The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis?

Abstract

Objective: To determine whether inappropriate subgroup analysis together with chance could change the conclusion of a systematic review of several randomised trials of an ineffective treatment.

Design: 44 randomised controlled trials of DICE therapy for stroke were performed (simulated by rolling different coloured dice; two trials per investigator). Each roll of the dice yielded the outcome (death or survival) for that "patient." Publication bias was also simulated. The results were combined in a systematic review.

Setting: Edinburgh. Main outcome measure--Mortality.

Results: The "hypothesis generating" trial suggested that DICE therapy provided complete protection against death from acute stroke. However, analysis of all the trials suggested a reduction of only 11% (SD 11) in the odds of death. A predefined subgroup analysis by colour of dice suggested that red dice therapy increased the odds by 9% (22). If the analysis excluded red dice trials and those of poor methodological quality the odds decreased by 22% (13, P=0.09). Analysis of "published" trials showed a decrease of 23% (13, P=0.07) while analysis of only those in which the trialist had become familiar with the intervention showed a decrease of 39% (17, P=0.02).

Conclusion: The early benefits of DICE therapy were not confirmed by subsequent trials. A plausible (but inappropriate) subset analysis of the effects of treatment led to the qualitatively different conclusion that DICE therapy reduced mortality, whereas in truth it was ineffective. Chance influences the outcome of clinical trials and systematic reviews of trials much more than many investigators realise, and its effects may lead to incorrect conclusions about the benefits of treatment.

Introduction

The sequence of events is all too familiar: banner headlines in the press announce the arrival of some miraculously effective new drug or treatment, but soon the early enthusiasm wanes as further research reveals that the benefits of treatment are rather more modest, that there are side effects, and that only some categories of patients are likely to benefit. Later, a systematic review of the evidence may suggest that the treatment was rather more effective than had been realised--as in the cases of, for example, streptokinase for acute myocardial infarction, aspirin for the prevention and treatment of vascular disease, tamoxifen and ovarian ablation for breast cancer--or that it was more hazardous than realised, as in the case of routine antiarrhythmic prophylaxis after myocardial infarction. Such systematic reviews may then be followed by large trials or "mega" trials to confirm or refute their findings.

Quantitative systematic reviews (meta-analyses) remain the best means of assessing the benefits of any health care intervention that has been tested in randomised controlled trials, particularly in areas in which many similar trials of the same intervention have been performed. Such reviews will form the basis of most of the systematic reviews in the Cochrane Collaboration's database. By combining the results of several studies, systematic reviews increase the numbers of outcome events available for comparison in the treatment and control groups and so reduce random errors. This is particularly important in assessment of treatments in which moderate random errors could obscure moderate, but potentially important, treatment effects. Like all powerful tools, systematic reviews should be handled with care as injudicious use may lead to damagingly incorrect conclusions. In the same way that chance can influence the results of individual clinical trials, chance can also influence the results of a systematic review of those trials to a surprising extent, particularly if relatively few patients and outcome events are included in the overall analysis and if inappropriate subgroup analyses are performed. In addition, publication bias--whereby studies with significant results are more likely to be published than those with null results--does exist and may increase the risk of inappropriate conclusions in systematic reviews that are restricted to published studies.

As part of a teaching exercise in clinical trials and systematic reviews we assembled a collaborative group of doctors; these trialists undertook a series of simulated randomised studies (DICE therapy), which, when combined, had sufficient statistical power to detect moderate treatment effects. We particularly wanted to know whether the combination of random error, inappropriate exclusion of studies from the review, inappropriate subgroup analysis, and publication bias could qualitatively change the conclusion of a systematic review of these trials from the right one (no evidence of benefit from DICE therapy) to the wrong one (DICE therapy is beneficial but only in particular circumstances).

Methods

We gave each participant in a practical class in statistics for the Edinburgh stroke course a red, green, or white dice and asked them all to write their name and the colour of the dice on a trial data form. We then asked participants to roll their dice a specified number of times to represent the number of patients in the treatment group of a randomised controlled trial. This number was written on the data form and each time the dice showed a "six" this was recorded as a "death" on the form. If any other number showed, a "survival" was recorded. The procedure was repeated for a control group of the same size. This was called the participant's first trial. Each participant then conducted the experiment again to simulate another trial of a different size (the participant's second trial). The sizes of the treatment and control groups varied from one participant to another. The largest trial comprised 200 "patients"--that is, 100 rolls of the dice for the treatment group and 100 for the control group--and the smallest comprised five in each group. The purpose of each trialist performing two trials was to simulate the effect of gaining clinical experience with an intervention. It was hypothesised that the intervention might be more effective and safer if administered by an experienced practitioner, and so an analysis of the second trials might yield clearer evidence of treatment effects.

We had told the participants that the behaviour of the dice might be unpredictable and that some of the dice might have a bias that altered over time so that the proportion of sixes would vary between the treatment series and the control series. This would result in a different number of deaths in the treatment and control groups, thereby simulating a treatment effect (or hazard). We had also told the participants that some of the dice might be free from any such bias--in other words, simulating a truly ineffective treatment, in which the difference between treatment and control was due to the play of chance alone. A prespecified hypothesis stated that different colours of dice might have different biases and therefore some colours might be "effective," some "hazardous," and some "ineffective," and so a subgroup analysis based on the colour of the dice would be performed. In fact, none of the dice had any bias, and all had the appropriate one in six chance of showing a six when rolled.

A second prespecified hypothesis was that trials of poor methodological quality could show different effects and so would be analysed separately. A trial was considered to be of poor methodological quality if there were important data errors--that is, if the name of the investigator or the colour of the dice was not properly recorded on the form.

Some people who perform systematic reviews favour inclusion of published trials only, believing that because these have been subject to formal peer review they are of higher quality than non-published trials. Since publication bias tends to favour publication of positive studies, systematic reviews restricted to published studies may yield biased results. To examine this effect, we simulated whether a trial had been published in the following way. The trial with the most significantly positive result was identified and regarded as the "hypothesis generating" trial. We then ordered the other studies in a random sequence using a random number generator. We then classifed each trial as having either a positive, negative, or null result on the basis of its O--E statistic--that is, the value of the observed number minus the expected number of outcome events. An O--E statistic of less than zero implied that fewer events than expected occurred in the treatment group, and so the trial was classified as positive; if the O--E statistic was greater than zero then more than the expected number of events occurred in the treatment group and so the trial was classified as

negative. If the O--E statistic was zero the trial was classified as null. Using rough estimates of the extent of publication bias, similar to those calculated by others for real trials, and using the random sequence described above, we selected the first 70% of the positive trials and the first 40% of the null or negative ones as those trials that were "published." An analysis restricted to published studies only therefore modelled the effect of publication bias (or of applying a methodological quality criterion that studies had to be published to be considered good enough to be included in the review).

We included all the eligible trials in a meta-analysis, using a standard method to derive an estimate of the overall effect of treatment. We then performed prespecified subgroup analyses based on the colour of the dice and the quality of the trials. Further subgroup analyses were also performed, on the basis of whether the trials were performed first or second and on whether they were published to see if systematic reviews based on all trials and those based on only published trials might reach qualitatively different conclusions.

Results

Of the 24 course participants who began the trials of DICE therapy, two were excluded from the study: one participant left the course early before completing his trials, and one submitted his form after the deadline. Two of the remaining 22 participants failed to write their name on the trial data form, and their four trials, along with six trials in which the colour of the dice was not stated, were regarded as being of poor methodological quality. The trial with the most positive result was trialist A's first trial, in which there were no deaths in 10 treated patients and six deaths in 10 control patients. The 93% (SD 33) reduction in the odds of death was significant (P=0.004). This was regarded as the hypothesis generating study--that is, the one that would have been published first.

Even though all of the dice were unbiased and should have shown a six on 16.7% of rolls--that is, a mortality in both the treated and the control group of 16.7%--we observed a surprisingly wide variation in mortality. Mortality ranged from 0% to 27% (median 16.0%) in the treatment group and from 0% to 60% (17.6%) in the control group.

The overall reduction in the relative odds of death was 11% (figure). This trend in favour of treatment was not significant (P>0.1) and if 95% confidence limits were used would be compatible with a reduction of 33%, or an increase of 11%, in the odds of death with treatment. It cannot be concluded that the treatment was ineffective because if the observed 11% reduction in death was confirmed by a large study it would represent a benefit of 16 lives saved per 1000 patients treated. This result would be clinically worthwhile, particularly as DICE therapy is cheap and widely practicable. The upper limit of the 95% confidence interval, however, suggested that routine use of DICE therapy could equally be associated with an excess of 16 deaths per 1000 patients treated.

The figure shows that DICE therapy with a red dice was associated with a non-significant 9% increase in mortality. A prespecified subgroup analysis was performed, which excluded trials that used red dice and those of poor methodological quality. This analysis suggested that therapy with other colours of dice was associated with a reduction of 22% in the odds of death, which almost reached significance (SD 13, P=0.09) (figure). If confirmed, this reduction would represent a benefit of 34 lives saved per 1000 patients treated.

To examine the effect of operators' experience, we compared the benefits in the first (learning) trials and those in the second (experienced) trials separately. The benefits of DICE therapy were greater when the analysis was restricted to experienced operators; the reduction of 24% (SD 15) in the odds of death was almost significant (figure).

An analysis restricted to published trials showed a reduction of 23% in mortality with treatment that was of borderline significance (SD 13, P=0.07). An analysis restricted to the published trials performed by experienced operators, however, showed that the reduction in mortality with treatment was larger (39%) and significant (SD 17, 2 P=0.02), with the absolute reduction being equivalent to 70 lives saved per 1000 patients treated (figure).

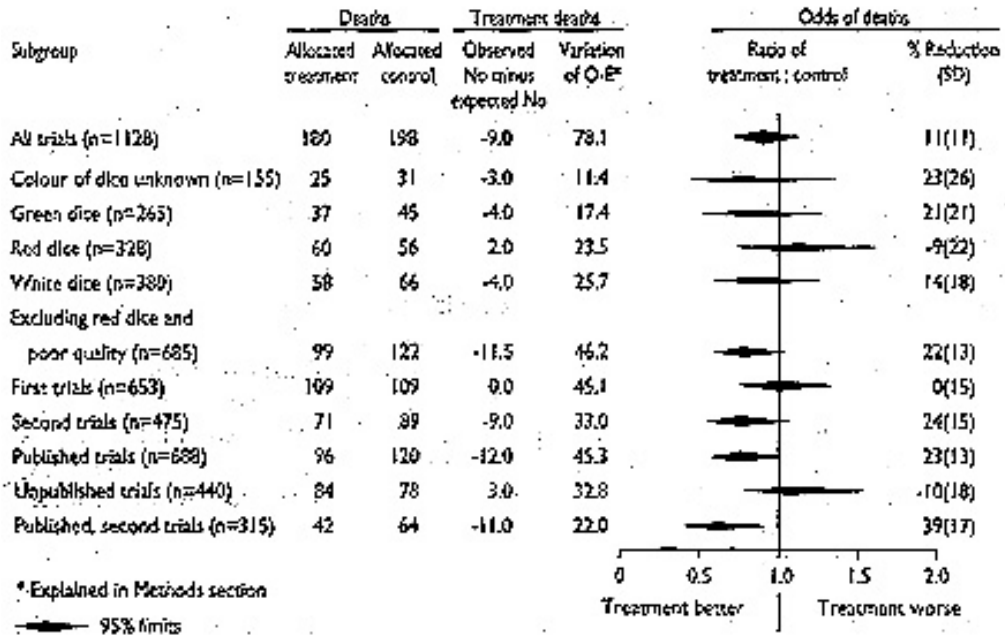

FIG--Pooled odds ratios for mortality for all trials and for each subgroup analysis performed

Discussion

BEWARE OF CHANCE

Chance does not get the credit it deserves. Most doctors admit that chance influences whether they win the Christmas raffle but underestimate the effect of chance on the results of any clinical trials they read about. Proponents of a particular drug (or operation) may claim that the trial evaluating it was conducted to high methodological standards, but such standards sometimes provide little defence against the havoc wrought by chance. We undertook this slightly tongue in cheek study to illustrate just how extraordinary the effects of chance can be and that well meaning attempts to perform inappropriate subgroup analyses in a highly restrictive systematic review may compound the problem still further.

This experiment could have been performed with computer simulation, but we preferred to use the human touch of individuals throwing dice. Throughout the experiment we tried to simulate the real world of clinical research, randomised controlled trials, and systematic reviews--in particular, the way that a single small trial with a very favourable result for a new therapy is reported and acts as a "hypothesis generator" for a series of other randomised controlled trials. These are often small, exploratory studies and so are highly susceptible to the play of chance, as illustrated by our results: treatment effects in our trials ranged from extreme benefit (complete protection against death in trialist A's first trial) to huge hazard (a fivefold excess of deaths in trialist M's first trial).

Some of the participants were convinced that their own dice was really loaded. Trialist A described his reaction to his first trial. At first he was a little surprised that no sixes (deaths) had occurred in the treatment group, but then he felt that by chance he should expect only one or two deaths anyway. When he started on the control group he rolled one six, followed by another and then a third. He said that his room felt eerily quiet as he rolled a fourth six: he had never rolled four sixes in a row in his life. By the time he had rolled the fifth, he was certain that the dice was loaded, and the sixth six only confirmed his belief that DICE therapy clearly had an effect. Similar experiences have been reported in the real world of clinical trials: Koudstaal et al discontinued their trial of thrombolysis in acute ischaemic stroke after the first two patients to be given thrombolysis both died. Although they calculated that the chance of this occurring was small, so is the chance of throwing six "sixes" out of 10 throws of a dice!

Given that all the dice were, in truth, unbiased, we had hoped that a meta-analysis of the results of all these trials would have shown no important difference between the treatment and control groups. It showed, however, a non-significant reduction in the odds of death of 11% and so did not exclude the possibility of an important clinical benefit, equivalent to 16 deaths avoided per 1000 patients treated. Equally the analysis could not exclude the possibility that treatment was associated with an extra 16 deaths per 1000 patients treated. Even after a review of the data from all the trials, however, the numbers of patients and outcome events remained relatively small (about 2250 patients and 380 deaths), which meant that random errors could have still significantly influenced the results.

BEWARE OF SUBGROUP ANALYSES

The potentially favourable result of our overall meta-analysis could be inflated by carrying out both prespecified and further subgroup analyses, one of which achieved significance. Although any number of such analyses could have been performed, we chose those which may have some parallel in clinical medicine. For instance, the different colours of dice could represent different drugs with similar effects--for example, different classes or doses of ß blocker or different settings of care. The first trials could represent trials performed soon after the introduction of a new technique--such as a surgical procedure--that required new skills to be developed, and the second trials those that were performed later. If the technique had a significant learning curve it might be expected that the results of the later trials would be significantly better than the early ones, and this may be used to justify analysing the first trials separately or even excluding them completely from the meta-analysis. In our analysis the results of the second trials did indeed seem to be better than those of the first trials, but this was due not to the experience of the investigators but to chance alone.

Subgroup analysis is therefore a little like a lucky dip: you never know what you might come up with. It is particularly hazardous when the overall estimate of the treatment effect is close to null and is not significant. In such cases it is often possible to select a group of trials that show a significantly greater than average treatment effect due purely to chance, leaving another group with less than the average treatment effect. Even if a particular subgroup analysis is prespecified (as with the colour of dice in our study) and seems to be clinically sensible, these safeguards offer no protection against the vagaries of chance. For example, a recent systematic review of nine trials (comprising 3719 patients) of the calcium antagonist nimodipine in patients with acute ischaemic stroke showed no evidence of overall benefit in those patients allocated to nimodipine (odds ratio for mortality approximately 1). A subgroup analysis, however, showed that patients treated within 12 hours of stroke did benefit (odds reduction for mortality 38%) whereas nimodipine increased mortality if started more than 12 hours after stroke onset. Is it biologically plausible that treatment within 12 hours is beneficial but later treatment is harmful? Such qualitative interactions are rare in medicine. Quantitative interactions are much more common (that is, treatment is beneficial in some categories of patients and less beneficial in others), and so the difference between the effects of early and late treatment with nimodipine may simply have been due to chance effects. Where the subgroup analyses have not been predefined, the risk exists that data have been trawled ("data-dredging") in search of a significant result, and so even greater caution is necessary in interpreting the results. Perhaps all readers should bear in mind the simple rule, suggested by Peto et al, that it is generally safer to base any conclusions on the overall results of a trial (or overview of several trials) than to emphasise the apparent benefits in particular subgroups (or subsets of trials).

BEWARE OF RESTRICTIVE REVIEWS

The results of this experiment also support the practice, favoured by many specialists in systematic reviews, of including all relevant truly randomised trials in a review rather than selecting a much smaller group of studies for inclusion. A broadly inclusive systematic review maximises the amount of data available for statistical analysis and minimises selection bias. The use of more restrictive inclusion criteria--for example, selecting studies of a particular dose of drug or studies rated above the threshold of some arbitrary methodological quality score--may lead some systematic reviews to become merely subgroup analyses from the beginning, which, naturally, increases the risk that chance will lead the conclusions of the review astray.

Some systematic reviews are restricted to published studies partly because the investigators believe that studies that have not been subjected to formal peer review may not be methodologically sound. However, given that publication bias exists and tends to lead to the publication of "positive" trials rather than "null" or "negative" ones, an analysis restricted to published studies may lead to an unduly optimistic assessment of treatment benefits, as we demonstrated. Such an analysis also reduces the number of trials (and hence data) included in a systematic review, which increases the likelihood of random error. Publication bias combined with well meaning but inappropriate subgroup analysis forms a dangerous mixture, which can lead to quite the wrong conclusion.

CONCLUSION: FACT OR FANTASY?

We conclude with a fictional account of how doctors reacted to this welter of information about DICE therapy. The uncritical enthusiasts (who read few journals but watch television) began treating all their stroke patients with DICE therapy soon after the first ecstatic reports in the media. More conservative doctors waited for a few more trials to be reported before cautiously using DICE therapy in a few patients. A systematic review that included only the published trials was then published; it showed probable benefit and persuaded a few more doctors to try DICE therapy. With an increase in use, however, disillusionment with the therapy began to appear, coinciding with some short reports of fatal side effects in a few patients. The manufacturer of green and white dice commissioned its own "methodologically better" review, restricted to the published trials of only green and white dice (since red dice were already known to be harmful) which were performed by experienced investigators. The strikingly positive result of this review was seized on by the manufacturer of green and white dice, who arranged a press conference to reveal the benefits. Trialist A appeared at the conference, attesting to the miraculous benefits of white DICE therapy in his first trial (conveniently ignoring the null results of his second trial). The pressure from patient groups and the manufacturer forced the Food and Drug Administration in the United States to grant a product licence for the use of green and white dice in carefully selected patients by doctors experienced in DICE therapy. Fortunately, the Cochrane Collaboration's review of all the randomised trials, which suggested that the benefits were modest or non-existent, were widely disseminated. After much debate, DICE therapy was withdrawn and some people whose relatives died after DICE therapy sued the manufacturer.

This fantasy is perhaps more common than we care to admit. It emphasises the need for large randomised controlled trials and the need to appraise critically the results of randomised trials or systematic reviews and be cautious in interpretating them. In diseases in which the outcome event is relatively uncommon--for example, early death in acute myocardial infarction or stroke--trials or systematic reviews with just a few thousand patients may yield apparently significant but unreliable results. Instead, trials (or reviews) involving several thousands of patients will be required. A recent example that highlights this problem is the use of aspirin to prevent pre-eclampsia in pregnant women. A large randomised trial that comprised over 9000 women failed to show any clear benefit for aspirin, whereas a previous systematic review of published trials that comprised only 394 women had suggested a significant reduction of 65% in the risk of preeclampsia with aspirin. In general, therefore, we suggest that the premise "Don't Ignore Chance Effects" (DICE) always be kept in mind.

Key messages

* The potentially extreme effects of chance on the results of individual clinical trials and on systematic reviews should never be underestimated
* In general, systematic reviews should include all truly randomised studies
* Applying excessively restrictive inclusion criteria limits the sample size and increases the risk of bias and random error
* The results of subgroup analyses should be interpreted with caution
* The conclusions of a systematic review should be based not on a selection of trials or on a particular subgroup but on all the randomised evidence.