Chapter 12

General discussion

In this thesis, I have re-examined the evidence for the treatment of depression and anxiety. By doing so, I aimed to answer questions about these treatments that have remained unanswered so far, in spite of the hundreds of trials that have been conducted over the past decades. My focus was, firstly, on examining the impact of reporting and citation biases on the literature, and secondly, on investigating clinical predictors of treatment response.

In this discussion, I will first briefly summarize the main findings of the different chapters. Then I will place my findings on reporting and citation bias in a broader perspective, discuss possible solutions to these issues and their effectiveness (insofar as these solutions have already been implemented), and consider further work that is necessary. Subsequently, I will focus on treatment efficacy, the clinical relevance of the true efficacy and safety of antidepressants, and what kind of research is needed now to improve outcomes for depressed and anxious patients.

Summary of main findings

In the first part of this thesis, I examined the presence of reporting and citation biases in the literature on the treatment of depression and anxiety. For several of these chapters, I used Food and Drug Administration (FDA) drug application packages. Because pharmaceutical companies are required to preregister these trials with the FDA and must submit their results, regardless of trial outcome, these packages contain a complete and unbiased overview of all pre-marketing trials conducted in the pursuit of approval for a specific drug for a specific indication.

Building upon previous research in depression, chapter 2 showed that study reporting bias, outcome reporting bias, and spin are present in the literature on antidepressants for the short-term treatment of anxiety disorders. According to the FDA, 72% of these antidepressant trials were positive, but 96% of published articles were positive. Of the negative trials, seven remained unpublished, three were published with outcome reporting bias, and three others were published with spin. Only three trials were clearly published as negative. As a consequence of these biases, the effect size of antidepressants for anxiety disorders had been overestimated by 15%, although this difference was not statistically significant.

Chapter 3 examined reporting bias for harm outcomes in antidepressant trials for anxiety as well as depression. No bias in the reporting of discontinuation rates was found. However, nearly two-thirds of journal articles did not mention serious adverse events (SAEs) at all. Of the articles that did mention SAEs, the majority contained discrepancies with the FDA or did not include any descriptions of the SAEs, including one article that failed to mention two suicides in the drug group. Together, these findings show that the published literature is an unreliable source of information, both regarding the efficacy of

antidepressants and regarding their safety (especially where SAEs are concerned).

In chapter 4, I investigated the phenomenon of pooling otherwise unpublished trials for publication. These pooled-trials publications were very common, but very few (12%) of these publications had as their primary aim to present data on the included trials' primary research question (comparing the efficacy of the drug to placebo), and even fewer (3%) presented efficacy data for the primary research question for individual trials. Even though the vast majority of these pooled-trials publications included one or more negative trials, only 3% had a negative conclusion. Consequently, pooled-trials publications flood the literature with positive results for secondary questions while obscuring the negative findings for the primary outcome.

In chapters 5 and 6 I examined spin (or positive focus) and citation bias in the literature on 5-HTTLPR. Within both the amygdala activation and the gene-environment interactions literature, articles with negative findings often presented positive conclusions in their abstract. Both truly positive articles and positively presented articles receive more citations than articles with negative findings and negative conclusions. This effect was stronger within the amygdala activation literature than the gene-environment literature, perhaps because of the greater controversy surrounding gene-environment interactions. These results show how negative findings, even when published, can remain relatively invisible, which contributes to a more positive view of the evidence base for an effect than is warranted.

Chapter 7 investigated the cumulative effect of reporting and citation biases on the evidence base for psychotherapy and antidepressants for depression. While study publication bias is a well-known phenomenon by now, this chapter demonstrates the pernicious cumulative effect of study publication bias, outcome reporting bias, spin, and citation bias. Starting with a cohort of 105 antidepressant trials, of which 52 were considered negative by the FDA, I finally found that only four articles unambiguously report that the antidepressant was not effective. Positive trials were also cited three times as frequently as negative trials (92 versus 32 citations on average). Each of these biases makes it more difficult to discover negative results, and together, they can render the vast majority of negative results virtually invisible.

In chapter 8, I studied whether the evidence is actually put into practice. Guideline committees critically appraise and synthesize all relevant evidence in order to arrive at treatment recommendations. These guidelines can be seen as representing the "state of the art" for any given topic. However, adherence to the guidelines for antidepressant initiation in children and adolescents was very poor. Physicians preferred citalopram, an antidepressant that has never been shown to be effective in these age groups, over the recommended fluoxetine, which has the best available evidence in favor of its efficacy and safety in young people. Starting doses were also higher than recommended, especially in teens, who were usually prescribed adult starting doses. These findings show that translation of the evidence base to clinical practice often fails, thus undermining the

goals of evidence-based medicine.

The second part of this thesis aimed to investigate who benefits from antidepressants. In chapters 9 and 10 I examined initial severity as a predictor of antidepressant response compared to placebo in anxiety disorders. In chapter 9, I used aggregate data from 56 clinical trials for GAD, SAD, OCD, PTSD, and panic disorder and found no evidence that greater initial severity was associated with a better response to antidepressants. However, because the use of trial-level data to investigate a patient-level characteristic, like initial severity, can be subject to the ecological fallacy, and because reduced power means that interactions may have been missed, I used individual participant data (IPD) to investigate this question in chapter 10. Here, I found that initial severity was associated with antidepressant response for GAD and panic disorder, but not for SAD, OCD, or PTSD. These findings suggest that the benefit of antidepressants is relatively small at low severity for GAD and panic disorder and the risk-benefit ratio may therefore be unfavorable for patients with mild GAD or mild panic disorder.

In chapter 11, I examined whether early improvement in individual depressive symptoms could enhance the predictive power of a model already containing early improvement in the total score. There was limited evidence that this was the case, which suggests that clinicians can gain about as much information about a patient's likelihood of responding or remitting to antidepressants from the total score. However, the predictive utility of this model was still relatively limited, particularly for outcomes after twelve weeks (rather than six weeks) of treatment, which suggests that we cannot predict a patient's likelihood of a good response with much certainty by two weeks of treatment.

Evidence-b(i)ased psychiatry[1]

In the first part of my thesis, I showed that reporting and citation biases are prevalent within the antidepressant and psychotherapy literature. These biases complicate assessment of the risks and benefits of these treatment options. While some biases, like study publication bias and outcome reporting bias, can and do affect the results of systematic reviews and meta-analyses [325], spin and citation bias, in principle, do not. However, the interpretation of meta-analytic results is not perfectly straightforward and objective: these interpretations in fact vary widely, perhaps dependent on readers' prior beliefs [501, 502]. Hence, meta-analyses are not a foolproof remedy against the effects of spin and citation bias.

The problem of study publication bias was first recognized over fifty years ago [36]. It is therefore disconcerting to realize how little progress appears to have been made in the decades following its first identification. Only recently have medical journals, funders, and

[1] With acknowledgment to Melander and colleagues [27], who coined the term "evidence-b(i)ased medicine".

governments begun to take measures to combat study publication and outcome reporting bias. The primary weapon in their arsenal is mandatory pre-registration of clinical trials. In 2004, the International Committee of Medical Journal Editors (ICMJE) made pre-registration a requirement for publication in ICMJE journals, an important first step [330]. However, many medical journals do not adhere to the ICMJE guidelines, and many others also accept retrospective registration [332, 503]. Retrospective registration, particularly when it takes place after ascertainment of the primary outcome, is virtually useless when it comes to preventing study publication and outcome reporting bias.

In 2007, the FDA Amendments Act (FDAAA) was signed into law. This Act requires prospective registration of "applicable clinical trials" and empowers the FDA to levy fines against non-compliant investigators [504]. Although the FDA has never yet exercised its power to penalize investigators for failure to register, this regulatory requirement may nevertheless prove to be more effective than the ICMJE requirement, since it applies to all applicable clinical trials, regardless of the publishing aspirations of the investigators, and has the potential force of the law behind it.

Importantly, the FDAAA also requires investigators to post a public summary of results within one year of completion of the trial, which would ensure that trial results are easily available even if the trial is never submitted for publication. However, compliance with this requirement is dramatically low: only 13% of all trials report results within the 12-month deadline and only 38% report them at any time [505]. Ironically, given the particular attention paid to selective publication by pharmaceutical industry, industry-sponsored trials are more likely to report results on time than academic trials, perhaps because pharmaceutical companies have more resources to bring to bear to meet this requirement.

Compliance may be low in part because penalties have never yet been enforced and in part because the requirement for results reporting conflicts with the requirement by many journals that results have not yet been disseminated (although many journals have since clarified their requirements to permit basic results reporting on sites like ClinicalTrials.gov) [506].

An unfortunate weakness of the FDAAA is that it only includes trials of drugs or medical devices, thus excluding trials of behavioral interventions like psychotherapy. Since bias is equally problematic in these trials [35], this is a regrettable omission. An additional, though legally inevitable, problem is that the FDA only has jurisdiction over trials with at least one center located in the United States or that are performed in the service of an application for marketing approval in the United States.

Mandatory requirements for prospective registration have the potential to eliminate publication bias and outcome reporting bias from the medical literature, but they require "constant vigilance" in order to succeed. Clinical trial registries must be examined thoroughly to detect trials that have been registered but not published. While particular

attention should be paid to trials that do not even report basic results, we should not be satisfied with trials remaining unpublished (and essentially invisible) otherwise. Basic results reporting on ClinicalTrials.gov is important, but it cannot be considered equivalent to publishing a trial in a high-impact medical journal that will be read by thousands of researchers and clinicians. Furthermore, unless peer reviewers and editors carefully scrutinize pre-registrations during the review process, it is likely that outcome switching will continue to occur.

The medical literature could also benefit from importing the concept of registered reports [507], an idea that originates within the field of psychology and that is more of a "carrots" (reward) approach compared to the "sticks" (punishment) of the FDAAA and the ICMJE policy.

Registered reports take the logical next step beyond pre-registration. A researcher designs a study, writes the introduction and methods section of the article that will eventually result from the study, and submits this to the desired journal. The study undergoes peer review and, if the research question is considered interesting and the design suitable, the study receives "in-principle acceptance", that is: if the study is conducted as proposed, it will be published by the journal regardless of the results. The researcher only performs the study after in-principle acceptance has been obtained.

Registered reports, hence, are a form of "results-free" reviewing, which ensures that papers are not rejected because of negative results. This is an aspect that is still missing in a system of prospective registration, which ensures that we are aware of all trials that have been conducted (and potentially of their results, if compliance with mandatory results reporting improves) but does nothing to prevent negative results from being published later than positive results (time-lag bias), in lower-impact journals than positive results (place of publication bias), in languages other than English (language bias) [46], and so on. It cannot even prevent non-publication altogether (although perhaps nothing, short of extremely strict enforcement of high penalties, is 100% guaranteed to prevent study publication bias).

Another benefit to registered reports is that it invites peer review of the study design before the study has been conducted, at a stage when fatal or minor flaws can still be rectified. The current system of peer review after completion of the study, on the other hand, is more reminiscent of the famous quote by Ronald Fisher [508]: "To consult the statistician after an experiment is finished is often merely to ask him to perform a post mortem examination. He can perhaps say what the experiment died of." The registered report format is difficult to adapt to all possible study types (e.g. secondary analyses of pre-existing epidemiological cohorts, exploratory studies), which may limit its implementation in some fields, but it is perfectly suited to clinical trials.

There is some reason to be cautiously optimistic about the future of the medical literature. Comparing the trials for newer antidepressants to those of older antidepressants (chapter

7), for example, we found that the newer, negative trials were much more likely to be published than older negative trials. This may, to some extent, be related to the extremely critical and focused attention that has been paid to biased reporting of antidepressant trials in particular, but other studies have also found an increase in publication rates in recent years [509].

If results reporting on sites like ClinicalTrials.gov is included, disclosure rates for industry-sponsored trials were >90% in recent years [510]. It is possible, therefore, that the problems that motivated our work in chapter 2, investigating study publication bias and outcome reporting bias in antidepressant trials for anxiety disorders, will largely disappear in the coming years, which would be a major step forward for evidence-based psychiatry.

Rectifying the existing literature, however, has proven to be a difficult task. None of the articles that we identified as biased have since been retracted or corrected [511]. Even the article reporting the results of Study 329, the trial of paroxetine for adolescent depression that has been embroiled in controversy ever since its first publication and that has been the focus of a dedicated campaign to correct the record, has neither been retracted nor corrected [512].

It is also unlikely that universal prospective registration will effectively combat all biases. In particular, it is unlikely to ameliorate the problems of spin and citation bias at all, and it is possible and perhaps likely that financial incentives to suppress negative results will lead pharmaceutical sponsors to seek out other methods, such as publishing negative results in journal articles that are not widely read or not even indexed in databases like PubMed or EMBASE. One of the negative trials in our analysis in chapter 7, for instance, was published in the Journal of Drug Assessment, which was not indexed in PubMed until very recently. We only discovered this publication because it happened to be mentioned in a review [513].

Although there is a widespread sense that negative results are more difficult to publish, and it might be argued that this is why this trial was published in a journal where it was virtually guaranteed to go unnoticed, this argument does not seem very plausible in the age of PLOS ONE, BMJ Open, and comparable journals, which publish research regardless of its clinical relevance and newsworthiness.

Conversely, pharmaceutical companies make an effort to publish positive results from their studies, by pooling trials, slicing and dicing the results, and publishing highly redundant articles that appear to serve primarily to keep a drug in the spotlight, as shown in chapter 4, in which we investigated pooled-trials publications. The added scientific value of these publications often seems limited, and instead these publications often seem to represent "minimal publishable units". For duloxetine, for instance, there are separate papers investigating the efficacy of duloxetine in African-American compared to Caucasian patients, and in Hispanic compared to Caucasian patients [174]. Both papers conclude that duloxetine works equally well regardless of race or ethnicity, and we can

wonder whether they would have been published if they had reached any other conclusion.

The medical literature, in this case, has been co-opted by pharmaceutical interests. However, medical journals are not just being duped, but are also complicit in these practices and have their own (financial) conflicts of interest, both directly (e.g. selling of reprints to pharmaceutical companies) and indirectly (e.g. maintaining or achieving a high impact factor through publishing highly citable papers) [514].

Biased and incomplete reporting of harm outcomes, which we investigated in chapter 4, also remains a problem that is less likely to be solved by universal prospective registration. To some extent, poor reporting of harms may be because researchers and clinicians are simply less interested in these outcomes. Most trials are powered to detect a significant difference in the primary outcome and are underpowered to study (rare) harm outcomes, which means the trials are not very informative in this regard. Nevertheless, full reporting is essential in order to enable meta-analyses, which can compensate for the small sample sizes of individual trials.

Word count limits may have been a good reason for the limited reporting of the many and diverse harm outcomes in the past, since tables of common adverse events alone may take up several pages. However, as most journals are now primarily electronic, authors can include the full information in the online supplemental information. With regard to serious adverse events, one important take-away from our study in chapter 4 is that causal attribution should play no role in the reporting of these events. Even events that are thought to be completely unrelated to the drug should be reported, since it is impossible to definitively establish causality in individual cases.

As important as the biased reporting of SAEs, however, is the complete absence of reporting on these events, which occurred in the majority of journal articles. Since regulatory authorities require investigators to monitor SAEs in drug trials, journals can be sure that these events were recorded and should require authors to report on the number and nature of SAEs.

Some initiatives, like accreditation or ratings, could encourage pharmaceutical companies to grant their negative results as much visibility as their positive results [515], perhaps following the example of the Access to Medicine Index [516]. This Index charts to what extent the largest pharmaceutical companies are working to make medication and vaccines more accessible to people in low- and middle-income countries. In this way, it makes visible which companies are doing well and which companies are doing poorly, and incentivizes companies to do better. A "Good Scientific Practice" index might similarly provide an incentive for companies to adhere to scientific best practices: to publish all studies, regardless of results; to stick to the statistical analysis plan; and to interpret results objectively and fairly, without spin. Like the Access to Medicine Index, such an index could be funded by non-profit foundations and governments.

Much more radical propositions, however, have also been put forward. In particular, there

is a case to be made for taking clinical trial programs out of the hands of pharmaceutical companies and placing them into the care of governmental institutions or universities [517]. The current situation is rife with conflicts of interest: pharmaceutical companies, who stand to benefit from a drug if and only if it is approved by the authorities and prescribed by physicians, carry the responsibility for conducting the trials that are meant to show that the drug is safe and effective. Strict regulation can mitigate this inherent conflict of interest, but complete elimination of pharmaceutical company sponsorship, though potentially expensive, might be the only comprehensive solution. This is, however, unlikely to happen, especially in this era of "public-private partnerships" [517].

Elimination of pharmaceutical sponsorship also does not solve the conflicts of interest of non-industry researchers. Academic researchers do not usually have the same financial incentives to suppress negative results as pharmaceutical companies do, although some do (e.g. developers of a psychotherapy approach who earn royalties from sales of the treatment manual). They may, however, have non-financial conflicts of interest, for instance due to developing the treatment in question, a concept known as "allegiance bias" in psychotherapy research [518].

Academic researchers also have clear career incentives to publish frequently and in high-impact journals [519]. These high-impact journals, however, are most interested in novel, surprising, counter-intuitive, and (usually) positive results, one possible reason why the retraction rate is higher in higher-impact journals [520]. Unfortunately, study results are the one aspect of a study that a researcher has no control over – at least not unless he or she engages in outcome switching, so-called p-hacking, or other questionable research practices. Career incentives to get published often and in "good" journals, therefore, are often misaligned with good science [321].

Even in the absence of any questionable research practices, this process can lead to the "natural selection of bad science", as researchers whose methods are less rigorous (e.g. smaller sample sizes) do better career-wise, while science itself suffers [521, 522]. However, it is at least in principle possible to re-align these career incentives to facilitate instead of hinder good science; this is not the case for industry research, which will always primarily serve commercial, rather than scientific interests.

Instead of rewarding scientists for results and emphasizing simple quantitative measures of success that can easily be "gamed", like the h-index [521], scientists must be rewarded for methodologically sound and clinically relevant research, regardless of the results [523]. As a first step, this may require greater openness with regard to data, methods, analysis code, and so on, in order to enhance reproducibility and other researchers' ability to appraise the work.

While methodological soundness is, to some extent, a subjective term, most research areas do have norms for what is considered adequate, which are sometimes formalized (e.g. in the form of reporting guidelines) and sometimes remain implicit. As argued

by Moore and colleagues, it is difficult, if not impossible, to judge whether research is "excellent" as opposed to merely adequate, but not as difficult to tell apart "sound" and "unsound" research [523]. Importantly, methodologically sound research should be informative (e.g. sufficiently powerful) regardless of whether the results reach the arbitrary threshold of "$p < 0.05$".

When negative results are no longer looked down upon, the need to spin results may also lessen. A re-orientation toward methodological soundness might also decrease the tendency to preferentially cite positive findings, although it seems likely that citation bias will be one of the more difficult biases to combat, since it is not amenable to relatively straightforward solutions like pre-registration and since it is probably human nature to prefer positive (and potentially actionable) findings. Chapters 5, 6, and 7 showed that both spin and citation bias are currently highly prevalent in the literature, both within observational research on the etiology of depression (chapters 5 and 6) and within the clinical trial literature (chapter 7).

The presence of spin, in particular, exemplifies the failure of peer review, since this problem is, in principle, easy enough to detect (especially in clinical trials), perhaps even easier than outcome reporting bias since it does not even require the reviewer to look up the clinical trial registration. Peer review, however, often fails [524]. Peer reviewers could, in theory, also be tasked with mitigating citation bias, but it is likely that peer reviewers are not aware of all the relevant literature themselves, so this is not a very feasible demand, except perhaps in very small research fields.

In many cases, the most effective defense against spin and citation bias might still be the performance of high-quality, unbiased meta-analyses, although this is not an actual solution but rather a harm-reduction method. This is also likely to be a more effective approach in fields where the evidence base is clear, homogeneous, and uncontroversial. In the literature on the 5-HTTLPR and stress interaction, for instance, several meta-analyses have been performed, reaching opposite conclusions [305, 306, 307, 308], and these meta-analyses only seem to have polarized the field further, rather than leading to consensus.

In this case, where meta-analyses conflict, researchers' prior beliefs probably play a large role in determining which meta-analysis they place their faith in. A recently published, very large, well-conducted collaborative meta-analysis (with a total sample size of 38,802) that found no evidence for an interaction between 5-HTTLPR and stress in the development of depression might finally put the debate to rest, however [525]. Even in cases where meta-analyses effectively establish a consensus, though, citation bias remains problematic to the extent that it incentivizes researchers, who are evaluated on the basis of citation metrics, to produce positive findings.

While the evidence base is often tainted by biases, it remains the best resource that we have to guide clinical decision-making. In chapter 8, however, I found that adherence

to evidence-based guidelines for antidepressant initiation in young people was very poor. Clinical practice is often resistant to change [365], but a major change in physicians' habits for antidepressant initiation in young people did take place in the first decade of the 21st century, since physicians abandoned paroxetine after it became embroiled in controversy and negative media attention due to treatment-emergent suicidality. However, they started prescribing citalopram instead, not fluoxetine. It is possible that the guideline, which was published in 2009, appeared too late to take full advantage of this period of shifting prescription habits to spur physicians on to switch to fluoxetine.

Another problem, specific to child and adolescent psychiatry, is that the vast majority of research has been performed in adults and there is little evidence specifically for pediatric patients, which may lead physicians to extrapolate from the (relatively solid) evidence in adults instead of working from the (relatively shaky) evidence in children and adolescents. However, such extrapolation is risky since children are not little adults. In the case of antidepressants, this is especially clear, as treatment-emergent suicidality appears to be an issue primarily in young people [190, 191].

This study, looking at antidepressant initiation in young people, is not an isolated example of poor adherence to guidelines [526], which shows that transfer of the evidence into practice is far from guaranteed. Many factors can affect the likelihood of guideline adherence, some of which are controllable by guideline developers (e.g. clarity of the recommendation), while others are not (e.g. a stable and uncontroversial evidence base) [361, 527].

It is clear, however, that guidelines must be actively disseminated and implemented, not just developed [365]. No matter how good the evidence base or how many measures are taken to combat bias, it is only when the evidence is actually put into practice that the ideals of evidence-based medicine are achieved.

The effectiveness of treatment

When the first selective serotonin reuptake inhibitors (SSRIs), such as fluoxetine and sertraline, were developed in the late eighties and early nineties, expectations were high: these antidepressants were received as wonder drugs [528]. In the quintessential book from this period, Listening to Prozac, Peter Kramer, a psychiatrist, related the stories of patients who did not just stop being depressed when they started taking fluoxetine (Prozac), but who became "better than well" [529]. Kramer speculated that these drugs might herald an era of "cosmetic psychopharmacology", in which healthy people would take psychotropic drugs to improve their personalities.

To some extent, Kramer was right: fully 12% of adult Americans took antidepressants in 2013 [24]. Among middle-aged women (40 – 59 years old), 23% took antidepressants in the period 2005 – 2008 [530], a percentage that has likely increased even further in the

decade since. Although depressive and anxiety disorders are relatively common in women, it is difficult to believe that a quarter of middle-aged women suffer from sufficiently severe and persistent depression or anxiety to warrant antidepressant treatment.

The SSRIs' image as wonder drugs, however, has also become tarnished, and the pendulum has swung the other way, with some now arguing that these drugs do not work at all or even make things worse [531, 532, 533]. The blame for the backlash against antidepressants can be placed, at least in part, on the pharmaceutical industry, since their practice of hiding unfavorable results enabled the overhyping of antidepressants and the subsequent letdown when both antidepressant efficacy and safety were revealed to be less impressive than the marketing had suggested.

However, antidepressant efficacy is unlikely to be zero. Our best estimate for the (unbiased) effect size of antidepressants compared to placebo for the acute treatment of depression and anxiety is a standardized mean difference (SMD) of around 0.30 – 0.35, based upon the meta-analysis by Turner and colleagues [19] and our own meta-analysis, included in this thesis (Chapter 2). This effect size might still have been overestimated slightly as a consequence of unblinding: some antidepressant-treated participants are likely to deduce that they are taking the active drug because they experience side effects. On the other hand, placebo-treated participants are also rather likely to experience adverse events that align with the anticipated side-effect profile of the drug, perhaps as a consequence of a "nocebo" effect. Placebo-treated participants in trials of tricyclic antidepressants (TCAs), for instance, experience more "typical" TCA side effects, such as vision problems and dry mouth, than placebo-treated participants in SSRI trials [534].

While some have suggested that the effect of antidepressants could be wholly explained by unblinding [535], this seems improbable, for a few reasons. First, some putative antidepressants do, in fact, fail in the course of development. Despite having side effects, these drugs cannot be shown to have antidepressant or anti-anxiety activity. Some are subsequently repurposed for other disorders (e.g. atomoxetine, which is now approved and used for attention-deficit hyperactivity disorder (ADHD) [536]; or flibanserin, a controversial drug recently approved by the FDA for hypoactive sexual desire disorder [537, 538]). Secondly, at least one approved antidepressant, reboxetine, has been shown to be less effective than SSRIs and is not even statistically significantly more effective than placebo, despite having worse tolerability than SSRIs [101].

Finally, although early work found that experiencing side effects was associated with better outcomes (suggesting that unblinding might play a role) [539], more recent studies have not confirmed any association between the proportion of patients with adverse events and trial outcome [540, 541]. Most recently, an IPD analysis found antidepressants to be effective both in patients with and without adverse events, and also found no association between adverse event severity and antidepressant efficacy, providing strong evidence against the hypothesis that unblinding explains the effect of antidepressants [542].

An effect size of around 0.3 cannot be considered large, but it is consistent with effect sizes found in other areas of psychiatry and general medicine [421]. The corresponding number needed to treat (NNT) is around 8 to 10, that is, out of every 8 to 10 patients treated with an antidepressant, one will have a better outcome than they would have had with placebo treatment.

It is, however, important to keep in mind that treatments do not actually have effect sizes; they only have effect sizes relative to a control condition. Consequently, the effect sizes of different treatments, and especially of drugs versus psychotherapy, are incomparable. Drug trials have excellent control conditions: double-blind placebo administration ensures that the only difference between the active and the control group is in the pharmacological action of the drug (apart from some possible bias due to unblinding).

Psychotherapy trials, on the other hand, often have very inadequate control conditions: many psychotherapy trials use wait-list controls and are also, by necessity, open or single-blind. Because of this, the effect size of psychotherapy versus wait-list is much larger than that of antidepressants versus placebo, at around 0.9 [543]. However, the effect size of psychotherapy versus other control conditions (e.g. care as usual, blinded pill placebo) in high-quality studies is much smaller, only around 0.22 [543, 79]. Direct comparisons between antidepressants and psychotherapy generally find that they are about as effective for depression and anxiety, with some exceptions (e.g. for dysthymia) [544].

Why are the effects of pharmacotherapy and of psychotherapy (compared to a good control condition) so small? In part, this is probably because some patients readily improve even in response to control conditions. In Chapter 9, for instance, we found that the pre-post effect size of placebo varied from 0.5 (for OCD) to 1.0 (for GAD). For depression, the pre-post effect size was 0.9 [69]. These pre-post effect sizes must be interpreted with caution, because it is impossible to separate true improvement from a mere regression to the mean effect [545], but they at least suggest that some patients do not require active treatment to improve.

In Chapters 9 and 10 I investigated baseline severity as a possible moderator of antidepressant efficacy, but I only found evidence that it was a significant moderator for GAD and panic disorder. This might imply that patients with mild GAD or panic disorder are more likely to improve spontaneously, although this must remain speculative: Chapter 10 (Figures 1 and 2) showed that the change in score in the placebo group actually increases with increasing severity (suggesting greater "spontaneous improvement"), but this may be due to regression to the mean or due to a floor effect at low levels of severity.

It remains unclear, therefore, which characteristics set apart those participants that are likely to respond adequately to control conditions, although we do have some knowledge of which patients show a relatively benign course even in the absence of treatment. Some of the characteristics that are associated with recovery include a short episode duration and mild symptoms [3, 546]. These characteristics motivate the practice of "watchful

waiting", which the Dutch guidelines recommend, for instance, for mild episodes of major depression with a duration of less than three months [74]. They are also applied equally to determine which patients can be assigned to low-intensity interventions like guided self-help, even though the (limited) available evidence suggests that these interventions are effective regardless of severity [547].

Among those who do not respond to control conditions or improve spontaneously, many also fail to respond to active treatment, whether antidepressants or psychotherapy, as reflected in the small effect sizes of these treatments. One explanation for this is that the etiology of depression and anxiety is likely to be heterogeneous *and* complex. These disorders can arise from purely organic causes (e.g. depression in Parkinson's disease [548]), but for most patients, they probably develop out of a complex interplay of genetic and other biological vulnerabilities, adverse early childhood experiences, inadequate or abusive parenting, trauma, poverty, personality disorders, childhood disorders like ADHD or autism, intellectual disability, and so on [549, 550].

Consequently, depression and anxiety are perhaps best conceptualized as the end stages of a very long illness process that more often than not has its roots in early childhood. It is therefore unsurprising that psychiatric treatments often prove inadequate in the face of this tangled knot of vulnerabilities. Worse, psychiatry is often forced to pit its treatments not just against these past vulnerabilities, but also against current, difficult to change, and deeply depressing, stressful, or anxiety-provoking life circumstances: unemployment, poverty, loneliness, abusive partners or family members, physical illness, caregiving responsibilities, single parenthood, and so on.

Seen in the light of our limited understanding of the brain [67], the deep roots of these disorders, and the adverse circumstances that many patients find themselves in, modest effect sizes are not surprising, especially considering that these treatment trials often only last six to twelve weeks. Our limited understanding of the brain also implies that revolutionary new treatments will probably be a long way off. Indeed, it could be argued that the treatment of depression and anxiety has not improved significantly since the development of the tricyclic antidepressants and benzodiazepines in the 1950s (although the SSRIs have better tolerability) and of cognitive (behavioral) therapy in the 1960s and 1970s. Instead, we have gained a diversity of approaches, all of which seem to be approximately equally effective [324, 544].

Although this proliferation of approaches has not resulted in the discovery of substantially more effective treatments, it could nevertheless be useful if some patients benefit from one particular treatment, while others benefit from a different treatment. The existence of such differential responses to treatments is a plausible hypothesis, given the heterogeneous etiology of depression and anxiety.

In the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial, only 37% of participants attained remission in the first treatment step, consisting of up to 14

weeks of citalopram [551]. This is somewhat worse than our own finding that 51% of participants attained remission after twelve weeks of treatment (chapter 11), which may be because STAR*D had less strict inclusion and exclusion criteria (e.g. with regard to episode duration and comorbid disorders). For participants who stuck with treatment through multiple treatment failures, though, the cumulative remission rate was 67% in STAR*D.

While far from perfect, this nevertheless suggests that some patients can respond to second-line treatments even if they do not respond to an SSRI. The trick, then, is to achieve better matching of patients to treatments. Although we also need to develop effective new treatments for the 33% of participants who did not remit even after four treatment steps, this is a much more daunting undertaking than matching patients to the type and intensity of treatment that is most likely to work for them.

Better matching is also important to reduce treatment dropout. Depressed and anxious patients are likely to become discouraged after a treatment failure and may give up on treatment altogether. Dropout is strongly associated with poor outcomes: in STAR*D, for instance, 36% of non-remitters dropped out of treatment within 8 weeks, as compared to only 7% of eventual remitters [552]. Cause and effect probably go in both directions: participants who experience no benefit are likely to drop out; participants who drop out will experience no further benefit from treatment. It is therefore important to minimize the duration of ineffective treatment and to maximize the likelihood that the first attempted treatment is successful.

Precision psychiatry: making the best of what we have

One possible approach to maximizing the likelihood that the first attempted treatment is successful is, of course, to provide a high-intensity combination treatment to all patients. Addition of an atypical antipsychotic to an antidepressant leads to higher response rates in depression, for instance [472, 553], and combining psychotherapy and antidepressants is more effective than either treatment alone for both depression and anxiety [554]. However, the likelihood of success must be balanced with the likelihood of harm, which implies that patients should not receive more treatment than they require. Furthermore, patients should not continue receiving an ineffective treatment any longer than necessary.

In Part 2 of this thesis, I studied several clinical characteristics that might distinguish between patients who will benefit from treatment and those who will not. In particular, I examined the initial severity of symptoms as a predictor of response to antidepressants. In my IPD meta-analysis, I found evidence that antidepressant efficacy increased with increasing severity for GAD and panic disorder, but not for the other anxiety disorders. This implies that, for GAD and panic disorder, the benefits of antidepressants are rather small at low severity and as such, antidepressants may not be preferred as first-line

treatment.

The limited benefit of antidepressants compared to placebo should not be taken as encouragement to "do nothing", however, since taking part in the placebo group of a clinical trial is a fairly intensive intervention in and of itself. In support of the effectiveness of this intensive clinical management, it has been found that a greater frequency of trial visits is associated with a better clinical response in both the placebo and antidepressant group [555]. It remains to be seen whether the limited benefit of antidepressants in patients with mild GAD and panic disorder is because of spontaneous recovery (independent of trial participation) or because intensive clinical management is sufficient for these patients.

For patients with SAD, OCD, or PTSD, I did not find evidence that antidepressant efficacy depends upon initial severity. This implies that patients with mild to moderate disorders receive about as much benefit from antidepressants as patients with severe disorders. While this is encouraging news for patients who choose to take antidepressants, there may nevertheless be good reasons to reserve antidepressants (as a first-line treatment) for patients with more severe disorders. In patients with milder disorders, for instance, the burden of the disorder might not be proportional to the burden of antidepressant side effects.

It is also possible that patients with mild disorders experience more benefit from psychotherapy than patients with severe disorders, which could be an argument in favor of providing psychotherapy instead. The available evidence, however, suggests that the effectiveness of psychotherapy compared to antidepressants does not depend upon severity, although this evidence is derived from depression, not anxiety disorders [80]. Even so, it is possible that psychotherapy alone might suffice for patients with mild disorders, whereas patients with severe disorders might require both antidepressants and psychotherapy. Furthermore, for OCD, the evidence suggests that exposure and response prevention (ERP) and other CBT approaches are more effective than antidepressants and hence should generally be preferred regardless of severity [556].

In Chapter 11, I showed that early improvement, within the first two weeks, is predictive of later response or remission in patients with depression, but not so predictive that a blanket recommendation to change treatment for patients who show no early improvement should be implemented. Examining individual symptoms rather than the total score alone did not make a marked difference. At week 6, 26% of patients who showed no early improvement had responded, and by week 12, 38% had. Even patients who showed early worsening had a similar probability of response.

While this represents a considerable drop in the likelihood of a good outcome, these numbers are far from negligibly small. Furthermore, there is no evidence that switching antidepressants is associated with better outcomes than continuing the same antidepressant, either in the case of non-response after a conventional treatment period (six weeks or longer) [470] or in the case of non-improvement after two to four weeks [557, 558].

This suggests that lack of early improvement does not identify a subset of patients who are resistant to that particular antidepressant, but rather a subset of patients who are difficult to treat in general. Many of these patients may need more intensive treatment (e.g. augmentation or combination with psychotherapy), but since such intensified treatment is associated with a greater risk of harm, it seems premature to initiate it after just two weeks of monotherapy, in a patient group that still has about a 40% probability of responding to an antidepressant alone.

It might, however, be possible to identify a group of patients with an indication for early intensification of treatment by combining information about early improvement with baseline information about the previous course of the disorder, the presence of comorbid disorders or physical illnesses, circumstances that precipitated the disorder, and so on. In our IPD meta-analysis, the necessity of combining patient data from multiple trials precluded the incorporation of such information, because it was not consistently assessed across trials.

The main priority for future research, however, must be to identify characteristics that predict response to a specific treatment. Unfortunately, head-to-head trials of specific treatments are relatively scarce [544, 108], while there is an abundance of (markedly less useful) trials of antidepressants versus placebo [59], or of psychotherapy compared to a control treatment like wait-list or treatment-as-usual (TAU) [543]. The relative scarcity of head-to-head data may hamper the search for moderators of treatment efficacy, since fine-grained questions require relatively large amounts of data. An additional problem is that the study populations of many of the available trials (both head-to-head and otherwise) are highly selected and may not be particularly representative of the actual population of depressed and anxious people seeking treatments [475, 559, 560].

However, enough head-to-head trials have been done in the past few decades to at least begin to move from merely predicting outcome (regardless of treatment) to predicting response to specific treatments or at least treatment classes (e.g. SSRIs vs. SNRIs, CBT vs. antidepressants), especially for depression [81]. It is also becoming easier to access individual participant data from these trials. GSK was ahead of the curve in this regard by initiating Clinical Study Data Request (CSDR), which was used to access individual participant data for chapters 10 and 11 and which has now been joined by eleven other companies, including Lilly and Bayer. Other companies like Johnson & Johnson have agreed to a data sharing platform in collaboration with Yale, the so-called Yale Open Data Access (YODA) project [561], and Pfizer also makes clinical trial data available through its own portal [562].

Using these IPD is still hampered by lengthy application procedures, cumbersome access systems, and the inability to combine data from sponsors that participate in different platforms [563, 564], but since CSDR, the oldest of these platforms, was only initiated in 2013, it is not surprising that there is still room for improvement. Individual participant data from National Institute of Mental Health (NIMH)-sponsored trials, including

STAR*D, can also be requested [565], and IPD meta-analyses of psychotherapy trials have also been conducted [79, 80, 566], although these generally require the cooperation of the primary study authors.

In future research examining moderators of treatment efficacy, it will be important to adhere to best practices for predictive analytics – in particular, this means that the performance of predictive models must be tested in an independent data set in order to prevent over-fitting [567], a requirement that is not always met. Ideally, any hypotheses derived from such research should subsequently be tested in a prospective randomized trial, in which participants are assigned to their (model-predicted) "best" treatment or not, but this has the disadvantage that it would significantly delay implementation of these models in clinical practice.

Importantly, any model that predicts response to treatments that are similar in terms of harms, costs, and time investment (e.g. CBT versus another psychotherapy of similar intensity) only needs to be better than chance to be at least a little useful. However, we may require better performance of models that aim to predict which patients require intensive treatment (e.g. combination therapy), since mistakenly assigning a participant to the intensive treatment exposes them to potential harms and will result in greater costs.

Predictive models should, whenever possible, be based on information that is easy to collect in clinical practice or that is already routinely collected [68]. The work in chapters 9, 10 and 11 of this thesis suggests that information about symptoms alone is unlikely to yield very accurate predictions, but other clinical information about, for instance, illness history, family history, basic sociodemographic information other than age and gender, and the like, could easily be added. Should such models still prove inadequate, other information could be added that is more difficult to collect, for instance from blood samples or ecological momentary assessment. Even neuroimaging (for instance, functional magnetic resonance imaging (fMRI)) could be used, although given the expense of neuroimaging, it seems unlikely that this will soon be applied to the full population of treatment-naïve patients. Hence, predictive models that use difficult-to-collect information should probably be preferentially tested in a patient population with a relatively poor prognosis (e.g. those presenting in secondary care), for whom the application of such methods might actually be cost-effective.

Besides their immediate utility in clinical practice, these predictive models, if sufficiently refined, may also have other uses. First, patients who respond to the same treatment may also have the same or a similar underlying etiology. The heterogeneity of depression and anxiety is considered one important reason why progress in discovering the underlying causes of these disorders has been slow [568, 569]. Identifying more homogeneous patient subgroups with predictive models may therefore aid efforts to clarify the etiology of depression and anxiety, which in turn may help to refine predictive models, potentially setting in motion a virtuous cycle. Secondly, identifying patients who respond well to

various treatments will eventually also identify those patients who do not respond well to any treatment. Once we can identify patients who are truly treatment-resistant, efforts to develop new treatments can be focused on this patient group rather than on the entire group of depressed or anxious patients, many of whom are already adequately served by existing treatments.

Concluding remarks

This thesis aimed to bring the evidence regarding the treatment of depression and anxiety to light. To do so, I first examined the impact of reporting and citation biases on the evidence base. This effort has helped to elucidate how the apparent efficacy and safety of treatments, especially antidepressants, has been inflated. Having clarified the safety and efficacy of treatments, the next step is to determine who benefits from (which) treatment, and my thesis examined several clinical predictors that could be used for this. However, this area of research is still very much in its infancy. The increasing awareness of bias, coupled with the increasing availability of individual participant data, will hopefully allow for the continued development of evidence-based, precision psychiatry in future research.