Stepped trials: Magnifying methodological muddles — the supernatant effect

Abstract

Stepped trials such as STAR*D suffer from a serious methodological flaw. In each stage of the trial one is dealing with a different population sample that has a different percentage of subjects with drug-treatable illness. This reality is magnified by the fact that the initial sample cannot be shown to be, nor validly assumed to be, homogenous. That is analogous to the positive protective value of a pathology test. Thyroid function tests have lower predictive value when done on a sample of patients in a general medical practice, but a higher predictive value when done by endocrinologists, because the incidence of illness in one sample is different compared to the other. This is directly analogous to the chance of a particular drug being efficacious, which depends on the incidence of treatable illness in the sample under consideration. Failure to recognise that simple fact invalidates most of the conclusions drawn from the STAR*D study.

Background

An aspect of the massive $35m STAR*D trial that I have not seen commented on is the issue of how the inevitable heterogeneity of the initial sample will magnify methodological muddles and produce discrepancies in the results of each progressive stage, or step. This is an explanation of this simple but important methodological flaw and mathematical artefact (the 'supernatant effect') which severely limits the validity of the findings from this study. The supernatant effect also illustrates why the concept of 'treatment resistant depression' is of little value and is fatally flawed.

First, a consideration of the logic and scientific methodology of RCTs in the psycho-pharmacology literature indicates that, from a purist scientific point of view, they suffer from serious limitations — the greatest of these is their impotence in regard to the questions of causality, or mechanism***:** these limitations are partly an inevitable consequence of the complexities of investigations of biological systems, but also a reflection of suboptimal design and methodology — this in turn is strongly influenced by the preponderance of drug company funded studies which require shorter, simpler, cost-effective results, and therefore adopt methods suited to satisfying the EMA and FDA, rather than methods aimed at elucidating other research questions using more suitable scientific approaches.

*****Causality and mechanism are the heart of real science

It is strange to be commenting on such a matter this far into the third millennium, and when the discipline has been doing RCTs for some 60 years (one might imagine that more improvement could have been achieved by now*****). I am not referring here to detailed objections concerning the appropriateness, or (in)sensitivity, of the rating scales used (their ratings of core depressive features like anergia and anhedonia are minimal, and ratings of functional outcomes are often non-existent), nor to the statistical analysis of the difference between active treatment and placebo (with most differences being so small that they are inconsequential), nor am I even asking for a proper conceptualisation and definition of what a placebo response really means: [spoiler] answer, not a lot.

*****To understand how little has been achieved, and how few answers are provided even by modern, supposedly better, RCTs, one only has to look at the recent Cipriani 21 ADs study [1]

A great deal has been written about this STAR*D trial both by the investigators themselves, more than 120 papers — which is puffery on steroids — and by

Address for correspondence: ken.psychotropical@gmail.com
Website: https://psychotropical.com

other commentators (PubMed, March 2019, returns 170 records with STAR*D in the title, and 350 that mention it). Anyone who has read all that bumf is probably now confined to a padded cell — this is academia gone mad, but a great source of income for the publishing houses.

It may remind you of the term 'intellectual masturbation'.

STAR*D consumed an estimated $35 million of public funds. It is, in part, an example of the problems that arise from political interference in the scientific process: Donald Klein DSc. stated (quoted with permission) in a note to members of my international MAOI expert group:

> I believe that it is not generally known that STAR*D was a contract rather than a grant. **It never had peer review of the protocol** by a group who could say don't do it and refuse to fund. As I recall it was one of four contracts that were put together hastily as a reply to accurate, but transient, Congressional criticism that NIMH was not doing any clinical investigation

It is entirely possible that, in the tsunami of writings about STAR*D, unbeknownst to me, someone else has made the same points as I make now: however, even if I was silly enough to try to read all 350 papers, I am sure I would be unable to obtain copies of many of them, without unaffordable financial outlay.

Moving on, before commenting further, I note that if this huge and mega-expensive trial ($35m Klein said) had shown **anything definite and useful** it would be possible to describe that concisely, and it would not have required library-shelves full of discussion — the very fact that it has generated so much bumf is strong presumptive evidence that its findings were doubtful and unhelpful.

If something clear and definite comes out of a study, then it can be summed up in one page — it is not going to take 120 papers to explain it. I do not suppose that I am the only one who finds this resonating with Shakespeare's sentiment:

> The lady doth protest too much, methinks

In a nutshell

The fundamental and elemental methodological flaw that I draw attention to, and that appears to have been overlooked, is simply this: given that a significant percentage of the initial sample are sure never to respond to drug treatment (because the diagnosis is either wrong, or they have a drug-non-responsive illness), then, inevitably, with each stage that removes patients that are responders, the remaining 'mis-diagnosed' patients will constitute a larger percentage of the remaining sample, and thus produce the artefact of making each successive treatment seem less effective. Such a flawed method is comparing apples with oranges. I call this the 'supernatant effect'.

As Rush observed [2]:

> Whether measured by HRSD17 or QIDS-SR16, each successive phase step was associated with a successively lower likelihood of both response and remission, **regardless of the type of medication** used at each step.'

The logic and validity of STAR*D, and all RCTs, assumes and requires a homogenous sample

Defining a homogenous sample of patients with 'biological' depressive illnesses, or 'drug-responsive' illness is difficult. The scientific logic and validity of such trials as STAR*D assumes and requires a homogenous sample, as indeed do all RCTs. Perhaps one could come to close a homogenous sample by looking at patients who have responded to ECT treatment. In contrast, recruiting patients

from primary care, and by advertisement, as STAR*D did, has to be about the worst possible way of approaching a homogenous sample of 'biological' depression, especially when using the unsatisfactory 'checklist-diagnosis' of the DSM.

As our eminent Australian commentator on matters 'endogenous', Professor Gordon Parker [3], has argued:

> Viewing major depression as a unitary [homogenous] entity — as against a non-specific domain diagnosis capturing heterogeneous expressions of depression — is a starting point for downstream non-specific results … Most antidepressant drug trials recruit out-patients and effectively exclude those with melancholic depression — the quintessential 'biological' depressive condition.

Exactly, thank you Gordon.

Thus, it is probable that such trials are recruiting a substantial percentage (possibly even a majority) of patients who do not have a 'drug-responsive' biological depressive illness. Response in such cases is substantially contributed to by regression to the mean in groups of patients who do not have a consistent and persistent biological depressive illness.

Furthermore, it may be noted that the so-called placebo response, which is a widely-touted concept — but an ill-defined one — is manifest in subjectively-rated symptoms, but not 'objectively' measured outcomes (like having a reduced blood pressure, or reduced uric acid concentration) [4-12].

As Parker states (I quote/paraphrase):

> The consequences of such flawed logic are that current foundations lack a firm base, and RCT practices generate limited valid evidence [13]. Baseline inflations for recruitment purposes, together with individuals' placebo and spontaneous improvement propensities, risk [magnified] regression to the mean, confounding responder status. Note that studies in the 1960s — in which antidepressants differentiated distinctly from placebos — were weighted to the melancholic subtype, and generated response rates of 60-70% to broad-spectrum antidepressant drugs, with placebo rates as low as 10% [3].

Preskorn is the only author I am aware of who has written about this supernatant effect phenomenon, he discusses another facet of this issue, relating specifically to SSRI trials [14].

This is a precise equivalent to the issue of the **sensitivity, specificity, and predictive value of laboratory tests**. The '**positive predictive power PPV**' of a test is dependent on the incidence of the condition in the population being tested. Akobeng explains this issue in part 1 on p.339 [15, 16].

The incidence of a condition in a population being blood-tested is analogous to the incidence of biological depression in the sample being treated

The incidence of the condition in question in the population being tested is analogous to the number of true cases biological depression in the sample being treated. This is a crucial concept to grasp, and I recommend anyone who is not confident they understand it to read Akobeng's article carefully. The PPV is analogous to the chance of responding to appropriate drug treatment — and just as PPV is dependent on the incidence of the condition in the population tested, so is response to treatment. If, as my extensive experience indicates, tranylcypromine works best in cases of melancholic and biological depression, then success in selecting these cases will determine the proportion who respond. If on the other hand 'atypical' cases are selected, with a lower preponderance of such features, then the success rate is lower — but that success rate is not

dependant on the efficacy of the drug in question: it is an artefact, the supernatant effect.

An analogy would be treating cases of presumed infection with antibiotics when a proportion of the cases in fact have a viral, not a bacterial, aetiology. If one did a stepped trial, and put people who failed to respond with the first antibiotic, onto a second different antibiotic, then it is immediately apparent that the sample generated in the second step would consist of a higher proportion of people who actually had a viral illness, which is not treatable by antibiotics — that problem would be further compounded in subsequent third and fourth steps using such a methodology. Obviously, the poor success rate of whatever antibiotic was used in subsequent steps would not be reflective of its ability to treat bacterial infections, it would merely reflect the fact that a greater proportion of the sample in question actually had a viral infection, artefactually reducing the apparent success rate: the supernatant effect.

Thus, the supernatant effect creates the self-fulfilling circular argument falsely 'showing' that MAOIs do not work, as happened in STAR*D

The mathematics involved in understanding STAR*D is precisely analogous and Bayesian reasoning can be utilized to produce a realistic estimate of the chance of treatment success — what all this tells us is that if one uses MAOIs for treating biological depression after only one failure (as was my habit), then one is inevitably going to have a higher success rate than if one uses them after three or more treatment failures (like in STAR*D). One might note that in the 1960s and 70s what is exactly what did happen, patients failed with the TCA then had either ECT or an MAOI — whereas now they are subjected to a range of half a dozen different antidepressants. Therefore, people are comparing what are essentially different patient samples over time.

All this is because the chance of treatment success (equivalent to the PPV) is largely dependent on the percentage of treatment-responsive patients in the sample in question, which inevitably becomes less with each stage in STAR*D.

Artefact, pure artefact — the supernatant effect

An additional consideration to add to this discussion would be that any practitioner who has an especially acute facility for diagnosing [drug-treatable] biological illnesses, as opposed to other depressive syndromes, will also have a higher success rate. Such a practitioner will effectively be treating a different population with a greater prevalence of true biological depression: hence that practitioner will have a higher success rate.

A more mundane aspect of the supernatant effect is the inevitable bias introduced in making specialist referrals. If a particular specialist has a reputation as being an expert in treating serious depressive illness, then it is reasonable to suppose they are somewhat more likely to receive suitable referrals, rather than people who obviously have major psychological issues, but may have similar levels of distress. I make this point particularly because people keep asking me, 'what percentage of your patients got better with Parnate' — the answer is, it depends on the type of referrals made to you, and how you then select them for treatment.

Modelling the 'supernatant effect'

If this mathematical 'artefact' is not understood and accounted for, it will distort the results and understanding of a study such as STAR*D: I call this phenomenon the 'supernatant effect'.

If we assume that 20% (and I would regard that as a conservative estimate) of the initial sample would never have responded (specifically) to drug treatment (i.e.

either did not have a depressive illness, or did not have a drug-responsive depressive syndrome), and if we then assume a certain modest percentage of remission with each of the four stages, it is immediately obvious that the percentage of inappropriately included, or 'non-responsive' patients will be an increasing percentage of the sample at each successive stage.

Someone with the appropriate skills could easily setup computer program to model this, with variable percentages and rates of drop out, to illustrate how ridiculous the comparison becomes once you get to the third or fourth stage of a stepped trial. It is so simple you could model one example in your head right now.

Therefore, to compare the relative success rates in different stages of the trial, or to suppose they reveal anything useful, is not methodologically correct.

Such modelling shows that the finding of a poorer response in later stages, e.g., with venlafaxine & mirtazapine, vs. tranylcypromine, is uninterpretable. The above argument is even more important than the other criticisms that have already been made concerning the invalid nature of the comparisons involving venlafaxine/mirtazapine and tranylcypromine in stage 4, relating to inadequate dosage, duration of treatment, different rating methodology (telephone interview only) etc. Indeed, there are so many problems related to assigning meaningfulness to those results that, had refereeing been thorough, they would never have been published.

I should also point out that the 'supernatant effect' illustrates why the concept of 'treatment resistant depression' is, from the standpoint of disease patho-physiology/treatment response, complete and utter nonsense — the above discussion illustrates that the patients who fall into that TRD category will depend largely on the order in which the different treatment choices are utilised.

References

1. Cipriani, A., T.A. Furukawa, and G. Salanti, *Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis.* Lancet, 2018. **391**(10128): p. 1357–1366 http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(17)32802-7/fulltext.
2. Rush, J.A. and B.S. Jain, *Clinical Implications of the STAR*D Trial.* Handb Exp Pharmacol, 2019. **250**: p. 51-99.
3. Parker, G. and V. Manicavasagar, *Modelling and Managing the Depressive.* Disorders: A Clinical Guide. Cambridge University Press, 2005., 2005.
4. Moerman, D.E. and W.B. Jonas, *Deconstructing the placebo effect and finding the meaning response.* Annals of Internal medicine, 2002. **136**(6): p. 471-476.
5. Macedo, A., M. Farré, and J.-E. Banos, *Placebo effect and placebos: what are we talking about? Some conceptual and historical considerations.* European journal of clinical pharmacology, 2003. **59**(4): p. 337-342.
6. Sumner, C.R., et al., *Does placebo response differ between objective and subjective measures in children with attention-deficit/hyperactivity disorder?* Postgrad Med, 2010. **122**(5): p. 52-61.
7. Wechsler, M.E., et al., *Active albuterol or placebo, sham acupuncture, or no intervention in asthma.* N Engl J Med, 2011. **365**(2): p. 119-26.
8. Hróbjartsson, A. and P.C. Gøtzsche, *Placebo interventions for all clinical conditions.* Cochrane database of systematic reviews, 2010(1).
9. Ramtvedt, B.E. and K. Sundet, *Relationships Between Computer-Based Testing and Behavioral Ratings in the Assessment of Attention and Activity in a Pediatric ADHD Stimulant Crossover Trial.* The Clinical Neuropsychologist, 2014. **28**(7): p. 1146-1161.

10. Hutchinson, P. and D.E. Moerman, *The Meaning Response, "Placebo," and Methods.* Perspect Biol Med, 2018. **61**(3): p. 361-378.

11. Dutile, S., T.J. Kaptchuk, and M.E. Wechsler, *The placebo effect in asthma.* Curr Allergy Asthma Rep, 2014. **14**(8): p. 456.

12. Macaluso, F.S., et al., *Factors Affecting Clinical and Endoscopic Outcomes of Placebo Arm in Trials of Biologics and Small Molecule Drugs in Ulcerative Colitis: A Meta-Analysis.* Inflamm Bowel Dis, 2018.

13. Parker, G., *Antidepressants on trial: how valid is the evidence?* Br J Psychiatry, 2009. **194**(1): p. 1-3.

14. Preskorn, S.H., *Declining differences in response rates with antidepressants versus placebo: a modest proposal for another contributing cause.* J Psychiatr Pract, 2013. **19**(3): p. 227-33.

15. Akobeng, A.K., *Understanding diagnostic tests 1: sensitivity, specificity and predictive values.* Acta Paediatr, 2007. **96**(3): p. 338-41.

16. Akobeng, A.K., *Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice.* Acta Paediatr, 2007. **96**(4): p. 487-91.