2. Future studies should use additional measures of average and variation (further faults)

There were several faults in the structure and analysis of the PACE report: numerical, statistical and logical. Some are major; some merely make us wonder about the clarity of thought employed. Here is our list:

1. The use of mean and variance for skewed data.

The SF-36 Physical Function scale is a 21 point scale, where each point is multiplied by 5 to give a score from 0 to 100 inclusive. It is designed to measure difficulty in functioning, as opposed to measuring the range of physical function in healthy folk, so when applied to a random sample of adults there is a heavy bias, or clumping effect, at the maximum, healthy score of 100. In such a biased distribution, statisticians arrange all the results in order and choose the score in the middle, the median, to determine the "average" value. For the SF-36 physical function the median would be 100 (in fact approximately 60% of the population have a score of 100).

To define normal function, the authors of the PACE trial used normative data from a study by Bowling[1] which warned of a heavily clumped distribution of scores. PACE decided to set the boundary for normal function at a score of 60 or above, using the mean (84) minus one standard deviation (s.d.) of 24.

As a comparison, the distribution for incomes in the UK in 2007/08[2] has comparable measure of bias, a mean of £26800 and a standard deviation of £29500. Using mean minus s.d. gives us a "normal" income of anything above minus £2700, which is as inappropriate to income as the SF-36 target score of 60 is to normality/recovery. There can be no statistical justification for accepting that the mean minus s.d. here somehow defines normality.

2. The use of statistics to determine boundaries.

It is reasonable to look at statistics to determine, say, typical heights of adult males. But it would not make sense to use data on adult male weights to determine the boundary for obesity. The pattern of adult male weights has changed over the years, and varies from country to country, but the concept of obesity should be independent of this – it is a matter of health.

In the same way, how can it make sense to use statistics to determine the scores that would determine recovery? It is particularly strange with the PACE trial because the boundaries that determined whether a patient was ill enough to be eligible for inclusion in the trial appear to have been decided and adjusted by using medical judgement. Surely an assessment of, say, recovery is a medical assessment, not a statistical one.

3. Moving the goalposts: inappropriate comparisons.

The patients in the PACE trial had a mean age of 38 with a standard deviation of 12 years. No further details were given about the nature of the distribution, but this could reflect a fairly even spread of patients aged 18 to 59, with just a few in their 60s.

Originally the data used to set the boundaries for the SF-36 Physical Function scores were from a set of adults of working age. Once the trial had begun, the new, much lower, thresholds were re-set by using a highly inappropriate sample: the sample now included people aged up to and over 85, which had a dramatic effect on the data. As you can imagine, the scores in the range 85+ (mean 39.3, s.d. 31.5) are far removed from that of someone in the 35 to 44 range (mean 93.9, s.d. 13.4).

4. Overlapping boundaries.

The final boundary for being within normal function on the SF-36 scale was a score of 60 or above. The final boundary for being ill enough to be eligible to be included in the trial was a score of 65 or less. So being ill enough to be included is apparently healthier than the bottom end of being within normal function: dropping one point (which is a worth 5 marks on the scale of 0-100) would still mean normal function.

5. Unnecessary criteria.

In the protocol for the PACE trial, the improvement target for Fatigue was a 50% reduction in fatigue score, or a score of 3 or less. Here the bimodal scoring system was used, where the scale runs from 0 to 11, the most severely ill patients scoring 11. Admission for the trial was set at a score of 6 or more, so anyone improving to a score of 3 or less had to have at least halved their baseline score; as such, the second condition (a score of 3 or less) was irrelevant, and makes us wonder how carefully these conditions were considered.

Very ill patients would have to reduce their score from 11 to 5, a drop of 6 points, to qualify.

6. The percentages that look similar but are very different.

In the protocol, the improvement target for the SF-36 score was a 50% increase in baseline score or a score of 75 or more, which sounds very similar to the target for Fatigue. But here the scale runs from 0 to 100, where a very ill patient would score zero. In this situation a patient scoring 10 would only have to improve by one point, to reach 15, to satisfy the condition. What is even worse, as 50% of zero is zero, a patient right at the bottom and staying there would also qualify.

This shows a remarkable lack of understanding of percentages.

7. The choice of primary efficacy measures.

When the Chalder Fatigue score was applied to people with ME/CFS, there was considerable clumping of values at the "most severely ill" part of the scale (a score of 11), despite the fact that the most seriously ill were not able to take part. This makes it unsuitable for use of a measure of "illness" for people with ME. The change to the Likert method of scoring the answers did not improve this problem: although technically it provided 4 levels of answer and score for each question, in reality it changed the assessment from an 12 point scale to a 23 point one, which is far too small. The clumping at the bottom prevents any real measure of deterioration, and shrinks the gap between the severity of the illness and normal health. It would be like a study of anorexia, using scales that did not go below 100 pounds. It needs several more questions appropriate to severe fatigue.

The SF-36 scale on the other hand clumps at the healthy end of the scale (a score of 100), preventing it from properly measuring the range of physical functioning for healthy people. This makes it unsuitable for determining a normal, healthy range of scores. It would be like measuring a person's physical fitness with a series of fairly undemanding tests. It has the effect of shortening the gap between the ill and the healthy. It is only a 21 point scale: simply multiplying each score by 5 to make it out of 100 does nothing but deceive.

In most schools in the UK, mathematics is taught in ability sets. Broadly speaking, there could be 5 such levels across one year group, each one about a year ahead of the one below it. Tests can be designed either to be diagnostic or to assess ability. A good assessment test would aim at getting a range of marks from 20% to 80% from the majority of students in one set. If a test designed for the mathematically weakest set were also to be taken by the strongest set, only a few in the strongest set would achieve 100%. In fact there would probably be a small overlap between the sets of marks. This is simply because the test would not give the students in the strongest set an opportunity to show the many additional skills that they have.

With a test to determine fatigue or physical function there is easily an enormous spread of "performance" from being super-fit and healthy, and being very ill with ME. It should be possible (but difficult) to produce a test to cover the full range, but this would need to be lengthy, with much more detail than either the Chalder Fatigue scale with 12 (or possibly 23) points and the SF-36 with 21. Such a scale would also need to extend much further at either end. The vast majority of participants, whether healthy or ill should be unable to score zero or 100%: any participant should have ample room to move up or down, and there should be as much of a spread of scores for the ill as for the healthy. Only then could we measure improvement and deterioration, and define levels. But of course, we would need to check these scores against objective data for them to prove their worth.

8. An absence of a control group.

Both of the key therapies (CBT and GET) took an assertive stance in persuading patients that their symptoms were not as bad as they thought, and that it was common for healthy people to have a number of similar symptoms. Both of the primary efficary measures asked patients to rate their perception of what they felt that they could do, and how badly they were affected. It is hardly surprising that changes were recorded. What would be surprising to those who have not experienced the illness is how small those changes were.

Yet there was no control group of, say, relaxation classes that also delivered a similar message. The other treatment group that offered a controversial form of pacing actually delivered a cautionary message.

It would be perfectly reasonable to postulate that the improvements claimed for CBT and GET in the PACE trial, only found in subjective measures and not confirmed by any improvement in objective data, were simply the result of patients modifying comments such as "very much worse" to "much worse" in the light of being told that they focussed too much on their symptoms.