



Size is everything – large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects

R.A. Moore^{a,*}, David Gavaghan^b, M.R. Tramèr^{a,1}, S.L. Collins^a, H.J. McQuay^a

^a*Pain Research, Nuffield Department of Anaesthetics, The Churchill, Oxford Radcliffe Hospital, Oxford OX3 7LJ, UK*

^b*Computing Laboratory, Wolfson Building, Parks Rd, Oxford OX1 3QD, UK*

Received 26 June 1998; accepted 9 July 1998

Abstract

Variability in patients' response to interventions in pain and other clinical settings is large. Many explanations such as trial methods, environment or culture have been proposed, but this paper sets out to show that the main cause of the variability may be random chance, and that if trials are small their estimate of magnitude of effect may be incorrect, simply because of the random play of chance. This is highly relevant to the questions of 'How large do trials have to be for statistical accuracy?' and 'How large do trials have to be for their results to be clinically valid?' The true underlying control event rate (CER) and experimental event rate (EER) were determined from single-dose acute pain analgesic trials in over 5000 patients. Trial group size required to obtain statistically significant and clinically relevant (0.95 probability of number-needed-to-treat within ± 0.5 of its true value) results were computed using these values. Ten thousand trials using these CER and EER values were simulated using varying group sizes to investigate the variation due to random chance alone. Most common analgesics have EERs in the range 0.4–0.6 and CER of about 0.19. With such efficacy, to have a 90% chance of obtaining a statistically significant result in the correct direction requires group sizes in the range 30–60. For clinical relevance nearly 500 patients are required in each group. Only with an extremely effective drug (EER > 0.8) will we be reasonably sure of obtaining a clinically relevant NNT with commonly used group sizes of around 40 patients per treatment arm. The simulated trials showed substantial variation in CER and EER, with the probability of obtaining the correct values improving as group size increased. We contend that much of the variability in control and experimental event rates is due to random chance alone. Single small trials are unlikely to be correct. If we want to be sure of getting correct (clinically relevant) results in clinical trials we must study more patients. Credible estimates of clinical efficacy are only likely to come from large trials or from pooling multiple trials of conventional (small) size. © 1998 International Association for the Study of Pain. Published by Elsevier Science B.V.

Keywords: Random variability; Placebo; Meta-analysis; Systematic review

1. Introduction

We know that random variation can occur, and expect it to occur in clinical trials. Studies examining how much random variation can contribute to total variation in clinical trials are rare. This paper is about the variability in patients' response to an intervention, whether the intervention is an experimental treatment or control.

If we decide on some indication of success of the treatment, such as relief of at least 50% of a symptom, then a proportion of patients will achieve success with the experimental treatment, and a proportion of patients will achieve success with the control. We use the phrase 'experimental event rate' (EER) to describe the proportion of patients achieving success (the event) with the experimental treatment, and the phrase 'control event rate' (CER) to describe the proportion of patients achieving success (the event) with the control treatment.

Variability in response rates will, with their magnitude, influence how many patients need to be studied to produce a high chance that a clinical trial will come to a statistically

* Corresponding author. Tel: +44 865 226161; fax: +44 865 226160; e-mail: andrew.moore@pru.ox.ac.uk

¹ Present address: Division d'Anesthésiologie, Département APSIC, Hôpital Cantonal Universitaire de Genève, CH-1211 Genève, Switzerland.

significant outcome. Often the results of a single clinical trial are taken into clinical practice, although clinical trials are not powered to measure the magnitude of a result as well as its direction. This paper sets out to investigate random variability in the setting of acute pain, using the number-needed-to-treat (NNT) (Cook and Sackett, 1995) as a marker of clinical relevance of an intervention.

1.1. Observation-success (event) rates vary

The medical literature contains many examples of clinical trials which reach different conclusions about how successful an intervention may be, or whether it works at all. In pain research, for instance, one study with tramadol concluded that it was an excellent analgesic (Sunshine et al., 1992) and another that it had no analgesic effect at all (Stubhaug et al., 1995). The reality is that the proportion of patients who respond to treatment, either with placebo or active therapy, varies, and the extent of that response also varies. Which of the tramadol papers was correct? This paper is about the causes of diversity of trial results, and the impact this has for meta-analysis.

Variation in event rates is seen in many areas of medicine (Soll and McQueen, 1992; Tramèr et al., 1995; Ali and Goetz, 1997) as well as in acute and chronic pain. For this paper we restrict our examples to acute pain. For example, with ibuprofen, there was a huge range in response rates for placebo and ibuprofen 400 mg in randomized, double-blind studies in patients with moderate or severe post-operative pain (Fig. 1). In individual trials between 0% and 60% of patients achieved at least 50% pain relief with placebo, and between about 10% and 100% with ibuprofen 400 mg.

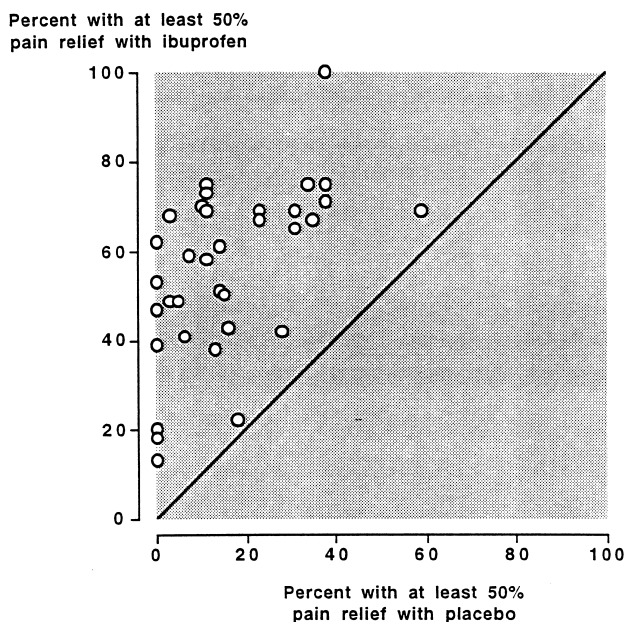


Fig. 1. Percentages of patients with at least 50% pain relief with placebo or ibuprofen 400 mg in randomized double-blind trials.

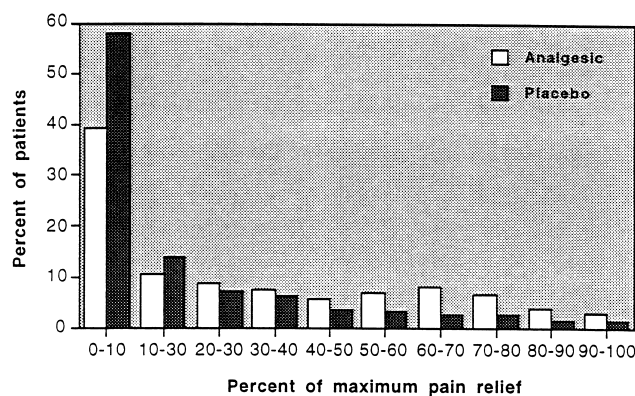


Fig. 2. Percent of maximum pain relief obtained in single-dose randomized double-blind trials in post-operative pain for 826 patients given placebo and 3157 patients given analgesics.

1.2. Source of event rate variability

What is going on? Attempts have been made to try to understand or explain this variability (Cooper, 1991), especially the variability in control event rate (Evans, 1974; McQuay et al., 1996).

1.3. Trial design

One obvious source is trial design. Could there be undiscovered bias despite randomization and the use of double-blind methods, which if true would undermine the confidence placed in clinical trial results?

Randomization controls for selection bias, and the double-blind design is there to control observer bias. Patients may know a placebo was one possible treatment, and investigators know the study design and active treatments; it has been suggested that this can modify patients' behaviour in trials (Gracely et al., 1985; Wall, 1993). Patients may have opportunities to communicate with each other. Doctors know the trial design when recruiting patients, which may be a source of bias (Bergmann et al., 1994). Nurse observers often spend most time with patients, and the nurse might be able to influence a patient's response by his/her demeanour based on experience of other patients' reactions. That would produce time-dependent changes in study results as has been seen before (Shapiro et al., 1954).

1.4. Population

The reason for large variations in control event rates with placebo may have something to do with the population studied – Scottish stoics versus Welsh wimps. There is little evidence for this, but there may be differences between men and women, or in response in different clinical settings (Moore and McQuay, 1997).

1.5. Environment

Another explanation may be the environmental situation

in which a trial is conducted. Inpatients in a nice hospital with a charming nurse might have a good response while outpatients filling in diaries alone at home might not (Ulrich, 1983). Other clinical or societal factors which we have yet to recognize may influence event rates.

1.6. Random effects

The observation is that an individual patient can have no pain relief or 100% pain relief. That is true whether they receive placebo or active treatment (Fig. 2) (McQuay et al., 1996). Clearly if we choose only one patient to have placebo and only one patient to have treatment, either or both could pass or fail to reach the dichotomous hurdle of at least 50% pain relief. The more patients who have the treatment or placebo, the more likely we are to have a result which reflects the true underlying distribution. But how many is enough for us to be comfortable that random effects can be ignored?

Until the full effects of the random play of chance are appreciated, we cannot begin to unravel effects of trial design, or population or environmental effects. The aim of this paper is to examine the effects of random chance, and we describe this by using trials of single doses of analgesics in acute pain of moderate or severe intensity, together with the implications of random chance for meta-analysis. The two issues are the impact of random effects both on the direction of an effect and on its magnitude.

The following section describes the origin of the data used to determine real control event rate (CER) in acute pain trials and the questions to be addressed by the calculations and simulations.

2. Methods

2.1. Data origins

We have had access to individual patient data from randomized, double-blind, single-dose evaluations of analgesics in over 5000 patients (McQuay et al., 1996; Moore and McQuay, 1997). In reviews we used published information from many hundreds of trials in acute pain (Moore et al., 1997). The strength of these studies is that they used standard methods of pain assessment in the same pain conditions, and, because they were randomized and double-blind, they were relatively free from known sources of bias.

This constitutes unique information which can be used to provide a very strong indication of the true underlying distribution of pain relief for placebo and active treatments. Three sets of data were used as the basis for mathematical modelling of acute pain studies: (1) individual patient data for analgesics and placebo in acute pain (Fig. 2); (2) placebo event rates in clinical trials in acute pain (Fig. 3); (3) group sizes commonly used in clinical trials in acute pain (Fig. 4).

2.2. Questions

Assume that we have a dichotomous outcome measure of at least 50% pain relief applied to standardized acute pain trials. If we knew the true control event rate (CER) with placebo and the true experimental event rate (EER) for some analgesic for this measure, then we would like to answer the following questions (throughout we will assume equal group sizes).

1. How unlikely are we to get the wrong answer (statistical significance)? For our assumed CER and EER, how many patients have to be studied to give a certain probability (0.5, 0.75, 0.9 and 0.95) of not getting the wrong answer (i.e. of getting a statistically significant result in the correct direction and rejecting the null hypothesis)?
2. How likely are we to get the correct answer (clinical relevance)? We also want to know the clinical relevance of any result (i.e. we want to know both the direction of the result and its magnitude). We have chosen NNT (McQuay and Moore, 1997) as a measure of clinical relevance, so we want to know how large the group size has to be to give a certain probability that the value of an NNT is credible within clinically acceptable bounds. In our case, we have chosen to define clinical relevance to be ± 0.5 , and our preferred value of certain probability will be 0.95. So to re-phrase the question using these concrete values: how many patients need to be studied for us to have a 95% chance that the NNT is within ± 0.5 of its true value?
3. Finally we want to know the effect of the random play of chance on clinical trials in which the true underlying CER and EER are known. This can be expressed as a plot of the probability distribution for CER against EER (L'Abbé et al., 1987) to indicate where a single trial conducted according to standard protocols and of conventional size is likely to lie just because of chance and excluding any environmental effects.

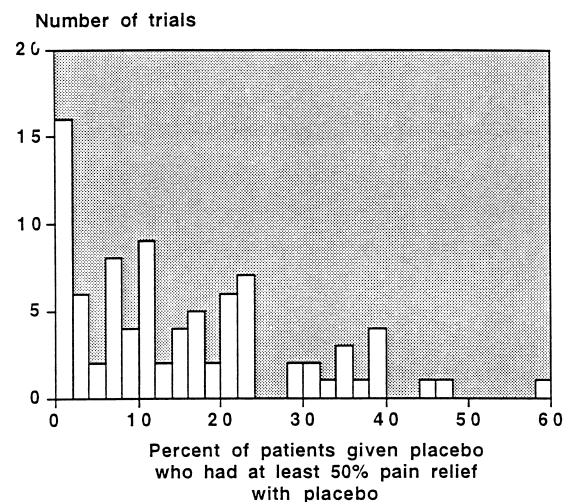


Fig. 3. Placebo response rates in 87 randomized double-blind trials in post-operative pain.

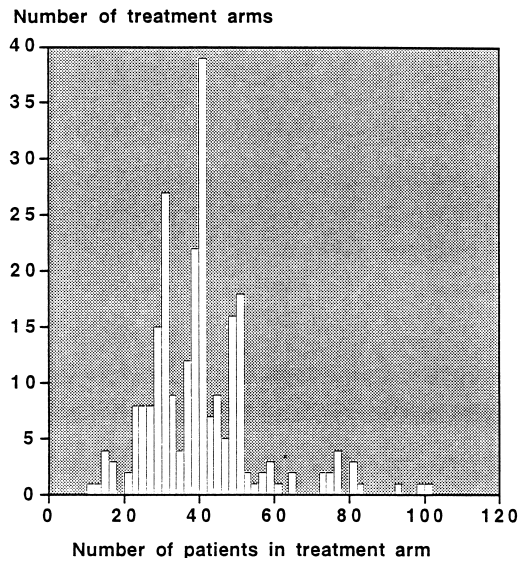


Fig. 4. Frequency of trial size in 244 treatment arms from 90 trials.

2.3. Calculations and simulations

2.3.1. Statistical significance

When dealing with binary data in clinical trials by far the most commonly used statistical method is the simple 2×2 contingency table with continuity correction, using the χ^2 -test. This method was therefore used to answer the question of how large does the group size, n , have to be to give a certain probability (0.5, 0.75, 0.9 and 0.95) of getting a statistically significant result in the correct direction (i.e. of not getting the wrong answer) for our assumed CER and EER. The level of significance was taken as 0.95 (corresponding to a P -value of 0.05), because this is the most commonly used value in clinical practice.

Let E be a discrete random variable representing the number of experimental events and let C be a discrete random variable representing the number of control events. Note that E and C are independent binomial random variables, so that for a group size of n , and an EER, say of p , and a CER of q , we can calculate the probability \Pr (Eqs. (1) and (2)):

$$\Pr(E=n_e) = \binom{n}{n_e} p^{n_e} (1-p)^{n-n_e} \quad (1)$$

$$\Pr(C=n_c) = \binom{n}{n_c} q^{n_c} (1-q)^{n-n_c} \quad (2)$$

where n_e, n_c are the observed numbers of experimental and control events respectively, and by independence for any pair (n_e, n_c) we have (Eq. (3))

$$\Pr(E=n_e, C=n_c) = \Pr(E=n_e)\Pr(C=n_c) \quad (3)$$

Let T be the total probability that we obtain a significant result for given values of n, p and q i.e. the probability that we get the ‘right answer’. Let (n_e, n_c) denote any pair (n_e, n_c) for which we obtain a significant result from the χ^2 test. Then (Eq. (4))

$$T = \sum_{(n_e, n_c)} \Pr(E=n_e, C=n_c) = \sum_{(n_e, n_c)} \binom{n}{n_e} p^{n_e} (1-p)^{n-n_e} \binom{n}{n_c} q^{n_c} (1-q)^{n-n_c} \quad (4)$$

In algorithmic form, this is achieved as follows.

1. Set the variable T to zero.
2. For each pair (n_e, n_c) such that $n_e > n_c$ (we have assumed the treatment is more effective than the control) determine whether n_e and n_c give a statistically significant result using the χ^2 -test.
3. If step 2 gives a statistically significant result, then calculate the probability of obtaining the pair (n_e, n_c) and add this probability to the sum T defined in 1.
4. Once steps 2 and 3 have been performed for all n_e from n to 0, and for all n_c from 0 to n_e the variable T will contain the desired probability.

This will be clearer with a concrete example and we will choose a small value of n for convenience (it should be noted that this particular example is very inaccurate, since the approximations associated with the χ^2 test are very poor for such small values of n , but it is useful for illustration). Suppose that the CER $q = 0.16$, the EER $p = 0.5$, and group size $n = 6$. The possible values for the number of experimental events are $n_e = 0, 1, 2, 3, 4, 5, 6$. However to obtain a significant result in the correct direction we must have a greater number of experimental events than control events, so we only need to test values of n_c up to the value of n_e . So, starting with $n_e = 6$, we check $n_c = 0, 1, 2, 3, 4, 5$ to see if any are significant, then for $n_e = 5$ we check $n_c = 0, 1, 2, 3, 4$ and so on. In this process we find that only the pairs (6,0), (6,1) and (5,0) are significant. The binomial probabilities of obtaining each of these pairs can then be calculated as above. The easiest to calculate is (6,0), since the probability of obtaining six events in the active group is $(0.5)^6$, and of obtaining 0 events in the control is $(0.84)^6$ and taking the product gives 5.49×10^{-3} . In total these three pairs give a probability $T = 0.0447$ for $n = 6$, which can be interpreted as follows: in a trial with six patients in each of the active and control groups, and a true EER of 0.5 and a true CER of 0.16, the probability of obtaining a significant result at the 0.95 significance level using the standard χ^2 test is just 0.0447 i.e. we would get the ‘right answer’ in fewer than 5% of such trials.

2.4. Clinical relevance

All of the above relates only to the group size necessary to have a certain probability of obtaining a statistically significant improvement of active over control in an RCT. But we want to know the clinical credibility of the intervention, using the NNT as the measure of clinical effectiveness. If we can give an accurate measurement of the NNT of an analgesic, then we can answer the more important question

for the practising clinician, which is not ‘does this intervention work?’, but rather ‘how well does this intervention work compared with other possible interventions?’ (Eq. (5))

If we again take the CER to be q and the EER to be p , then the NNT is defined to be

$$\text{NNT} = \frac{1}{p - q} \tag{5}$$

or the reciprocal of the absolute risk reduction. Given these values of p and q , and a particular group size n , we wish to find the probability that the observed NNT will be within ± 0.5 of the true NNT. The NNT is again a discrete (although non-integer) random variable dependent on the absolute risk, which is in turn dependent on the variables E and C , the number of experimental and control events. If we let $Z = E - C$ (so that Z can take values between $-n$ and $+n$), then the $\text{NNT} = n/(E - C) = n/Z$. We wish to obtain the probability that a particular observed NNT will be within ± 0.5 of the true NNT, that is

$$\begin{aligned} \Pr\left(-0.5 \leq \text{NNT} - \frac{1}{p - q} \leq +0.5\right) &= \Pr\left(-0.5 \leq \frac{n}{Z} - \frac{1}{p - q} \leq +0.5\right) \\ &= \Pr\left(-0.5 + \frac{1}{p - q} \leq \frac{n}{Z} \leq +0.5 + \frac{1}{p - q}\right) \\ &= \Pr\left(\frac{2 - p - q}{2(p - q)} \leq \frac{n}{Z} \leq \frac{2 + p - q}{2(p - q)}\right) \\ &= \Pr\left(\frac{2n(p - q)}{2 + p - q} \leq Z \leq \frac{2n(p - q)}{2 - p + q}\right) \tag{6} \\ &= \Pr(n_{z_l} \leq Z \leq n_{z_u}) \\ &= \sum_{n_z = n_{z_l}}^{n_z = n_{z_u}} \Pr(Z = n_z) \tag{7} \end{aligned}$$

where n_{z_l} and n_{z_u} are the nearest integers above and below the left and right hand sides respectively of the inequality in Eq. (6). Since we know that the probability distributions of E and C are binomial, we can obtain the probability distribution of Z as follows (Eq. (8))

$$\Pr(Z = n_z) = \Pr(E - C = n_z) = \Pr(E = n_z + n_c, C = n_c) =$$

$$\sum_{n_c = 0}^{n_c = n - n_z} \binom{n}{n_z + n_c} p^{n_z + n_c} (1 - p)^{n - (n_z + n_c)} \binom{n}{n_c} q^{n_c} (1 - q)^{n - n_c} \tag{8}$$

Substituting this into Eq. (7) gives the required probability that the NNT will be within clinical accuracy.

2.5. Simulations to investigate the distribution of CER and EER

We want to know the likely spread of CER and EER over the range of probable group sizes due to random chance alone. This spread can be shown as a L’Abbé plot of CER against EER (L’Abbé et al., 1987). To do this 10 000 randomized controlled trials will be simulated, each of size n and consisting of an active arm with EER p , and a control arm with CER q . The group size n will also be randomly generated from a normal distribution with mean 40, standard deviation 15 and minimum group size 10. These simulations can also be used as a convenient check on the theoretical results for the NNT derived above, so a calculation of the NNT is included in the simulation algorithm so that the proportion of simulated NNTs within ‘clinical accuracy’ can be derived. The algorithm for the simulation is then as follows:

1. Generate the group size n from the normal distribution (if $n < 10$ then repeat).
2. For each of the n patients in the control group generate a random number, r say, uniformly distributed between 0 and 1. If $r < q$ then add 1 to the number of control events. This will result in a simulated value of the total number of control events, say n_c , and calculate the observed CER as n_c/n .
3. Repeat step 2 for the EER (so now use $r < p$) etc. to obtain the observed EER as n_e/n . (Note: the expected number of control events will be $n \times q$ and of experimental events $n \times p$ each time we do this, as required.)
4. Calculate the NNT of this simulated trial from the observed EER and CER obtained in steps 2 and 3. If NNT is infinite then do not include this trial when calculating the mean NNT.
5. Repeat steps 1–4 10 000 times. Calculate the NNT from these 10 000 trials and count the number of simulated trials in which the simulated NNT value is within ± 0.5 of the true NNT.

Note that for any given group size n there are $(n + 1)^2$ possible pairs $(n_e/n, n_c/n)$ of observed EER and CER. By plotting the frequency with which each of these pairs occurs within a particular unit of area (we choose to divide the region into squares of side 0.1) as a two-dimensional function of the CER and EER, we can obtain a measure of the likely distribution of CER and EER as the group size is varied.

3. Results

3.1. Statistical significance

The results obtained for statistical significance are summarized in Table 1, where we give the group size necessary to have probabilities of 0.5, 0.75, 0.9 and 0.95 of obtaining a

Table 1

Number of patients required in each group for a statistically significant result at the 0.95 level

Probability	Experimental event rate							
	0.30	0.35	0.40	0.45	0.50	0.60	0.70	0.80
0.50	83	34	26	19	17	13	10	7
0.75	137	55	40	32	27	21	14	10
0.90	200	79	57	44	34	29	20	13
0.95	244	95	68	53	41	33	23	16

Group sizes required to obtain a probability of 0.5, 0.75, 0.9 and 0.95 of obtaining a statistically significant result from the χ^2 test with a CER of 0.16 and EERs from 0.3 to 0.8.

statistically significant result at the 0.95 level from the χ^2 -test with a CER of 0.16 and EERs from 0.3 to 0.8. The calculations for small values of n are not accurate, but are included for completeness. The normal approximations on which the χ^2 is based become increasingly accurate for $n > 10$.

Most common analgesics have EERs in the range 0.4–0.6, and it is clear that to have any degree of certainty of obtaining a statistically significant result we must use group sizes of at least 40. If we expect an analgesic to have an EER falling within the most common range of 0.4–0.6, then to have a 90% chance of obtaining a statistically significant result in the right direction we should choose group sizes in the range 30–60 (depending on our confidence in the efficacy of the drug), and for a 95% chance we should choose a group size in the range 30–70. It is clear that, even for the more potent common analgesics with EERs of 0.5, a group size of 20 would only be expected to indicate a significant improvement over placebo half of the time, suggesting that trials with so few patients are a waste of time and money, and should be considered ethically unsound.

3.2. Clinical relevance

Table 2 gives the approximate group size necessary to have a probability of 0.5, 0.75, 0.90 and 0.95 of obtaining a clinically relevant value of the NNT again with a CER of 0.16 and with EERs from 0.3 to 0.8. Comparing this with the group sizes needed (Table 1) to obtain a statistically significant result, we can see that except for very high EERs, we

Table 2

Number of patients required in each group for a clinically relevant number needed to treat

Probability	Experimental event rate					
	0.30	0.40	0.50	0.60	0.70	0.80
0.50	>500	200	50	20	10	<10
0.75	>500	>500	150	60	25	10
0.90	>500	>500	320	110	50	20
0.95	>500	>500	470	180	80	40

Group sizes required to obtain a probability of 0.5, 0.75, 0.9 and 0.95 of obtaining a clinically relevant NNT (NNT within ± 0.5 of true value) with a CER of 0.16 and EERs from 0.3 to 0.8.

need about 10 times as many patients in each group for clinical relevance. For the commonest value of EER of 0.5 we can now answer the question that we posed: we need nearly 500 patients in each group to have a probability of 0.95 of obtaining the NNT to within ± 0.5 of its true value. Only when we have an extremely effective drug with an EER in excess of 0.8 will we be reasonably sure of obtaining a clinically credible NNT with the commonly used group size of around 40 patients per treatment arm.

3.3. Simulated trials

Fig. 5 shows the results of simulating 10 000 trials, each with underlying CER = 0.16 and EER = 0.5, using randomly varying group sizes generated from a normal distribution with mean 40 ± 15 , but with minimum group size restricted to 10. This gives a distribution of group sizes very similar to that observed in practice for acute post-operative pain (Fig. 1). The frequency with which each possible pair of EER and CER values occurred was counted and this was used to construct a two-dimensional contour plot (Fig. 5)

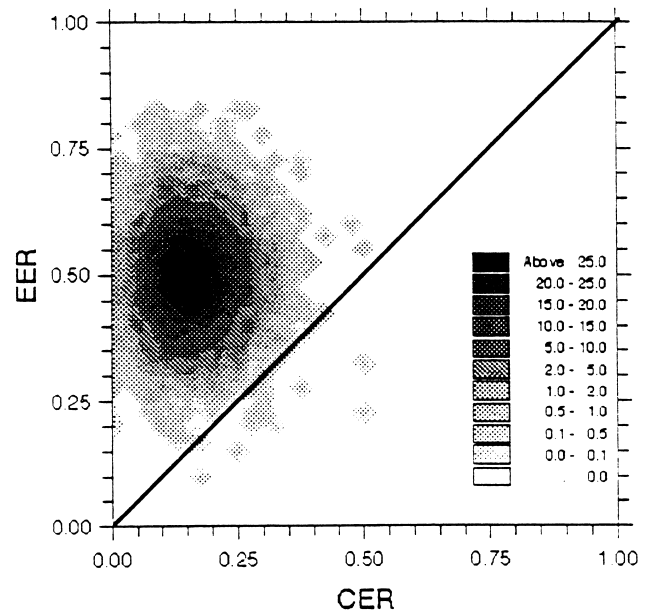


Fig. 5. Two-dimensional L'Abbe plot of the probability density for trials in acute post-operative pain.

Table 3

NNT accuracy – percentage of 10 000 simulated NNTs which were within ± 0.5 of the true NNT

Group size	Experimental event rate		
	0.40	0.50	0.60
25	26.2	36.7	56.7
50	27.5	50.6	72.6
100	37.7	60.8	87.6
200	54.7	81.1	95.9
300	63.1	88.5	98.6
400	71.4	93.3	99.4
500	73.9	95.3	99.8

Percent of trials in which the simulated NNT was within ± 0.5 of the true NNT value, with increasing group size, and with a CER of 0.16 and EERs between 0.4 and 0.6.

which shows the density per unit area (as a percentage of the total number of trials simulated) of these pairs.

3.4. Simulated NNTs

This simulation was also used to check the theoretical calculations of the probability of a trial producing a clinically relevant NNT by counting the number of times our simulated NNT values are within ± 0.5 of the true NNT value out of the 10 000 simulations. For CER = 0.16 and values of EER between 0.4 and 0.6, Table 3 gives the percentage of trials in which the simulated NNT was within ± 0.5 of the true NNT value. The results in Table 3 confirm that if we want to obtain a measure of the NNT which is of acceptable clinical accuracy, we should use group sizes of a minimum of 500 patients in both the active and control arms.

4. Discussion

This paper shows that size is everything. The variability in the response rates to both placebo and active treatments means that if we want to be sure of getting the correct (clinically relevant) result in both direction and magnitude in clinical trials of analgesics we must study more patients than the conventional 40 patients per group, a number chosen to be confident of not getting the wrong answer in direction only.

This variability in the response rates to both placebo and active treatments has been recognized before, and was blamed on either flaws in trial design and conduct, or on non-specific effects of placebo (Evans, 1974). We contend that much of this variability is due to random chance alone, and we need not search for abstruse causes. This variability is the likely cause of the two discordant reports of tramadol's efficacy (Sunshine et al., 1992; Stubhaug et al., 1995). It also justifies clinical conservatism, the caution necessary before taking the results of a single (small) trial into practice. That single small trial is unlikely to give accurate

results. A trial with group sizes of 40 could have NNT values between one and nine just by chance, when the true value was three.

Most clinical trials of analgesics are performed to demonstrate statistical superiority over placebo, and are powered to be confident of not getting the wrong answer. To achieve this, group sizes of about 40 patients are used; 95% of the time this will yield the desired statistical superiority over placebo, given a useful intervention like 400 mg of ibuprofen (Table 1). But to reach a clinically credible estimate of efficacy, defined as a NNT within ± 0.5 of the true value, we need 10 times as many patients (Table 2).

Acute pain trials with 1000 patients are never done. Credible estimates of clinical relevance which examine both magnitude and direction of an effect will only come from pooling multiple trials of conventional size. The sting in the tail here is that those estimates also need data on 1000 patients, or more than 10 trials, to achieve this credibility (Table 2).

Comparing Figs. 1 and 5, all the points on Fig. 1 fall within the variability predicted due to random chance alone. No other explanation is necessary. Only when we have substantial data should we investigate other possible influences such as pain model (Moore and McQuay, 1997), population studied and nebulous environmental factors.

These principles almost certainly apply in other clinical settings (Soll and McQueen, 1992; Tramèr et al., 1995; Ali and Goetz, 1997). Powering trials for statistical significance is arguably not good enough, because the magnitude of the clinical effect will still be uncertain. Clinically useful trials also need clinically useful outcomes, as well as trial size big enough to allow us to be confident about magnitude of the effect. We need to know what degree of improvement on a particular scale matters to the patient (Guyatt et al., 1998). This is quite a challenge to the way we do clinical trials at present, where the focus is on the minimum size necessary for statistical significance.

Acknowledgements

The work was supported by Pain Research funds, European Union Biomed 2 contract BMH4 CT95 0172 and NHS Research and Development Health Technology Assessment Programme 94/11/4. DG was supported by an MRC Career Development Fellowship and MRT by the Swiss National Scientific Foundation. European Union Biomed 2 contract BMH4 CT95 0172. NHS Research and Development Health Technology Assessment Programme 94/11/4.

References

- Ali, M.Z. and Goetz, M.B., A meta-analysis of the relative efficacy and toxicity of single daily dosing versus multiple daily dosing of aminoglycosides, *Clin. Infect. Dis.*, 24 (1997) 796–809.

- Bergmann, J., Chassany, O., Gandiol, J., Deblois, P., Kanis, J.A. and Segrestaa, J. et al., A randomised clinical trial of the effect of informed consent on the analgesic activity of placebo and naproxen in cancer pain, *Clin. Trials Meta-Anal.*, 29 (1994) 41–47.
- Cook, R.J. and Sackett, D.L., The number needed to treat: a clinically useful measure of treatment effect, *Br. Med. J.*, 310 (1995) 452–454.
- Cooper, S.A., Single-dose analgesic studies: the upside and downside of assay sensitivity. In: M.B. Max, R.K. Portenoy, E.M. Laska, (Eds.), *The design of analgesic clinical trials (Advances in Pain Research and Therapy Vol. 18)*, Raven Press, New York, 1991, pp. 117–124.
- Evans, F.J., The placebo response in pain reduction. In: J.J. Bonica, (Ed.), *Advances in Neurology*, Vol. 4. Raven Press, New York, 1974, pp. 289–296.
- Gracely, R.H., Dubner, R., Deeter, W.R. and Wolskee, P.J., Clinicians' expectations influence placebo analgesia, *Lancet*, 1 (1985) 43.
- Guyatt, G.H., Juniper, E.F., Walter, S.D., Griffith, L.E. and Goldsmith, R.S., Interpreting treatment effects in randomised trials, *Br. Med. J.*, 316 (1998) 690–693.
- L'Abbé, K.A., Detsky, A.S. and O'Rourke, K., Meta-analysis in clinical research, *Ann. Intern. Med.*, 107 (1987) 224–233.
- McQuay, H., Carroll, D. and Moore, A., Variation in the placebo effect in randomised controlled trials of analgesics: all is as blind as it seems, *Pain*, 64 (1996) 331–335.
- McQuay, H.J. and Moore, R.A., Using numerical results from systematic reviews in clinical practice, *Ann. Intern. Med.*, 126 (1997) 712–720.
- Moore, A., Collins, S., Carroll, D. and McQuay, H., Paracetamol with and without codeine in acute pain: a quantitative systematic review, *Pain*, 70 (1997) 193–201.
- Moore, R.A. and McQuay, H.J., Single-patient data meta-analysis of 3453 postoperative patients: oral tramadol versus placebo, codeine and combination analgesics, *Pain*, 69 (1997) 287–294.
- Shapiro, A.P., Myers, T., Reiser, M.F. and Ferris, E.B., Comparison of blood pressure response to veriloid and to the doctor, *Psychosom. Med.*, 16 (1954) 478–488.
- Soll, J.C. and McQueen, M.C., Respiratory distress syndrome. In: J.C. Sinclair, M.E. Bracken, (Eds.), *Effective care of the newborn infant*, Oxford University Press, Oxford, 1992, pp. 333.
- Sunshine, A., Olson, N.Z., Zigelboim, I., DeCastro, A. and Minn, F.L., Analgesic oral efficacy of tramadol hydrochloride in postoperative pain, *Clin. Pharmacol. Ther.*, 51 (1992) 740–746.
- Stubhaug, A., Grimstad, J. and Breivik, H., Lack of analgesic effect of 50 and 100 mg oral tramadol after orthopaedic surgery: a randomized, double-blind, placebo and standard active drug comparison, *Pain*, 62 (1995) 111–118.
- Tramèr, M., Moore, A. and McQuay, H., Prevention of vomiting after paediatric strabismus surgery: a systematic review using the numbers-needed-to-treat method, *Br. J. Anaesth.*, 75 (1995) 556–561.
- Ulrich, R.S., View through a window may influence recovery from surgery, *Science*, 224 (1983) 420–421.
- Wall, P.D., *Pain and the placebo response. Experimental and theoretical studies of consciousness*. CIBA Foundation Symposium 174, Wiley, Chichester, 1993, pp. 187–216.