deliberately wrote "by extension", because the disturbing findings from Heres and colleagues [673] were later replicated in a much larger study examining head-to-head trials in general medicine [674]. The authors of the latter study concluded from the data that "The literature of head-to-head RCTs [randomised controlled trials] is dominated by the industry. Industry-sponsored comparative assessments systematically yield favorable results for the sponsors, even more so when noninferiority designs are involved" [674]. These studies thus clearly indicate that pharmaceutical companies have the capabilities (or possibilities) to create the "scientific evidence" that best suits their commercial interests. In the following sections I will detail how sponsors get (or at least try to get) the results they want. To that end, I will outline methodological biases in clinical trials and then provide an account of reporting biases.

Methodological Biases

Why is the quality (and credibility) of evidence for the effectiveness of most medical interventions so dismayingly poor? In my view the two main reasons are the serious methodological limitations of most clinical trials and the lenient criteria for drug approval adopted by regulatory agencies. Scientists at the European Medicines Agency (EMA) evaluated 111 successive applications submitted from September 1997 to May 2001 to their agency [169]. In 49% of applications, the EMA objected the quality of long-term safety data, in 42% they noted a lack of adequate randomised controlled trials, in 38% they objected the robustness of methodology, in 33% they criticised the selected patient population, in 29% the choice of outcomes, in 18% the insufficient long-term follow-up data, in 17% the inadequate duration of treatment, and so on. However, the only major methodological limitation that was independently related to the agency's decision to approve or reject an application was the lack of adequate randomised controlled trials [169]. The other limitations did not seem to influence their decision to approve or reject a new drug application, including quality of long-term safety data, robustness of methodology, the selected patient population, choice of outcomes, and duration of treatment. Given that this analysis dates a few years back, perhaps standards have improved? Unfortunately, this is not the case.

The results of a recent analysis indicate that, overall, the methodological quality of clinical trials conducted for regulatory approval of new drug applications has arguably even decreased in more recent years. Zhang and colleagues examined the methodological characteristics of pivotal trials supporting new treatments approved by the FDA [675]. While in 1995–1997 altogether 94% and 79% of trials were randomised and double-blind, respectively, in 2015–2017 these rates dropped to 82% and 68%. Moreover, in 1995–1997 altogether 44% of pivotal trials had a clinical outcome (e.g. cardio-vascular events), but in 2015–2017 only 23% had so, while the rate of the less stringent surrogate outcomes (e.g. cholesterol levels) increased from 48% in 1995–1997 to 59% in 2015–2017. Likewise, the rate of active comparators decreased from 44% in 1995–1997 to 29% in 2015–2017, while the rate of

uncontrolled trials (i.e. neither active nor placebo comparator) increased from 9% in 1995–1997 to 18% in 2015–2017. The only positive development was an increase in both median sample size (277 patients in 1995–1997 vs. 467 in 2015–2017) and median trial duration (11 weeks in 1995–1997 vs. 24 weeks in 2015–2017). As recently summarised by Drs Kesselheim and Avorn, both highly respected professors of medicine at Harvard Medical School,

> "In recent years, under steady pressure from the pharmaceutical industry and the patient groups it funds, the FDA has progressively lowered its standards of effectiveness and safety required for drug approvals. New drugs are now more likely to be supported by fewer studies and less adequate clinical trial designs than in the past. Worse, more than half of new drugs are now approved based on what's called surrogate endpoints—changes in the body measured by lab tests that may not reflect clinical benefit—rather than requiring that the drug affect how a person feels, functions or survives". [676]

You may rightly object that the main issue detailed above is the use of surrogate outcomes and uncontrolled trial designs. Thus, all should be fine if researchers and regulatory agencies would adhere to clinical outcomes assessed in randomised controlled trials, shouldn't it? Unfortunately, this is not true. The double-blind randomised controlled trial is widely considered as the gold standard to determine efficacy as well as tolerability and safety of medical interventions, for it has good internal validity (refers to the degree of confidence that the causal relationship being tested is trustworthy and not influenced by other factors). However, due to narrowly defined (unrepresentative) patient populations, extensive monitoring and short treatment duration, all of which considerably deviate from routine practice, the external validity of most clinical trials is poor (refers to the extent to which results from a study can be generalised to other situations or patient populations). Many double-blind randomised controlled trials also have other serious methodological limitations, implying that their results are systematically biased and no meaningful conclusions can be drawn from the data, even when objective clinical outcomes are

assessed [29, 62, 84, 661, 677, 678]. A list of common methodological limitations is provided in Table 4.1.

Most clinical trials are of very short duration and sample size is modest [165, 675], making it impossible to determine sustained treatment benefits and to detect rare adverse drug reactions [679]. Small sample size also implies low statistical power, which reduces the chance to find a true treatment effect but also produces both inflated treatment effects and false-positive results [680, 681]. Most trials have extensive exclusion criteria and preselect those patients assumed to respond best to the medication, especially younger male patients without comorbid (concomitant) medical conditions [682]. Many placebo-controlled trials are inadequately blinded (or blinding is not ascertained), meaning that investigators and patients may correctly identify whether they receive the active drug or inert placebo [683]. This is an important issue, for unblinding is associated with stronger effects on subjective outcomes like quality of life or mental health ratings [684, 685]. When a new drug is compared to another active drug, often an inferior comparator drug is chosen, the

Table 4.1 Common methodological limitations in double-blind randomised controlled trials

Problem	Examples
Inadequate samples	Unrepresentative patient population due to restrictive selection criteria; sample size too small
Inadequate trial duration	Only acute treatment trials; no long-term trials and post-treatment follow-up
Poor comparators	Only placebo control; inferior active comparator; too high or too low dosed comparator drug
Inadequate randomisation	Inadequate generation of randomised sequence; Treatment allocation not concealed
Unblinding of both participants and investigators	Unblinding due to lack of side effects in placebo group; unblinding due to drug-specific side effects in active controlled trials
Poor outcomes	No clinical outcomes; only subjective outcome measures; only surrogate outcomes
Inadequate harm assessment	No reporting of severe adverse events and discontinuation due to adverse events; only reporting of common adverse events; unsystematic and unstructured harm assessment; inconsistent coding of adverse events; no grouping of adverse events

dose of the comparator drug is too high (so that the sponsor's drug appears safer and better tolerated) or too low (so that the sponsor's drug appears more effective) [62, 84].

Contrary to efficacy outcomes, adverse events are typically assessed in an unsystematic and unstructured way by simply asking patients whether they experienced any unwelcome medical events since the last study visit [686]. In trial publications, often only the most common adverse events are reported, while no information on severe adverse events and discontinuation due to adverse events is given at all [687]. A specific adverse event is sometimes coded with different (and inadequate) terms which leads to misrepresentation and underestimation of its true prevalence rate [688]. Adverse events are rarely grouped by anatomic or physiological system, which further limits the detection of harm signals and significant adverse drug effects [689]. Due to unsystematic assessment, inadequate recording and poor reporting, common adverse drug effects can be systematically underestimated and, occasionally, missed altogether [29, 87, 687]. Finally, the identification of rare but serious adverse drug reactions is almost impossible in clinical trials, even when they have more than 1000 participants [690], a sample size unusually large in general medicine and especially in psychiatry [165, 675]. For these various reasons, Healy and Mangin also referred to clinical trials as "the gold standard way to miss adverse events" [691].

In sum, trial protocols, especially industry-sponsored trials, are typically designed in a way that they produce the best possible outcome for the sponsor's drug. These strategies compromise not only the internal validity of a trial, but also (and perhaps in particular) its external validity. The biases that they create are often systematic, that is, in favour of the sponsor's drug, resulting in overestimation of benefits and underestimation of harms. A main consequence of these various limitations is that it can be difficult (some might say impossible) to determine whether a drug shown to be safe and effective in a clinical trial also works in real-world routine practice. That is, generalisations of clinical trial results outside the narrowly defined study population may be invalid and the sustainability of treatment effects beyond the acute treatment phase is often uncertain [679, 692]. It follows that various drugs approved by drug regulators as safe and effective were in fact neither safe nor truly effective outside the

restricted and tightly controlled experimental setting [171, 173, 376, 458]. Therefore, when a new drug is introduced into the market, "the amount of information on benefits and risks, especially long term, is relatively small, and often based on highly selected populations with respect to age, comorbidities, use of concomitant medications, and other factors" [690].

I will now revisit these issues in more detail as they pertain specifically to antidepressant trials. I will guide you through these trial characteristics step by step. I'll start with limitations relevant to efficacy estimates and then turn to limitations relevant to safety/tolerability estimates.

Methodological Biases Distorting Efficacy Estimates

Antidepressant trials have myriads of (serious) methodological limitations [14, 102, 144, 693–695]. The first crucial aspect is the size and composition of the study sample. The average sample size in antidepressant trials is just about 224 participants [141], which is small but sufficiently large to reliably detect a minimally important treatment effect. But are the effects measured in these samples generalisable? So let's look at the selection of trial participants. Ideally, the study sample is representative of the broader patient population that will use the investigated drug in clinical practice, for it makes little sense to demonstrate efficacy and safety in a narrowly defined study population that is very untypical of the average patient being prescribed the drug in real-world routine practice. Unfortunately, this is exactly the case in antidepressant trials.

Trial participants are carefully preselected by applying very restrictive selection criteria. Most trials enrol psychiatric outpatients or people from the community recruited through advertisements, but neither psychiatric inpatients (those with mostly severe clinical depression) nor primary care patients (the largest group of antidepressant users). In addition, antidepressant trials commonly exclude participants with depression severity below or above a certain cutoff, participants with bipolar and psychotic features, participants with substance abuse or dependence, participants with acute suicidal ideation, as well as participants with comorbid (concomitant) mental disorders and general medical conditions [144, 188].

As you may easily recognise, these stringent selection criteria result in very narrowly defined and unrepresentative patient populations. Several studies have consistently shown that between 78% and 88% of patients who seek treatment in primary care and psychiatric outpatient clinics would be excluded from antidepressant trials due to these restrictive selection criteria [696–698]. Given that patients treated in psychiatric hospitals (inpatient clinics) very frequently have comorbid mental and general medical conditions and often are acutely suicidal, almost all psychiatric inpatients would arguably be excluded from a typical placebo-controlled antidepressant efficacy trial.

Another alarming finding is that selection criteria in antidepressant trials have become yet more restrictive over time, thus trial participants are even less representative in more recent studies [188]. While on average 84% of treatment-seeking patients would be excluded from antidepressant trials published between 1995 and 2009, this rate grew to 91% based on selection criteria applied in trials published from 2010 to 2014 [697]. Finally, it appears that these restrictive inclusion and exclusion criteria introduce a systematic bias. According to results of the STAR*D study, patients typically excluded from efficacy trials have a poorer treatment outcome than the unrepresentative participants preferably selected into these studies (response rates were 39% vs. 52% and remission rates 25% vs. 34%) [696]. In another analysis it was shown that the large group of patients with depression typically excluded from antidepressant trials due to restrictive selection criteria are more chronically ill [699], a patient group often unresponsive to antidepressants and thus commonly referred to as "treatment resistant" [227, 700].

A very common, almost universal, design feature in antidepressant trials is the so-called placebo run-in phase (also referred to as placebo washout) [166, 194]. The placebo run-in phase puts all participants on placebo before randomisation and typically lasts about a week. It serves two main purposes. First, many participants enrolled in antidepressant trials are already on an antidepressant and thus need to be withdrawn from this drug before they can be randomised to either the investigational drug, an active comparator, or placebo. Second, participants who improve significantly in the placebo run-in phase are typically excluded from the trial. By consequence, placebo run-in (washout) phases likely induce a

systematic bias in favour of the drug. In an older meta-analysis, the effect size for active drug against placebo was 0.50 in trials with placebo run-in and 0.41 in trials without placebo run-in, but this difference was statistically not significant [701]. However, this study was based on a small set of studies, thus lacking statistical power. In addition, as can be seen from the surprisingly high effect sizes (0.50 and 0.41, respectively), the dataset was unrepresentative, for the average treatment effect size in antidepressant trials is considerably lower (about 0.3) [17, 57, 141]. In a subsequent analysis based on a much larger and representative dataset, the effect sizes in trials with and without placebo run-in were 0.31 and 0.22, respectively, and this difference was statistically significant [13]. It is thus reasonable to conclude that placebo run-in (washout) results in inflated efficacy estimates.

Another common but problematic design feature in antidepressant trials is the permission of rescue medication, that is, sedative-hypnotic drugs such as benzodiazepines. Between 30% and 40% of antidepressant trials, including the influential STAR*D study, allowed the comedication with sedative-hypnotic drugs [141, 191]. However, these figures are most likely grave underestimates of the true rate, for use of comedication is often not reported in trial publications. According to Walsh and colleagues, only 60% of antidepressant trial reports stated explicitly whether comedication was permitted or not, and in these trials the rate of comedication was 84% [166]. Likewise, Dr Healy noted that comedication with sedative-hypnotic drugs (typically benzodiazepines) was a standard design feature in SSRI premarketing trials [9]. This certainly confounds the effects of the investigational drug. But then, why would antidepressant trial protocols permit the use of other psychotropic drugs when their main objective is to evaluate the efficacy of a specific psychotropic drug? The answer is simple and straightforward. Many antidepressants, especially the activating agents, frequently cause insomnia, nervousness, and agitation, which can be alleviated with sedative-hypnotics.

So comedication is permitted in many, likely even most, antidepressant trials. The fundamental question now is how many participants in a trial eventually received this rescue medication. If the rate is high, then the issue is serious, given that a doctor's decision to additionally prescribe a sedative-hypnotic drug is certainly non-random. The few data available

indeed indicate that the majority of participants randomised to activating antidepressants are co-medicated with sedative-hypnotic drugs, whereas participants randomised to sedating antidepressants less often receive comedication [241]. This comes as no surprise, for the whole idea of permitting comedication with sedative-hypnotics was to mitigate the common side effects of activating antidepressants [9]. This design feature thus clearly compromises the internal validity of many antidepressant trials, for sedative-hypnotics not only alleviate antidepressant side effects, thus inflating tolerability/safety estimates, in fact they also treat depression, for anxiety, insomnia, and agitation are also common depression symptoms [1, 702]. That is, the treatment effects of antidepressants and sedative-hypnotics are necessarily confounded, but it is not clear whether this bias is systematic, since patients in the placebo group may also benefit from comedication.

It is well established in general medicine that unblinding of investigators or outcome assessors, also referred to as observer bias, produces exaggerated efficacy estimates in subjective outcome measures [685, 703–706]. Given that ratings of depression severity, and by consequence their transformation into response and remission rates, are inherently subjective (i.e. not based on objective clinical tests), unblinding is a serious issue in antidepressant trials [707]. This is particularly true since antidepressants, to varying degree, can cause marked side effects that are detectable by the clinical investigators who make the outcome assessments. This unblinding issue is well known for decades (but still largely ignored) and calls into question the integrity of the double-blind procedure in antidepressant trials.

Back in 1967, Dr Leyburn wrote in the *Lancet* "Patients who come into the consulting-room for assessment, perhaps for the sixth time and rather bored with the whole thing, but with their mouths so dry that one can hear their tongues scraping and clicking about in their mouths, are likely to be taking, say, amitriptyline, rather than the placebo" [400]. In 1993, Fisher and Greenberg likewise wrote that the double-blind procedure is deficient in placebo-controlled antidepressant trials [708], a conclusion also drawn by Even and colleagues in 2000 [707]. According to the latter authors, "This raises troublesome questions. For example, have all antidepressants consistently demonstrated their efficacy? Would the

defects in design of therapeutic trials have smoothed out differences in strength of the available antidepressants? Might truly blind trials enable us to discriminate between efficacious and inefficacious antidepressants?" [707].

Because there are only very few truly double-blind antidepressant trials, and because unblinding is rarely ascertained in psychiatric drug trials [683, 684], we won't be able to answer these fundamental questions. But we know that unblinding is mostly due to the detection of side effects and the drugs' psychotropic effects, especially sedation and activation [400, 707, 708]. We can further assume that, due to treatment expectations, unblinding will result in more favourable outcome ratings in active treatment groups. To establish an association between unblinding and inflated efficacy estimates, we need to answer the following questions. How often is the blind broken in antidepressant trials? And how strongly are efficacy estimates affected by unblinding?

A few studies have examined how reliable clinical investigators can identify treatment allocation in trials of older antidepressants (tricyclics and MAOIs) for various indications and found that investigators (outcome assessors) were able to correctly guess the active drugs in about 80–90% of cases, and patients in roughly 70–80% of cases [707, 708]. Even less studies assessed the integrity of the double-blind in trials of new-generation antidepressants. A rare exception is the Depression Hypericum Trial, a 8-week three-arm trial that compared the efficacy of hypericum perforatum (St John's Wort) and sertraline against placebo [151]. If patients and clinicians were effectively blinded, stochastics (probability theory) dictates that, by chance, rates of correct guesses should be 33% in each group. However, at the end of 8 weeks, the proportion of patients guessing their treatment correctly was 55% for sertraline, 29% for hypericum, and 31% for placebo, a difference that was statistically significant. The probability of clinicians correctly guessing treatment allocation was 66% for sertraline, 29% for hypericum, and 36% for placebo, again a statistically significant difference. The findings from the Depression Hypericum Trial thus demonstrate that many clinicians, and to a lesser extent also patients, were able to correctly guess sertraline treatment, but not hypericum and placebo treatment.

According to Baethge and colleagues, only 1.8% of antidepressant trials provide an assessment of blinding [684]. Pooled across trials in schizophrenia and affective disorders, 58% and 70% of patients and investigators, respectively, correctly guessed active treatment. Finally, in a recent trial of sertraline against placebo in primary care (PANDA study), 46% of participants on sertraline thought they were taking the active drug compared to 19% of participants on placebo [152]. Thus, 81% of placebo-recipients correctly guessed that they were on placebo, demonstrating that the blind was broken in a substantial portion of participants. The literature reviewed so far thus clearly indicates that unblinding is a serious issue in antidepressant trials. I will now detail if this methodological limitation biases the trial results systematically.

As in general medicine, unblinding most likely also inflates efficacy estimates in antidepressant trials. According to Baethge and colleagues, correct guessing of treatment assignment in schizophrenia and affective disorder trials was correlated with higher treatment effect sizes [684]. Khan and colleagues examined all sorts of depression treatments, including antidepressants and psychotherapy, and found that unblinded trials produced larger treatment effects (relative to placebo) than blinded trials for any treatment modality, but most pronounced in combination therapy (i.e. antidepressants and psychotherapy combined) [709]. Given that most antidepressants can cause marked side effects, an effectively blinded trial is basically impossible when inert placebo pills are used in the control group. A few tricyclic trials therefore used active placebos, that is, placebos that cause side effects comparable to some of the tricyclic side effects (especially dry mouth). A meta-analysis of these active placebo-controlled trials produced a pooled effect size much smaller than that typically found in trials with inert placebos [710]. Thus, taken together, these findings strongly indicate that unblinding introduces a systematic bias in favour of antidepressants, thus producing inflated efficacy estimates [18].

A last issue that warrants scrutiny is the handling of study dropouts (i.e. participants who discontinue treatment prematurely and thus terminate the trial). It is well known that when information on an outcome variable is missing, this may lead to a significant distortion of results when missing values are not adequately addressed [161, 711]. Even in

short-term antidepressant trials of 8 weeks duration, the dropout rate is roughly 30% [163, 175]. That is, almost a third of participants stops the treatment prematurely and thus their outcome at the end of the trial is unknown. This is problematic, since a loss of 20% or more can cause biased efficacy estimates and limits the generalisability of results [144]. In clinical trials, the intention to treat (ITT) analysis is standard practice now [712]. It requires that all participants randomised to treatment must be analysed, and not only those participants that completed the trial (referred to as per protocol or completer analysis). ITT increases the external validity of trial results, for in real-world routine practice it is common that patients discontinue treatment prematurely. But since the treatment outcome of study dropouts is unknown, these data must be imputed.

The most common statistical method in ITT analyses is the Last Observation Carried Forward (LOCF), which "is a data imputation process used in longitudinal repeated-measures clinical trials in which the last obtained data entry is substituted for any subsequent missing data, in an attempt to minimize the problem of dropout-associated missing data" [163]. For instance, if a participant stops a 8-week antidepressant trial prematurely at week 2 (let's say due to side effects) with a Hamilton depression score of 20 points, this last measure (observation) will be projected (carried forward) to be his/her 8-week treatment outcome. LOCF became the preferred method during the 1990s and was applied in about 80–90% of all antidepressant trials in the late 1990s and early 2000s [712]. That is, the efficacy of most new-generation antidepressants (especially SSRIs and SNRIs) was evaluated with LOCF method. More recently, however, the rate of LOCF fell to about 50% as it was increasingly replaced by more adequate methods [712]. But what's the issue with LOCF?

The LOCF method has serious limitations if the timing and reason of dropout differs between treatment groups, for it assumes that a given depression score at time of discontinuation would remain unchanged until the end of the trial [144]. This is of course a false assumption, for spontaneous remission and regression towards the mean (extremely high scores are often inflated due to random error and thus decline over time when repeatedly measured) will result in a reduction of average

depression scores independent of treatment [177]. If patients on placebo drop out earlier due to a felt lack of efficacy than patients on active drug, it's very likely that they discontinue with higher depression scores, even though many would have improved considerably until the end of the trial had they continued participation. The timing of dropout is seldom reported in antidepressant trials, but it's well established that participants receiving placebo more often discontinue treatment due to lack of efficacy than participants receiving antidepressants [367]. LOCF thus likely introduces systematic bias in favour of active treatment and thus inflates efficacy estimates [163, 695].

This assumption has been empirically confirmed. Siddiqui and colleagues [161] compared LOCF to the Mixed-Effect Model Repeated Measure (MMRM) model, a newer, more accurate method that predicts missing outcome scores based on all available data, including symptom trajectories from other participants (i.e., the average decline of scores over time for participants with similar scores). They ran a simulation study and an analysis based on phase III trials submitted to the FDA as part of a new drug application. First, "The simulation studies demonstrate that LOCF analysis can lead to substantial biases in estimators of treatment effects and can greatly inflate Type I error rates of the statistical tests, whereas MMRM analysis on the available data leads to estimators with comparatively small bias". A Type I error indicates that an estimated effect reached statistical significance when there likely is no true effect [48, 713]. Second, "analysis of 48 clinical trial datasets obtained from 25 New Drug Applications (NDA) submissions of neurological and psychiatric drug products, MMRM analysis appears to be a superior approach in controlling Type I error rates and minimizing biases, as compared to LOCF" [161]. That is, the widespread application of LOCF in antidepressant trials during the 1990s and early 2000s has most likely resulted in various false-positive results, meaning that in some trials efficacy estimates became statistically significant even though true treatment effectiveness was uncertain.

Methodological Biases Distorting Safety/ Tolerability Estimates

Safety refers to the adverse effects of a drug (also termed harms or side effects), whereas tolerability represents the degree to which adverse effects can be tolerated by patients. Per convention, adverse effects occurring in at least 10% of people are considered "very common", those affecting 1% to 10% "common", those affecting 0.1% to 1% "uncommon", those affecting 0.01% to 0.1% "rare", and those affecting less than 0.01% "very rare". The average sample size in antidepressant trials is 224 and most trials last merely 6–8 weeks [141]. As detailed above, this is sufficient to measure short-term efficacy of a drug, but insufficient to reliably detect even common adverse effects and to establish long-term safety [679, 692]. As detailed by Berlin and colleagues [690], with a sample size of 1000 participants there is a 82% chance to statistically detect an adverse drug effect that increases a harm event from 5% baseline risk to 10% during treatment (common adverse effect). Thus, even with such a large sample size rarely seen in antidepressant trials, there is a 18% chance to miss a common adverse drug effect. If a drug increases an adverse event rate from 1% to 2% (also falling into the rubric of common adverse effects), then with a sample size of 1000 there is only a small chance of 17% to statistically detect it. In that case, it would require a sample size of 5000 participants to detect it with a probability of 80%. When a drug increases the risk of an adverse event from 0.1% to 0.2% (uncommon adverse effect), then with a sample size of 1000 there is a meagre 5% chance to detect it, with a sample of 5000 participants the chance would be 7%, with a sample of 10,000 it would be 17%, and only with a sample of 50,000 it would be 79%. Thus, even if we pool the results from 10 trials with a sample size of 224 each, the resulting total sample size of 2240 participants will not generate enough statistical power to detect uncommon, let alone rare and very rare, adverse drug effects.

But even a large trial with a sample size of say 1000 participants won't guarantee that common adverse drug effects are statistically detected, for inadequate assessment and analysis of adverse events is a serious issue in randomised controlled trials [29, 686, 688, 689, 691, 714]. In most

antidepressant trials, adverse event assessments fully rely on spontaneous patient reports prompted through open-ended questions, that is, unstructured and unsystematic assessments. This can lead to a considerable underestimation of both frequency and severity of side effects, especially when patients are not comfortable discussing sensitive adverse events such as sexual dysfunction [29, 275, 335, 691].

The pharmaceutical companies seeking regulatory approval for their SSRI drugs already observed in the phase I trials (the first small, uncontrolled trials conducted in humans as part of a new drug application) that over 50% of healthy volunteers developed sexual dysfunction after SSRI exposure [334]. The companies realised this was a serious tolerability/safety issue, and therefore sexual dysfunction was avoided (or concealed) as much as possible in subsequent trials. That is, systematic assessment of sexual dysfunction did deliberately not take place in phase II and III trials (unlike phase I trials that are conducted in small samples of healthy volunteers to assess drug safety and dosing, phase II and III trials are conducted in larger clinical samples with the specific condition the drug is supposed to treat, and assess efficacy, safety, and tolerability). When sexual dysfunction was spontaneously reported by patients, it was commonly ascribed to the underlying condition, that is, the depressive disorder. And sometimes, clinical investigators were even instructed by the trial sponsor not to enquire about sexual dysfunction [9]. The unsystematic assessment and inadequate recording of adverse events thus allowed the companies to profess sexual dysfunction rates of less than 5% in phase II and III trials. A rate of less than 5% for sexual dysfunction was also the figure given in the initial SSRI drug labels [334]. How seriously did these official rates underestimate the true prevalence of treatment-emergent sexual dysfunction with antidepressants? Let's have a look.

In the pivotal premarketing placebo-controlled clinical trials of fluoxetine, treatment-emergent sexual dysfunction was recorded in merely 1.9% of participants receiving fluoxetine, but in postmarketing (postapproval) trials, based on systematic assessment with questionnaires, rates as high as 75% were reported. With respect to SSRIs as a class, spontaneous reports of sexual dysfunction produced rates of 2% to 7%, but these rates rose to 55% when systematically enquired via questionnaires [275]. Finally, according to a recent meta-analysis focusing exclusively on

clinical trials with a systematic assessment of sexual dysfunction, the rates are even higher for various SSRI drugs, being around 70% to 80% for fluoxetine, paroxetine, citalopram, sertraline, and venlafaxine (the latter is an SNRI), but only about 12% in placebo groups [336].

Treatment-emergent suicidality was also evident right from the beginning when the first SSRIs were clinically tested in humans. The new onset (occurrence) of suicidal ideation and behaviour on fluoxetine was also a main reason why the German drug regulators first refused to approve Eli Lilly's new drug application [9]. It was also quite clear that treatment-emergent suicidality was linked to fluoxetine's activation syndrome, that is, disinhibition, agitation, anxiety, nervousness, and akathisia. For Eli Lilly it was thus prerequisite to eliminate these side effects in antidepressant trials, which is why it (and other companies seeking approval for SSRIs and other activating antidepressants) by default permitted the comedication with sedative-hypnotic drugs. The companies further obfuscated the risk of treatment-emergent suicidality by systematically misrecording suicidal events [9, 29, 322, 323, 715]. For instance, suicidal events occurring in the lead-in phase (i.e. before randomisation) were counted as events in the placebo group, suicidal events leading to treatment discontinuation were not listed as adverse events, and discontinuation due to suicidality was often miscoded as discontinuation due to lack of efficacy. Some suicides and suicide attempts were not coded as serious adverse events but simply as study dropouts, and events clearly described as suicidal ideation or behaviour on case report forms were misrepresented by coding them as "emotional lability" or "worsening depression". Together these unethical and fraudulent practices led to a systematic underestimation of the risk of treatment-emergent suicidality in antidepressant trials. Although the drug regulators spotted most of these deceptions in the new drug applications for the SSRIs and SNRIs, they led the pharmaceutical companies get away with it and granted approval [9, 715, 716].

Finally, adverse events are inconsistently coded, commonly divided into multiple subcategories, and rarely grouped by anatomic or physiological system [29, 688, 689]. That is, the very same adverse event is frequently coded with different terms (e.g. akathisia interchangeably as agitation, nervousness, or restlessness), while events belonging to the same syndrome are commonly coded with different subcategories (e.g.

sexual dysfunction specifically as abnormal ejaculation, reduced libido, impotence, or anorgasmia). These methodological limitations impede the detection of harm signals. Imagine a clinical trial where 100 people were randomised to an antidepressant and 100 to placebo. In the antidepressant group, 9 patients developed akathisia, whereas in the placebo group there was only 1 such adverse event. According to a Chi-square test, this difference is statistically significant ($p < 0.05$) and would suggest that the antidepressant causes akathisia. However, if the 9 akathisia events are coded as nervousness in 3 cases, agitation in 3 cases, and restlessness in 3 cases, none of these adverse events would significantly differ from placebo and thus it would appear that the antidepressant does not cause akathisia or any of these coded adverse events. An important harm signal would thus go unnoticed. To account for this, lumping techniques were developed (i.e. grouping by anatomic or physiological system), but they are rarely used. It is therefore difficult or almost impossible to statistically detect adverse drug effects in modestly sized short-term trials when they are not very common. Along with the unsystematic assessment of adverse events detailed above (i.e. spontaneous self-reports), these biases corroborate (or amplify) the systematic underestimation of antidepressants' harm potential.

Although statistical analyses often lack the power (due to small sample sizes and low event rates) to reliably detect differences between treatment groups in subcategorised adverse event rates, such tests are frequently performed [686]. By consequence, these tests don't demonstrate statistically significant between-group differences even when the rate is considerably larger in one treatment group (e.g. 6% vs. 2%) [714, 717]. What's worse is that these statistically non-significant differences are often erroneously interpreted as no difference [718], even though researchers should know that "absence of evidence is not evidence of absence" [719]. Just because a difference in adverse event rates is statistically not significant does not indicate that there is no difference. It simply means that the sample was not large enough to draw reliable (or conclusive) statistical inferences from the data.

As detailed above, comedication with sedative-hypnotic drugs may bias the efficacy estimates of antidepressants. However, permitting the use of sedative-hypnotics in antidepressant trials has another important

implication. Since insomnia, agitation, and anxiety can also be symptoms of depression and withdrawal symptoms in participants who were on antidepressants before being randomised to placebo, sedative-hypnotics are also frequently used in placebo groups [241]. This may inflate the rate of specific adverse events in the placebo group, for sedative-hypnotics also have side effects, for example drowsiness and dizziness [720]. On the other hand, comedication may also mitigate some depression symptoms in the placebo group, for example insomnia and agitation. Therefore, the use of sedative-hypnotics will not necessarily change the rate of any adverse event. But since tolerability of an antidepressant is determined by comparing the rate of treatment discontinuation due to adverse events in the antidepressant group to that recorded in the placebo group, the use of sedative-hypnotics may bias this group difference. In any case, comedication certainly lowers the incidence rate of specific antidepressant side effects such as insomnia, agitation, nervousness, and anxiety, for these symptoms are effectively alleviated through the administration of sedative-hypnotics [9, 720, 721].

The narrow selection of younger participants without complicated illness and comorbid medical conditions is a standard feature of clinical trial protocols [682]. I already discussed that this may inflate the real-world effectiveness of antidepressants. It may, however, also bias safety estimates. Serious adverse drug reactions are more common in older people with various chronic medical conditions, often in interaction with other prescription drugs [722, 723]. Depression is very common in older people with comorbid chronic medical conditions [724] and antidepressant use by consequence is highest in this vulnerable patient population [725]. The patients most likely to be prescribed antidepressants in real-world practice are thus exactly those people at highest risk of adverse drug reactions. The safety and tolerability of antidepressants in these vulnerable patients is largely unknown though, since clinical trials preferably select younger patients without comorbid chronic medical conditions. However, given that frail patients (i.e. old adults with various chronic medical conditions on multiple medications) are more susceptible to adverse drug effects, the safety of antidepressants is certainly poorer in this high-risk population than in the patients typically included in antidepressant trials [300].

Discontinuation Trials, Placebo Response, and Other Issues

As you might remember, I did not discuss the evidence from antidepressant discontinuation trials for relapse prevention (assumed to assess long-term prophylactic effects) in the section on the long-term efficacy of antidepressants in depression. This is due to serious methodological limitations and systematic biases in these trials, which is why they are presented here.

In relapse-prevention (discontinuation) trials, participants are first treated open-label with an antidepressant (commonly for about 3–6 months), but it is important to note that many participants were already taking antidepressants (sometimes for years) before entering the actual treatment trial. Participants who by the end of the open-label acute treatment phase stably improved on the investigational drug (mostly defined as being in remission) enter the double-blind placebo-controlled maintenance phase. At the beginning of this second phase, participants are randomly assigned either to remain on the drug or to have the antidepressant rapidly discontinued (in most studies abruptly) and replaced by an inert placebo pill. Double-blind means that both patients and clinical investigators ought not to know whether someone was put on placebo or whether active treatment was continued. The blinded placebo-controlled trial phase commonly lasts about 6–12 months; there are only a few small trials for older drugs that lasted 24 months or longer [726]. The primary outcome in these trials is the resurgence of clinically relevant depression symptoms (defined as relapse), which is commonly based on a cut-off score on a depression rating scale such as the HDRS [726–728]. The main finding from discontinuation trials is that over an average observation period of 12 months, about 20% of participants maintained on antidepressant compared to 40% of those switched to placebo experience a relapse, yielding a rate ratio of 2 and a number needed to treat of 5 [214].

These figures are so impressive that some leading psychiatric academics consider antidepressants "one of the most effective of all drugs" [22]. However, as I previously noted about this subject, "as researchers, we should not be seduced into believing that a drug is highly effective simply

because a specific trial protocol has consistently produced impressive treatment effects, as these effects could be the result of a flawed trial protocol" [214]. And in the case of relapse-prevention (discontinuation) trials, there is indeed compelling scientific evidence that the protocol is seriously flawed and the results thus inconclusive, probably even misleading [9, 12, 214, 215, 228, 230, 235, 237]. Let me explain.

First, only patients who remitted during the acute treatment period (which is typically a minority of all patients) are randomised to either continued antidepressant use or abrupt discontinuation. The results of the randomised maintenance phase thus apply only to a particular subgroup of patients with a good short-term treatment outcome, but not to those who experience spontaneous recovery or those with a poor response to acute treatment. Second, the outcome in the double-blind randomised maintenance phase is merely a re-assessment of the unblinded acute-phase outcome (i.e. sustained response is assessed in acute treatment responders). Third, because participants were already treated open-label (unblinded) in the acute phase, they may instantly recognise when they are randomised to placebo and abruptly taken off the active drug. Various participants (and by consequence the investigators) are thus most likely unblinded. These three serious limitations systematically bias the results in favour of maintenance therapy and thus lead to inflated efficacy estimates [12, 215, 230, 234, 729].

Most important, however, is the fact that the outcome in relapse-prevention (discontinuation) trials is confounded, since many (sometimes most) relapses in the placebo group occur shortly after discontinuation of the antidepressant and are thus most likely withdrawal reactions [214, 228, 235, 237, 239, 729, 730]. It is well established that abrupt discontinuation of antidepressants can cause withdrawal syndromes, both acute and protracted, that often mimic a depression relapse or that may trigger a depression relapse (e.g. due to stressful physical withdrawal symptoms) [238, 345, 347, 731–735]. As a result, relapse-prevention (discontinuation) trials cannot differentiate between a true relapse, that is the recurrence of a genuine depression episode, and the consequences of a neurophysiological adaptation to prolonged drug exposure (pharmacodynamic effect) causing severe mental and physical withdrawal symptoms after abrupt/rapid discontinuation (also referred

to as oppositional tolerance) [217, 223, 351]. What may seem a benefit of continued antidepressant treatment (i.e. a long-term prophylactic effect) could very well be construed as an adverse treatment effect (i.e. iatrogenic harm) [214, 239]. Therefore, discontinuation trials cannot demonstrate that antidepressants truly prevent depression relapses [214, 228, 230, 235, 237, 238]. Whether continuing antidepressant use beyond the acute treatment phase relative to abrupt/rapid discontinuation prevents relapses or rather the occurrence of withdrawal reactions is still fiercely debated, but our recent analysis of relapse prevention (discontinuation) trials submitted to the FDA indicates that it is most likely the latter [239].

Various authors argued that the placebo response (i.e. observed improvements in placebo groups) has significantly increased in antidepressant trials over time and that this is a main reason for the modest/poor efficacy estimates of new-generation antidepressants [736–738]. The placebo response has indeed increased during the 1980s, mostly due to the broadening of the diagnostic criteria for depression (leading to the inclusion of many people with milder conditions) and changes in trial designs (with the advent of large multi-centre trials with longer duration and fixed dosing) [166, 192, 736]. But what about increased placebo response during the 1990s and 2000s? A comprehensive analysis based on published and unpublished trials by Furukawa and colleagues showed that since the early 1990s the placebo response remained largely constant [192]. By contrast, Khan and colleagues found evidence for increasing placebo response during the 1990s and 2000s, but they also showed that the average drug–placebo difference remained unchanged and that the rate of positive trials (i.e. statistically significant drug–placebo differences) has even increased, which argues against the hypothesis that an increasing placebo response prevents the demonstration of efficacy [739]. Moreover, when Furukawa and colleagues re-analysed the data by Khan and colleagues, they found no increase in the placebo response after controlling for changes in trial designs [740]. To further complicate matters, the most recent analysis even suggests that the placebo response slightly decreased from 2001 to 2015 [741]. In any case, there is no consistent evidence that the placebo response has increased since the mid-1990s and

no evidence at all that a higher placebo response is associated with smaller efficacy estimates or a higher rate of negative trials.

Another popular argument is that the improvement seen in placebo groups (i.e. observed placebo response) is mostly due to the placebo effect [11, 742]. By contrast, others argued that the placebo effect in antidepressant trials is trivial or inexistent [743]. The truth most likely lies somewhere in between these extreme positions [144]. However, it is quite clear that most apparent improvements observed in placebo groups (and by consequence also in antidepressant groups) are due to spontaneous remission, regression to the mean, and unspecific treatment effects (e.g. regular contact with a physician, clinical management, and comedication with sedative-hypnotic drugs) [177]. As demonstrated in many other medical fields, what has often been misconstrued as a genuine placebo effect is much better explained by other factors [744, 745]. Thus, as I outlined elsewhere, "it follows that the placebo effect in antidepressant trials is largely (though not entirely) a methodological artefact, and that the symptom reduction seen in placebo recipients is mostly due to both regression to the mean and spontaneous remission" [177].

Last but not least, there is ongoing controversy about the most popular scale to assess depression in clinical trials, the Hamilton Depression Rating Scale 17-item version (HDRS-17) [746, 747]. Various authors suggested that the HDRS-17 has poor validity and may underestimate antidepressant efficacy, for the scale is not unidimensional and may capture antidepressant side effects (e.g. insomnia, gastrointestinal symptoms, agitation, sexual dysfunction) [748, 749]. However, thus far there is no convincing evidence that alternative scales that more specifically assess core depression symptoms, for instance the Bech scale (HDRS-6) or the Montgomery-Asberg Depression Rating Scale (MADRS), generate significantly higher efficacy estimates, especially in severe depression [17, 258]. With respect to patient-centred outcomes, that is, quality of life and social functioning, effect size estimates again do not differ meaningfully from the HDRS-17 effect size [750, 751]. Moreover, it is also important to stress that simply because antidepressants may aggravate some depression symptoms (e.g. sleep problems, psychomotor agitation, sexual dysfunction, loss of appetite), this by no means legitimates the

exclusion of these symptoms in the assessment of depression [752]. Instead of removing such symptoms from a depression rating scale, one should rather wonder why we call a drug an antidepressant when in fact it worsens (or causes) various established depression symptoms [18, 753].

The quality of the depression ratings obtained through clinician-administered interviews (e.g. HDRS-17, MADRS) is another methodological limitation. An analysis of HDRS-17 assessments showed that "interviews were brief and cursory and the quality of interviews was below what would be expected in a clinical drug trial" [754]. Based on a small study, Kobak and colleagues suggested that antidepressants may fail to demonstrate efficacy due to these low-quality interviews [755]. However, the evidence is inconsistent, and a subsequent study by Khan and colleagues found the exact opposite [756]. According to their study, significant drug-placebo differences were only detected in trials where traditional semi-structured (low-quality) interviews were conducted, but not when a stringent (high-quality) interview technique was applied (i.e. structured interview guide with audiotaping and rater applied performance scale). It is thus debatable whether low-quality outcome ratings introduce systematic bias. But given that self-report instruments (quality of life, depression) produce comparable or even smaller effect sizes as the common clinician-administered rating scales (i.e. HDRS-17, MADRS), it is highly unlikely that lack of efficacy is due to low-quality interviews. Low-quality clinician outcome ratings may even inflate efficacy estimates, possibly due to unblinding of clinical investigators [18]. In this respect it is also important to note that patient self-reports of depression assessed with questionnaires such as the Beck Depression Inventory (BDI) produce significantly smaller effect sizes than clinician rating scales such as the HDRS-17 [147, 757, 758].

Selective Reporting and Spin

As I have outlined above, antidepressant trials are marred with methodological limitations, of which various seem to result in inflated efficacy estimates and underestimation of harms. Despite these systematic biases in the design and conduct of antidepressant trials, about half of all

placebo-controlled trials failed to demonstrate efficacy [57, 175]. This is, however, not the impression a physician gets when he/she consults the scientific literature, where almost all trial publications report positive results [174]. How is this possible? How can the scientific literature paint such a false and misleading picture of the actual scientific evidence? The answer is as simple as it is shocking: the trial data are misrepresented and selectively reported. Before I go into detail on how the evidence on the efficacy and safety of antidepressants is systematically biased in the scientific literature, I will briefly outline how clinical trial results are misreported in general medicine.

The scientific evidence consistently shows that about 20–50% of clinical trials remain unpublished and trials with positive results are about 2 to 5 times more likely to get published. The primary outcome reported in the published article is discrepant to the pre-specified primary outcome in about 30–40% of all trial publications, and roughly 40–60% of all negative primary outcomes (i.e. pre-specified primary outcomes that failed to demonstrate efficacy) are not reported in trial publications (i.e. journal articles). Moreover, about 30–50% of all negative primary outcomes are misrepresented as positive in the published article [85, 86, 89, 90, 668, 759, 760]. In addition, safety outcomes and (serious) adverse events are inadequately described and massively underreported in the scientific literature [90, 677, 687, 761, 762]. According to a comprehensive analysis by Golder and colleagues, only 36% of all adverse events are reported in trial publications and 54% of all publications provide no information on adverse events at all [87]. The authors thus concluded "There is strong evidence that much of the information on adverse events remains unpublished and that the number and range of adverse events is higher in unpublished than in published versions of the same study" [87]. Even deaths, the most serious adverse events, are not reported in most trial publications [665, 763].

How do we know about these issues? By comparing the trial results reported in journal articles to other sources, including results posted on trial registries, reviews provided by the drug regulators, internal industry documents released through litigation, and clinical study reports available to some medical authorities and researchers. Clinical study reports are very comprehensive documents of many hundred (sometimes thousands) pages written by the trial sponsors. They come the closest to the

raw data and regulatory agencies base their drug reviews on these extensive documents, for full raw data are property of the trial sponsors and not even regulators have access to them. The problem is that, with very few exceptions, clinical study reports are publicly unavailable. Pharmaceutical companies don't publish them and except for drug regulators and a few other health authorities (e.g. research ethics committees), it is very difficult or almost impossible to get access to them. The detailed results provided by the clinical study reports are thus rarely known to the public unless a pharmaceutical company is required to release them through litigation. Trial registries such as ClinicalTrial.gov are another important data source, albeit much less detailed and complete as the clinical study reports [665]. Since 2007, with a few exceptions, trial sponsors, including both industry and non-industry, are mandated to publish clinical trial results in a publicly accessible registry within one year of trial completion. Unfortunately, sponsors poorly comply with these legal requirements. According to a recent analysis, only 41% of trial results were reported within the 1-year deadline and 64% had results submitted at any time; 36% of trial results were thus not reported in the trial registry [764].

You may wonder how it could be that eminent medical academics selectively report outcomes, conceal (serious) adverse events, and if that doesn't help to create a positive message about the drug, prefer not to publish the trial? That is, why do so many leading academics (often professors of medicine) behave in such unscientific (and unethical) ways? Although there are certainly various reasons, including professional and personal interests, the two most important factors arguably are that, first, in most cases the authors don't analyse the data themselves, and second, they actually don't even write the articles [29, 459, 657, 659, 765–767]. The data from industry-sponsored trials are the property of the sponsor and analysed in-house by the company's own statisticians or else by a contract-research organisation. And most articles are ghostwritten, that is, they are largely (sometimes entirely) written by a medical communication firm hired by the company's marketing department, and not by the eminent medical academics listed as "authors".

The next question is, what to make of these findings? We can safely draw three main conclusions. First and foremost, the evidence is clear

and compelling that the efficacy and safety of medical interventions is significantly overestimated in the scientific literature. Therefore, trial results reported in journal articles are arguably the least reliable and most incomplete source. Second, trial registries can provide valuable information, for they allow to evaluate whether publications fully report all prespecified trial outcomes. Sometimes registries also allow to access trial results that were not reported in journal articles, but results posted in trial registries are often incomplete or lacking altogether. Third, clinical study reports are certainly the most reliable and most comprehensive data source, but they are not publicly available and often inaccessible. And that's why evidence-based medicine largely fails, for the scientific literature (i.e. evidence from peer-reviewed journal articles) is the cornerstone of clinical decision making in modern healthcare. Respected medical authorities such as Cochrane who provide systematic reviews relevant to clinical decision making strongly rely on publications in scientific journals. When the evidence is unreliable and biased due to selective reporting, so are the overall assessments of efficacy and safety of medical interventions. This serious issue has not gone unnoticed, and several EBM experts called for a careful reevaluation of the E in EBM (Evidence-Based Medicine) [29, 377, 378, 650, 768]. According to Jefferson and Jorgensen,

> "So, should we ignore evidence from journal articles? If steps are not taken urgently to address the situation, then 'probably' would be our answer. By the law of Garbage In Garbage Out, whatever we produce in our reviews will be systematically assembled and synthesised garbage with a nice Cochrane logo on it. One major problem is our ignorance of the presence of garbage, as its invisibility makes its distortions credible and impossible to check. This is how some of us happily signed off a Cochrane review with findings which had been completely and invisibly subverted by reporting bias". [653]

Selective Reporting in Antidepressant Trials

The most extreme, though not necessarily the most pernicious, form of selective reporting is publication bias, meaning that trials with favourable

results are published, often multiple times, whereas trials with unfavourable results will never see the day of light and thus remain unknown to physicians and the public. About a third of antidepressant trials for adult depression remain unpublished [57, 174]. And, of course, trial sponsors (mostly pharmaceutical companies) do not decide at random whether they publish a trial or not. They intentionally, and almost exclusively, publish trials with positive results [174], of which some are published multiple times as part of repeated pooled analyses [56, 769]. Let's start with a telling example.

The German Institute for Quality and Efficiency in Health Care (IQWiG) conducts health technology assessments to determine whether statutory health insurance should cover the costs of a new prescription drug. The health technology assessment report for Pfizer's reboxetine (the first selective norepinephrine reuptake inhibitor for depression) was impeded by Pfizer for not providing a complete list of all unpublished trials as requested by IQWiG [82]. The institute had received data from 3 published trials, but based on secondary publications reporting results for subsamples and other outcomes, IQWiG knew that the main efficacy results of many reboxetine trials were never published. Pfizer first refused to provide these data, but after long negotiations finally decided to cooperate and provided data for 10 unpublished short-term efficacy trials. Thus, according to IQWiG, Pfizer published only 3 of 13 efficacy trials (23%) of its antidepressant reboxetine and data on altogether 74% of trial participants remained unpublished. This is unethical on its own right, but the real scandal became only apparent when IQWiG compared the results of the published and unpublished trials. In the few published trials, reboxetine was superior to placebo and equally effective as SSRI comparator drugs. However, when IQWiG included the data from the many unpublished trials, reboxetine was no better than placebo and inferior to the SSRIs. Put differently, Pfizer did only publish a small subset of trials where reboxetine was superior to placebo and not inferior to SSRIs but tried to hide the majority of trials where its drug not only failed to beat placebo but also lost to the SSRIs. Based on the full data from all trials, the authors therefore concluded "Reboxetine is, overall, an ineffective and potentially harmful antidepressant. Published evidence is affected

by publication bias, underlining the urgent need for mandatory publication of trial data" [82].

Unfortunately, reboxetine is no exception and Pfizer's selective reporting of favourable trial results is standard operating procedure in the pharmaceutical industry. A comprehensive analysis by Turner and colleagues showed that according to the scientific literature (i.e. journal articles), 94% of antidepressant trials for adult depression are positive. However, based on the FDA's evaluation of trial data submitted to them for marketing approval, only 51% of antidepressant trials are positive. How is this possible? In total 74 placebo-controlled efficacy trials were submitted by pharmaceutical companies to the FDA as part of a new drug application for 12 new antidepressant drugs eventually approved between 1982 and 2004. In total 38 trials were positive (51%), and all but one of these (97%) were published. But there were also 36 trials, that is about half of all trials submitted to the FDA, with questionable or negative results. And this is where it gets really dirty. Of the 12 trials with questionable results, 6 (50%) were not published and 6 (50%) were published as positive. You think it can't get worse than this? Well, it does. Among the 24 negative trials, only 3 (12%) were published as negative, whereas 5 (21%) were published as positive and 16 (67%) were not published. That's why in the scientific literature almost every antidepressant trial appears positive when in reality just about half truly are. Resulting from this selective reporting of antidepressant trials, the efficacy of new-generation antidepressants was inflated by 32% in the scientific literature [57].

But how can a pharmaceutical company publish a trial as positive when the results were negative (i.e. no significant drug-placebo difference on the primary efficacy outcome)? Unfortunately, this is quite easy as there are multiple ways how the drug manufacturers can cheat [29, 56, 57, 60, 459, 767]. For instance, a company can decide to publish the more favourable per protocol analysis instead of the more conservative (but more accurate) intention to treat analysis. They can also report the results for a study subsample from selected study sites instead of the full study population. Or else they can switch the primary outcome when a statistically significant effect could be demonstrated on a secondary outcome or a newly created outcome measure. The leading academics commonly listed as "authors" on these publications are perhaps not even

aware of the extent of fraud they are indirectly supporting (and lending their badge of scientific excellence), for, as detailed above, in most cases they neither analysed the data nor did they actually write these articles [29, 78, 657].

That is, most antidepressant trials with negative results are distorted and presented as positive or else are simply not published [56, 57, 174]. But still many unpublished trials sooner or later appear in the scientific literature. They just rarely report the primary efficacy outcomes. The data from negative trials are often pooled to answer a different question by presenting data on a secondary outcome that do not reveal that the drug failed to beat placebo and are by and large positively framed (e.g. by focusing on selective safety outcomes) [174, 769]. This constant production of favourable publications is no longer research conducted in the spirit of advancing scientific knowledge but mere marketing. As bluntly stated by Spielmans and colleagues, "Such redundant publications add little to scientific understanding" [769]. It further indicates that the pharmaceutical industry actively (and efficiently) manages the scientific literature in order to advance its commercial interests (i.e. expanding markets and increasing prescription rates) [651, 658].

But selective reporting is not restricted to efficacy data. It equally affects safety data. Maund and colleagues compared the adverse events reported in clinical study reports of duloxetine trials for depression to those reported in the published journal articles for the same trials [83]. They found that in each trial, a median of 406 treatment-emergent adverse events were not reported in the journal articles. The total number of treatment-emergent adverse events reported in journal articles was less than half the number reported in the clinical study reports. Hughes and colleagues compared result summaries posted in a mandatory trial registry to the corresponding information provided in journal articles for the same trials. In 35 duloxetine trials, the trial registry listed a total of 11 deaths and 4 suicides; all (100%) were reported in the corresponding journal articles. However, of 40 suicidal events reported in the trial registry, only 33 (82.5%) were reported in the corresponding journal articles, and of 27 events of treatment-emergent psychiatric symptoms, only 21 (77.8%) were reported in journal articles. For the 7 sertraline trials listed in the trial registry, the situation was even worse. Of 11 deaths reported

in the trial registry, none (0%) was reported in the journal articles. No suicides were reported in both trial registry and journal articles. But of 5 suicidal events and 11 treatment-emergent psychiatric symptoms listed in the trial registry, again none (0% each) was reported in the corresponding journal articles [763].

Wieseler and colleagues assessed a large sample of antidepressant trials for depression (including bupropion, duloxetine, mirtazapine, reboxetine and venlafaxine) and various non-psychiatric drug trials for other conditions (e.g. diabetes and asthma). They compared the completeness of safety data reported in the clinical study reports to the corresponding data published in journal articles. Mortality, adverse events, and serious adverse events were completely reported in 100%, 92%, and 88% of clinical study reports, but only in 30%, 21%, and 24% of journal articles [665]. So, just like in general medical interventions [87], there is clear evidence that safety outcomes are underreported in antidepressant trials. The main question now is whether this incomplete information introduces a systematic bias in favour of the drugs comparable to the inflated efficacy estimates detailed above [57].

De Vries and colleagues compared safety evaluations provided by the FDA to the data presented by the drug manufacturers in the corresponding journal articles [81]. The risk of discontinuation due to adverse events in antidepressant groups compared to placebo groups was similar for FDA evaluation and journal articles (in depression the risk is about 2 times higher with antidepressants compared to placebo), suggesting that tolerability is not subject to reporting bias. Likewise, according to the comprehensive analysis of reboxetine trials by IQWiG detailed above, the risk of adverse events was similar for published and unpublished data [82]. However, while in the few published trials the increased risk of adverse events did not reach statistical significance, in the much larger database of unpublished trials the same effect estimate was statistically highly significant (due to increased statistical power). Moreover, while the few published trials suggested that discontinuation due to adverse events was no more likely with reboxetine than with placebo (suggesting the drug is well tolerated), according to the unpublished data the risk was about 2.5 times higher with reboxetine (indicating that the drug is not well tolerated). When both published and unpublished data were pooled,

the risk was about 2 times higher with reboxetine and the effect estimate was statistically highly significant. So according to these findings, selective reporting of antidepressant trials can indeed lead to distorted and exaggerated tolerability estimates.

There is also evidence that the underreporting of serious adverse events leads to systematically inflated safety estimates in antidepressant trials. According to De Vries and colleagues, there were discrepancies in the number of serious adverse events between the FDA evaluation and the corresponding journal articles in 43% of trials. In 78% of these discrepant cases, the published data (journal articles) led to a smaller or reversed drug–placebo difference and thus a systematically more favorable drug–placebo comparison [81]. Sharma and colleagues analysed the clinical study reports of 70 antidepressant trials (including duloxetine, fluoxetine, paroxetine, sertraline, and venlafaxine) obtained from European drug regulation agencies with a total of 18,526 patients. 16 deaths occurred in these trials, of which four deaths were misreported by the drug company, all systematically in favour of the antidepressant. For instance, "A patient receiving venlafaxine (trial 69) attempted suicide by strangulation without forewarning and died five days later in hospital. Although the suicide attempt occurred on day 21 out of the 56 days of randomised treatment, the death was called a post-study event as it occurred in hospital and treatment had been discontinued because of the suicide attempt" [323]. Moreover, of 62 suicide attempts, 27 events (44%) were misreported as "emotional lability" or "worsening depression" in the treatment-emergent adverse event tables, although in patient narratives or individual patient listings they were clearly identified as suicide attempts. Likewise, 32 of 63 suicidal ideation events (51%) were again misreported as "emotional lability" or "worsening depression". As detailed in the section "Methodological biases distorting safety/tolerability estimates", this misreporting and miscoding of suicidal events was a deliberate (and nefarious) tactic of the pharmaceutical companies to conceal the suicidality harm signal in antidepressant trials [9, 29].

The amount of selective reporting is even worse in paediatric antidepressant trials. The majority of these studies remain unpublished, and in the few that were published, the sponsoring pharmaceutical companies distorted and selectively reported the outcome data [29, 58, 290, 322,

770]. Not only were efficacy outcomes selectively reported, but harm outcomes as well. Especially treatment-emergent suicidality was systematically underreported and deliberately obfuscated on a large scale. The outcome data of antidepressant trials in children and adolescents was so terribly manipulated, misreported, and misrepresented that the Lancet Editors felt compelled to write an article titled "Depressing research", where they stated

> "It is hard to imagine the anguish experienced by the parents, relatives, and friends of a child who has taken his or her own life. That such an event could be precipitated by a supposedly beneficial drug is a catastrophe. The idea of that drug's use being based on the selective reporting of favourable research should be unimaginable … The story of research into selective serotonin reuptake inhibitor (SSRI) use in childhood depression is one of confusion, manipulation, and institutional failure … In a global medical culture where evidence-based practice is seen as the gold standard for care, these failings are a disaster. Meta-analysis of published data supports an increasing number of clinical decisions and guidelines, which in turn dictate the use of vast levels of health-care resources. This process is made entirely redundant if its results are so easily manipulated by those with potentially massive financial gains". [771]

Creating the Right Marketing Message for Antidepressants

A comprehensive analysis by Healy and Cattell [78] showed that a large number of articles on sertraline published between 1998 and 2000, including the vast majority of clinical trials from various therapeutic areas, were sponsored by Pfizer (manufacturer of sertraline) and written by a medical communication firm. The latter information was available to the authors due to an internal Pfizer document released through litigation, for on most publications the involvement of the medical communication firm was not disclosed (which is a violation of publication ethics). Most importantly, all ghostwritten trial publications were favourable to Pfizer's sertraline. The academics listed as authors on these articles had a large number of publications and the articles also appeared mostly in

high-impact journals and had a high citation rate. By contrast, articles on sertraline not sponsored by Pfizer and not prepared by a medical communication firm often reported negative findings (mostly safety issues), were typically published in low-impact journals, and the authors had a relatively small publication output. Healy and Cattel thus concluded "The profile of the articles reported here suggests that the background of certain authors may have increased the possibility of the company's publications appearing in the most prestigious journals. Specific journals seem to have been targeted. The combination of distinguished journal, distinguished author, an efficient distribution system and sponsored platforms appears to have led to an impact on the therapeutics domain greatly in excess of 50% of the impact of the rest of the literature on sertraline" [78].

From a commercial perspective, selective reporting of favourable results in journal articles allegedly written by leading academics clearly pays off for the pharmaceutical companies. They can be published in top-tier journals and are massively disseminated due to their high citation rates. Indeed, positive antidepressant trials are much more cited than the very few published trials with negative results [81]. The reach and impact of positive trials is further increased through multiple publications of the same trial results [56, 769]. And to make sure that the right marketing message is firmly established in the scientific literature, namely antidepressants being effective (at least based on the published articles), the pharmaceutical companies also heavily produce meta-analyses that synthesise these selectively reported positive results over and over again. Between January 2007 and March 2014, that is in roughly 7 years, an incredible number of 185 meta-analyses of antidepressant trials for depression were published, of which 54 (29%) were authored by industry employees and altogether 147 (79%) of meta-analyses had some link to industry (sponsored by industry or authored by industry employees or academics with financial relationships to industry) [772].

Publishing an abundance of favourable efficacy data is one way to disseminate the right marketing message; ignoring safety issues is another way to make sure that medical organisations and prescribers receive only positive information about a drug. It is thus worthwhile contrasting the 185 meta-analyses on the efficacy of antidepressants published between

2007 and early 2014, most sponsored or otherwise supported by the pharmaceutical industry, with meta-analyses specifically focusing on important safety issues relevant to public health and clinical decision making. So what about treatment-emergent suicidality and withdrawal syndromes, two prominent topics discussed in detail in the chapter "Conflicts of interest in medicine" that were fiercely debated for decades (for critical overviews, see for example [351, 357, 715, 773])? Let's have a look.

The first case reports highlighting and discussing antidepressant withdrawal were published soon after the introduction of the first antidepressants in the early 1960s [774, 775], but it took almost 40 years until the first randomised controlled trial, sponsored by Eli Lilly, was published [735]. The first systematic review followed in 2015 [347] and the first meta-analysis in 2019 [345], both conducted by researchers without industry-ties. As regards treatment-emergent suicidality, this was first prominently discussed in the early 1990s after the introduction of fluoxetine [326, 776], followed by a meta-analysis conducted by Eli Lilly in 1991 attempting to settle any doubts [777]. Then there were a few non-industry sponsored meta-analyses in the early and mid-2000s (e.g. [327, 330, 778, 779]), including the comprehensive FDA-analysis that led to the suicidality safety warning (referred to as black box warning) in children and adolescents [321, 780]. Between 2007 and early 2014, however, to the best of my knowledge there was only one other meta-analysis, that is the FDA analysis that led to the expansion of the suicidality black box warning to also include young adults [324]. In sum, while 185 meta-analyses on the efficacy of antidepressants were published between January 2007 and March 2014, during the same period there was not one meta-analysis on withdrawal syndromes and only one meta-analysis on treatment-emergent suicidality. Even very common side effects such as treatment-emergent sexual dysfunction are rarely studied. As far as I am aware, there were only two meta-analyses of treatment-emergent sexual dysfunction published during the period 2007–2014, namely, a study by Serretti and Chiesa from 2009 [336] and another by Reichenpfader and colleagues from 2014 [781].

Spin is another pernicious issue in the reporting and interpretation of antidepressant trials. Spin is defined as "a specific reporting that fails to

faithfully reflect the nature and range of findings and that could affect the impression that the results produce in readers, a way to distort science reporting without actually lying … Reporting results in a manuscript implies some choices about which data analyses are reported, how data are reported, how they should be interpreted, and what rhetoric is used. These choices, which can be legitimate in some contexts, in another context can create an inaccurate impression of the study results … It is almost impossible to determine whether spin is the consequence of a lack of understanding of methodologic principles, a parroting of common practices, a form of unconscious behavior, or an actual willingness to mislead the reader. However, spin, when it occurs, often favors the author's vested interest (financial, intellectual, academic, and so forth)" [667]. The most basic depiction of spin is the standard conclusion stated in almost every single positive antidepressant trial that the investigated drug was effective, safe, and well tolerated, even when efficacy estimates were marginally small, some adverse events considerably higher with antidepressants, and treatment discontinuation due to adverse events significantly increased compared to placebo. I will now provide two compelling examples of how spin manifests in antidepressant trials and how it contributes to the exaggeration of efficacy and minimisation of harms. I deliberately chose two governmentally sponsored trials to illustrate that spin is not exclusively an issue in industry-sponsored trials.

The Depression Hypericum Trial tested hypericum perforatum (St John's Wort) and sertraline against placebo [151]. Both active drugs failed to beat placebo on the primary efficacy outcome, the mean change in HDRS-17 total score from baseline to 8 weeks. The rates of full response at week 8 were 23.9% for hypericum, 24.8% for sertraline, and 31.9% for placebo, with no statistically significant between-group difference. In addition, there were five secondary efficacy outcomes (a self-report measure of depression, one measure of disability, one measure of global functioning, and two measures of general illness severity). Hypericum failed to separate from placebo on all of them and sertraline on four of them. That is, one weak but statistically significant difference was found between sertraline and placebo on one of the measures of general illness severity (the Clinical Global Impressions-Severity scale). Nevertheless, the results were quite clear and consistent overall. Both hypericum and sertraline

failed to conclusively improve depression, disability, global functioning, and general psychopathology in comparison to placebo.

But still, the authors, oddly enough, concluded in the main text "According to available data, hypericum should not be substituted for standard clinical care of proven efficacy, including antidepressant medications and specific psychotherapies, for the treatment of major depression of moderate severity" [151]. This conclusion does not logically follow from the data. In this trial St John's Wort was indeed not effective, but so was sertraline, the "standard clinical care of proven efficacy". If the authors judge St John's Wort ineffective in this patient population (which they obviously did), then they must also conclude that sertraline is ineffective. Moreover, their conclusion did not logically follow from the broader scientific literature. Considering all studies, of which many were already available when the Depression Hypericum Trial was published, St John's Wort is just as effective as standard antidepressants and also superior to placebo, though, as with antidepressants in general, the effect size is small [203]. Thus, according to all available data, the only appropriate conclusion would be that standard antidepressants are no better than St John's Wort. If the authors, which had extensive financial ties to manufacturers of antidepressants, including shares in Pfizer (the manufacturer of sertraline), consider St John's Wort ineffective in major depression, then so are standard antidepressants like sertraline.

The second example is the TADS trial, a governmentally sponsored 12-week randomised treatment trial evaluating the efficacy of fluoxetine and cognitive-behavioural therapy (CBT) against placebo in adolescents with depression [782]. Although the study was sponsored by the NIMH, many authors had received research support and honoraria for serving as consultants and/or speakers for Eli Lilly, the manufacturer of fluoxetine. The study was also supported by an unrestricted educational grant from Eli Lilly. Two primary efficacy outcomes were prespecified; first, the continuous score on the Children's Depression Rating Scale-Revised, and second, response (much or very much improved) based on the Clinical Global Impressions scale. The two secondary efficacy outcomes were the Reynolds Adolescent Depression Scale and the Suicidal Ideation Questionnaire-Junior High School Version. Let us first look at the summary of the results as stated in the abstract, often the only part of a paper that busy clinicians have the time to read.

> "Compared with placebo, the combination of fluoxetine with CBT was statistically significant (*P*=.001) on the Children's Depression Rating Scale-Revised. Compared with fluoxetine alone (*P*=.02) and CBT alone (*P*=.01), treatment of fluoxetine with CBT was superior. Fluoxetine alone is a superior treatment to CBT alone (*P*=.01). Rates of response for fluoxetine with CBT were 71.0% (95% confidence interval [CI], 62%–80%); fluoxetine alone, 60.6% (95% CI, 51%–70%); CBT alone, 43.2% (95% CI, 34%–52%); and placebo, 34.8% (95% CI, 26%–44%). On the Clinical Global Impressions improvement responder analysis, the 2 fluoxetine-containing conditions were statistically superior to CBT and to placebo. Clinically significant suicidal thinking, which was present in 29% of the sample at baseline, improved significantly in all 4 treatment groups. Fluoxetine with CBT showed the greatest reduction (*P*=.02). Seven (1.6%) of 439 patients attempted suicide; there were no completed suicides. Conclusion: The combination of fluoxetine with CBT offered the most favorable tradeoff between benefit and risk for adolescents with major depressive disorder". [782]

So the authors stressed that fluoxetine combined with cognitive-behavioural therapy (CBT) was more effective than placebo, fluoxetine alone, and CBT alone on the first primary outcome (Children's Depression Rating Scale-Revised). They also emphasised that fluoxetine alone was more effective than CBT alone on both primary outcomes and that fluoxetine (alone and in combination with CBT) was more effective than placebo on the second primary outcome (Clinical Global Impression scale). However, they did not mention that fluoxetine was not significantly better than placebo on the first primary outcome (Children's Depression Rating Scale-Revised). Neither did they state that fluoxetine failed to beat placebo on the two secondary efficacy outcomes, the Reynolds Adolescent Depression Scale and the Suicidal Ideation Questionnaire-Junior High School Version. Instead they mentioned that suicidal thinking significantly improved in all treatment groups and that fluoxetine with CBT showed the greatest reduction. You might rightly argue that the authors only mentioned statistically significant results, which is what clinicians are mostly interested in. Okay, fair enough. But in that case, why did the authors not mention that various adverse events

were significantly more frequent in fluoxetine-treated patients compared to CBT and placebo? Let's look a bit closer at these safety data.

According to spontaneous adverse event reporting, there were significantly more treatment-emergent events of self-harm (including self-injurious and suicidal behaviours) in patients treated with fluoxetine compared to non-fluoxetine treated patients (including CBT alone and placebo). The rate of treatment-emergent self-harm was 10.2% in patients treated with fluoxetine compared to 4.9% in patients not treated with fluoxetine and the difference was statistically significant ($p < 0.05$). The rates of self-harm for fluoxetine alone was 11.9% as compared to 5.4% for placebo, but due to lack of statistical power, this difference was statistically not significant [782]. Moreover, rates of suicide attempts were 2.8% for fluoxetine treatment (with or without CBT) and 0.4% for non-fluoxetine treatment (CBT alone or placebo). The authors claimed that the numbers were too small for statistical comparison, but according to my own calculation the difference fell just short of statistical significance according to a two-tailed Fisher's exact test ($p = 0.064$) and are thus concerning in view of the significantly increased risk of treatment-emergent self-harm. In addition, 14.8% of patients treated with fluoxetine (with or without CBT) and 4.5% of patients not treated with fluoxetine (CBT alone or placebo) experienced a treatment-emergent psychiatric adverse event (mostly mood dysregulation and insomnia, which are known side-effects of fluoxetine). This difference is statistically highly significant according to a two-tailed Fisher's exact test ($p < 0.001$).

In sum, in the abstract, the TADS authors emphasised significant efficacy outcomes for fluoxetine (alone and in combination with CBT) but did not mention that fluoxetine alone failed to beat placebo on three of four efficacy outcomes (of which one was a primary outcome). Moreover, they did not mention that the rates of treatment-emergent self-harm and other psychiatric adverse events were significantly higher in patients treated with fluoxetine than in patients not treated with fluoxetine. Clinicians simply skimming the abstract may thus understandably gain the false impression that fluoxetine alone is both effective and safe in adolescents. This false impression was reinforced in the conclusions of the main text, where the authors claimed "The effectiveness outcomes were clear and the clinical implications straightforward … Fluoxetine alone

was effective, but not as effective as fluoxetine with CBT" [782]. This statement is problematic, for there was no definite statistical evidence for the efficacy of fluoxetine against placebo on one of two primary outcomes. Conclusive statistical evidence of effectiveness would imply that fluoxetine was significantly better than placebo on both primary outcomes. Fluoxetine also failed to demonstrate efficacy on both secondary outcomes, that is, patient self-reported depression and suicidal thinking. Moreover, the data clearly indicate that fluoxetine treatment was associated with increased rates of self-harm and other psychiatric adverse events, which was not mentioned in the abstract and in the conclusions of the main text. Finally, and worthy of note, at the naturalistic 36-week follow-up reported in another publication, the response rates for fluoxetine combined with CBT, fluoxetine alone, and CBT alone did not differ (86%, 81% and 81%). That is, although CBT alone was less effective than fluoxetine (alone or in combination with CBT) in the acute placebo-controlled 12-week phase, at week 36 it was just as effective as medication, indicating that psychotherapy, quite understandably, takes a bit longer to work than medication. Moreover, at week 36 there were significantly more suicidal events in patients treated with fluoxetine alone (14.7%) as compared to combination therapy (8.4%) or CBT (6.3%) [783]. According to these long-term outcomes, fluoxetine alone seems not indicated in adolescents with major depression due to increased risk of self-harm.

Paroxetine Study 329

Nowhere else became the deleterious impact of selective reporting and spin coupled with aggressive off-label promotion (i.e. promotion for an unapproved condition) more evident than in antidepressant prescribing for paediatric depression [289, 771, 784]. A particularly revealing (and shocking) case in point is the study 329, a paroxetine trial in adolescents with depression sponsored by its manufacturer GlaxoSmithKline. Various books and articles had been written about this infamous, fraudulent trial that served as an infomercial to promote off-label paroxetine prescribing in youth [29, 322, 770, 785, 786]. The trial even has its own Wikipedia

entry, https://en.wikipedia.org/wiki/Study_329. Before I will go into detail of why this study is a prime example of fraud in industry-sponsored antidepressant trials, it is important to stress that we would never have known the full extent of this scandal if GlaxoSmithKline had not been pressured to release internal documents and provide free access to the clinical study report (which comes close to the raw data) through litigation. The original article by Keller and colleagues on the 8-week acute phase results reported that paroxetine (93 participants), but not imipramine (95 participants), was significantly better than placebo (87 participants) on four of eight efficacy outcomes. Rates of withdrawal from the study because of adverse events were 9.7% for paroxetine, 31.5% for imipramine, and 6.9% for placebo. The article further reported 11 serious adverse events in the paroxetine group, 5 in the imipramine group, and 2 in the placebo group. 5 suicidal and self-injurious adverse events were reported for paroxetine, 3 for imipramine, and 1 for placebo. The authors concluded that "Paroxetine is generally well tolerated and effective for major depression in adolescents" [787].

The documents released through litigation, including the clinical study report, and a comprehensive re-analysis of the data by independent academics, tell of a completely different story [29, 322, 770, 785]. The Keller et al. article was largely ghostwritten by a medical communication firm in close collaboration with GlaxoSmithKline's marketing department. Most "authors" listed on the paper had financial relationships with GlaxoSmithKline (mostly honoraria for serving on advisory board and speakers' bureau), which were not declared in the published article. The two primary outcomes and the five secondary outcomes designated in the study protocol were all negative, that is, paroxetine failed to beat placebo on any of the prespecified efficacy outcomes. All four efficacy outcomes demonstrating statistical significance in the Keller et al. article were introduced post-hoc by GlaxoSmithKline after dredging the data (also referred to as p-hacking). The two prespecified primary outcomes that failed to demonstrate efficacy were reported in the article but presented as if they were secondary outcomes. Of the five prespecified secondary outcomes (which also failed to demonstrate efficacy), only two were reported in the article, the others were omitted. Thus, in short, paroxetine unequivocally failed to demonstrate efficacy. It is only through concealing prespecified

outcomes and creating new ones post-hoc that GlaxoSmithKline could give a false impression of some questionable efficacy [322, 770, 785].

But GlaxoSmithKline also deceived on a large scale to present paroxetine as safe and "generally well tolerated" [787]. The comparator drug imipramine was dosed way too high, so that it caused a lot of side effects and discontinuation due to adverse events (the latter at an incredibly high rate of 31.5%) [29, 770]. As detailed above, overdosing a comparator drug is a common strategy so that the sponsor's drug looks safer and better tolerated in comparison [62, 84]. In addition, many adverse events were miscoded and misreported, including reasons for premature treatment discontinuation. Contrary to the rate of discontinuation due to adverse events of 9.7% for paroxetine reported by Keller et al., the independent re-analysis of the data by Le Noury et al. showed a rate of 15.0% for paroxetine [322], that is, about twice the rate for placebo (6.9%). According to the clinical study report, the rate of serious adverse events (mostly suicidal and self-injurious events) were 11.8% for paroxetine and 2.3% for placebo, a statistically significant difference [785]. Very concerning was also the misrepresentation of suicidal and self-injurious behaviours. These adverse events were mostly miscoded as "emotional lability" and some events listed in the appendix were not included. Contrary to 5, 3, and 1 events for paroxetine, imipramine, and placebo reported by Keller et al., the clinical study report stated 7, 3 and 1 events, and the re-analysis by Le Noury et al. found 11, 4, and 2 events [322]. That is, paroxetine use was related to a clear excess of suicidal and self-injurious behaviours. According to my own calculation, the rate was significantly higher with paroxetine compared to placebo based on the data given by both the clinical study report (7.5% vs. 1.1%) and Le Noury et al (11.8% vs. 2.3%). In this respect it's also important to mention the number of severe psychiatric adverse events (including but not limited to suicidal and self-injurious behaviours) reported in the re-analysis. Le Noury et al found 32 severe psychiatric adverse events for paroxetine (among 93 participants) compared to 4 for imipramine (among 95 participants) and 6 for placebo (among 87 participants), a difference that is clinically meaningful and statistically highly significant (my own calculation).

In conclusion, the ghostwritten report of study 329 by GlaxoSmithKline (i.e. Keller et al., 2001) stated that paroxetine was effective and generally well tolerated in adolescents with depression [787]. However, careful examination of the clinical study report and a comprehensive re-analysis of the raw data by independent academics showed that paroxetine was not only ineffective, but harmful [322, 785]. GlaxoSmithKline applied a variety of fraudulent and unethical strategies to misrepresent the efficacy and safety of paroxetine, including a comparator drug dosed way too high, selective reporting of efficacy outcomes, post-hoc creation of new outcomes, and both underreporting and miscoding of (severe/serious) adverse events. In addition, GlaxoSmithKline intentionally withheld data from a second paroxetine trial for adolescent depression that also failed to demonstrate efficacy and safety (study 377). This second trial was completed in 1998, that is, long before the results of study 329 were published in 2001. When pooled together, these two trials showed that paroxetine was completely ineffective and associated with a significantly increased rate of suicidal behaviour compared to placebo [58, 786]. To withhold the data of both trials from drug regulators (which would have immediately noted these issues), the company did not seek regulatory approval for paroxetine in adolescent depression [788].

In an internal GlaxoSmithKline document titled "Seroxat/Paxil Adolescent Depression: Position piece on the phase III clinical studies", the marketing department gave recommendations on how to deal with the two negative adolescent trials. "Effectively manage the dissemination of these data in order to minimize any potential negative commercial impact", the document states. And further, "It would be commercially unacceptable to include a statement that efficacy had not been demonstrated, as this would undermine the profile of paroxetine" [788]. So GlaxoSmithKline clearly knew that paroxetine should not be used in adolescents, but the company remained silent about lack of efficacy and increased risk of suicidal behaviour. Quite the contrary, the company exploited the distorted Keller et al. publication to aggressively promote off-label use of paroxetine for adolescent depression, knowing that it was neither effective nor safe [29, 322, 770, 785]. In a memorandum to its sales representatives, the company stated "This 'cutting edge,' landmark study is the first to compare efficacy of an SSRI and a TCA with placebo

in the treatment of major depression in adolescents. *Paxil* demonstrates REMARKABLE Efficacy and Safety in the treatment of adolescent depression" [785]. This message was further disseminated at conferences and in the media by GlaxoSmithKline's key opinion leaders, influential academic psychiatrists on the company's payroll [29, 770].

Fortunately, in mid-2003, drug regulators issued a safety warning and stressed that paroxetine should not be used in children and adolescents due to treatment-emergent suicidality [789]. Based on their evaluation, the UK Committee on the Safety of Medicines (CSM) concluded that there is "a clear increase in suicidal behaviour versus placebo" [786]. As summarised by McGoey and Jackson,

> "It seems unarguable, then, that for five years, GSK [GlaxoSmithKline] deliberately failed to disclose clinical trial data which provided evidence that Seroxat should not be prescribed to under-18s. Given that, in 1999 alone, 32 000 prescriptions for Seroxat had been issued to children in the UK, it is clear that in the time-lag between the completion of the relevant clinical trials (1998) and the CSM's warning notices (2003), tens of thousands of under-18s were prescribed a drug that was unlikely to work, and which carried an unacceptable risk of a serious, indeed fatal, adverse reaction. We do not know how many, if any, under-18s actually committed suicide between 1998 and 2003 as a result of taking Seroxat, but given the large number of children involved, it is certainly possible that deaths occurred which could have been avoided by prompt disclosure of this information". [786]

Years later, in 2012, GlaxoSmithKline pleaded guilty and was fined US$3 billion by the US Department of Justice for large-scale healthcare fraud, including illegal promotion of paroxetine for unapproved adolescent depression, creating misleading journal articles making unsubstantiated and/or false representations or statements about safety and efficacy of paroxetine, and hiding paroxetine trials that had negative findings [790]. But what about the fraudulent publication of study 329 by Keller et al., which has in total 808 citations as of June 2021? You would certainly think that a scientific journal has the ethical and legal obligation to retract a fraudulent research article. Well, you err. Despite several requests

and unequivocal evidence that the article misreports and misrepresents the efficacy and safety of paroxetine, including a legal conviction by the US Department of Justice, the *Journal of the American Academy of Child and Adolescent Psychiatry* who published the article, refuses to retract it [29, 770, 791]. In fact, the American Academy of Child and Adolescent Psychiatry, the owner of the journal, deliberately turns a blind eye to this nefarious study. As stated by Dr. Doshi, "No correction, no retraction, no apology, no comment" [791]. By consequence, the false and misleading findings of study 329 remain in the scientific literature and the article is still widely cited, not only as a prime example of scientific fraud, but also as "evidence" that paroxetine is effective and generally well tolerated in adolescents with major depression.

In sum, the efficacy and safety of antidepressants is systematically misrepresented in the scientific literature due to methodological biases, selective reporting, and spin. It is therefore almost impossible to evaluate the drugs' true treatment effects, especially in real-world routine care. The chapter has also shown that the pharmaceutical industry, psychiatric associations, and eminent academics play a major role in this pervasive distortion of the scientific evidence. This leads us directly to the next chapter, "Conflicts of interest in medicine".