Bandolier Professional

Independent evidence-based health care

# THE IMPORTANCE OF SIZE

## On socks

The sock-drawer problem seemed to be a good place to start when thinking about size (or numbers), especially on cold, dark, winter mornings. When we think about size (or numbers) in the sock drawer or any other context, the numbers we need to answer a question depend on the precise question asked. Socks can help us get our brains around it (and help us look clever at dinner parties).

### The scenario

We have 20 pairs of socks: 10 pairs of red socks (20 red socks in total) and 10 pairs of white socks (20 white socks in total). The problem is that they are all mixed up in my sock drawer, and it is dark, so I can't see the colour of the socks I am taking out of the drawer. We can ask a number of subtly different questions.

Q1: How many socks do I have to take out to be sure that I have a pair of socks?
A1: Two socks, so long as I don't care what colour they are. I may have two red socks, two white socks, or one white and one red. Whichever, I have two socks, and by many definitions that constitutes a pair.

Q2: How many socks do I have to take out to be sure that I have a pair of socks the same colour?
A2: Three socks, so long as I don't care what colour they are. I may have three red socks, three white socks, two white and one red, or one white and two red. Whichever, I have two socks of the same colour.

Q3: How many socks do I have to take out to be sure that I have a pair of red socks?
A3: Twenty-two socks, because it could be that just by chance I pull out all the 20 white socks first, and then have to pull out two of the remaining 20 red socks.

But that is unlikely. So how many socks do I need to pull out to be 95% sure (or 90%, or 75%, or 50%) of having a pair of red socks? The number of socks I need to pull out rises sharply with the confidence I need to put on the answer.

These answers to the questions will depend on the proportion of red and white socks. The numbers will change as the proportion of red socks in the drawer changes.

## Clinical trials

A clinical trial is two sock drawers, with different proportions of socks in each, and where we try to find out the proportion of red socks by randomly choosing socks from each drawer until we have enough to answer the question.

But what is the question? You have several choices here, including these:

- Are there more red socks in drawer 1 than drawer 2?
- How big is the difference in the proportion of red socks between drawer 1 and drawer 2?
- What proportion of red socks is in drawer 1?

Let us think of ibuprofen as an analgesic. In clinical trial terms these sock drawer questions are roughly like asking whether ibuprofen is a better analgesic than placebo, by how much is it better, and what proportion of people with pain get that pain relieved by ibuprofen? Different answers to different questions, and the numbers of socks or patients needed to answer them differs.

## Random chance, quality, and size

These are all related topics. This essay therefore examines aspects of all of them. First it looks at the effects of random chance that we can get just by rolling dice. Then it examines how likely our results are to change, depending on the amount of information we have available, and finally it reverts to the answers to the socks drawer, or ibuprofen question.

## Rolling dice

"*There is much luck in the world, but it is luck. We are none of us safe*". So said EM Forster nearly 100 years ago. ***Bandolier*** is constantly astonished in its travels by people who appreciate the importance of chance, in, say, winning a lottery or avoiding a car accident, but not in clinical trials. Perhaps it is all down to the way statistics are taught. We should forget probabilities and p-values, and acquaint ourselves with more relevant information, notably how much data do we need to be sure that an observation is not likely to occur just by chance.

Why are people impressed with p-values? The cherished value of 0.05 merely says that a result is not more likely to have occurred by chance than 1 time in 20. Most of us have played Monopoly or other games involved with throwing dice. We will have experienced that throwing two sixes with two dice happens relatively often, yet the chance of that is about 1 time in 36.

Look at it another way. If you were about to cross a bridge, and were told that there was a 1 in 20 chance of it falling down when you were on it, would you take the chance? What about 1 in 100, or 1 in 1000? That p-value of 0.05 also tells you that 1 time in 20 the bridge **will** fall down.

The dice analogy is pertinent, because there are now (at least) two papers that look at random chance and clinical trials, reminding us how often and how much chance can affect results. An older study actually used dice to mimic clinical trials in stroke prevention [1], while a more recent study [2] used computer simulations of cancer therapy.

## DICE 1

In this study [1] participants in a practical class on statistics at a stroke course were given dice and asked to roll them a specified number of times to represent the treatment group of a randomised trial. If six was thrown, this was recorded as a death, with any other number a survival. The procedure was repeated for a control group of similar size. Group size ranged from 5 to 100 patients.

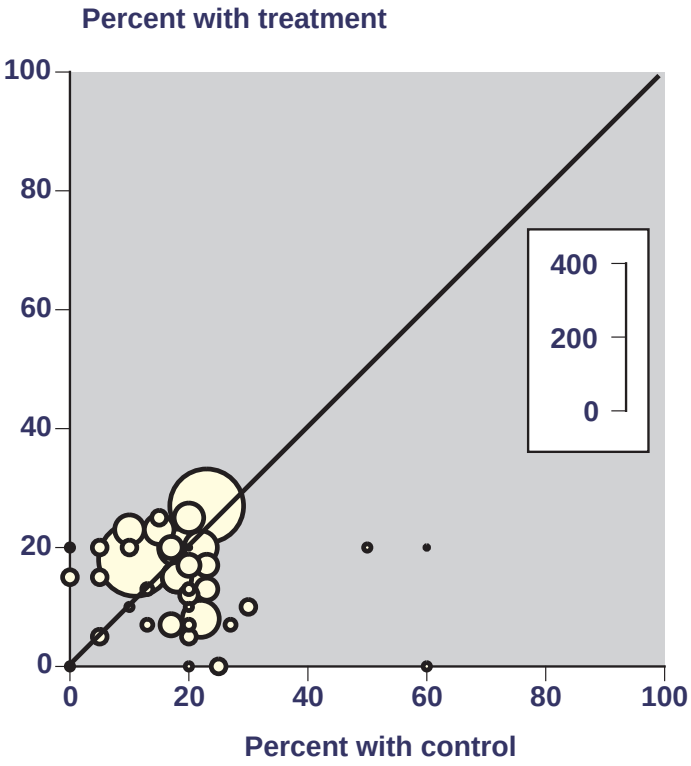
The paper gives the results of all 44 trials for 2,256 "patients". While the paper does many clever things, it is perhaps more instructive to look at the results of the 44 trials. Since each arm of the trial looks for the throwing of one out of six possibilities for standard dice, we might expect that the rate of events was 16.7% (100/6) in each, with an odds ratio or relative risk of 1.

Figure 1 shows a L'Abbé plot of the 44 trials. The expected result is a grouping in the bottom left, on the line of equality at about 17%. Actually, it is a bit more dispersed than that, with some trials far from the line of equality.

The odds ratios for individual trials are shown in Figure 2. Two trials (20 and 40 in total) had odds ratios statistically different from 1. That's one time in every 22 trials, what we expect by chance.

The variability in individual trial arms is shown in Figure 3, where the results are shown for all 88 trial arms. The vertical line shows the overall result (16.7%). Larger samples come close to this, but small samples show values as low as zero, and as high as 60%.

**Figure 1: L'Abbé plot of DICE 1 trials**




The overall result, pooling data from all 44 trials, showed that events occurred in 16.0% of treatments and 17.6% of controls (overall mean 16.7%). The relative risk was 0.8 (0.5 to 1.1) and the NNT was 63, with a 95% confidence interval than went from one benefit for every 21 treatments to one harm for every 67 treatments (Table 1).

Many of the experimental DICE trials were quite small, with as few as five per group. The smaller trials, with 40 per group or less, actually came up with a statistically significant result (Table 1). The NNT here was 19 (11 to 98).

## DICE 2

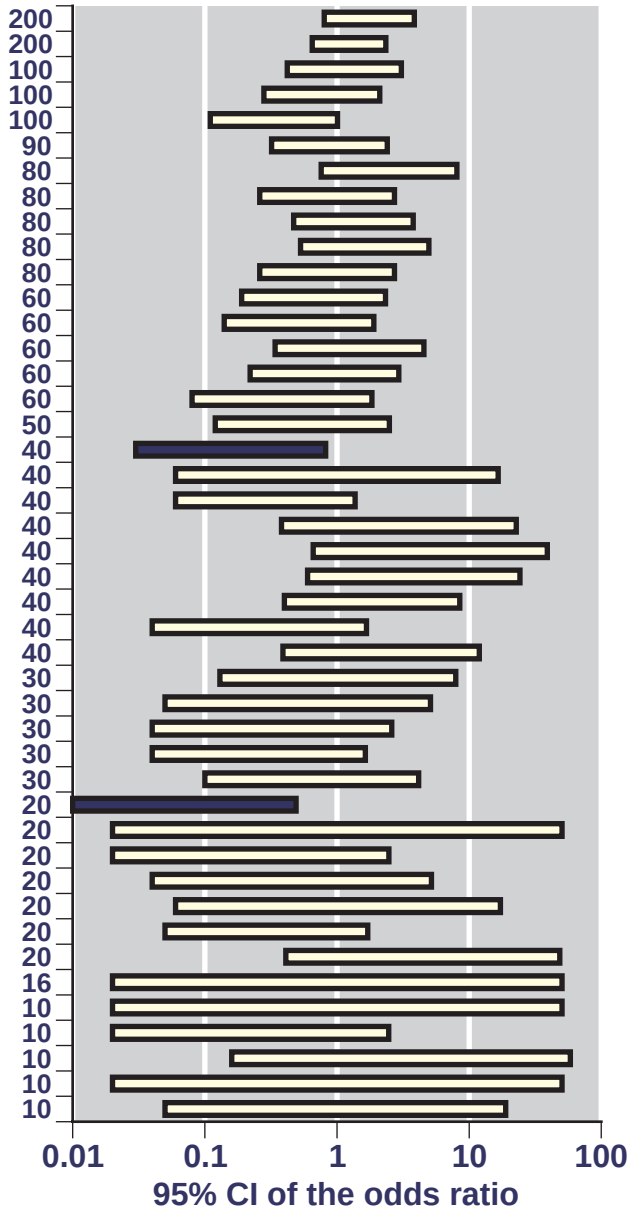
Information on the time between randomisation and death in a control group of 580 patients in a colorectal cancer trial was used to simulate 100 theoretical clinical trials. Each time the same 580 patients were randomly allocated to a theoretical treatment or control group, and survival curves calculated [2].

Four of the trials artificially generated had statistically significant results. One was significant at the 0.003 level (1 in 333) and showed a large theoretical decrease in mortality of 40%.

**Table 1: Meta-analysis of DICE trials, with sensitivity analysis by size of trial**

| | Number of | | Outcome (%) wih | | | |
|---|---|---|---|---|---|---|
| | Trials | Patients | Treatment | Control | Relative risk (95% CI) | NNT (95% CI) |
| All trials | 44 | 2256 | 16.0 | 17.6 | 0.8 (0.5 to 1.1) | 62 (21 to -67) |
| Larger trials (>40 per group) | 11 | 1190 | 19.5 | 17.8 | 1.1 (0.9 to 1.4) | -60 (36 to -16) |
| Smaller trials (<40 per group) | 33 | 1066 | 12.0 | 17.3 | 0.7 (0.53 to 0.94) | 19 (11 to 98) |

**Figure 2: Odds ratios for individual DICE studies, by number in "trial". Filled bars were statistically significant**

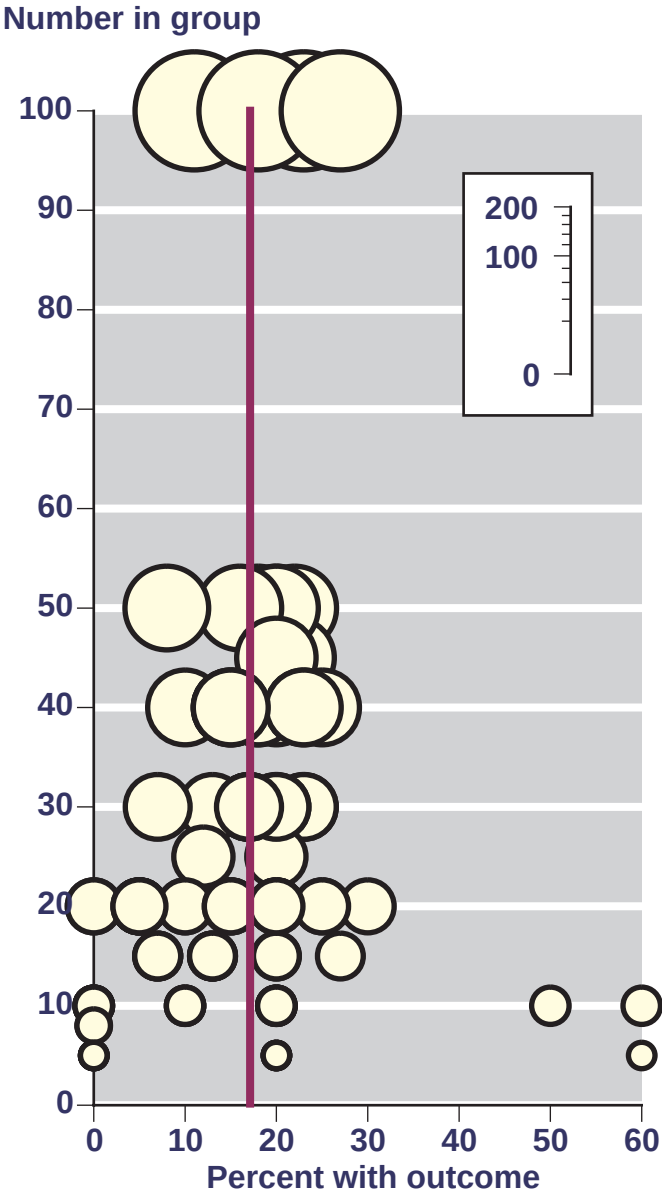



Subgroup analysis was done for this trial by randomly allocating patients to type A or type B, and doing this 100 times. Over half (55%) of the subgroup analyses showed statistical significance between subgroups. The extremes of results were no difference between subgroups, to a result with high significance of 0.00005 (1 in 20,000). In another trial that had bare statistical significance, four of 100 simulated subgroups had statistical significance at the 1 in 100 level.

## Comment

What does all this tell us? It emphasises that the random play of chance is a factor we cannot ignore, and that small trials are more prone to chance effects than larger ones. And it is not just an effect seen in single trials. Even when we pool data from small trials just from rolling dice, as in DICE 1, a meta-analysis can come up with a statistically significant effect when there was none.

**Figure 3: Percentage of events in each trial arm of DICE "trials"**




High levels of statistical significance can be generated just by the random play of chance. DICE 2 found levels of statistical significance of 1 in 333 for at least one simulated trial, and 1 in 20,000 for a subgroup analysis of that trial.

Not only do we need well-conducted trials of robust design and reporting, we also need large amounts of information if the size of a clinical effect is to be accurately assessed. The rule of thumb is that where the difference between control and treatment is small we need very large amounts. Only when the difference is large (an absolute risk increase or decrease of 50%, affecting every second patient) can we be reasonably happy with information from 500 patients or fewer.

When we see differences between trials, or between responses to placebo, the rush is often to try and explain the difference according to some facet of trial design or patient characteristic. Almost never does anyone ask how likely the difference is to occur just by the random play of chance.

## Quality and size

Study architecture and size are often linked. Two Danish researchers looked for large clinical trials with at least 1,000 patients together with meta-analyses of small trials [3]. They were asking the sensible question about how possible discrepancies between large trials and meta-analyses could be affected by methodological quality.

They found 14 meta-analyses, pulled all the original papers, subjected those to quality review, and examined outcomes in terms of odds ratios. They then used the ratio of the odds ratio in the large randomised trial to that from the meta-analysis of small trials to produce a "ratio of odds ratios" as the final outcome. When the ratio of odds ratios was significantly less than 1, that indicated that small trials with particular quality criteria exaggerated the effect of an intervention compared with the large trial.

The quality criteria they tested for were generation of the allocation sequence, allocation concealment, double blinding, and withdrawals or dropouts. The relevant criteria are in Table 2.

### Results

They used 23 large trials and 167 small trials with 136,000 patients. Compared with large trials, small trials with inadequate generation or allocation concealment of the randomisation sequence, or those that were not adequately double blinded over-estimated the effect of treatment (Table 3). When methodological quality was compared in large and small trials, inadequate generation of the randomisation sequence and inadequate double-blinding caused overestimation of the treatment effect (Table 4), and much the same was found for a similar analysis of small trials alone.

Quality scoring using the Oxford system [4], perhaps one of the most commonly used scoring systems in systematic reviews, produced sensible results. Small trials with lower quality scores over-estimated treatment effects compared

**Table 2: Quality criteria tested**

| Quality feature | Adequate | Inadequate |
|---|---|---|
| Generation of the allocation sequence | computer-generated random number or similar | not described |
| Allocation concealment | central independent unit, sealed envelope, or similar | not described, or open table of random numbers |
| Double blinding | identical placebo or similar | not described, or tablets versus injection not double dummy |
| Withdrawals or dropouts | number and reasons for drop outs | not described |

**Table 3: Comparison of large trials with small trials with different quality criteria**

| Common comparator | Comparison | Ratio of odds ratios (95%CI) |
|---|---|---|
| Large trials | Small trials with **inadequate** generation of allocation sequence | **0.46 (0.25 to 0.83)** |
| Large trials | Small trials with **adequate** generation of allocation sequence | 0.90 (0.47 to 1.76) |
| Large trials | Small trials **inadequate** allocation concealment | **0.49 (0.27 to 0.86)** |
| Large trials | Small trials **adequate** allocation concealment | 1.01 (0.48 to 2.11) |
| Large trials | Small trial with **inadequate** or no double blinding | **0.52 (0.28 to 0.96)** |
| Large trials | Small trial with **adequate** or no double blinding | 0.84 (0.43 to 1.66) |
| Large trials | Small trials with **inadequate** follow up | 0.72 (0.30 to 1.71) |
| Large trials | Small trials with **adequate** follow up | 0.58 (0.32 to 1.02) |

When the ratio of the odds ratios is less than 1, it indicates that the feature (inadequate blinding, for example) exaggerates the intervention effect

**Table 4: Comparison of adequate versus inadequate quality criteria in large and small trials**

| Common comparator | Comparison | Ratio of odds ratios (95%CI) |
|---|---|---|
| Adequate | **Inadequate** generation of allocation sequence | **0.49 (0.30 to 0.81)** |
| Adequate | **Inadequate** allocation concealment | 0.60 (0.31 to 1.15) |
| Adequate | **Inadequate** or no double blinding | **0.56 (0.33 to 0.98)** |
| Adequate | **Inadequate** follow up | 1.50 (0.80 to 2.78) |

When the ratio of the odds ratios is less than 1, it indicates that the feature (inadequate blinding, for example) exaggerates the intervention effect

with large trials. Small trials with higher quality scores did not. With both large and small trials, treatment effects were exaggerated with low versus high quality scores.

## Cumulative meta-analysis

Confirmation that our estimate of the effect of treatment can be heavily dependent on size comes from a study from the USA and Greece [5]. Researchers looked at 60 meta-analyses of randomised trials where there were at least five trials published in more than three different calendar years. They were in either pregnancy and perinatal medicine or myocardial infarction.

For each meta-analysis trials were chronologically ordered by publication year and cumulative meta-analysis performed to arrive at a pooled odds ratio at the end of each calendar year. The relative change in treatment effect was calculated for each successive additional calendar year by dividing the odds ratio of the new assessment with more patients by the odds ratio of the previous assessment with fewer patients. This gives a "relative odds ratio", in which a number greater than 1 indicated more treatment effect, and one less than 1 indicates less treatment effect.

The relative odds ratio can be plotted against the number of patients included. The expected result is a horizontal funnel, with less change with more patients, and the relative odds ratio settling down to 1.

### Results

In the paper, the two graphs for pregnancy / perinatal medicine and myocardial infarction showed exactly this expected pattern, but are just impossible to reproduce here. Below 100 patients the relative odds ratios varied between 0.2 and 6. By the time 1000 patients were included they were between 0.5 and 2. By 5,000 patients they settle down close to 1. The 95% prediction interval for the relative change in the odds ratio for different numbers for both examples is shown in Table 5.

When evidence was based on only a few patients there was substantial uncertainty about how much the pooled treatment effect will change in the future. With only 100 patients randomised, additional information from more trials could multiply or divide the odds ratios at that point by three.

### Table 5: 95% prediction interval for relative change in odds ratio for different numbers of accumulated patients randomised

| | Fixed effect prediction interval for relative change in odds ratio | |
|---|---|---|
| Number of patients | Pregnancy/perinatal | Myocardial infarction |
| 100 | 0.32 - 2.78 | 0.18 - 5.51 |
| 500 | 0.59 - 1.71 | 0.60 - 1.67 |
| 1000 | 0.67 - 1.49 | 0.74 - 1.35 |
| 2000 | 0.74 - 1.35 | 0.83 - 1.21 |
| 15000 | 0.85 - 1.14 | 0.96 - 1.05 |

## Comment

At first look this is all complicated pointy-head stuff, but actually it's no more than simple common sense. If trials are not done properly, they might be wrong. If trials are small, they might be wrong. To be sure of what we know we need large data sets of high quality, whether from single trials or meta-analyses. The corollary is that if we have small amounts of information, or information of poor quality, the chance of that result being incorrect is substantial, and then we need to be cautious and conservative.

## Variability and size

Clinical trials should have a power calculation performed at the design stage. This will estimate how many patients are needed so that, say, 90% of studies with X number of patients would show a difference of Y% between two treatments. When the value of Y is very large, the value of X can be small. More often the value of Y is modest, or small. In those circumstances, X needs to be larger, and more patients will be needed in trials for them to have a hope of showing a difference.

Yet clinical trials are often ridiculously small. ***Bandolier***'s record is a randomised study on three patients in a parallel group design. But when are trials so tiny that they can be ignored? Many folk take a pragmatic view that trials with fewer than 10 patients per treatment arm should be ignored, though others may disagree.

### Random play of chance

The degree of variability between trials of adequate power is still large, because trials are powered to detect that there is a *difference* between treatments, rather than *how big* that difference is. The random play of chance can remain a significant factor despite adequate power to detect a difference.

Figure 4 shows the randomised, double blind studies comparing ibuprofen 400 mg with placebo in acute postoperative pain [7]. The trials had the same patient population, with identical initial pain intensity and with identical outcomes measured in the same way for the same time using standard measuring techniques. There were big differences in the outcomes of individual studies.

Figure 5 shows the results of 10,000 studies in a computer model based on information from about 5,000 individual patients [8]. Anywhere in the gray area is where a study could occur just because of the random play of chance. And for those who may think that this reflects on pain as a subjective outcome, the same variability can be seen in other trial settings, with objective outcomes.

## How much information is enough?

While it is relatively easy to demonstrate that inadequate amounts of information can result in erroneous conclusions, the alternative question, how much information we need to avoid erroneous conclusions, is more difficult to answer.

**Figure 3: Trials of ibuprofen in acute pain that are randomised, double blind, and with the same outcomes over the same time in patients with the same initial pain intensity**

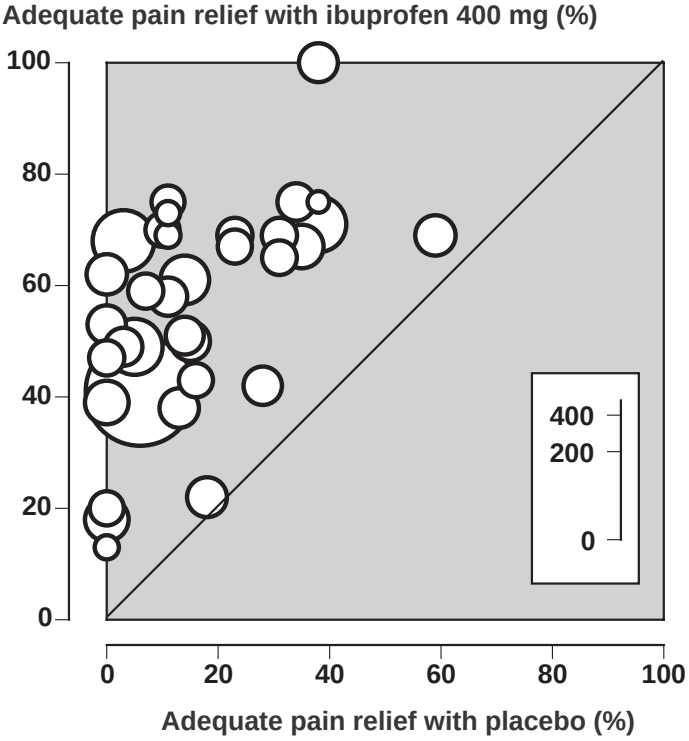
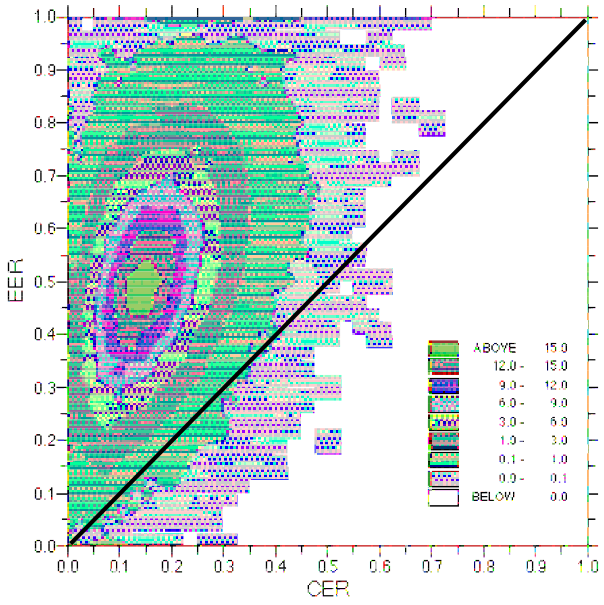



**Figure 4: Computer model of trials of ibuprofen in acute pain. Intensity of colour matches probability of outcome of a single trial**




*In Figure 4 CER (control event rate) is equivalent to placebo and EER (experimental event rate) to ibuprofen in clinical trails*

It depends on a number of things. Two important issues are the size of the effect you are looking at (absolute differences between treatment and control), and how sure you want to be.

A worked example using simulations of acute pain trials [8] gives us some idea. Using the same 16% event rate as in DICE 1 as the rate with controls (because it happens to be what is found with placebo), it looked at event rates with treatment of 40%, 50% and 60%, equivalent to NNTs of 4.2, 2.9 and 2.3. The numbers in treatment and placebo group were each simulated from 25 patients per group (trial size 50) to 500 patients per group (trial size 1000). For each condition 10,000 trials were simulated and the percentage where the NNT was within ±0.5 of the true NNT counted.

The results are shown in Table 6. With 1000 patients in a trial where the NNT was 2.3, we could be 100% sure that the NNT measured was within ±0.5 of the true NNT; all trials of this size would produce values between 1.8 and 2.8. In a trial of 50 patients where the NNT was 4.2, only one in four trials would produce an NNT within ±0.5; the true value is between 3.7 and 4.7, and three-quarters of trials (or meta-analyses) of this size would produce NNTs below 3.7 or over 4.7.

The study also shows that to be certain of the size of the effect (the NNT, say), we need ten times more information than just to know that there is statistical significance.

## Comment

A single small trial done to show statistical significance can mislead for no reason other than the random play of chance. Right trials can be wrong.

**Table 6: Effect of size and size of effect on confidence of treatment effect**

| | Percent events with treatment | | |
|---|---|---|---|
| | 40 | 50 | 60 |
| NNT | 4.2 | 2.9 | 2.3 |
| Group size | | | |
| 25 | 26 | 37 | 57 |
| 50 | 28 | 51 | 73 |
| 100 | 38 | 61 | 88 |
| 200 | 55 | 81 | 96 |
| 300 | 63 | 89 | 99 |
| 400 | 71 | 93 | 99 |
| 500 | 74 | 95 | 100 |

With control the event rate was 16

At least

| | |
|---|---|
| 50% | within +/- 0.5 |
| 80% | within +/- 0.5 |
| 95% | within +/- 0.5 |

## Examples from the literature

In this section we will choose a number of examples of how lack of size could, or has, made a difference to decision-making. All come from meta-analysis, or randomised trials, or both. They are offered not as criticisms, but as important lessons about how we can be misled if we are insufficiently vigilant.

## Gastrointestinal bleeding and NSAID

Epidemiological studies associating NSAID use and upper GI problems and published in the 1990s were reviewed and the data pooled [9] to give a much clearer picture of risks. To be included studies had to:

- Be case control or cohort studies on non-aspirin NSAIDs
- Include data on bleeding, perforation, or other serious upper gastrointestinal tract event resulting in hospital admission or referral to a specialist
- Have data to calculate relative risk.

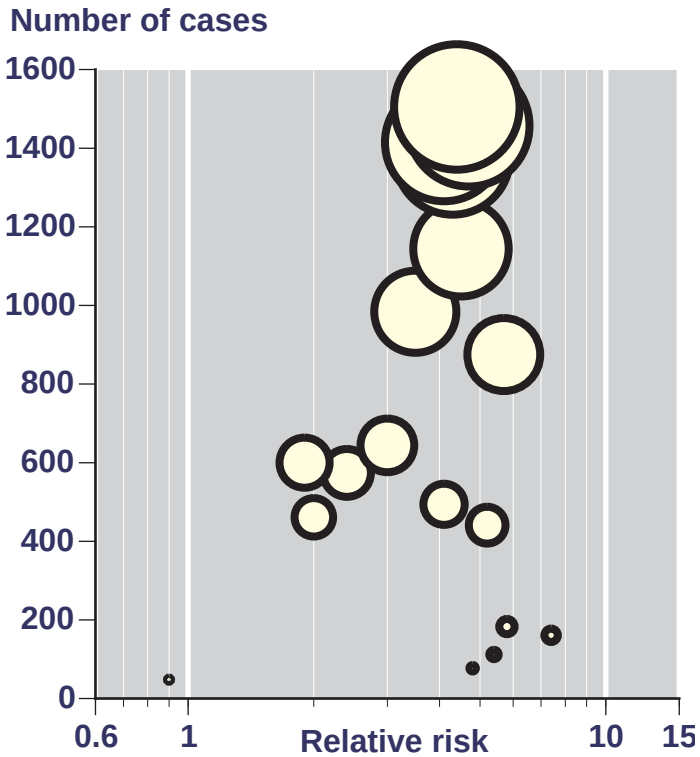
### Results

Eighteen studies were found. All had specific definitions of exposure and outcome and similar ascertainment for comparison groups. All but two attempted to control for potential confounding factors, like age, sex, history of ulcer or concomitant medicines.

The main results were that, compared with nonusers, NSAID users had a higher risk of upper GI bleed when they were current NSAID users and used a higher dose. People with a history of ulcer or with a previous bleed who took NSAIDs were at much greater risk than those with no history of ulcer who took NSAIDs. Older folk who took NSAIDs were at greater risk than under 50s who took NSAIDs.

In this set of high quality epidemiological studies there was a clear effect of size on the estimate of relative risk of upper gastrointestinal bleed with NSAID. The pooled estimate was 3.8 (3.6 to 4.1). With fewer than 1000 cases, the results of individual studies was highly variable (Figure 5). With fewer than 200 cases the variation was extreme, with point estimates from NSAIDs causing bleeds nearly eight times more frequently to relative risks below 1, suggesting that NSAIDs were protective against gastrointestinal bleeding.

**Figure 5: Effect of size of study in determining overall relative risk of GI bleed, NSAID users compared with nonusers**




## Aspirin and stroke

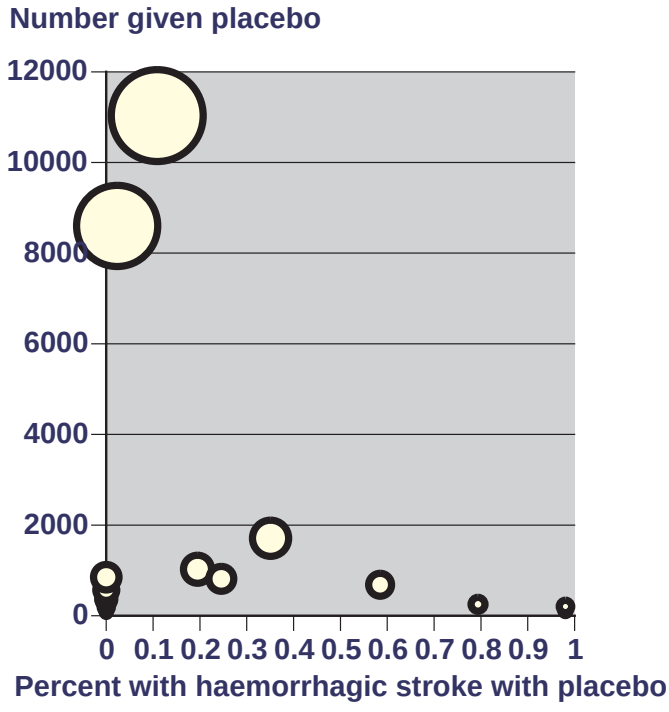
Because some studies have suggested that it may increase the risk of haemorrhagic stroke, and so a meta-analysis of studies has been done [10] to try and measure this risk. Authors performed extensive searching for randomised studies of aspirin versus control and which had stroke as an outcome. To be included studies had to have:

- Random allocation to aspirin or control.
- No intervention difference other than use of aspirin.
- Duration of at least 1 month.
- Information on stroke subtypes.

### Results

They found 16 trials with over 55,000 subjects. The mean aspirin dose was 273 mg/day (range 75 - 1500 mg/day)

**Figure 6: Estimate of rate of haemorrhagic stroke with placebo in randomised trials of aspirin**



and the mean duration of treatment was 37 months (1 to 72 months). The study was predominantly in white men (88% men, 99% white) with a mean age of 59 years.

The studies differed in size from 60 to over 11,000 patients. The rate at which haemorrhagic stroke was found with placebo (Figure 6) also varied. Overall the rate with placebo was 0.12%, but varied from 0% to almost 1%. The two largest studies had similar, low, rates of haemorrhagic stroke, and all the variability was in small studies with few actual cases. Even in this large analysis of 55,000 people, there were only 108 cases of haemorrhagic stroke.

## Magnesium

In 1991 a meta-analysis came up with very positive conclusions about using magnesium after a heart attack [11]. To investigate the effect of intravenous magnesium on mortality in suspected acute myocardial infarction results from 1301 patients in seven randomised trials were examined. There were 25 (3.8%) deaths among 657 patients allocated to receive magnesium and 53 (8.2%) deaths among 644 patients allocated control. This represented a 55% reduction in the odds of death (p less than 0.001) with 95% confidence intervals ranging from about one third to about two thirds. 70 of 648 patients allocated magnesium compared with 109 of 641 controls had serious ventricular arrhythmias, suggesting that magnesium reduced the incidence, though the definition varied among trials. The conclusion was that intravenous magnesium therapy may reduce mortality in patients with acute myocardial infarction. It also concluded that further large scale trials to confirm (or refute) these findings were desirable.

**Figure 7: Mortality with control in randomised trials of intravenous magnesium**

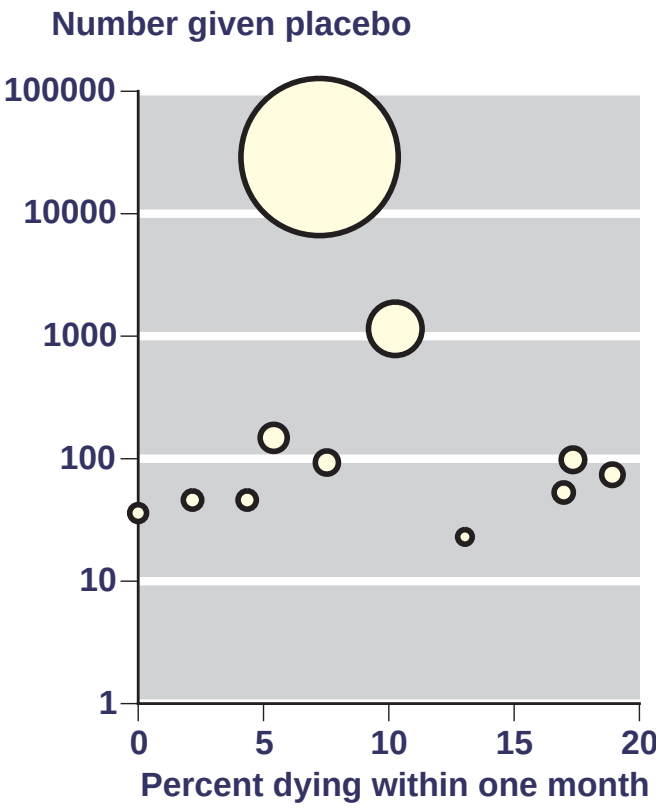



**Table 7: Magnesium meta-analysis and RCTs**

| Study | Deaths/total | | Relative risk 95% CI |
|---|---|---|---|
| | Magnesium | Control | |
| Meta-analysis | 25/657 | 53/644 | 0.5 (0.3 to 0.7) |
| LIMIT-2 | 90/1159 | 118/1157 | 0.8 (0.6 to 1.0) |
| ISIS-4 | 2216/29011 | 2103/29039 | 1.1 (1.0 to 1.1) |

When the trials became available, they were not supportive of the findings of the meta-analysis.

ISIS 4 was a huge trial (56,000 patients) and offered the opportunity of comparing very large trials with meta-analysis. A BMJ editorial was highly critical of meta-analysis [12] and especially emphasised that results of meta-analysis exclusively based on small trials should be distrusted because "several medium-sized trials of high quality seem necessary to render results trustworthy".

Figure 7 makes the point for magnesium, and shows the percentage of deaths within one month with control. The larger studies had control rates of about 7%. The small studies with about 100 patients or fewer had rates varying from no deaths (when it is hard to show any effect) to 20%. The actual numbers of events and patients, together with the relative risks, are shown in Table 7. Overall, in 66,000 patients, intravenous magnesium gas a relative risk of 1.03 (0.97 to 1.08) for mortality at one month. It didn't work.

The reason that the original meta-analysis suggested that it did was not necessarily bad meta-analysis, but small numbers, as the authors themselves pointed out [11].

## Nicotine replacement therapy

A Cochrane review [13] is typically thorough because the Cochrane Tobacco Addiction Group has its own ongoing register of trials that is being constantly updated. Included were randomised trials in which NRT was compared to placebo or no treatment, or where different doses of NRT were compared. Excluded were trials not reporting cessation rates or with follow-up of less than six months. The main outcome measure was abstinence from smoking after at least six months of follow-up. The most rigorous definition of abstinence for each trial was used, with biochemically validated rates if available.

### Results

Numbers needed to treat for nicotine replacement versus placebo or no treatment controls depended on the size of trials included in an analysis. Table 8 shows that, for patch or gum, the NNT was higher (worse) for studies with more than 500 patients than those with fewer than 500 patients, statistically so for gum.

Figure 8 gives us a clue to why that is. It plots the absolute risk increase obtained by subtracting placebo quit rate from nicotine quit rate for gum or patch, and plotting the absolute risk increase against the number of patients in the trial.

**Figure 8: Difference between nicotine patch or gum and placebo patch or gum for smoking cessation after at least six months**

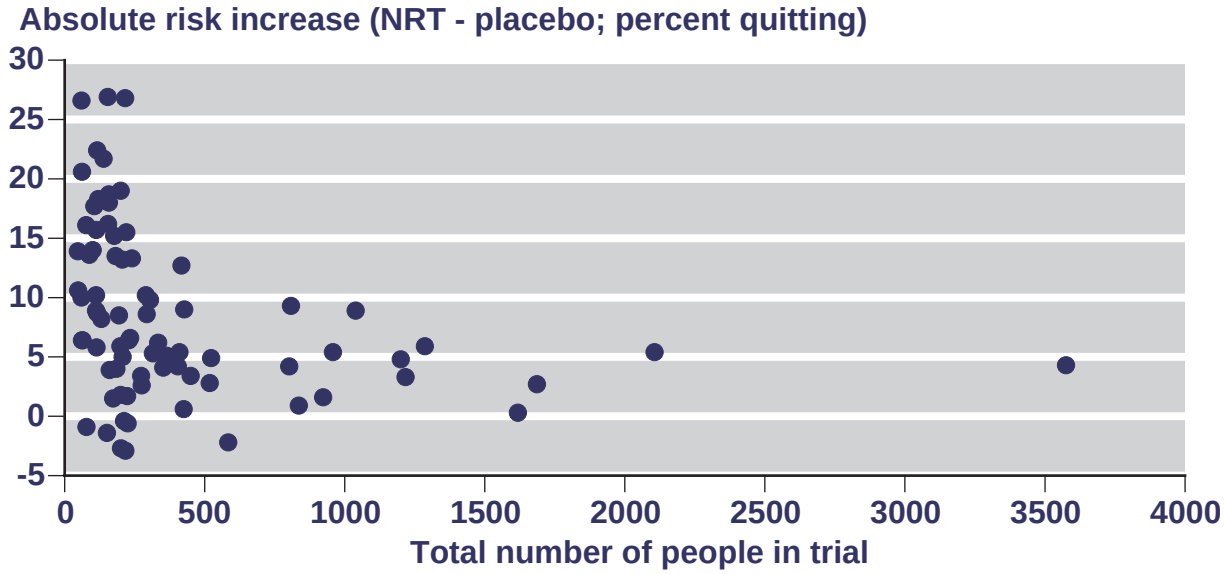



**Table 8: NNTs obtained with larger or smaller trials of nicotine patch and gum**

| | | Number of | | Percent quitting with | | |
|---|---|---|---|---|---|---|
| NRT type | Trial size | Trials | Patients | NRT | Placebo | NNT 95% CI |
| Patch | >500 | 9 | 11,170 | 13 | 8 | 18 (15 - 22) |
| | <500 | 22 | 4,508 | 17 | 9 | 13 (10 - 17) |
| Gum | >500 | 7 | 8,509 | 13 | 9 | 25 (18 - 38) |
| | <500 | 41 | 8,197 | 25 | 15 | 11 (9 - 13) |

What we see is that, for trials of fewer than 500 patients the absolute risk increase runs from -5% to almost 30%, with many trials having absolute risk increases of 10% or more For trials of more than 500 patients the spread is much less, from about -3% to 10%, and absolute risk increases were of the order of 0% to 5%.

## Overall comment

Early in the reign of Augustus, Dionysius of Halicarnasus commented that "history is philosophy from examples". We think of evidence in much the same way, in seeking examples from the archaeology of medicine to learn what constitutes good science and what bad, perhaps leavened here and there with a bit of real philosophy and science. Theory tells us that randomisation is good, and examples from reviews frequently confirm it. Yet we are condemned to relearn the lessons because so many systematic reviews include trials whose architecture potentially misleads.
Cynics might say that much decision-making in healthcare is done on small amounts of inadequate information. They may be right, but knowing that that information may be misleading is still helpful, because we know that we need to examine what we do in practice to check that it conforms with what we thought we started out with. Suspending belief is not an option.

The simple fact is that none of this is particularly new. The idea that larger studies are more reliable has been known for some time [14], and suggestions have been put forward about defining the amount of information required before a result can be known with accuracy [5, 15].

Suggested answers to the sock problem included turning the light on, having only red socks, keep them in different drawers, and washing them together so that they all go pink. With clinical trials it is a bit different, but the one that we can use is that of turning the light on. For any problem we need to be clear what the question is, and apply what is known about the amount of information needed to answer it.

We should also be crystal clear that small studies, even if impeccably performed and reported, can give the wrong answer just because of the random play of chance.

## References:


1 CE Counsell et al. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? BMJ 1994 309: 1677-1681.

2 M Clarke, J Halsey. DICE2: a further investigation of the effects of chance in life, death and subgroup analyses. International Journal of Clinical Practice 2001 55: 240-242.

3 LL Kjaergard & C Gluud. Reported methodologic quality and discrepancies between large and small randomised trials in meta-analyses. Annals of Internal Medicine 2001 135: 982-989.

4 AR Jadad et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Controlled Clinical Trials 1996 17: 1-12.

5 RA Moore et al. Size is everything - large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. Pain 1998 78: 209-16.

6 JP Ioannidis & J Lau. Evolution of treatment effects over time: empirical insight from recursive metaanlyses. Proceedings of the National Academy of Sciences 2001 98: 831-836.

7 Collins SL et al. Single dose oral ibuprofen and diclofenac for postoperative pain (Cochrane Review). In: The Cochrane Library, Issue 4, 2002. Oxford: Update Software.

8 RA Moore et al. Size is everything - large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. Pain 1998 78: 209-16.

9 S Hernández-Diaz, LA García Rodriguez. Association between nonsteroidal anti-inflammatory drugs and upper gastrointestinal tract bleeding and perforation: An overview of epidemiological studies published in the 1990s. Archives of Internal Medicine 2000 160: 2093-2099.

10 J He, et al. Aspirin and risk of hemorrhagic stroke. A meta-analysis of randomized controlled trials. JAMA 1998 280: 1930-1935.

11 Teo KK, et al. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. BMJ. 1991 303:1499-503.

12 M Egger, GD Smith. Misleading meta-analysis: lessons from "an effective, safe, simple" intervention that wasn't. British Medical Journal 1995 310: 752-4.

13 C Silagy et al. Nicotine replacement therapy for smoking cessation (Cochrane Review). In: *The Cochrane Library,* Issue 1, 2001. Oxford: Update Software.

14 MD Flather et al. Strengths and limitations of meta-analysis: larger studies may be more reliable. Controlled Clinical trials 1997 18: 568-579.

15 JM Pogue, S Yusuf. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. Controlled Clinical Trials 1997 18: 580-593.