An introduction to power and sample size estimation

S R Jones, S Carley and M Harrison

Emerg. Med. J. 2003;20;453-458
doi:10.1136/emj.20.5.453

STATISTICS

An introduction to power and sample size estimation

S R Jones, S Carley, M Harrison

Emerg Med J 2003;**20**:453–458

The importance of power and sample size estimation for study design and analysis.

OBJECTIVES

1 Understand power and sample size estimation.

2 Understand why power is an important part of both study design and analysis.

3 Understand the differences between sample size calculations in comparative and diagnostic studies.

4 Learn how to perform a sample size calculation.

- (a) For continuous data
- (b) For non-continuous data
- (c) For diagnostic tests

POWER AND SAMPLE SIZE ESTIMATION

Power and sample size estimations are measures of how many patients are needed in a study. Nearly all clinical studies entail studying a sample of patients with a particular characteristic rather than the whole population. We then use this sample to draw inferences about the whole population.

In previous articles in the series on statistics published in this journal, statistical inference has been used to determine if the results found are true or possibly due to chance alone. Clearly we can reduce the possibility of our results coming from chance by eliminating bias in the study design using techniques such as randomisation, blinding, etc. However, another factor influences the possibility that our results may be incorrect, the number of patients studied. Intuitively we assume that the greater the proportion of the whole population studied, the closer we will get to true answer for that population. But how many do we need to study in order to get as close as we need to the right answer?

WHAT IS POWER AND WHY DOES IT MATTER

Power and sample size estimations are used by researchers to determine how many subjects are needed to answer the research question (or null hypothesis).

An example is the case of thrombolysis in acute myocardial infarction (AMI). For many years clinicians felt that this treatment would be of benefit given the proposed aetiology of AMI, however successive studies failed to prove the case. It was not until the completion of adequately powered "mega-trials" that the small but important benefit of thrombolysis was proved.

Generally these trials compared thrombolysis with placebo and often had a primary outcome measure of mortality at a certain number of days. The basic hypothesis for the studies may have compared, for example, the day 21 mortality of thrombolysis compared with placebo. There are two hypotheses then that we need to consider:

1 The null hypothesis is that there is no difference between the treatments in terms of mortality.

2 The alternative hypothesis is that there is a difference between the treatments in terms of mortality.

In trying to determine whether the two groups are the same (accepting the null hypothesis) or they are different (accepting the alternative hypothesis) we can potentially make two kinds of error. These are called a type I error and a type II error.

A type I error is said to have occurred when we reject the null hypothesis incorrectly (that is, it is true and there is no difference between the two groups) and report a difference between the two groups being studied.

A type II error is said to occur when we accept the null hypothesis incorrectly (that is, it is false and there is a difference between the two groups which is the alternative hypothesis) and report that there is no difference between the two groups.

They can be expressed as a two by two table (table 1).

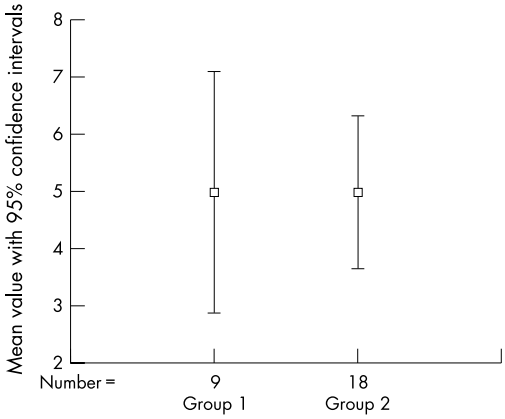
Power calculations tell us how many patients are required in order to avoid a type I or a type II error.

Table 1 Two by two table

		Actual truth	
		Treatment benefit	No treatment benefit
Clinical trial result	Treatment benefit	Correct result	Type I error False positive result
	No treatment benefit	Type II error False negative result	Correct result

See end of article for authors' affiliations

Correspondence to:
Dr S R Jones, Emergency Department, Manchester Royal Infirmary, Oxford Road, Manchester M13 9WL, UK;
steve.r.jones@bigfoot.com



Frequency table

Group 1	Group 2
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
	1
	2
	3
	4
	5
	6
	7
	8
	9

Figure 1 Change in confidence interval width with increasing numbers of subjects.

The term power is commonly used with reference to all sample size estimations in research. Strictly speaking "power" refers to the number of patients required to avoid a type II error in a comparative study. Sample size estimation is a more encompassing term that looks at more than just the type II error and is applicable to all types of studies. In common parlance the terms are used interchangeably.

WHAT AFFECTS THE POWER OF A STUDY?

There are several factors that can affect the power of a study. These should be considered early on in the development of a study. Some of the factors we have control over, others we do not.

The precision and variance of measurements within any sample

Why might a study not find a difference if there truly is one? For any given result from a sample of patients we can only determine a probability distribution around that value that will suggest where the true population value lies. The best known example of this would be 95% confidence intervals. The size of the confidence interval is inversely proportional to the number of subjects studied. So the more people we study the more precise we can be about where the true population value lies.

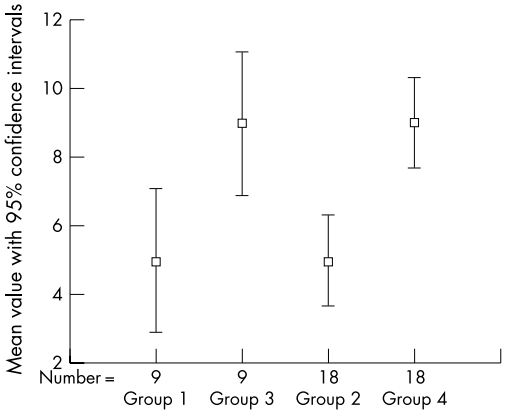
Figure 1 shows that for a single measurement, the more subjects studied the narrower the probability distribution becomes. In group 1 the mean is 5 with wide confidence intervals (3–7). By doubling the number of patients studied (but in our example keeping the values the same) the confidence intervals have narrowed (3.5–6.5) giving a more precise estimate of the true population mean.

The probability distribution of where the true value lies is an integral part of most statistical tests for comparisons between groups (for example, *t* tests). A study with a small sample size will have large confidence intervals and will only show up as statistically abnormal if there is a large difference between the two groups. Figure 2 demonstrates how increasing the number of subjects can give a more precise estimate of differences.

The magnitude of a clinically significant difference

If we are trying to detect very small differences between treatments, very precise estimates of the true population value are required. This is because we need to find the true population value very precisely for each treatment group. Conversely, if we find, or are looking for, a large difference a fairly wide probability distribution may be acceptable.

In other words if we are looking for a big difference between treatments we might be able to accept a wide probability distribution, if we want to detect a small difference we will need great precision and small probability distributions. As the width of probability distributions is largely determined by how many subjects we study it is clear that the difference sought affects sample size calculations.



Frequency table

Group 1	Group 2	Group 3	Group 4
1	1	5	5
2	2	6	6
3	3	7	7
4	4	8	8
5	5	9	9
6	6	10	10
7	7	11	11
8	8	12	12
9	9	13	13
	1		5
	2		6
	3		7
	4		8
	5		9
	6		10
	7		11
	8		12
	9		13

Figure 2 Effect of confidence interval reduction to demonstrate a true difference in means. This example shows that the initial comparison between groups 1 and 3 showed no statistical difference as the confidence intervals overlapped. In groups 3 and 4 the number of patients is doubled (although the mean remains the same). We see that the confidence intervals no longer overlap indicating that the difference in means is unlikely to have occurred by chance.

Factors affecting a power calculation

- The precision and variance of measurements within any sample
- Magnitude of a clinically significant difference
- How certain we want to be to avoid type 1 error
- The type of statistical test we are performing

When comparing two or more samples we usually have little control over the size of the effect. However, we need to make sure that the difference is worth detecting. For example, it may be possible to design a study that would demonstrate a reduction in the onset time of local anaesthesia from 60 seconds to 59 seconds, but such a small difference would be of no clinical importance. Conversely a study demonstrating a difference of 60 seconds to 10 minutes clearly would. Stating what the "clinically important difference" is a key component of a sample size calculation.

How important is a type I or type II error for the study in question?

We can specify how concerned we would be to avoid a type I or type II error. A type I error is said to have occurred when we reject the null hypothesis incorrectly. Conventionally we choose a probability of <0.05 for a type I error. This means that if we find a positive result the chances of finding this (or a greater difference) would occur on less than 5% of occasions. This figure, or significance level, is designated as $p\alpha$ and is usually pre-set by us early in the planning of a study, when performing a sample size calculation. By convention, rather than design, we more often than not choose 0.05. The lower the significance level the lower the power, so using 0.01 will reduce our power accordingly.

(To avoid a type I error—that is, if we find a positive result the chances of finding this, or a greater difference, would occur on less than α% of occasions)

A type II error is said to occur when we accept the null hypothesis incorrectly and report that there is no difference between the two groups. If there truly is a difference between the interventions we express the probability of getting a type II error and how likely are we to find it. This figure is referred to as $p\beta$. There is less convention as to the accepted level of $p\beta$, but figures of 0.8–0.9 are common (that is, if a difference truly exists between interventions then we will find it on 80%–90% of occasions.)

The avoidance of a type II error is the essence of power calculations. The power of a study, $p\beta$, is the probability that the study will detect a predetermined difference in measurement between the two groups, if it truly exists, given a pre-set value of $p\alpha$ and a sample size, N.

The type of statistical test we are performing

Sample size calculations indicate how the statistical tests used in the study are likely to perform. Therefore, it is no surprise that the type of test used affects how the sample size is calculated. For example, parametric tests are better at finding differences between groups than non-parametric tests (which is why we often try to convert basic data to normal distributions). Consequently, an analysis reliant upon a non-parametric test (for example, Mann-Whitney U) will need more patients than one based on a parametric test (for example, Student's *t* test).

SHOULD SAMPLE SIZE CALCULATIONS BE PERFORMED BEFORE OR AFTER THE STUDY?

The answer is definitely before, occasionally during, and sometimes after.

In designing a study we want to make sure that the work that we do is worthwhile so that we get the correct answer and we get it in the most efficient way. This is so that we can recruit enough patients to give our results adequate power but not too many that we waste time getting more data than we need. Unfortunately, when designing the study we may have to make assumptions about desired effect size and variance within the data.

Interim power calculations are occasionally used when the data used in the original calculation are known to be suspect. They must be used with caution as repeated analysis may lead to a researcher stopping a study as soon as statistical significance is obtained (which may occur by chance at several times during subject recruitment). Once the study is underway analysis of the interim results may be used to perform further power calculations and adjustments made to the sample size accordingly. This may be done to avoid the premature ending of a study, or in the case of life saving, or hazardous therapies, to avoid the prolongation of a study. Interim sample size calculations should only be used when stated in the a priori research method.

When we are assessing results from trials with negative results it is particularly important to question the sample size of the study. It may well be that the study was underpowered and that we have incorrectly accepted the null hypothesis, a type II error. If the study had had more subjects, then a difference may well have been detected. In an ideal world this should never happen because a sample size calculation should appear in the methods section of all papers, reality shows us that this is not the case. As a consumer of research we should be able to estimate the power of a study from the given results.

Retrospective sample size calculation are not covered in this article. Several calculators for retrospective sample size are available on the internet (UCLA power calculators (http://calculators.stat.ucla.edu/powercalc/), Interactive statistical pages (http://www.statistics.com/content/javastat.html).

WHAT TYPE OF STUDY SHOULD HAVE A POWER CALCULATION PERFORMED?

Nearly all quantitative studies can be subjected to a sample size calculation. However, they may be of little value in early exploratory studies where scarce data are available on which to base the calculations (though this may be addressed by performing a pilot study first and using the data from that).

Clearly sample size calculations are a key component of clinical trials as the emphasis in most of these studies is in finding the magnitude of difference between therapies. All clinical trials should have an assessment of sample size.

In other study types sample size estimation should be performed to improve the precision of our final results. For example, the principal outcome measures for many diagnostic studies will be the sensitivity and specificity for a particular test, typically reported with confidence intervals for these values. As with comparative studies, the greater number of patients studied the more likely the sample finding is to reflect the true population value. By performing a sample size calculation for a diagnostic study we can specify the precision with which we would like to report the confidence intervals for the sensitivity and specificity.

As clinical trials and diagnostic studies are likely to form the core of research work in emergency medicine we have concentrated on these in this article.

POWER IN COMPARATIVE TRIALS

Studies reporting continuous normally distributed data

Suppose that Egbert Everard had become involved in a clinical trial involving hypertensive patients. A new antihypertensive drug, Jabba Juice, was being compared with bendrofluazide as a new first line treatment for hypertension (table 2).

As you can see the figures for $p\alpha$ and $p\beta$ are somewhat typical. These are usually set by convention, rather than changing between one study and another, although as we see below they can change.

Table 2 Egbert writes down some things that he thinks are important for the calculation

What is the null hypothesis?	That Jabba Juice will be no more effective than bendrofluazide in treating new presentations of hypertension.
What level do we want to avoid a type I error at? (pα)	We set this to 0.05
What level do we want to avoid a type II error at? (pβ)	We set this to 0.8
What is the "clinically important difference" we want to detect?	For this study we want to detect a minimum 10 mm Hg difference between treatments.
What type of data and analysis are likely?	Continuous normally distributed data. To be analysed using a *t* test
What is the standard deviation of blood pressure in this group of patients?	From other studies we know that the standard deviation is 20 mm Hg.

A key requirement is the "clinically important difference" we want to detect between the treatment groups. As discussed above this needs to be a difference that is clinically important as, if it is very small, it may not be worth knowing about.

Another figure that we require to know is the standard deviation of the variable within the study population. Blood pressure measurements are a form of normally distributed continuous data and as such will have standard deviation, which Egbert has found from other studies looking at similar groups of people.

Once we know these last two figures we can work out the standardised difference and then use a table to give us an idea of the number of patients required.

$$\text{Standardised difference} = \frac{\text{difference between the means}}{\text{population standard deviation}}$$

The difference between the means is the clinically important difference—that is, it represents the difference between the mean blood pressure of the bendrofluazide group and the mean blood pressure of the new treatment group.

From Egbert's scribblings:

$$\text{Standardised difference} = \frac{10\text{ mm Hg}}{20\text{ mm Hg}} = 0.5$$

Using table 3 we can see that with a standardised difference of 0.5 and a power level (pβ) of 0.8 the number of patients required is 64. This table is for a one tailed hypothesis, (?) the null hypothesis requires the study to be powerful enough to detect either treatment being better or worse than the other, so we will need a minimum of 64×2=128 patients. This is so that we make sure we get patients that fall both sides of the mean difference we have set.

Table 3 How power changes with standardised difference

	Power level (pβ)			
Sdiff	0.99	0.95	0.90	0.80
0.10	3676	2600	2103	1571
0.20	920	651	527	394
0.30	410	290	235	176
0.40	231	164	133	100
0.50	148	105	86	64
0.60	104	74	60	45
0.70	76	54	44	33
0.80	59	42	34	26
0.90	47	34	27	21
1.00	38	27	22	17
1.10	32	23	19	14
1.20	27	20	16	12
1.30	23	17	14	11
1.40	20	15	12	9
1.50	18	13	11	8

Sdiff, standardised difference.

Another method of setting the sample size is to use the nomogram developed by Gore and Altman[2] as shown in figure 3.

From this we can use a straight edge to join the standardised difference to the power required for the study. Where the edge crosses the middle variable gives an indication as to the number, N, required.

The nomogram can also be used to calculate power for a two tailed hypothesis comparison of a continuous measurement with the same number of patients in each group.

If the data are not normally distributed the nomogram is unreliable and formal statistical help should be sought.

Studies reporting categorical data

Suppose that Egbert Everard, in his constant quest to improve care for his patients suffering from myocardial infarction, had been persuaded by a pharmaceutical representative to help conduct a study into the new post-thrombolysis drug, Jedi Flow. He knew from previous studies that large numbers would be needed so performed a sample size calculation to determine just how daunting the task would be (table 4).

Once again the figures for pα and pβ are standard, and we have set the level for a clinically important difference.

Unlike continuous data, the sample size calculation for categorical data is based on proportions. However, similar to continuous data we still need to calculate a standardised difference. This enables us to use the nomogram to work out how many patients are needed.

p_1=proportional mortality in thrombolysis group =12% or 0.12

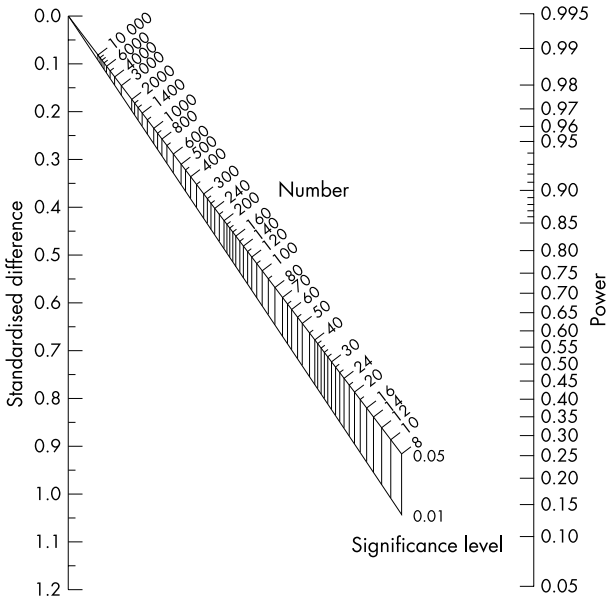


Figure 3 Nomogram for the calculation of sample size.

Table 4 Sample size calculation

What is the null hypothesis?	That adding Jedi Flow will be no more effective than thrombolysis alone in improving the mortality rate in acute MI.
What level do we want to avoid a type I error at? (pα)	We set this to 0.05
What level do we want to avoid a type II error at? (pβ)	We set this to 0.8
What is the "clinically important difference" we want to detect?	3%
What is the mortality rate using thrombolysis alone?	12%

$$\text{Standardised difference} = \frac{p_1 - p_2}{\sqrt{[P(1-P)]}}$$

p_2=proportional mortality in Jedi Flow group =9% or 0.09 (This is the 3% clinically important difference in mortality we want to show).

P=(p_{1+} p_2)/2=

$$P = \frac{p_1 + p_2}{2} \qquad P = \frac{0.12 + 0.09}{2} = 0.105$$

$$\text{Standardised difference} = \frac{0.12 - 0.09}{\sqrt{(0.105(1-0.105))}} = \frac{0.03}{0.30} = \quad 0.1$$

The standardised difference is 0.1. If we use the nomogram, and draw a line from 0.1 to the power axis at 0.8, we can see from the intersect with the central axis, at 0.05 pα level, we need 3000 patients in the study. This means we need 1500 patients in the Jedi Flow group and 1500 in the thrombolysis group.

POWER IN DIAGNOSTIC TESTS

Power calculations are rarely reported in diagnostic studies and in our experience few people are aware of them. They are of particular relevance to emergency medicine practice because of the nature of our work. The methods described here are taken from the work by Buderer.[3]

Dr Egbert Everard decides that the diagnosis of ankle fractures may be improved by the use of a new hand held ultrasound device in the emergency department at Death Star General. The DefRay device is used to examine the ankle and gives a read out of whether the ankle is fractured or not. Dr Everard thinks this new device may reduce the need for patients having to wait hours in the radiology department thereby avoiding all the ear ache from patients when they come back. He thinks that the DefRay may be used as a screening tool, only those patients with a positive DefRay test would be sent to the radiology department to demonstrate the exact nature of the injury.

He designs a diagnostic study where all patients with suspected ankle fracture are examined in the emergency department using the DefRay. This result is recorded and then the patients are sent around for a radiograph regardless of the result of the DefRay test. Dr Everard and a colleague will then compare the results of the DefRay against the standard radiograph.

Missed ankle fractures cost Dr Everard's department a lot of money in the past year and so it is very important that the DefRay performs well if it be accepted as a screening test. Egbert wonders how many patients he will need. He writes down some notes (table 5).

For a diagnostic study we calculate the power required to achieve either an adequate sensitivity or an adequate specificity. The calculations work around the standard two by two way of reporting diagnostic data as shown in table 6.

To calculate the need for adequate sensitivity

$$TP + FN = z^2 \times \frac{(SN\,(1-SN))}{W^2} = 1.96^2 \times \frac{(0.95(1-0.95))}{0.05^2} = 3.842 \times \frac{0.0475}{0.0025}$$

$$N(sN) = \frac{TP + FN}{P} \qquad = \frac{72.998}{0.3}$$

To calculate the need for adequate specificity

If Egbert were equally interested in having a test with a specificity and sensitivity we would take the greater of the two, but he is not. He is most interested in making sure the test has a high sensitivity to rule out ankle fractures. He therefore takes the figure for sensitivity, 243 patients.

Table 6 Two by two reporting table for diagnostic tests

		Gold standard	
		Positive	Negative
Clinical trial result	Test positive	a True positive	b False positive
	Test negative	c False negative	d True negative

Sensitivity =a/a+b. Specificity =d/b+d

Table 5 Everard's calculations

What is the null hypothesis?	That the DefRay will not be more than 90% sensitive and 70% specific for detecting ankle fractures
What is the lowest sensitivity that is acceptable?	95% (call it SN)
What is the lowest specificity that is acceptable?	80% (call it SP)
What do you want the confidence intervals to be?	5% for sensitivity (Call it W)
How many patients in the study will have the target disorder? (In this case ankle fractures in Egbert's population of patients)	30% (Call it P)

For purposes of calculation W, SN, SP, and P are expressed as numbers between 0 and 1, rather than as percentages.

$$FP + TN = z^2 \times \frac{(SP\,(1 - SP))}{W^2}$$
$$= 1.96^2 \times \frac{(0.80\,(1 - 0.80))}{0.05^2}$$

$$= 3.842 \times \frac{0.16}{0.0025}$$

$$N\,(sp) = \frac{FP + TN}{(1 - P)} \quad = \frac{245.888}{(1 - 0.3)}$$

CONCLUSION

Sample size estimation is key in performing effective comparative studies. An understanding of the concepts of power, sample size, and type I and II errors will help the researcher and the critical reader of the medical literature.

QUIZ

(1) What factors affect a power calculation for a trial of therapy?

(2) Dr Egbert Everard wants to test a new blood test (Siththastic) for the diagnosis of the dark side gene. He wants the test to have a sensitivity of at least 70% and a specificity of 90% with 5% confidence levels. Disease prevalence in this population is 10%.

- (i) How many patients does Egbert need to be 95% sure his test is more than 70% sensitive?
- (ii) How many patients does Egbert need to be 95% sure that his test is more than 90% specific?

(3) If Dr Everard was to trial a new treatment for light sabre burns that was hoped would reduce mortality from 55% to 45%. He sets the pα to 0.05 and pβ to 0.99 but finds that he needs lots of patients, so to make his life easier he changes the power to 0.80.

(i) How many patients in each group did he need with the pα to 0.05 and pβ to 0.80?

(ii) How many patients did he need with the higher (original) power?

Quiz answers

(1) See box.

(2) (i) 2881 patients; (ii) 81 patients

(3) (i) about 400 patients in each group; (ii) about 900 patients in each group

ACKNOWLEDGEMENTS

We would like to thank Fiona Lecky, honorary senior lecturer in emergency medicine, Hope Hospital, Salford for her help in the preparation of this paper.

Authors' affiliations

S R Jones, North Manchester Hospital, Manchester, UK
S Carley, Royal Bolton Hospital, Bolton, UK
M Harrison, North Staffordshire Hospital, UK

REFERENCES

1 **Driscoll P**, Wardrope J. An introduction to statistics. *J Accid Emerg Med* 2000;**17**:205.
2 **Gore SM**, Altman DG. How large a sample. In: *Statistics in practice*. London: BMJ Publishing, 2001:6–8.
2 **Buderer NM**. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Acad Emerg Med* 1996;**3**:895–900.

PostScript

LETTERS

If you have a burning desire to respond to a paper published in *EMJ*, why not make use of our ''rapid response'' option?
Log on to our web site (www.emjonline.com), find the paper that interests you, and send your response via email by clicking on the ''eLetters'' option in the box at the top right hand corner.
Providing it isn't libellous or obscene, it will be posted within seven days. You can retrieve it by clicking on ''read eletters'' on our homepage.
The editors will decide as before whether to also publish it in a future paper issue.

Emergency management of diabetic ketoacidosis

I read with great interest the paper by Hardern and Quinn.[1] However, it seems that there is some misinterpretation of the quoted papers.

In the introduction, the authors write ''DKA is a potentially fatal metabolic disorder''. It would be useful to point out that it is only life threatening in its most severe stage: coma.[2] They summarise the pathogenesis of coma in DKA on page 212: ''Severe acidosis has adverse effects on many organs, espcially the brain ...''. Indeed, according to Alberti,[3] very low blood pH is the immediate cause of coma: the glycolytic enzyme phosphofructokinase is inactivated by decreasing pH and, thus, the glucose utilisation in brain cells is impaired.

On page 212, the authors write ''It may, therefore, seem appealing to give bicarbonate as treatment for the metabolic acidosis that occurs in DKA. There is no evidence to support this'' with quotations of Lutterman, Lever, Morris and Viallon. Morris and Viallon did not observe comatose patients. Lutterman has given sodium bicarbonate to four comatose patients, all recovered to full alertness. The same result has been observed by Lever in 27 patients with ''deep coma'' and he did not observe adverse reactions to this treatment. Where is there a published report on a similar number of comatose patients with DKA, with zero lethality, without sodium bicarbonate, and without increase of the low blood pH?

On the same page, they quote the paper of Hale: ''no metabolic benefits from bicarbonate administration''. However, in the group with bicarbonate administration, the blood pH increased within 120 minutes to 7.23, whereas in the group without bicarbonate it only increased to 7.12. In a comatose patient, such a difference can be life saving.

V Rosival
Department of Clinical Biochemistry, Dérer's Hospital, Limbová 5, SK-833 05 Bratislava, Slovakia; rosivalv@hotmail.com

References

1 **Hardern RD**, Quinn ND. Emergency management of diabetic ketoacidosis in adults. *Emerg Med J* 2003;**20**:210–13.
2 **American Diabetes Association**. Hyperglycaemic crises in patients with diabetes mellitus. *Diabetes Care* 2003;**26**(suppl 1):S109–17.
3 **Alberti KGMM**, Zimmet P, DeFronzo RA, *et al*. *International textbook of diabetes mellitus*. 2nd edn. Chichester: Wiley, 1997:1218.

NICE head injury guidelines: cost implication for a district general hospital (''six scans to six figures'')

This observational study aimed to determine the cost implication for a typical size district general hospital (DGH) adopting the recently published NICE head injury (HI) management guidelines.[1] The author reviewed the records of all patients (4688) attending York District Hospital emergency department (60 000 new patients/year) in April 2003 to identify those patients who had suffered a head injury (393). April was chosen because it represented an average month both for attendances and skull radiograph and CT head scan requests. Data were collected from a combination of the triage nurse assessment and the doctor's notes, followed by retrospective application of the NICE HI guidelines.

Adherence to the guidelines would have resulted in requests for 61 adult and 18 child CT head scans compared with the six actually conducted for trauma during this period. This is probably a conservative estimate. That patients were only placed in the ''need CT head'' group if a specific indication for a scan was recorded may account for why 20.1% HI patients would have been scanned compared with the NICE estimate of 29.3%.

NICE estimates that about 58% of HI patients in the UK have skull radiographs taken. The figure was 22.6% in this study. Local guidelines[2] were not followed in every case but it seems that projected savings from a reduction in skull radiographs may be less than NICE predicts. Similarly, only 2.3% of HI patients were admitted for observation compared with the NICE estimate of 14%. Persisting symptoms, intoxication, or lack of supervision would have prohibited discharge of any of the admitted cases after a ''normal'' CT scan. Of the patients meeting NICE CT head criteria, 75% attended out of hours, implying that on-call radiologists would be interpreting scans after midnight on most nights.

The Canadian CT head rule that the guidelines are adopted from excluded children.[3] However, NICE recommends application of guidelines to both adults and children the same, including scanning all patients who vomit more than once. It surprised the author that only 7 of 149 (3.5%) children vomited more than once after HI, most fulfilling CT head criteria on a dangerous mechanism of injury and witnessed loss of consciousness (table 1).

The commonest mode of injury was a fall (47.8%), followed by assault (23.7%), ''accident'' (20.4%), and road traffic accident (5.6%). This is generally representative of previous UK HI studies,[4] except that there were marginally fewer road traffic accidents in this cohort. This subgroup accounts for a greater proportion of moderate to severe HIs and may partly explain the comparatively low number of CT scans requested. It is the aetiological differences between the UK and Canadian populations[2] (11% assault, 43% road traffic accident) that have raised questions about the validity of applying the Canadian CT head rule guidelines to the UK population.

Based on NICE pricing[1] (skull radiograph £26, CT head £160), this study suggests that a typical DGH adopting NICE HI guidelines would have net increased imaging costs alone

Table 1 Criteria on which patients would have had a CT head scan performed if NICE head injury guidelines had been applied

	Adults	Children	Total
GCS<9 at any time	1	0	1
GCS<15 at two hours after injury	6	0	6
Open skull fracture	0	0	0
Depressed skull fracture	0	0	0
Vomiting more than once	15	7	22
Focal neurological deficit	3	0	3
Any sign of base of skull fracture	2	0	2
Seizure	4	0	4
Retrograde amnesia >30 min	3	1	4
Symptomatic HI re-attendees	4	0	4
Coagulopathy (+witnessed LOC or anterograde amnesia)	1	0	1
Age >64 years (+witnessed LOC or anterograde amnesia)	7	0	7
Dangerous mechanism of injury(+witnessed LOC or anterograde amnesia)	24	12	36
Total	70*	20*	90*

*Some patients met more than one criterion, therefore the total number of criteria exceeds the total number of CT scan requests.

of over £9000/month and £110 000/year. Admission costs may not change significantly. Allied to radiology downtime, this would probably prove prohibitive to full implementation of the guidelines in many areas.

J B Lee
Leeds General Infirmary, Great George Street, Leeds LS1 3EX, UK; docjasonlee@hotmail.com

References

1 **National Institute for Clinical Excellence**. NICE head injury guidelines. http://www.nice.org.uk/.
2 **The Royal College of Radiologists**. Making the best use of a department of clinical radiology. *Guidelines for doctors*, 4th edn. London: RCR, 1998.
3 **Steill IG**, Wells GA, Vandemheen K, *et al*. The Canadian CT head rule for patients with minor head injury. *Lancet* 2001;**357**:1391–6.
4 **Kay A**, Teasdale G. Head injury in the United Kingdom. *World J Surg* 2001;**25**:1210–20.

Cost effective approach for emergency department investigation of deep vein thrombosis

I read with interest the paper by Kilroy *et al*.[1] Wells' criteria of pretest probability (PTP) has recently been validated in a large randomised control trial[2] for excluding pulmonary embolism. Kilroy *et al*[1] admit that using PTP the risk stratification was successful, it was not quite as discriminatory as Wells' original data.

Wells' study[2] has important implications for practice because it shows that combination of a low PTP, and a negative D-diner test safely excluded pulmonary embolism in outpatients, obviating further investigations in 40% of patients.[2] However, the occurrence of deep vein thrombosis in up to 20% of patients with a "high" PTP score and negative D-dimer test emphasises the point that the D-dimer test cannot be used in isolation.[2]

SimpliRED D-dimer assay in the study by Kilroy *et al*[1] had a low sensitivity. All D-dimer assays differ and clinicians should know the diagnostic performance of the test used in their own institutions.

Compression ultrasonography is by no means a cheap investigation as the authors perceive.[1] Plethysmography can be used as an alternative investigation for the diagnosis of deep vein thrombosis. Digital photoplethysmography can be used as an useful cheap tool to exclude deep vein thrombosis safely in an emergency department, thus reducing pressure on the radiology department.[3] Table 1 shows the sensitivity and specificity of using computed strain gauge plethysmography.

Based on our experience we would like to conclude that we can fulfil our assumption of good practice by achieving a negative predictive value of 100% by combining PTP, a modern D-dimer test, and either digital photoplethysmography or computed strain gauge plethysmography to exclude deep vein thrombosis in an emergency department. However, we believe further randomised control trials are necessary to test this hypothesis.

R Sinharay, F Pressley
Department of General Internal Medicine, Royal Gwent Hospital, Cardiff Road, Newport NP20 2UB, UK; ranjitsinharay@hotmail.com

References

1 **Kilroy DA**, Ireland S, Reid P, *et al*. Emergency department investigation of deep vein thrombosis. *Emerg Med J* 2003;**20**:29–32.
2 **Wells PS**, Anderson DR, Rodger M, *et al*. Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and D-dimer. *Ann Intern Med* 2001;**135**:98–107.
3 **Tan YK**, da Silva AF. Digital photoplethysmography in the diagnosis of suspected lower limb DVT; Is it useful? *Eur J Endovasc Surg* 1999;**18**:71–9.

Table 1 Results of ultrasonography (US) and computed strain gauge plethysmography (CSGP) on all 23 D-dimer positive patients with suspected deep vein thrombosis

Results	Number (%)	US	CSGP
True negative	16(69)	−ve	−ve
True positive	1(4)	+ve	+ve
False positive	4(17)	−ve	+ve
False negative	2(8)	−ve	+ve

BOOK REVIEW

Major incident management system

T J Hodgetts, C Porter. (Pp 94; £30.00). BMJ Books, London, 2002. ISBN 0-7279-1614-9

This book will become an essential companion to all those individuals, doctors, and paramedics, who are involved in the immediate management of a major incident at scene. To those not directly involved it provides a valuable insight into its basic theory.

Presented in a high quality loose leaf binder, the pages are laminated and all sections are colour coded, it is the authors' intention that the book is taken to and used at the scene of an incident. As clearly stated on the front cover this is an "aide memoire" and therefore there is little in the way of explanation, some prior knowledge of major incident management is therefore required. The book is written in a clear, didactic style with the salient points arranged in "bullet point" fashion. The diagrams are colourful, clear, and informative although several of the flow charts appeared a little too busy.

The sections follow in logical order from the actions of the first ambulance crew on scene, the setting up of command centres, modes of communication and terminology, triage and treatment, and finally scene evacuation. Much of the written information is augmented by the clever use of symbols placed in the adjacent margin thus making the information rapidly accessible. The final two sections of the book provide action cards for all ambulance and medical personnel designed to be distributed at scene, again this is given in a clear, succinct fashion with diagrams illustrating the hierarchy of command.

Clearly this book is aimed at the ambulance service and those doctors acting as medical incident officers and although does not have an impact directly on a hospital's major incident procedure, it is useful in understanding what happens before the patient's arrival at hospital.

C J Blakeley

CORRECTION

An authors' error occurred in this article by Dr Jones and others (2003;**20**:453–8)

We wish to clarify the following points in relation to the section on diagnostic sample size calculations.

(1) The method described is used to calculate the sample size required to estimate an expected level of sensitivity or specificity with a predefined degree of precision. If the researcher wishes to ensure that a particular test has a sensitivity or specificity higher than a predetermined level then an alternative method should be used.[1]

(2) The language used in the section on diagnostic tests may be misleading. The method described should use "expected" levels of sensitivity or specificity (SN or SP) to calculate the sample size, rather than the "lowest acceptable" level.

(3) The method described should not be used if there are fewer than five subjects in any of the cells of the 2×2 table.

We would like to sincerely thank Steve Goodacre for his assistance in recognising the problem and for his assistance in resolving it.

Reference

1 **Arkin CF**, Wachtel MS. How many patients are necessary to assess test performance? *JAMA* 1990;**3**:895–900.