

THE CREDENTIALS OF SCIENTIFIC EVIDENCE: REBALANCING THE EPISTEMOLOGICAL SCALES

9.2.1 Generalisability, external validity

Abstract

This commentary, part of a series critiquing the hegemony of **Randomized Controlled Trials (RCTs)**, addresses the critical problem of ‘external validity’ in medical research. It argues that the stringent inclusion criteria required to construct ‘optimal’ RCTs inevitably create artificial test populations that are unrepresentative of real-world patients — most patients treated in clinical practice would be ineligible for the trials that generate their treatment guidelines. The text details specific factors that compromise generalizability — including setting, selection bias, and the use of proxy outcomes — and critiques the reliance on **Contract Research Organizations (CROs)** employing inadequately trained raters. It asserts that without demonstrated external validity RCTs are clinically all of uncertain and marginal meaning, leading to guidelines that are discordant with the realities of patient care.

Introduction

It is a *sine qua non* that the usefulness of RCTs is predicated on the assumption that the results can be expanded to cover a wider population (i.e. have ‘external validity’) — but which individuals? and in exactly what population?¹ External validity has never been well demonstrated and is often doubtful, as Sir Michael Rawlins pointed out in his Harveian oration [1]. The necessities of constructing a sample for an optimal RCT tend to make the test population imperfectly representative of the general population in which the treatment will subsequently be used — contrasting requirements that are difficult, or impossible, to achieve. As an example, asthma is a clinically heterogenous group with different clinical phenotypes, making it similar to depression — **only about 5% of patients being treated for asthma would be eligible for the RCTs** which use the GINA guidelines [2]. Likewise, it has been estimated that **75% of cases treated for depression would not be eligible for the relevant trials** [3].

The short duration of most RCTs, contrasted with the usually lengthy duration of real-world treatment, is also problematic. Other factors relevant to generalizability are discussed in detail by Rothwell [4], here is his table 1: ‘Main issues that can affect external validity and should be addressed in reports of the results of randomised controlled trials or systematic reviews and considered by clinicians’:

1. Setting of the trial Health-care system (Country, Recruitment from primary, secondary, or tertiary care, Selection of participating centres, Selection of participating clinicians)
2. Selection of patients (Methods of prerandomisation diagnosis and investigation, Eligibility criteria, Exclusion criteria, Placebo run-in period, Treatment run-in period, “Enrichment” strategies, Ratio of randomised patients to eligible nonrandomized patients in participating centres, Proportion of patients who declined randomisation)
3. Characteristics of randomised patients (Baseline clinical characteristics, Racial group, Uniformity of underlying pathology, Stage in the natural history of their disease, Severity of disease, Comorbidity, Absolute risks of a poor outcome in the control group)

¹ Many, indeed most, human populations are both heterogenous and unstable over time

PsychoTropical Commentaries 9.2.1 External validity

4. Differences between trial protocol and routine practice (Trial intervention, Timing of treatment, Appropriateness/relevance of control intervention, Adequacy of non-trial treatment—both intended and actual, Prohibition of certain non-trial treatments, Therapeutic or diagnostic advances since trial was performed)
5. Outcome measures and follow-up (Clinical relevance of surrogate outcomes, Clinical relevance, validity, and reproducibility of complex scales, Effect of intervention on most relevant components of composite outcomes, **Identification of who measured outcome**, Use of patient-centred outcomes, Frequency of follow-up, Adequacy of the length of follow-up)
6. Adverse effects of treatment (Completeness of reporting of relevant adverse effects, Rates of discontinuation of treatment, Selection of trial centres and/or clinicians on the basis of skill or experience, Exclusion of patients at risk of complications, Exclusion of patients who experienced adverse effects during a run-in period, Intensity of trial safety procedures)

Trials of depression rarely adequately address those issues that Rothwell lists. One stands out as especially relevant to the drug-company sponsored trials carried out by **contract research organisations (CRO)** — such organisations have for many years carry out almost all AD trials — that is item in Number 5, **'Identification of who measured outcome'**. These persons are not identified in clinical trial reports and are generally not experienced clinicians interviewing patients, they are persons of unknown qualifications and experience standing in front of a patient they know little about and have never treated (because they are not doctors) with a clipboard ticking squares against a list of rote questions². One doubts that is going to result in an accurate assessment of symptoms of depression, feelings of suicidality, or sexual side-effects unless one is living in a fantasy world. Furthermore, there is not just the issue of who measured it, but also the issue of inter-rater reliability and whether the measure that they are using is even a discriminating or valid one in relation to the symptoms of illness (cf. criticism of the HDRS [5]) and cf. the null hypothesis.

That difficulty creates considerable obstacles for AD trials.

Prophylaxis and the NNT: a special case

When RCTs are used to guide prophylactic treatment for large numbers of 'patients' — who are healthy individuals in the population, who may never experience the projected harm — many individuals are placed at risk of experiencing adverse drug effects when there is an imprecisely quantified chance that they will experience the projected harm or morbidity. That fundamentally changes the risk benefit ratio of treatment (the so-called **number needed to treat or NNT**) because you may have to treat twenty patients with statins over a long period of time in order to prevent one MI or CVA [6] — presuming one believes the NNT is a useful or meaningful measure [7]. Unfortunately, NNT is an overrated and simplistic measure of restricted use and validity, to the extent that Murad et al. suggest **'The number needed to treat (NNT) has statistical and methodological limitations that make it unhelpful in the context of developing clinical practice guidelines and policy decisions'** [7]. Yet we have in operation numerous health policies predicated on such methodologies which affect millions of 'patients', cf. the ME policy reversal by NICE [8], perhaps the most egregious RCT-related mistake this century — so far.

² One should note here that there have been many instances where even the recruited authors of studies, whose names have been placed on the paper after the ghostwriting is completed, have been refused access to the raw data that has been used to compile the report — they cannot reliably attest to the meaningfulness of assessments of suicidality, or any other symptom for that matter.

PsychoTropical Commentaries 9.2.1 External validity

The same general point applies to EBM, because, in 30 years, it has not been demonstrated that its application results in improved real-world outcomes when compared to usual clinical care. As Berg states, 'Evidence-based practice is out of touch with developments within philosophy of science' [9].

References

1. Rawlins, M., *De testimonio: on the evidence for decisions about the use of therapeutic interventions*. Lancet, 2008. **372**(9656): p. 2152-61.
2. Travers, J., et al., *External validity of randomised controlled trials in asthma: to whom do the results of the trials apply?* Thorax, 2007. **62**(3): p. 219-23.
3. Blanco, C., et al., *Generalizability of clinical trial results for major depression to community samples: results from the National Epidemiologic Survey on Alcohol and Related Conditions*. J Clin Psychiatry, 2008. **69**(8): p. 1276-80.
4. Rothwell, P.M., *Factors that can affect the external validity of randomised controlled trials*. PLoS Clin Trials, 2006. **1**(1): p. e9.
5. Byrne, D., et al., *Evaluating the psychometric structure of the Hamilton Rating Scale for Depression pre- and post-treatment in antidepressant randomised trials: Secondary analysis of 6843 individual participants from 20 trials*. Psychiatry Res, 2024. **339**: p. 116057.
6. Ridker, P.M., et al., *Number needed to treat with rosuvastatin to prevent first cardiovascular events and death among men and women with low low-density lipoprotein cholesterol and elevated high-sensitivity C-reactive protein: justification for the use of statins in prevention: an intervention trial evaluating rosuvastatin (JUPITER)*. Circ Cardiovasc Qual Outcomes, 2009. **2**(6): p. 616-23.
7. Murad, M.H., et al., *The number needed to treat: it is time to bow out gracefully*. Journal of Clinical Epidemiology, 2025. **187**: p. 111924.
8. Gillman, P.K., *Monoamine oxidase inhibitors: A paradigm of poor science*. Journal of Psychopharmacology, 2025. **0**(0): p. 02698811251381762.
9. Berg, H., *Is evidence-based practice justified?—A philosophical critique*. Journal of evaluation in clinical practice, 2024. **30**(5): p. 855-859.