J. Inherit. Metab. Dis. 24 (2001) 291–298

Rare diseases and the assessment of intervention: What sorts of clinical trials can we use?

B. WILCKEN
The Children's Hospital at Westmead, Locked Bag 4001, Westmead, New South Wales 2145, (Sydney), Australia. E-mail: bridgetw@chw.edu.au

Summary: There is increasing emphasis on the importance of practising evidence-based medicine. Randomized controlled trials are the standard way to assess the benefits of an intervention, and observational studies are not usually accorded much weight; the results are likely to be considered misleading. For rare diseases, there are great difficulties in obtaining adequate evidence for interventions or for the benefits of early diagnosis. This is because the disorders are not only very rare but also have variable expression, may have very long courses, and have incompletely known late effects; and surrogate end-points often have to be used. Randomized controlled trials are usually impossible because of inadequate power, and because there are preconceived notions of the effects of treatments already in use. The adoption of the best possible design for observational trials, formation of a central registry of such trials, and a greater general appreciation of the problems that rare diseases pose will help in obtaining the best possible evidence for the effects of interventions.

Evidence-based medicine was one of the catch-cries of the 1990s, and still is. From just 18 published articles found in Medline to address this subject in the early 1990s, the numbers rose inexorably year by year: 77 in 1995, then 239, 661, 1079, and most recently a massive 1557 in 1999. A randomized controlled trial, and better still a systematic review of randomized controlled trials, is the accepted standard for assessing clinical effectiveness, and anything that falls too far below this standard is in danger of being discounted. This leaves those of us who diagnose and care for patients with rare diseases in a double difficulty. Not only is it difficult to get evidence of diagnostic or therapeutic efficacy, but such evidence as can be obtained is often regarded as too unreliable to be taken into account at all.

One of the first trials of treatment in a rare inborn error of metabolism, phenylketonuria, (McKusick 261600) was undertaken by Professor Horst Bickel and his colleagues, and reported in a preliminary communication in the *Lancet*

(Bickel et al 1953). Their patient was 2 years old, 'an idiot, unable to stand, walk, or talk'. She was placed on a specially prepared low-protein diet and over a few months improved markedly. Then (without the mother's knowledge) 5 g per day of phenylalanine was added back into the diet. Within 6 h she started to bang her head as formerly, and within days she had lost all the ground previously gained. To test this further, she was admitted to hospital, where the experiment was repeated (with her mother's permission), with similar results. Professor Bickel had performed a study with single-blind and open-label phases. The conclusion was that 'In this child at least, the benefits of a low-phenylalanine intake seem unequivocal'. There have been no randomized trials of treatment (versus no treatment) of phenylketonuria.

A recent editorial in the *New England Journal of Medicine* was entitled 'Randomized trials or observational tribulations' (Pocock and Elbourne, 2000) and commented on two articles showing no difference in estimated treatment effects between randomized and nonrandomized trials, a finding disputed in the editorial on several grounds. This was by no means the first time that such a question has been investigated (e.g. Sacks et al 1982). Because of the many biases that can arise in observational studies, most people would agree that, where possible, a randomized trial is the preferred model for clinical trials. Although different questions require different trial methodologies, the hierarchy of evidence is generally agreed to be

- Randomized controlled trials, and their derivatives (systematic reviews of RCTs)
- Controlled observational studies
- Uncontrolled studies
- Expert opinion.

It is unfortunate that scientists and clinicians dealing with the very rare diseases often seem to be locked into the bottom rung of this hierarchy.

The problems with assessing intervention in inborn errors of metabolism are several. The disorders are usually very rare and, despite being largely monogenic, are in reality complex diseases with very variable expression, which complicates the use of historical controls. There is often a very long course of the illness, with long-term rather than short-term complications. On top of all this, monitoring must often be by surrogate measures, and use surrogate end-points. It is important not only to be able to assess current treatment options but also to be able to assess the effectiveness of early or presymptomatic treatment. How do we achieve the levels of evidence we need for these endeavours?

Very often there has been treatment that is believed to work, albeit not as obviously as is the case with phenylketonuria. If we believe in the efficacy of the treatment, then collaborating in a randomized controlled trial becomes ethically difficult or impossible. Then, too, there is the problem of desperation on the part of parents. There may be no known treatment for a progressive disease. When something of promise comes along, it is very hard to persuade parents to agree to randomization, especially when the time course is likely to be long. This arose with X-linked adrenoleukodystrophy (McKusick 300100) and the use of the 'Lorenzo's oil' (glyceryl trioleate and glyceryl trierucate), which was thought to pre-

vent fatty acid chain lengthening and thus reduce the accumulation of very long-chain fatty acids (Rizzo et al 1989). No randomized trial could be undertaken. The regimen with a low-fat diet and Lorenzo's oil did lower the circulating levels of very long-chain fatty acids, but data available now from observational studies suggest that the therapy is probably ineffective in preventing progression. However, that still remains uncertain (Alger et al 2000). The same scenario may occur with the newer treatments for lysosomal storage disorders.

Statistical power is a major difficulty. Underpowered trials are poorly regarded. One publication suggests that they usually have, 'poor design, poor randomization, ill-defined end-points, poor supervision, inexperienced researchers, ...' (Griffiths 1997). Yet here is a major problem for most of us. This can be illustrated by considering one newborn screening topic: is early diagnosis of congenital adrenal hyperplasia (McKusick 201910) beneficial? One aim of such screening is to prevent death in male babies with a severe salt-losing phenotype during an adrenal crisis. Detection of a 50% increase in deaths in the unscreened, when compared with screened babies (and surely the percentage would be less), would require 2 500 000 in each arm of the trial—a trial unlikely to be funded (Figure 1). There are of course other proposed benefits of such screening, but a firm proof of efficacy is bedevilled by the comparative rarity (Edwards et al 1997). But she also notes that, theoretically, publication of the results of a small underpowered trial with no likelihood of a statistically significant result may disturb the previous equipoise—that is, the

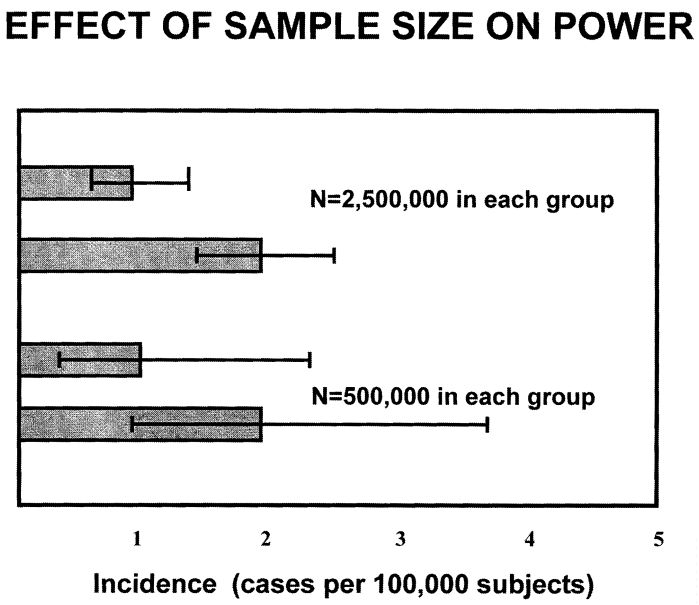

Figure 1 In a disorder with a frequency of 1 : 50 000, demonstration of a 50% reduction in the occurrence of an end-point with 95% confidence would require 2 500 000 subjects in each arm of a study. With 'only' 500 000 in each arm, there is an overlap in the confidence intervals

uncertainty of the value of an intervention—so that further trials would be unethical. However, a small trial may be all that can be managed with available numbers, and others have argued that some information is better than none (Lilford et al 1995).

The necessity of using surrogate end-points in chronic diseases is a common problem, not specific to inborn errors of metabolism. In some cases we know a great deal. For example, we have a reasonable idea about plasma levels of phenylalanine and the outcome in treated phenylketonuria. One of the best studies (Smith et al 1990) showed that mean IQ standard deviation scores in patients were similar to those of controls if mean blood phenylalanine levels were maintained below 400 μmol/L during the first four years of life, but declined progressively with higher average phenylalanine levels. This makes it possible to use blood phenylalanine levels when assessing interventions in children. Even so, we are aware now that measures of brain phenylalanine might be more relevant (Koch et al 2000). But we have not progressed very far with other disorders. We do not know the levels of plasma leucine that ensure a good outcome in maple syrup urine disease (McKusick 248600). Nor do we know the level of homocysteine, either free or bound, that we should be aiming for to maintain the lowest possible risk of thromboembolism in cystathionine β-synthase deficiency (McKusick 236200), or indeed whether some other analyte would be a more important modifier of risk. We suspect that plasma very long-chain fatty acids, elevated in X-linked adrenoleukodystrophy, are not useful analytes to measure for monitoring the effects of treatment. Treatments that reduce the elevated levels do not appear to modify outcome, although, as alluded to earlier, uncertainty remains (Alger et al 2000).

Many inborn errors are chronic diseases with unpredictable but often long courses. The effects of therapy may only be evident after many years. Examples of inborn errors where the start of intervention may precede a clinical end-point by up to 20 years or even more include childhood familial hypercholesterolaemia (McKusick 143890), homocystinuria (cystathionine β-synthase deficiency), and any mild organic acidaemia. This sort of time course makes it hard to embrace enthusiastically a suitably designed trial, and especially a randomized controlled trial. And the course may be quite unpredictable. Once again we turn to X-linked adrenoleukodystrophy as an example. Here, about half the affected boys will have a devastating course with severe physical disability, dementia, and death during the first to second decade. The other half will have a much more prolonged and benign course, with more or less intact survival to adulthood, the final phenotype still not being fully known (Moser et al 2000). It has not so far proved possible to distinguish between these phenotypes at an early age. This is perhaps an extreme example, but many inborn errors have sufficiently unpredictable courses to make the evaluation of a trial of treatment very difficult.

A further problem is that the late effects of an inborn error are likely to remain unknown for a prolonged period. For many disorders, the long-term outlook is no doubt still unknown. Methylmalonic aciduria, (methylmalonyl-CoA mutase deficiency; McKusick 251000), perhaps the most frequent of the classical organic acidurias, was first described in the late 1960s (Morrow and Barness 1968). Thus the early and life-threatening symptoms have been well known for over 30 years;

but the first publication revealing that renal failure was likely to occur in survivors of infantile presentations came only recently (Walter et al 1989). It is still not clear whether this is universal in serve cases, and whether mild congenital methylmalonic aciduria could lead to renal failure in the long term. Dozens of other examples could be given. When the long-term outlook is still uncertain, and some complications as yet uncovered, the effects of therapy are hard to judge. Perhaps the relative newness of our specialty rivals the rarity of our disorders in making it difficult to use the best tools of evidence-based medicine.

What sorts of studies could be used with these unpredictable and rare inborn errors of metabolism? The problems with lesser-order studies (i.e. those that do not incorporate double-blinded randomization of subjects and controls) have been well-ventilated. Observational studies require some control subjects, but without randomization it is difficult to control for confounding and bias. For example, researchers are interested in the outcome, and are not blinded. Subjects are also interested in the outcome and self-selection may be an especial problem. Use of historical controls presents particular difficulties because not only will management have altered and presumably improved over time, but more importantly, we very often do not know the natural history accurately. One excellent study of natural history was that of cystathionine β-synthase deficiency (Mudd et al 1985). This study was conducted by postal questionnaire to physicians caring for patients with inborn errors of metabolism. Data were collected on 629 patients from 114 individual physicians. Most of the physician responders were caring for only one to three such patients. Awareness of the disorder was at that time not generally high, and there might well be a bias towards the severe end of the spectrum, with only the most obvious cases being diagnosed in some regions. Another very useful study of natural history also had inbuilt and unavoidable bias towards the severe. Pitt and Danks (1991) studied the outcome of 51 never-treated adults with PKU. This showed, *inter alia*, that 6% had an IQ of 68 or greater; but of course it tells us nothing of the whole spectrum at the mild end of the range, as PKU patients with a relatively normal IQ may well go unrecognized throughout life. To improve the knowledge of natural history by retrospective study there need to be especially strict diagnostic criteria, and all known cases from a centre should be reported if they fulfil the criteria, even though some data may be missing. Data from regional centres with good diagnostic facilities may be preferred over those from referral centres, or at least should be analysed separately, as the patient base may be more comprehensive. These precautions will not ensure lack of bias, but will reduce it to a minimum.

Screening can provide otherwise hidden information. Natural history may be illuminated, although newborn screening, for example, is usually only undertaken when some form of intervention is intended. But it can enable the study of mild variants, and expose ascertainment bias, such as occurs when patients being investigated because of symptoms are found to have a rare genetic disorder, which is then thought to be the cause of the symptom. This was exemplified in the case of histidinaemia (McKusick 235800), which was initially thought to result in developmental delay and speech defects (Ghadimi and Partington, 1967). Only with the availability of both newborn screening and family studies did it become clear

that this enzyme deficiency is likely to be benign (Coulombe et al 1983). A new chapter is being written now with the advent of tandem mass spectrometry. Already, many screening programmes have found atypical cases of medium-chain acyl-CoA dehydrogenase deficiency (Andresen et al 2000; Carpenter et al 2000; Lindner et al 2000). In addition, an unexpectedly large incidence of two other disorders previously though extremely rare—short-chain acyl-CoA dehydrogenase deficiency and 3-methylcrotonyl-CoA carboxylase deficiency—has been found (Roscher et al 2000; Wilcken et al 2000). The clinical significance of these sorts of cases is at present unknown.

Despite all these difficulties, good studies of intervention are possible in rare diseases. While a randomized controlled trial could probably still be performed, for instance, to investigate the continued use of diet in adult males with PKU, many centres would nowadays find that unethical. There are few other inborn errors where randomization, for all the reasons discussed above, would be feasible. Observational studies with historical controls could be performed in a number of instances. Other designs are possible, depending on the question to be answered. What is needed is that trials, either multicentre or small trials, have the very best design possible, that protocols are well reviewed at the outset, and that there is some central body with which the trial can be registered. Review of trial design by an experienced clinical trials centre would ensure that the most reliable information was obtained. Registration of a trial would ensure first that, where agreeable to the investigators, others could know of a trial planned or in progress, and second that at the conclusion all of the data that resulted could be accessed, whether or not publication was achieved. To answer treatment questions for individual patients, '*N*-of-1' randomized trials are easily designed. These are indicated when effectiveness of the treatment is in doubt, when there is quick onset and offset of treatment effects, where there is a measurable treatment target and, of course, when the patient is keen to do the trial. A good example of the helpfulness of such a trial was a report of the effectiveness of benzoate and imipramine in a patient with late-onset nonketotic hyperglycinaemia (Wiltshire et al 2000).

There are many studies waiting to be done. An obvious one already mentioned is a study of the effectiveness and efficiency of tandem mass spectrometry in newborn screening. Does early identification improve outcome, and for which disorders? It is not realistic to wait for the answers before some screening programmes embark on testing, as without large numbers of participating programmes no answers can possibly be forthcoming. But it is surprising that no coherent plan has yet emerged to study this. The establishment of some central resource would perhaps encourage cooperative action. Many questions are being asked repeatedly: for example, the place of carnitine therapy in medium-chain acyl-CoA dehydrogenase deficiency (McKusick 201450) and other fatty acid oxidation defects, the severity of galactose restriction needed in galactosaemia (McKusick 230400), and so forth. Most could be answered. Is there a possibility of a central body to encourage appropriate trials, and to register them? Who could take this on? A web-site for the registry of trials is not such a difficult concept, and could easily be linked to the web-sites of the various societies—the SSIEM, SIMD, JSIMD, ASIEM. Possibly the Cochrane Col-

laboration will move more rapidly towards dealing with the problem of rare diseases and trials. Certainly, publicizing the problems can only help to dispel the idea that the only questions that ought to be addressed are those that can be answered with randomized controlled trials.

ACKNOWLEDGEMENT

I am most grateful to Dr Jennifer Peat and Dr Katrina Williams for helpful discussion and advice.

REFERENCES

Andresen BS, Dobrowoloski SF, O'Reilly L, et al (2000) The mutational spectrum in the MCAD gene of newborns identified by prospective tandem MS screening for 'diagnostic' acyl-carnitines in blood spots differs from that observed in clinically affected patients. *J Inherit Metab Dis* **23**(supplement 1): 12.

Alger S, Green A, Kohler W, Sokolwski P, Moser H (2000) Proceedings of the 4th International Workshop of the Adrenolekodystrophy International Research Group, September 1998. *J. Inherit Metab Dis* **23**: 449–452.

Bickel H, Gerrard J, Hickmans EM (1953) Influence of phenylalanine intake on phenylketonuria *Lancet* **ii**: 812–813.

Carpenter KH, Wiley V, Sim KS, Hammond J, Heath D, Wilcken B (2000) Newborn screening for MCAD deficiency: the New South Walse experience. *J Inherit Metab Dis* **23**(supplement 1):16.

Coulombe JT, Kammerer BL, Levy HL, Hirsch BZ, Scriver CR (1983) Histidinaemia part III: Impact: a prospective study. *J Inher Metab Dis* **6**: 58–61.

Edwards SJ, Lilford RJ, Braunholtz D, Jackson J (1997) Why underpowered trials are not necessarily unethical. *Lancet* **350**: 804–847.

Ghadimi H, Partington MW (1967) Salient features of histidinemia. *Am J Dis Child* **113**: 83–87.

Griffiths M (1987) 'Underpowered' trials. *Lancet* **350**: 1406–1407.

Koch R, Moats R, Guttler F, Guldberg P, Nelson M Jr (2000) Blood–brain phenylalanine relationships in persons with phenylketonuria. *Pediatrics* **106**: 1093–1096.

Lilford RJ, Thornton JG, Braunholtz, D (1995) Clinical trials and rare diseases: a way out of the conundrum *Br Med J* **311**: 1621–1625.

Lindner M, Zschocke J, Schulze A, et al (2000) Tandem mass spectrometry detects mild MCAD deficiency with negative phenylpropionic acid test. *J Inherit Metab Dis* **23**(supplement 1): 125.

Morrow G, Barness LA (1968) Methylmalonic aciduria. A newly discovered inborn error. *Ann Intern Med* **69**: 633–635.

Moser HW, Bexman L, Lu SE, Raymond GV (2000) Therapy of X-linked adrenoleukodystrophy: prognosis based upon age and MRI abnormality and plans for placebo-controlled trials. *J Inherit Metab Dis* **23**: 273–277.

Mudd SH, Skovby F, Levy HL, et al (1985) The natural history of homocystinuria due to cystathonine β-synthase deficiency. *Am J Hum Genet* **37**: 1–31.

Pitt D, Danks D (1991) The natural history of untreated phenylketonuria over 20 years. *J Paediatr Child Health* **27**: 189–190.

Pocock SJ, Elbourne DR (2000) Randomized trials or observational tribulations? *N Engl J Med* **342**: 1907–1909.

Rizzo WB, Leshner RT, Odone A, et al (1989) Dietary erucic acid therapy for X-linked adrenoleukodystrophy. *Neurology* **39**: 1415–1422.

Roscher A, Liebl B, Fingerhut, Olgemoller B (2000) Prospective study of MS-MS newborn screening in Bavaria, Germany. Interim results. *J Inherit Metab Dis* **23**(Suppl. 1): 4.
Sacks H, Chalmers TC, Smith H Jr (1982) Randomized versus historical controls for clinical trials. *Am J Med* **72**: 233–240.
Smith I, Beasley MG, Ades AE (1990) Intelligence and quality of dietary treatment in phenylketonuria. *Arch Dis Child* **65**: 472–478.
Walter JH, Michalski A, Wilson WM, Leonard JV, Barrett TM, Dillon MJ (1989) Chronic renal failure in methylamalonic acidaemia. *Eur J Pediatr* **148**: 344–348.
Wilcken B (2000) Two years of routine newborn screening by tandem mass spectrometry. *J Inherit Metab Dis* **23** (Suppl. 1): 4.
Wiltshire EJ, Poplawski NK, Harrison JR, Fletch JM (2000) Treatment of late-onset non-ketotic hperglycinaemia: effectiveness of imipramine and benzoate. *J Inherit Metab Dis* **23**: 15–21.