10. Future studies should aim for a more rigorous scientific structure (details)

There is a commonly held misconception that experiments in pure science are much more clear-cut than in medicine. The truth is that if you are delving into the unknown, in both areas there is a lot of uncertainty: in fact, if you are dealing with the quantum world, uncertainty lies at the heart of the subject, and in astronomy we cannot even get out a tape measure and check our estimated distances.

What is essential is that results are examined with a critical eye, looking for consistency and validity. No matter how cherished the theory, no matter what the anecdotal evidence suggests, no matter what spin scientists put on their own experience, we must follow the results. Even Einstein found that difficult when quantum theory became established, and probability rather than classical calculation came to the fore. Known for the comment "(God) does not play dice", the facts eventually forced him to accept that he was wrong.

Richard Feynman gave out a lot of good advice: something that is pertinent here is "The first principle is that you must not fool yourself, and you are the easiest person to fool."

ⓐ All studies should incorporate a number of objective assessments, preferably ones which have had a "dry run" to bed down before forming baseline assessments:

> The PACE trial has no excuse for this failure. The illness and the therapies have been studied for many years. Only objective measurement, such as the use of pedometers/actometers, can really convince others of true success: qualitative measures, through questionnaires etc., can add understanding and depth to these results, but cannot be used as the sole primary measures.
>
> I know this is true because I conducted a survey and asked people what they thought (yes, I'm being ironic!).

ⓑ Any attempt to measure fatigue in patients with ME/CFS must leave an appropriate habituation period before testing and treatment commences:

> Just getting to a centre, meeting new people and understanding new procedures takes energy. If we are to measure a proper baseline, there will need to be two or three sessions before a study properly starts. During that time, patients can familiarise themselves with the journey and the requirements. This also gives a good interval to become familiar with equipment such as actometers, and with the sometimes unusual wording of questionnaires. On the PACE trial, for all groups, there was a marked improvement in the first 12 weeks: it is impossible to assess how much of this would vanish with an introductory period, which leads to the question of how much of the small average overall improvements were due to becoming familiar with the situation.

ⓒ Any statistical analysis should go to great lengths to give a fair and easily-understood picture of the results:

> Statistics has two main purposes. One is to take a large and complex mass of data and present a fair and balanced summary of the situation. The people doing that must fully understand the nature of the processes and measurements that they are using as well as the medical situation. It is no different from making a summary of a book: the person doing that must have a good command of English and, equally, a thorough understanding of the book's contents. It is not good enough to simply take these things on trust.
>
> The second purpose is to carry out checks and assess whether results are appropriate: this is the job

for a statistician, but it has to be one who can clearly communicate the assumptions and implications of any analysis with the authors of a trial.

ⓓ Any changes to previously agreed protocols should only take place in special circumstances, and should be checked thoroughly for anomalies and dilution of standards:

We have already covered the fact that in the Chalder Fatigue Scale it was possible for a patient to be accepted as being ill enough for the trial, then, under the changed criteria, be declared within normal functioning despite there being no change, or even a small deterioration in that person's score. This effect was hidden by the decision to change the scoring systems from Bimodal to Likert.

What we cannot understand is how similar changes to the rules governing the Physical Function scores on the sf-36 scale slipped past both the large team of authors and the independent scrutineers at the Lancet. On this scale, low scores means poor physical functioning: the scores are out of 100 and go in steps of 5, which is effectively a 20 point scale. At the start of the trial, patients were required to have scored 60 or less to be considered sufficiently ill for inclusion. Later the entry requirement was lifted to 65 or below to increase uptake, so there must have been a significant number of patients in the trial scoring 65 to make up the numbers. It was agreed in the protocol that the target for a positive outcome should be raised as a result of this change from 70 to 75. Bafflingly, they subsequently changed the target measure, and defined normal range now to be a score of 60 or above. So not only did the original patients scoring 60 now qualify as being within the normal range, but the significant number joining the trial to make up numbers were above the target before the therapies even began.

We are told that 30% of the patients in the CBT group and 28% of those in the GET group were within *normal range* at the end of the trial, but not how many were within that range even before starting of the trial.

It is no wonder that actions like this either have promoted suspicion that weak results were being massaged, or have created a loss in confidence in the quality of the study and the scrutineering process. Here is a quote from Fiona Godlee, Editorial Director of Medicine, BioMed Central: "*Protocol publication allows easier comparison between what was originally intended and what was actually done. It reduces the potential for "data dredging" - where associations are sought or stumbled upon during data analysis rather than hypothesised a priori. It also reduces the potential for unacknowledged or post-hoc revision of the study aims, design, or planned analyses. Such practices are not only detrimental to the advancement of medical research, they are ethically unsound since they may result in patients receiving inappropriate care*."

Of course, in PACE only some of the changes to the protocol occurred after the data started to become available, but the fact that the FINE trial were already discovering very weak results, and the fact that here were people, regarded by many as experts on ME/CFS, suddenly making radical changes to weaken their original agreed standards, leave many of us feeling very uncomfortable about the integrity of the whole process. What is it that suddenly made them lower their expectations for therapies that they had been promoting for so very many years?

ⓔ Reports must be carefully scrutinised for errors in logic due to initial bias:

In a report published in the British Journal of Psychiatry 2002, the authors found that patients who were members of a support group were more ill than the average of those in their study, and that they showed less improvement to the treatment given. In their introduction they stated that that they "*tested the hypotheses that variables indicative of rigid illness beliefs ... membership of a support group advocating exercise avoidance .... would predict poor response to this kind of treatment*." The PACE trial perpetuated this attitude in their protocol document: "*Predictors of of a negative response to treatment found in previous studies include ... membership of a self-help group.*"

Of course it would be just as valid to have the hypothesis that membership of a support group is likely to have started a patient with ME/CFS along the road to improvement, and the results would

have equally supported that hypothesis. The logical error of course lies in the untested assumptions about the behaviour of support groups, and these behaviours of course will be very varied.

If the latter hypothesis was indeed the case, the report would have created an unfortunate bias in ME/CFS treatment centres against support groups. Without evidence, it was wrong to do so.

The report by Risdale et al. in the BJ of General Practice, January 2001, found that CBT and counselling were equally effective at improving fatigue in patients with ME/CFS with broadly similar results to those of the PACE trial on a less severely fatigued group. This would suggest that support groups may well be able to provide an initial step along that road, reducing the contribution that further input would produce.

ⓕ **Unusual anomalies or outliers should not be included in the averaging process to give misleading results:**

As explained in the section on averages, the mean and standard deviation are strongly affected by extreme or unbalanced values, and would cease to give meaningful results. This then has a knock-on effect for many of the usual statistical techniques employed to assess confidence levels etc.

Equally it is possible for assessments to have wells or hidden boundaries where measurements "get stuck", as appears to be happening at the bottom end of the Chalder Fatigue Scale.

Both of these effects indicate that a reliance on standard statistical techniques is inappropriate. Just because something can be calculated does not mean that it is sensible to do so.

ⓖ **There are clear differences between *statistical significance* and the general use of the word *significant*:**

If a sample is large enough, quite small percentage differences can be statistically significant. That does not mean that the change itself is meaningful. If you were tossing a coin, the usual expectation is to get 50% heads. Getting 54 heads out of 100 would not be unusual (i.e. statistically significant), whereas getting 540 out of 1000 would be extremely unusual (statistically significant), but it would only mean that the coin was very slightly weighted. A similar thing happens with measurements of, say, fatigue levels. A change from 50% energy to 54% energy would be undetectable in everyday living, but in a large enough sample such a change could become statistically significant.

Doctors need to have the confidence to determine for themselves along with their patients as to what amount of improvement is really meaningful.

ⓗ There are dangers in mistaking statistical correlation with cause and effect:

There is a moderate positive correlation between the numbers of consonants in the name of the month and the amount of rainfall in Crowborough, East Sussex (Pearsons 0.51, Spearmans rank 0.43): months that have less consonants, like May, have less rain, but months like December, with more consonants, are wetter. Simply reducing the number of consonants in December is unlikely to reduce the rainfall.

Yet there are instances in studies where correlation has been taken to "prove" cause and effect, and a clear illustration of that is where correlation is found between mental illness and ME/CFS. A study by Harvey et al. looked at 3035 fifty-three year olds, where 37 reported that they had ME/CFS. Three of these were subsequently excluded as they had other serious conditions - meningitis, chronic active hepatitis and schizophrenia. Out of the remaining 34, the usual proportion had suffered from mild or moderate psychiatric disorders between the ages of 15 and 32, and the usual proportion had not suffered any. But, instead of there only being 2 in the group who had suffered from severe psychiatric disorder, there were 6. This is statistically speaking a strong enough result to link psychiatric disorders with ME/CFS, but, of course, it rests on a supposition that the data is valid. Despite excluding three from the sample because they had other disorders, the authors accepted self-

diagnosis of ME/CFS on the basis that "*clinical experience suggests that it is uncommon for a patient to complain of CFS or ME and not to have sufficiently severe symptoms to warrant the diagnosis"*.

So this study perpetuated the idea of there being a link between psychiatric disorders and ME/CFS mainly on the self-diagnosis of four people who had previously suffered from major psychiatric disorders, but it used sophisticated statistical techniques to mask the unsound nature of the original data.

Weak correlations like this are usually indicative of some other hidden factor that has not been spotted. It may be that people with major psychiatric disorders may have less support from family and friends, reducing their opportunity to recover, and so increasing their proportion in the general ME population: or it may be that their treatment and lifestyle had adversely affected their immune system: or perhaps the drugs used may have a side effect increasing their susceptibility to ME/CFS. Perhaps you can come up with more possibilities.

ⓘ **Calculations are not always appropriate:**

We are aware of how impressive a calculated assessment can be, but there are times when it is inappropriate to use statistics. One application that appears to be becoming more common was used in the PACE trial to determine a measure of clinical significance. It was decided that half the standard deviation of the baseline scores of the patients in the Chalder Fatigue Scale and in the SF-36 Physical Functioning scale would act as such measures (a change in score of 2 and 8 respectively). We have already covered the fact that the standard deviation is an unreliable measure when data is heavily skewed, but this process is flawed in principle. It is easy to show this with two analogies. If you were overweight and joined a slimming club, would you expect your target weight loss over a certain period to be determined by the spread of weights of the other members in the club? If so, and if the other members are all around the same weight, there will not be much variation between you, so that will mean that you are in for an easy ride.

How would you feel if your grade in a key exam was not determined by your performance, but by the variation in performances of the other candidates? Schools could easily manipulate grades by entering candidates of various abilities: a lot of weak candidates would make it easier for the others to pass.

Now the authors of the PACE trial decided to use the Chalder and SF-36 scales, and the authors also controlled the range of scores that determined whether patients were eligible for the trial. Those decisions restricted the variation of entry scores, which was subseqently used to determine the size of a clinically significant change. That doesn't make sense. People running slimming clubs use experience and dietary understanding to determine targets. Examiners use years of experience and written criteria to determine pass marks. Surely doctors are able to determine what constitutes a clinically significant change? And looking at the natural variations in the Chalder scores from our survey suggests very clearly that a change of 2 points does not tell you anything significant.

ⓙ **A study which compares a group receiving treatment A with one receiving A+B is inherently unsound:**

Ernst and Lee have published a full account of the ways in which this sort of design could lead to false positive results, particularly if the improvements are small and measured through subjective assessment. Just to select one factor in the PACE trial: those in the CBT+SMC group received at least a dozen extra sessions than those in the SMC-only group.

We do not understand how this approach was approved.

The general standard of many of the studies associated with ME/CFS in the past has been poor, and faulty conclusions have been endlessly recycled and repeated, acquiring an utterly false respectability. It is time for all such studies to raise their standards and for their reports to go through a much more rigorous scrutiny.