On the Randomisation Gallery:
Random Reflections on a Popular Method in Medical Research

Jochen Hardt
Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Mainz, Germany

Abstract

Objective: Randomized controlled trials (RCTs) and meta-analysis are increasingly utilized methods in medicine. There is, however, substantial critique on both from the theoretical as well as from the practical side.

Methods: Seven points on RCTs and three on meta-analysis were chosen to summarize the current critique. Studies critically reviewing the quality of meta-analyses in various fields of medicine were narratively reviewed.

Results: The typical RCT is short, performed on a small group of selected patients without severe comorbidity. It is expensive and favors drugs over other forms of therapies. Furthermore, neither patients nor doctors much like RCTs, they feel restricted in their choice of treatments. Finally, it can be problematic to obtain informed consent in disorders affecting decision making, e.g. acute psychosis, severe depression or dementia. The evaluation of existing meta-analyses varied between acceptable and poor, good judgments were not often found.

Conclusion: The current uncritical use of RCTs and meta-analyses in medicine is not appropriate. Doctors using meta-analyses simply thinking to gather information efficiently are often unaware that there is evidence for bias.

Keywords: Randomized controlled trials; Meta-analysis; Quality assessment

Received: April 07, 2016; Accepted: May 26, 2016; Published: June 05, 2016

A Poem

If a shy, inexperienced, junior doctor were forced by a pitiless, selfish director to perform a nonsensical, randomized clinical trial over months and months, in the context of an overtaxing, crowded and unceasing clinical routine, spending evening after evening in the clinic, modulating patients’ diagnoses to allow them to take part in the trial, just in order to survive, trying to collect the data in time, knowing that it is expected of him to prove that this particular drug or treatment is better than a placebo, and if this game were continued to be played, constantly repeated for new generations of shy and inexperienced junior doctors, accompanied by the waning and rising financial support from public grants, which are actually sledgehammers, maybe, one day, a spirit of good sense will fly through the long corridors of the clinics and find a way into some scientific journal to shout ‘stop’.

Since that is not the way it is, however. A young, enthusiastic researcher, honest and intelligent, is proud to take part in the scientific community; his boss, an experienced senior researcher, takes all the time necessary to supervise him, to negotiate between his needs and those of the pharmaceutical industry; without any manner of personal interest, he tries to find out the very truth; finally, he offers assistance while writing down the results, respecting the work his junior colleague has done, integrating all the new and old knowledge in order to give each patient an optimal and individual kind of treatment.

Since that is the way it is, sometimes patients lean their faces against the clinic walls and, falling into the stupor of trust in modern science, one or another starts crying without knowing it [1].

Background

The discussion on randomized controlled trials (RCT) seems to be polarized. On the one hand side, performing RCTs is an important part of the Zeitgeist. Some methodological purists regard RCTs as the only way to evaluate causal relationships in medicine. Even just the word RCT almost seems to be a key required to publish
a result [2], for example, randomized patients into four groups, only to give each of them a questionnaire about quality of life. As a result, they could compare means of the four subgroups. If they had given all four questionnaires to all subjects, they would have had more statistical possibilities to analyze their data. However, the study could not have been labeled ‘randomized’[1]. Only few exceptions where observational data may be valid are made by the purists, e.g. when rare side effects of a treatment are reported. Evidence-based medicine has emerged from its "anti-authoritative" initial phase into an established standard [3]. Renowned journals now require their submissions to fulfill the CONSORT or the STROBE statements [4, 5]. On the other hand, there is a group of people who are rather skeptical about randomization generally. ‘If an intervention shows an effect, it can be seen anyway’, is essentially how they argue. According to their opinion, any effort on an RCT is a lost investment because it yields no more than what can be observed anyway without randomization.

Both such extreme views are wrong. The introductory poem is meant to illustrate that. Obviously, the idea behind randomization is a good one. Handling potential confounders by distributing them among certain conditions is obviously clever. When seeds are randomized onto plots of fields, substances onto cells or even animals, there is not much discussion about it. Problems occur when the subjects of randomization are humans [6]. We have our own will, which may interfere with the randomization. The aim of the present paper is to summarize critique on the randomized study from a practical point of view. It will be divided into seven points.

To examine the first two points, the Cochrane library was searched in June 2014 utilizing the key phrase "chronic back pain". This medical condition was chosen because its treatment includes drugs as well as non-drug measurements. Out of 61 reviews found, 26 were excluded (11 did not primarily deal with back pain, 8 were protocols, 4 were withdrawn, 3 were not RCTs). The remaining 35 were screened for the number of patients per group and for length of follow-up in the included individual studies. Up to four treatment arms were coded. If there were more arms, those with the smallest numbers of patients were ignored. If a control group was present, it was always included. Out of a total of 650 studies, 68 dublets were deleted, leaving 582 studies. To check for time trends, studies were grouped into three periods: before 1990, 1990-99, 2000 and later.

Seven Critique Points on the RCT

1) The typical RCT is short. There are various reasons for this, e.g. researchers want to obtain results fast, or patients would hardly stay in a trial that went on for years, etc. However, many disorders that modern medicine faces are long-lasting or recurrent, for example most mental disorders or chronic pain. So, there is a discrepancy between the results obtained by RCTs and the individual needs that patients later have. Of the 582 studies in Cochrane reviews dealing with chronic back pain, 511 reported the duration of a follow-up. The median duration was 3 months.

If one would arbitrarily define that a study on a chronic condition, as back pain usually is, should have a follow-up of minimally 6 months, 60% out of the present 511 studies would be excluded. The distribution of the duration is displayed in Figure 1. Studies lasting several years are rare. But the treatment patients later receive often will last years, or even decades. Fortunately, there was a significant time trend (Kruskal-Wallis test, p<0.001). Studies conducted before 1990 were shorter than those in the 1990s or after the millennium (median duration 1, 3 and 3 months respectively).

2) The typical RCT is based on a small sample size. Of course, large trials exist, particularly in cardiology. However, the average sample sizes in various studies are small [7]. There was a time when the pharmacological industry made use of small samples. Instead of giving one research group a lot of money to test a new drug, grant funds were divided into several groups and each of these conducted a small trial. The result then was a distribution of effects, and out of these, the best were selected for advertisements. This time is (almost) over, since trials now have to be registered prior to publication. Still, trials are often planned with small samples already or they have recruitment problems later on, so that they become too small in the end. In the 582 studies in Cochrane reviews dealing with chronic back pain, a total of 1348 groups were identified. The median sample size per group was 36. The distribution of the group sample sizes is displayed in Figure 1. Again, there is a time trend (Kruskal-Wallis test p<0.001), studies conducted before 1990 were smaller than those in the 1990s or after the millennium (median group sizes 25, 41 and 38 respectively).

3) The typical RCT is performed on selected patients without severe comorbidity. Serious comorbidities may partially mask the effects of the treatment under investigation. Therefore, inclusion and exclusion criteria are specified to increase internal validity. Later, though, the treatment will be used for all patients. In extreme cases, the participants of the RCT and the final recipients of the treatment differ so strongly that the effectiveness of the treatment for the latter group may be poor [8, 9], for example,
reported about a trial where only 0.5% of the patients seen in a clinic fulfilled the inclusion criteria. To transfer the results of this RCT to the remaining 99.5% should obviously be viewed critically.

4) The typical RCT favors drug therapy. The three key elements, randomization, double blinding, and creating an appropriate control condition is usually easy to realize in drug trials—verum and placebo can be produced in the form of similar pills and distributed randomly to the patients. Regarding various other forms of therapies, e.g., psychotherapy, surgical interventions or lifestyle changes like dieting, there are more difficulties when trying to randomize, to blind or to find an appropriate control condition. Concerning the therapy of mental disorders, we currently see an effect of an increase of pharmacological treatments [10]. Psychotherapy studies are simply not always feasibly as RCT [11]. Focusing too strongly on the method ‘RCT’ leads to favoring of certain forms of therapy, not necessarily those best for patients. For example, by running a register for controlled trials, the Cochrane collaboration supports drug therapy, as do journals by publishing only studies using an RCT when treatments are to be compared.

5) Neither patients nor doctors much like RCTs, they feel impaired in their freedom of choice of treatments. They feel that an optimal treatment should be individualized—which is not possible within the regime of a randomized trial. Even if aspects of the individual treatment are not evidence-based, the regime of the trial is often perceived as a limitation of their freedom of treatment choice. Clinicians in psychiatry or chronic pain treatment are familiar with this problem. One can argue that from an economic standpoint, in times of small budgets, comforting patients and their doctors at the expense of possibly inefficient treatments is a questionable luxury. Regardless of what standpoint one may take, everybody conducting an RCT should be aware that some colleagues involved may be against it and may circumvent the regime. Even strong measures of control will only partly counteract. A related problem is attrition [12]. For example, Bassler et al. [13] randomized 132 patients either to in- or outpatient psychotherapy. It turned out that the attrition rate in this study was about 50%. Interestingly, Bassler et al. asked the patients before randomization about their preference—to be treated as in- or outpatient. Analyses of attrition basically showed that patients stayed in the trial whose preference were met by randomization. The others probably went elsewhere. From a patient’s view, this is a consequence of their legitimate efforts to take control over their fates, but from the view of the researcher, randomization became absurd in this case.

A further aspect is blinding or double blinding. Rosenthal [14] demonstrated that expectations can strongly influence our perception. Therefore, theory tells us to conceal the treatment form from both patient and doctor. This is not always easy. “Any doctor fool enough to think that patients are passively obedient should recall that memorable clinical trial in which the subjects broke the code by discovering that when they threw their tablets down the lavatory, the drug floated but the placebos sank” [15].

A drug for a mental disorder or a pain killer would have to be very poor if the truth is not discovered by at least some patients. With the doctors, it is even worse. Who really believes it is possible to conceal the new and the standard drug or placebo from an experienced colleague? Side effects alone will soon tell who has received what. Some ideas to deal with this problem exist. The placebo can be made bitter or side effects like getting a dry mouth can be installed. However, such methods are rarely applied.

6) The typical RCT is expensive. Costs vary strongly depending on the design [16, 17] but an amount of Euro 4000 or more per patient is not unusual—only for the research, not including the treatment costs. Pressure to reduce costs has led to the phenomenon that currently more and more studies are being conducted in developing countries [15, 18] where it is unclear whether the probands would ever be able to buy the substances they have tested. This is not only politically incorrect, but may also lead to erroneous results. In addition, the high costs of RCTs also promote drug therapy. Large pharmaceutical companies are often better able to cover such costs than surgeons, physiotherapists or psychotherapists, to name a few.

Finally, it can be problematic to obtain informed consent in disorders affecting decision making, e.g., acute psychosis, severe depression or dementia.

To sum up, there is often a discrepancy between the ideal RCT and the ones that are performed in practice. Redmond et al. [19] compared 333 published articles with 227 corresponding protocols that were submitted to the ethics committee beforehand. Of 2966 outcomes, 870 (29.3%) were previously reported differently. Some protocols defined outcomes not reported on in the articles, and some articles reported outcomes not defined in the protocols. Clearly, the researchers changed their mind after having conducted the study. This may have well justified reasons, or it may not. In addition, the idea to change traditional medicine into evidence-based medicine by relying on RCTs alone will remain a dream for a long time. Kazdin [20] made an example calculation for psychotherapy with children and adolescents. There were about 550 types of psychotherapy and about 200 clinical disorders in DSM-IV applicable to children and adolescents. Testing all therapies for all disorders would result in 110,000 clinical trials—an impossible enterprise. Extrapolating these numbers to all of medical practice is difficult, because naturally not every therapy would be considered for every disorder. However, there are many more diseases and constantly new therapies are emerging. Hence, all of us would need to participate in various trials, some of them lasting decades, to return optimal results.

Meta Analysis

Is meta-analysis (or systematic review, the two are not differentiated here) the solution to overcome the problems with RCTs as cited above? Meta-analysis searches are done to summarize information efficiently and effectively and increasing the precision of estimates by pooling. However, critique has been raised against its current practise as well. It can be summarized into three key points.

1) Study selection. Usually, electronic searches yield many results, but only few studies meet inclusion criteria. One question is whether to include as many studies as possible into the meta-analysis or to focus only on “the best” ones—however that is defined. Results of the two strategies can differ strongly, as it was the case for evaluating the effects of homeopathy [21]. Meta-analysis itself is essentially an observational study and therefore sensitive against selection bias [22]. An extreme example of a
selection problem displayed a meta-analysis by Maniglio [23], who selected only four studies out of 20,502. These four were themselves meta-analyses; hence Maniglio called his paper a systematic review of reviews. Obviously, wide screening criteria were combined with narrow selection criteria. The opposite method is generally considered to be more adequate [24]

2) Once the studies are selected, data needs to be pooled. Here, it is often criticized that pooling can be equated to combining apples and oranges [27]. In fact, different researchers usually use different measures. Combining them by calculating effect sizes and pooling can even be regarded as an advantage. A treatment that shows success in various outcome measures would generally be considered to be superior to one demonstrating success for only a single outcome criterion. The crux of such a view is that outcome parameters and therapies that are to be compared should not be confounded. And that is rarely the case. If, for example a certain therapy would traditionally use one outcome criterion, and another therapy a second, the one using the criterion more sensitive to change would tend to look better than the other.

3) Pooling is not always easy and can be performed differently [28]. Among other aspects, there is a discussion on whether studies should be treated as fixed or random effect. Logically, it is sound to treat them as random, because often one finds heterogeneity in the results without being able to explain it. But the method currently most used in that case is the DerSimonian et al. [29] which puts relatively small weights on large studies, so that statisticians sometimes prefer to use a fixed effect model.

Satisfaction with existing meta-analysis

There are some studies examining the quality of meta-analyses in specific fields. Johnson et al. [30] examined 33 meta-analyses on how physical exercise can lower blood pressure (BP). They conclude that the "meta-analyses have contributed less than ideally to our understanding of how exercise may impact BP, or how these BP effects may be moderated by patient or exercise characteristics. Haidich et al. [31] examined 499 meta-analyses of various drug interventions. They found that about 80% were not all-inclusive. Most had a narrow scope and focused on particular agents. This was especially true for meta-analyses with industrial sponsoring (Odds Ratio=5.28 for industrial sponsored versus not). Goodyear-Smith et al. [32] examined five meta-analyses for a depression screening. They were performed by two different author groups and "consistently reached completely opposite conclusions. Justification for inclusion or exclusion of studies was abstruse" (p: 1). Ipser et al. [33] examined 27 meta-analyses on anxiety disorders. The authors claimed that the overall quality was acceptable. However, particularly weak points were reporting bias and criteria used for the assessment of study quality–two quite important aspects.

Minelli et al. [34] performed a review of 120 meta-analyses exploring genetic associations. Since the publication of the first meta-analysis for a genetic association in 1993, the number of meta-analyses in genetics reported on Medline has risen to more than 100 per year since 2006. The authors found some improvement in newer meta-analyses compared to those performed before 2000, but they concluded that “the quality of the handling of specifically genetic factors is disappointingly low and does not seem to be improving” (p: 1338).

Lundh et al. [35] rated the quality of 117 systematic reviews from pediatric oncology on a 1-7 point scale. Except for Cochrane reviews (which received a good rating: Median=6), the overall quality was poor: The median was only 1. Particularly poor were aspects such as “avoiding bias” and “validity assessed appropriately”. Similar results were obtained in urology [36], oral mucositis [37] retinopathy screening in diabetes [38] likely many other areas. There is, however, serious critique about the use of rating scales to assess the quality of meta-analyses. Herbison et al. [39] tested 45 scores to assess quality of meta-analyses and could not find any that could be considered optimal. Rather, the authors point out that different scores identify different studies as good.

Additionally, it could be demonstrated that financial ties to drug companies influence the evaluation of meta-analyses [40-43]. The numbers reported by industry-sponsored vs. non-industry-sponsored reviews were mostly similar, but the conclusions that were drawn from them tended to be more positive in the industry sponsored meta-analyses.

It is sad to summarize that only one of the groups that analyzed the quality of systematic reviews in a certain medical field was satisfied, the others claim poor or even very poor quality for most meta-analyses. Twenty years have passed since Feinstein [44] criticized meta-analysis as “alchemy for the 21st century”. Forty years have passed since Cochrane [45] wrote his book about health services, which promoted the use of randomized controlled trials with which meta-analyses came into medicine. Some improvement can be seen in the intervening time, as the example of back pain and most conclusions of authors examining meta-analyses demonstrate—but it is not much. In particular, we have to realize that meta-analysis is often performed poorly, and also be abused for industrial gain.

Why not simply return to observational studies?

The general critique of observational studies is that there may be confounders biasing the results. Such critique is justified; a well-known example is the 25-year long research about the protective role of antioxidative vitamins in the development of coronary artery diseases. A large observational study found an effect; a
large RCT did not [46]. However, randomization only protects against direct confounding. Wermuth et al. [47] described indirect confounding that can occur in observational and randomized studies as well. They have clearly worked out what can happen when we marginalize over some variables and condition on others. Both are always the case in a randomized study—its basic principle is to marginalize over all not considered variables. Specifying inclusion/exclusion criteria exactly mean conditioning on other variables. Some studies exist where similar treatments evaluated via RCT and observational studies [48-52]. Contrary to the example of antioxidative vitamins, they did not confirm the idea of a large overestimation of the effects in observational studies. However, some differences did exist. Interestingly, it was not the treatment groups that differed strongly, but the control groups in observational studies showed poorer prognosis than those from RCTs [48].

**Limitation**

Some of the critique points presented here do not solely apply to the RCT, but can constitute a problem in observational studies as well.

**Conclusion**

This paper should not be misunderstood as a plea to return to poor observational studies. We need good studies, randomized as well as observational. It has been stated a long time ago already that randomised and observational studies have much in common [53]. Admittedly, avoiding bias may be considerably harder in observational than in randomised studies [54]. Newly developed and highly complex statistical methods may help to do so [55]. The idea that even a poor RCT is better than any observational study is an over-simplification. It does not reflect the variety of research questions. Evidence-based principles have become established in medicine. They are even at risk of becoming a new form of under-questioned authority—which is strange considering that a reason for the introduction of evidence-based medicine was to overcome unjustified opinions of old authorities [56]. We have to be aware that simply relying on meta-analyses or RCTs is not a guarantee for high evidence, and that doing so carries a risk of drawing false conclusions. Meta-analyses in journals that allow critique and discussion should be preferred to those from journals that do not do so—Cochrane reviews in particular have a high quality on average. Unfortunately, that does not mean that every single Cochrane review is good, as the number of withdrawn reviews suggest. From the viewpoint of psychotherapy research, meta-analyses and RCTs will probably be viewed as methods too simple for research in the future, anyway. There is a tendency to look for the effects of various components and their interplay in psychotherapy, and the suggestion has been made to give up the "horserace" philosophy, where various treatment arms are compared like two pills [57-59]. Maybe other fields in medicine will follow.

**Acknowledgement**

The author thanks Tamara Brian and Alfred Weber for their support.

**Conflict of Interest**

The author declares no conflict of interest.
References


This article is available in: http://headache.imedpub.com/