Epidemiology and Psychiatric Sciences (2013), **22**, 105–109.
doi:10.1017/S2045796013000048

EDITORIAL

Pragmatic trials of complex psychosocial interventions: methodological challenges

G. Dunn*

Centre for Biostatistics, Institute of Population Health, University of Manchester, Manchester, UK

Having first introduced the pragmatic health care trial, the discussion then focuses on a selected list of technical problems that are important for the design, analysis and inference from such trials. The first is lack of independence of participants' outcomes do to clustering either arising from a cluster randomized design or to the way treatment is delivered (therapist and group effects). The second and third concern the implications of non-adherence to treatment and subsequent loss to follow-up, particularly, when non-adherence is associated with missing outcome data. Finally, it is argued that pragmatism and a desire for a scientific explanation should not be regarded as mutually exclusive.

First published online 5 March 2013

Key words: intention-to-treat-analysis, pragmatic trials, psychosocial interventions, statistics.

Introduction: health care trials

This Editorial will primarily be concerned with trials for complex interventions rather than the (relatively) more straightforward drug trials. As we all know, the randomized controlled trial is regarded as the 'gold standard' (in terms of methodological rigour) against which other forms of evaluation are to be assessed. In a traditional clinical trial the experimental conditions are usually competing therapies and the experimental subjects are individual patients. However, these days (in what might be referred to as a health care trial) the experimental conditions could be competing ways of providing a health care service and the experimental subjects may not necessarily be patients. They might, for example, be care providers, including psychiatrists, psychologists or nurses, managers or units of health care provision (clinics, wards or hospitals, for example). Spitzer *et al.* (1975) distinguish two types of health care trial: a *health service trial,* in which one assesses the mechanisms (or records) of health care provision and a *patient care trial,* in which one assesses conventional therapies but the clinical outcome variables are augmented by socio-personal data that is, the patient care trial is similar to the traditional trial but is distinguished from it through the use of the non-clinical outcome measures. In summary, both health service and patient care trials, the outcomes may include use of medical and other services, hospital admission, administrative problems, family burden, social functioning, quality of life, and so on. They might also include estimates of cost.

The key component of both the patient care and the health service trial is *random allocation* of participants (or clusters of participants) to competing interventions. Randomization serves three important roles. First, it is an impartial method of allocation of participants (patients; clinicians; services in which the intervention is being implemented) to the competing interventions. Second, it will tend to balance the intervention groups in terms of the effects of extraneous variables that might influence the outcome of an intervention. One might argue that it would be more effective to match or stratify on the basis of the extraneous variables. Stratification can be an important component of trial design, but it cannot cope with the extraneous variable(s) that no one has thought of. A stratified trial should still have random allocation within the strata. The third role for randomization is that it guarantees the validity of a subsequent statistical test of significance. If there are no intervention effects (i.e. the null hypothesis is true) then, apart from unforeseen or uncontrolled biases, the observed group differences must be the result of randomization (chance). One can simply ask 'What is the probability that the results have arisen solely as a result of randomization?' and decide whether the data are consistent with the null hypothesis accordingly. Perhaps more importantly, randomization (at least when everyone complies with their random allocation and provides complete outcome data) ensures intervention-effect estimates that are not subject to selection effects (confounding). These issues will be further clarified below.

* Address for correspondence: Professor G. Dunn, Centre for Biostatistics, Jean McFarlane Building (1st floor), Oxford Road, Manchester, M13 9PL, UK.

(Email: graham.dunn@manchester.ac.uk)

What makes a health care trial pragmatic?

Consider the spectrum of experimental investigations ranging from clinical laboratory studies and experimental medicine to explanatory clinical trials and then on to pragmatic health service and patient care trials. Schwartz & Lellouch (1967) introduced the distinction between *explanatory* and *pragmatic* approaches to clinical trial design, the former focused on evaluating efficacy under relatively tightly controlled conditions and using a relatively homogeneous and carefully defined population of trial participants, while the latter is aimed at evaluating (cost-) effectiveness of interventions in as realistic circumstances as possible. The pragmatic approach demands (or, at least, expects) heterogeneity among trial participants, heterogeneity in adherence to and the way in which the intervention is delivered. It is aiming to ask 'Does the intervention work in the 'real' world?' The implication of approaching a randomized trial from a pragmatic point of view is that we are expecting to increase the *external validity* of its findings – we are expecting our inferences to be *generalizable* to other settings and to other populations. An explanatory trial, on the other hand, is likely to have high *internal validity* – we can be confident that our inferences are valid for this particularly well-specified and controlled intervention on this precisely defined population of participants – but we do not have a clue whether the findings will be replicated in other populations under different circumstances. So, as we move from a more explanatory approach (it is a matter of degree) to a more pragmatic one we are aiming to increase generalizability, that is, to increase our ability to predict the relative performance of our intervention in a wide variety of clinical settings. However, we have to acknowledge that this admirable approach will come at a cost – we will have to accept that as we become more pragmatic we are increasingly likely to be losing experimental control. We are increasing external validity but putting internal validity at risk. This lack of control is the source of several technical challenges for the design, analysis and interpretation of pragmatic trials and a selection of these will be the focus of the rest of the present Editorial. There will be no attempt to be comprehensive.

Heterogeneity in delivery of the intervention and associated clustering effects

We start with a technical problem that will be fairly familiar to readers – lack of independence of, say, patient outcomes arising from the nested (clustered) nature of the data. This is exemplified by the cluster-randomized trial in which, for example, different therapists or services are randomly allocated to the alternative interventions. Therapists will be heterogeneous in the context of their training, experience and skills. Different services will vary in their clinical and administrative efficiency. The nested (clustered) nature of the trial can be allowed for in the sample size/power calculations at the design stage, and in the analysis of the trial outcomes (Donner & Klar, 2000). If clustering is ignored at the design stage the trial will be under-powered. If clustering is ignored at the analysis stage the results will be spuriously precise (apparently statistically-significant when they are not). What is less well known is that the same problems arise in trials in which therapy or other health care intervention is delivered by heterogeneous therapists, irrespective of whether the trial involves a clustered design. A particularly interesting example arises when one arm, say, of a two-armed trial, involves therapy being delivered to groups of patients (group therapy) while in the other there is just treatment as usual (i.e. no therapy in this sense) or a therapy (either similar to the first, or quite different) delivered individually. Here, in the prior sample size calculations and in the analysis of the trial results, adjustments need to be made differently in the two arms of the trial. One implication is that 1 : 1 allocation is unlikely to be the most efficient design (see Roberts & Roberts, 2005) typically, one increases the sample size to allow for clustering and in this case this increase will only apply to the clustered group therapy arm.

Non-adherence and intention-to-treat (ITT): analyse as randomized

Not everyone in a trial, particularly a less controlled, large multicentre pragmatic trial, will receive the intervention to which they have been allocated. This may not be the fault of the participant – it certainly does not necessarily imply delinquency on their behalf, but could be a clinical necessity (as determined by the treating clinician) or an administrative failure. Similarly, service providers may fail to fully implement an intervention as intended by the trial investigators. Even the most pragmatic of investigators might wish to explain, say, the apparent ineffectiveness of a trial intervention by asking whether one possible explanation might be non-adherence to treatments or inadequate implementation of service reforms.

Creed *et al.* (1997) describe what we will refer to as a Day Care Trial to compare the effects of day and inpatient treatment of acute psychiatric illness. Randomization produced 93 inpatients and 94 day patients. We then have the following sequence of events, illustrating that there are two types of non-compliance

https://doi.org/10.1017/S2045796013000048 Published online by Cambridge University Press

with (or non-adherence to) the randomized treatment allocation:

> 'Eight were excluded because of diagnosis or early discharge, leaving 89 inpatients and 90 day patients. Five randomized inpatients were transferred to the day hospital because of lack of beds, and 11 randomized day patients were transferred to the inpatient unit because they were too ill for the day hospital.' (Creed *et al.* 1997, p. 1382).

Other, more familiar examples will involve patients simply not turning up for their allocated therapy ('no shows', as Bloom, 1984, refers to them). Of course, non-adherence may not be all or none – some patients will turn up for a few sessions of therapy and then drop out. The pragmatic approach to the analysis of the outcome of the trial under these circumstances is to evaluate the effect of random allocation (i.e. the offer of the treatment or service) to evaluate whether the decision to offer a treatment, service or other intervention produces results that are unlikely to be explained by chance. This is the ITT approach – that is analysis as randomized. Here, no attempt is made to evaluate the effect of receiving the prescribed treatment. The great advantage of this approach is that it is evaluating the effectiveness of a treatment decision, that effectiveness being influenced by uptake and fidelity to therapy under test. The second great advantage is that it does not involve any special pleading or statistical adjustments dependent on untestable assumptions that may be open to challenge. If, however, we feel that we can defend a few necessary assumptions then we might wish to *supplement* (*not* replace) the ITT analysis with an attempt to estimate the average intervention effects in those participants who complied with their allocation.

So, how do we estimate the effect of actually receiving day care in the above trial? We could compare the outcomes in those patients who were both allocated to and received day care with the outcomes in those patients who were both allocated to and received hospital care (the so-called Per Protocol estimate). Another option would be to ignore randomization altogether and compare the outcomes in those who received day care with the outcomes in those receiving hospital care (the As Treated estimate). Both the Per Protocol and As Treated estimates are subject to potential selection effects (confounding) and are likely to be invalid. A third option is the ITT estimate in the subgroup of patients who always comply with their treatment allocation – that is, the estimate of the Complier-Average Causal Effect or CACE (Bloom, 1984; Sommer & Zeger, 1991; Angrist *et al.* 1996). This is not subject to confounding, and is estimated assuming that the following assumptions hold: (1) there are three groups of patients – those who always comply with their allocation (compliers) and two types of non-compliers (those who are admitted to hospital regardless of randomization and those who receive day care, again regardless of randomization); and (2) there is no effect of random allocation on the outcomes in the two types of non-complier. Under these assumptions the CACE is estimated by the ITT effect on outcome (the arithmetic difference between two means, say) divided by the ITT effect on receipt of day care (estimated as the arithmetic difference between two proportions). In a cluster-randomized or similarly-nested trial non-adherence may be a little more complicated and far less straightforward to deal with. There may be non-adherence at the level of the clusters (services fail to or inadequately implement a planned reform) or at the individual level (patients fail to turn up for their therapy). Detailed discussion is well beyond the scope of the present editorial, and we refer the interested reader to Jo *et al.* (2008*a*, *b*) and to Schochet & Chiang (2011).

Convinced pragmatists might react in horror to our attempts to carry out an explanatory analysis based on treatment receipt (and they often do – particularly if they are what I think of as traditional (conservative) trial statisticians!). 'Stick to ITT' they would say, which might be fine if we were to observe complete follow-up data. We firmly believe, however, that all pragmatic trials should have an explanatory component (we will return to this later). However, even the committed pragmatist – the committed ITT estimator – should acknowledge that there are situations in which we should carefully consider the implications of non-adherence to treatments or failure of implement more complex interventions. For example, if we have missing outcomes, and having a missing outcome is associated with previous treatment dropout (or switching treatments or failure to implement a health care intervention), then ignoring treatment receipt in the analysis might lead the pragmatist's ITT estimates seriously astray. This we cover in the following section.

Loss to follow-up (missing outcomes)

Most readers will be aware of the possible havoc created by missing outcomes for the estimation of intervention effects. The problem arises from the fact that missing data are unlikely to arise completely at random. It is not just a case of losing precision (or statistical power) but a potential source of serious bias. In my experience of analysing the results of trials of psychological interventions over the last few years, the most powerful predictor of loss to follow-up is frequently the participant's failure to adhere to their allocated therapy (e.g. they fail to turn up for their allocated therapy, or drop out of therapy prematurely – see, for example, patterns of missing outcome data in the

https://doi.org/10.1017/S2045796013000048 Published online by Cambridge University Press

ODIN trial – Dunn *et al.* 2003). One relatively simple way of eliminating or reducing the bias in ITT estimates that arises from non-compliance is the use of inverse probability weights in which the weights are the reciprocal of the probability of providing an outcome measurement, conditional on both randomization and treatment received (e.g. the use of these weights to adjust for loss to follow-up in the ODIN trial, see Dowrick *et al.* 2000). However, this approach may not be based on valid assumptions concerning the missing data mechanism. Missingness, for example, may be related to *latent* compliance class – as in CACE estimation – rather than on observed treatment status (a missing data mechanism labelled as 'Latently Ignorable' by Frangakis & Rubin, 1999). Here, one possible approach is to simultaneously model the effects of randomization and latent compliance class on provision of non-missing outcomes and on the outcome, itself. This will provide a valid CACE estimate which can then, in turn, lead to a valid ITT estimate via the relationship $ITT = P_C \times CACE$, where P_C is the estimated proportion of latent compliers in the trial. Again, technical details of estimation of intervention effects involving the use of missing data models assuming latent ignorability are well beyond the scope of the present Editorial, and interested readers are referred to Dunn *et al.* 2005 for further details (again, illustrating the methods using ODIN data). The important point is that in order to obtain a valid ITT effect one has to first estimate the CACE and then calculate the ITT effect from this.

Explanatory models: learning more from pragmatic trials

It is unfortunate that the words *pragmatic* and *explanatory*, as applied to clinical trials, appear to imply mutually exclusive or polar appositives. Pragmatism and understanding of mechanisms are not mutually exclusive. Good complex intervention trials, for example, should be able to answer *both* pragmatic questions (Does it work? Is it cost-effective?) *and* explanatory ones (How does it work? Why does it *not* appear to work?). It is important that these trials explicitly consider how and why the treatments and other interventions work clinically or economically. There is no reason why improving both the design and analysis of a trial to answer the explanatory questions of scientific interest should in any way compromise its ability to answer the pragmatic one. At its best the complex intervention trial will be a sophisticated health care experiment designed to test the theories motivating the intervention and also help understand the underlying nature of the clinical problem being treated. Psychological treatment trials, for example, almost always involve the collection of a very rich set of multivariate outcomes. Rarely is it satisfactory to insist that there should be one simple primary outcome. Although these complex intervention trials may be large, it would be a mistake to routinely aim to make them simple.

However, there is much scope for improvement in the way in which the multivariate outcomes are analysed and interpreted. Through the intelligent analysis of theoretically motivated and clinically face-valid mediating and/or moderating effects, we can obtain important insights about the way in which psychotherapies, say, do or do not work – through the linkage of intervening variables in a chain of potential causal effects. The most informative trials will involve the careful choice of potential mediators and associated outcomes, together with novel designs to maximize our ability to evaluate the alternative causal clinical and economic pathways. Investigators frequently try to answer these explanatory questions – but too often the methods used to answer them are naïve and potentially flawed. The problems mainly arise from the fact that the intermediate outcomes of randomization (adherence to the intervention, the fidelity of the delivery of the intervention, potential mediators of the intervention's effects) are not under the direct control of the investigator. The naïve and potentially flawed approaches to the analysis of mediation and other forms of process evaluation are based on the assumption that there are no hidden common causes of these outcomes other than the intervention – that is, there is no hidden confounding of the effects of the intermediate outcome(s) on the final one. Usually these assumptions are implicit (i.e., not acknowledged by the investigators – in fact, investigators are frequently unaware that they are making them and equally ignorant of their implications). The challenge is to carry out the explanatory analysis in a defensible and valid manner. It is not an easy task. Some progress has been made using modern statistical and econometric methods (particularly the use of instrumental variables) – see, for example, Dunn & Bentall (2007) and Emsley *et al.* (2010) but real improvements will only come with the development of new trial designs focused on both efficacy and mechanisms evaluation. Methods to answer the explanatory questions need to be considered fully at the design stage; at present, they are too often considered only as an afterthought (often a *post mortem*, a desperate attempt to understand why the trial has failed to demonstrate the anticipated benefits of the intervention).

Conflict of Interest

None.

https://doi.org/10.1017/S2045796013000048 Published online by Cambridge University Press

Financial Support

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

References

Angrist JD, Imbens GW, Rubin DB (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* **91**, 444–455.

Bloom HS (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* **8**, 225–246.

Creed F, Mbaya P, Lancashire S, Tomenson B, Williams B, Holme S (1997). Cost effectiveness of day and inpatient psychiatric treatment: results of a randomized controlled trial. *British Medical Journal* **314**, 1381–1385.

Donner A, Klar N (2000). *Design and Analysis of Cluster Randomised Trials in Health Research*. Edward Arnold: London.

Dowrick CF, Dunn G, Ayuso-Mateos J-L, Dalgard OS, Page H, Lehtinen V, Casey P, Wilkinson C, Vázquez-Barquero J-L, Wilkinson G, the ODIN Group (2000). Problem-solving treatment and group psycho-education for depression: a multicentre randomized controlled trial. *British Medical Journal* **321**, 1450–1454.

Dunn G, Bentall R (2007). Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statistics in Medicine* **26**, 4719–4745.

Dunn G, Maracy M, Dowrick C, Ayuso-Mateos JL, Dalgard OS, Page H, Lehtinen V, Casey P, Wilkinson C, Vázquez-Barquero JL, Wilkinson G, and The Outcomes of Depression International (ODIN) Group (2003). Estimating psychological treatment effects from an RCT with both non-compliance and loss to follow-up. *British Journal of Psychiatry* **183**, 323–331.

Dunn G, Maracy M, Tomenson B (2005). Estimating treatment effects from randomized clinical trials with non-compliance and loss to follow-up: the role of instrumental variable methods. *Statistical Methods in Medical Research* **14**, 369–395.

Emsley R, White IR, Dunn G (2010). Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Statistical Methods in Medical Research* **19**, 237–270.

Frangakis CE, Rubin DB (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment noncompliance and subsequent missing outcomes. *Biometrika* **86**, 365–379.

Jo B, Asparouhov T, Muthén BO (2008*a*). Intention-to-treat analysis in cluster randomized trials with noncompliance. *Statistics in Medicine* **27**, 5565–5577.

Jo B, Asparouhov T, Muthén BO, Ialongo NS, Brown CH (2008*b*). Cluster randomized trials with treatment noncompliance. *Psychological Methods* **13**, 1–18.

Roberts C, Roberts S (2005). The analysis of clinical trials with clustering effects due to treatment. *Clinical Trials* **2**, 152–162.

Schochet PZ, Chiang HS (2011). Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Educational and Behavioral Statistics* **36**, 307–345.

Schwartz D, Lellouch J (1967). Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Disease* **20**, 637–648.

Sommer A, Zeger SL (1991). On estimating efficacy from clinical trials. *Statistics in Medicine* **10**, 45–52.

Spitzer WO, Feinstein AR, Sackett DL (1975). What is a health care trial? *Journal of the American Medical Association* **233**, 161–163.