EVALUATING OUTCOMES OF NEWBORN SCREENING PROGRAMS

Bridget Wilcken

New South Wales Newborn Screening Program,
The Children's Hospital at Westmead, Sydney, Australia

Abstract. Newborn screening is a medical intervention. For every program, there should be evidence of its effect and effectiveness. The four questions to be addressed, very broadly, are: What is the effectiveness for case-finding (sensitivity, specificity, and positive predictive value)? What are the benefits of early detection versus clinical detection? What harm results from the program? Are the costs reasonably balanced in relation to benefits? Ideally, there would be randomized controlled trials (RCTs) of screening for each disorder. In practice, power calculations reveal that for very rare disorders this is not feasible. The numbers of screened and unscreened babies required would be huge, and trials would last for decades. There have only been RCTs of newborn screening for cystic fibrosis (birth prevalence 25-40 per 100,000 in Caucasians). No such trials were ever attempted for hypothyroidism, with a similar birth prevalence, and it may not now be ethical to mount one. Instead, lower orders of evidence must be used. Double-blind randomized controlled trials should be planned for not-so-rare disorders where possible. Where it is not feasible, careful planning and collection of data, plus the use of both historical controls and contemporaneous controls from other regions may have to suffice. To introduce programs with no plans for full evaluation is not best practice. Evaluation of outcomes of all kinds, not simply of case-finding, must be mandatory. Data for case-finding should be collected actively, with systematic searching for missed cases. Data about benefits need to be collected in well-planned long-term studies, although short-term benefits are also valuable. Good studies of harm, mainly from false positive results, are urgently needed. The problem of costs and benefits is difficult, and a "reasonable balance" rather than positive cost/benefit ratio seems desirable.

INTRODUCTION

Newborn screening is a medical intervention. It is mandated by law in some countries, but is completely voluntary in others, and available evidence suggests that where a program is well-established, both approaches are associated with a high coverage of close to 100 percent. Newborn screening tests are generally very inexpensive when considered on an individual basis, but a program is expensive to run for a government organization and it is important to know if the money is well spent. While early detection by newborn screening seems intuitively likely to be very advantageous for a number of disorders, it is surprising how little hard evidence exists for clinical advantages in many cases. As for any medical intervention, it is very important that we have proof of both validity and utility.

To evaluate outcomes in newborn screening, there need to be answers to a number of questions:

- How effective is case finding?
- What are the benefits of early detection over clinical detection?
- What are the harmful effects? Are the benefits achieved greater than the perceived harm?
- Are costs reasonably balanced in relation to benefits?

Of these questions, the second is the major one that needs to be considered and often is not. There is an urgent need for proper documentation of benefit. Of course, when a new program is being considered a simultaneous review of the likely answers to all these questions is advisable, but without evidence from pilot studies there are obvious problems in coming to any conclusions. Very often theoretical problems raised turn out to be those not actually encountered when a program has been put in place and run for some time, and other unforeseen problems may arise. An example of this was seen with screening for cystic fibrosis, when likely problems were canvassed by a committee (Neonatal screening for cystic fibrosis, 1983). While some of the

possible problems never eventuated [eg inability of the test to detect patients with pancreatic sufficiency (Wilcken *et al*, 1995)] others arose [eg inadvertent detection of carriers (Massie *et al*, 2000)]. In the United Kingdom, a general consideration of newborn screening including the possible introduction of tandem mass spectrometry (MSMS) triggered the commissioning of two separate health technology assessments. These came to completely different conclusions as to the course to be followed (Pollitt *et al*, 1997; Seymour *et al*, 1997).

EFFECTIVENESS OF CASE FINDING

The sensitivity (the ability of a test to detect all those with the disorder in question) and specificity (the ability to classify correctly those who do not have the disorder) are generally recorded by newborn screening programs, although there are often problems in interpreting those data. To arrive at a true value for the sensitivity there must be a systematic search for "missed" cases, and this should be considered a vital part of the screening program as a whole. The specificity may be easier to calculate correctly, but the positive predictive value (PPV, the likelihood that a positive result indicates a true-positive case) is unfortunately often not considered. This value varies according to the prevalence of a disorder in the community, but may be very much more important in practice than the value for specificity. Consider two geographic areas, with 500,000 babies tested in each: in one, the birth prevalence of the disorder sought was 1:100,000, and in the other 1:10,000. If 500 positive results were obtained in each instance the specificity would be 99% in each. But the likelihood of a positive result indicating a case (the PPV) would be 10% for one area, and only 1% for the other. This factor would make a difference in the perceived value of the program and the costs of case finding. A recent paper examined USA programs for phenylketonuria, galactosemia, biotinidase deficiency, congenital hypothyroidism and congenital adrenal hyperplasia, for the years 1993 and 1994 (Kwon and Farrell, 2000). There was an apparent sensitivity of 100% for most disorders, although there was no available data about how missed cases might be discovered, and the specificity was uniformly above 99%. The PPV's however ranged from 0.5 to 6%, with more than 50 false-positive results for each true positive.

There is a constant balancing act between keeping a high (perfect?) sensitivity on the one hand, and avoiding a high false-positive rate on the other. A "reasonable" sensitivity may vary from disorder to disorder, and depend on the value assigned to missed identification. It may be considered completely unacceptable to risk missing even one case of classical PKU, as the consequences are so severe, but quite reasonable to risk missing a baby with well-compensated hypothyroidism

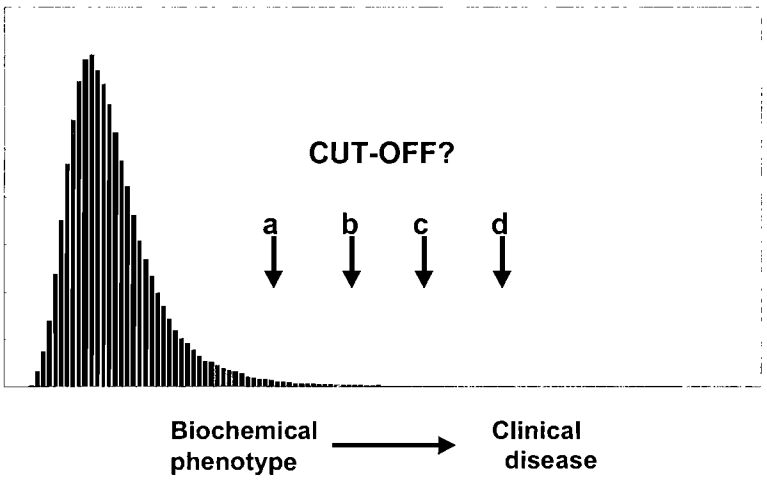

Fig 1. A typical frequency distribution curve for any analyte. Variation in the selected cut-off point will alter inversely the relationship between sensitivity and specificity. Sensitivity will decrease as the cut-off point moves from point **a** to point **d**. (More cases will be detected with a cut-off point at **a** than at **d**). Specificity is highest with a cut-off point at **d** (few if any false positive results will occur) and decreases as the cut-off point is moved to **a**.

due to an ectopic thyroid, as overt hypothyroidism may not occur for some time, and the adverse effects of delayed diagnosis would likely be very small. Moreover, it is not always easy to define what is or is not a case. For example, when does hyperphenylalaninemia define a "case"? Hyperphenylalaninemia is defined as a blood phenylalanine level of 120 umol/L or above (Scriver and Kaufman, 2001), but this is not a cut-off level used by the vast majority of screening programs. Treatment is seldom started in babies whose blood phenylalanine levels are persistently below 300 umol/L, and the definition of a case detected by screening varies a good deal. The recent experience of screening for medium-chain acyl-CoA dehydrogenase deficiency also underlines the difficulties in defining a "case" (Carpenter *et al*, 2001). So the assigning of cut-off levels for each assay is difficult, as there is usually a continuum from the biochemical phenotype to the clinical disease (Fig 1). There are many other aspects of effective case-finding, including timeliness of all of the aspects of testing and follow-up. These, while of great importance in the process of screening, are not considered here.

CLINICAL BENEFITS OF EARLY DETECTION

The core business of newborn screening is to ensure preclinical diagnosis of a condition if this benefits the baby. Therefore we MUST be able to evaluate clinical outcomes, and this requires formal study.

The hierarchy of study designs for evaluating effectiveness is well known:

- Randomized controlled trials
- Experimental studies without true randomization
- Controlled observational studies – cohort studies and case-control studies
- Observational studies without control groups
- Expert opinion, the most unreliable indicator

Unfortunately in newborn screening, for a variety of reasons, formal studies have in the main not been attempted. This is largely because of two problems (Wilcken, 2001). Studies of screening for very rare disorders would require huge numbers to obtain the necessary power to find a specified difference in outcome, and secondly, the follow-up time required may in some cases be very long.

Evidence for benefit in newborn screening

What has been achieved in newborn screening? There have only been randomized controlled trials of newborn screening for cystic fibrosis, and nothing else (Farrell *et al*, 2000; Chatfield *et al*, 1991). The Wisconsin trial, from 1985 to 1990 was very well designed, and has reported results now to 13 years, demonstrating benefits in nutrition (Farrell *et al*, 2000), but not so far reporting on respiratory status. Any alteration in life expectancy will not be known for many years. For some disorders there is a clear-cut benefit of preclinical diagnosis by newborn screening, and no randomized trials could now reasonably take place. This is true for PKU, congenital hypothyroidism [although the magnitude of the benefit is not clear (Tillotson *et al*, 1994)], homocystinuria, even if sensitivity is poor, maple syrup urine disease, especially the milder forms, and perhaps sickle-cell disease (Lees *et al*, 2001). There has also been, more rarely, evidence of no benefit from early diagnosis. This was so for histidinemia, which was tested for quite widely in the 70's and even in the 80's, but was then shown to be a benign disorder. Screening for neuroblastoma was also shown not to be beneficial, in that most of the detected cases were those with good prognosis (Woods *et al*, 1996).

Evaluating outcome

In general, for a baby to be diagnosed with a disorder that would eventually need treatment is likely to be beneficial. There is indeed some evidence of the benefit of preclinical diagnosis for many disorders, but it is not always clear how much benefit. For congenital adrenal hyperplasia, for example, many observational studies have suggested that there are fewer deaths of salt-losing boys, a reduced incidence of salt-losing crisis, and a reduction in incorrect sex assignment (Brosnan and Brosnan, 2000; Thil'en *et al*, 1988), and this is widely accepted. To prove some of these outcomes by randomized controlled trial would require 2.5 million babies in each arm of a trial, a trial surely unlikely to be attempted (Wilcken, 2001). For sickle cell disease, a recent Cochrane review concluded that while there had been no trials of newborn screening there was evidence of benefit from the non-trial literature (Lees *et al*, 2001). (The authors, however, suggested that systematic reviews of early intervention should be considered).

Although there are substantial difficulties outlined above, it is vital to be able to evaluate clinical outcome. There is no point in attempting newborn screening if benefit cannot be established for each program undertaken. The sort of outcome measures that should be taken into account include the neuropsychological result, medical problems, and hospitalization data, as well as the costs associated with these. Finding an adequate control population against which to measure outcome is difficult in the absence of a randomized controlled trial. Depending on the disorder involved and the extent of the screening

program, this may involve the use of historical controls, contemporaneous controls from different geographic areas, or, if screening coverage is incomplete, contemporaneous controls from the same geographic area. All of these have the high likelihood of different forms of bias, which for any individual approach could render the data uninformative, and it is therefore important to use all the evidence available, and if possible make use of more than one type of control group. Fortunately, for some disorders, like PKU, the overall benefit of screening is quite obvious. The lack of formal evidence of benefit fits well the "self-evident evidence paradox" (Pollitt *et al*, 1999) that the more effective an intervention (in this case screening and early treatment) the fewer the scholarly publications likely to be devoted to it. Another problem in assessing outcome is the use of surrogate end-points. These may be well established for some disorders [eg blood phenylalanine levels in PKU (Smith *et al*, 1990)], but not helpful in others [eg galactosemia, where there is no correlation between galactose-1-phosphate levels during treatment and outcome (Waggoner *et al*, 1990)].

Examples where published evidence is insufficient

For a number of disorders commonly included in neonatal screening programs there is doubtful or insufficient evidence of benefit in the literature. That is not to say that there may not be a benefit from early diagnosis, but simply that there is insufficient published evidence about this. The many disorders in this category include glucose-6-phosphate dehydrogenase deficiency and congenital toxoplasmosis, and these provide good illustrations of the general problems facing screeners. Glucose-6-phosphate dehydrogenase deficiency is one of the commonest known enzymopathies, with the highest prevalence in tropical and sub-tropical areas of Asia, Africa and the Middle East, and areas of the Mediterranean. While the most common clinical features are neonatal jaundice, and acute hemolysis triggered by certain drugs, infections or fava beans, the vast majority of G-6-PD deficient individuals are asymptomatic throughout life. The aim of newborn screening for this disorder is to prevent attacks of acute hemolysis. Screening is unlikely to prevent neonatal jaundice, and although this may also be associated with kernicterus, with consequent significant morbidity and mortality, it is unlikely that this complication would be prevented by newborn screening but rather by improved pediatric practice. In Singapore, the prevalence of kernicterus had declined sharply to very low levels before neonatal screening for G-6-PD deficiency was instituted (Joseph *et al*, 1999). It is suggested that parental education about the avoidance of trigger substances will prevent acute hemolytic episodes, and that seems likely, although there is little published evidence about this. Because G-6-PD deficiency is extremely common in some areas, screening is likely to produce a relatively high false positive rate, and the follow-up required could compromise other screening activities. Careful studies seem to be needed to evaluate whether there is sufficient benefit actually realised for this sort of screening. Similarly, screening for congenital toxoplasmosis seems likely to be useful, but there is little confirmatory evidence. Infection during pregnancy leads to neonatal infection in about 80%, and about 80% of infected babies can be detected by screening tests. Of those babies, at least 80% will be asymptomatic, but risk long-term complications of chorioretinitis and developmental delay (Guerina *et al*, 1994). Early treatment probably prevents long-term sequelae, but there seems some doubt about this, and the treatment, with three drugs for one year, is not trivial. Again, further studies seem to be indicated, to explore the effects of screening on outcome.

Benefit *vs* harm

There has been little clear-cut demonstration of lasting harm from newborn screening programs. Concerns that early diagnosis of a serious condition would harm parent-child interaction have not been substantiated (Boland and Thompson, 1990; Helton *et al*, 1991). Most of the harm of newborn screening is likely to come from three sources:

- false positive results
- problems with definition of a case
- unwanted carrier detection

The effects of false positive results have been investigated in only a few studies. While some seem to show that lingering concerns about the child's health may persist for a significant period (Sveger and Thelin, 1981), studies with some form of control group have cast doubt on this (Sorensen *et al*, 1984). Case definition is likely to be a real concern however. The "new" newborn screening by tandem mass spectrometry detects with unexpected frequency inborn errors previously thought very rare (Zitkovik *et al*, 2001). One example is methylcrotonyl CoA carboxylase deficiency. So far it is not possible to certain in an individual case whether or not it is necessary to introduce early dietary and other management. The child has certainly been "labelled" and the family must be managed most sensitively to reduce any harm that will have flowed from the screening exercise. A similar situation is seen in unwanted carrier detection, where a DNA test is introduced as a second-line test in a screening program. At present this occurs with cystic fibrosis and medium-chain acyl-CoA dehydrogenase deficiency (Wilcken *et al*, 1995; Carpenter *et al*, 2001), but we can be sure that with the

expansion of DNA testing and the development of streamlined technology this will be an increasing problem in the future.

Evaluating costs

This is an important but very specialized area, and will not be explored in this paper. Areas to consider are the *incremental* costs of introducing a new program in a region where newborn screening is already taking place, and the costs associated with false positive results, as well as the costs of actual case-finding. There are major ethical questions involved in an examination of costs. The costs of treatment and cost savings may depend very much on overall outcomes, as a longer life may in some disorders mean longer, and therefore more costly treatment.

CONCLUSIONS

There is often such an emphasis on the process of newborn screening, and the appropriate quality control of that, that the bigger picture may be somewhat neglected for all but the longest and best-established screening programs (US newborn screening system, 2000). The core business of newborn screening is to produce a benefit to the baby tested. Therefore it is essential to be able to evaluate *clinical* outcomes. If randomized controlled trials are not possible, as is often the case, then national or multicenter prospective trials of promising programs are needed, with the best design possible, so that useful data can be produced. The newborn screening community cannot lag behind in the current endeavours to ensure that interventions produce the beneficial effects that they are supposed to provide.

REFERENCES

American Academy of Pediatrics, Newborn Screening Task Force. Serving the family from birth to the medical home-Newborn screening: a blueprint for the future. *Pediatrics* 2000;106(suppl):383-427.

Boland C, Thompson NL. Effects of newborn screening for cystic fibrosis on reported maternal behaviour. *Arch Dis Child* 1990; 65: 1240-4

Brosnan CA, Brosnan PG. Methodological issues in newborn screening evaluation with special reference to congenital adrenal hyperplasia. *J Pediatr Endocrinol Metab* 2000; 13: 1555-62

Carpenter K, Wiley V, Sim KG, Heath D, Wilcken B. Evaluation of newborn screening for medium chain acyl CoA dehydrogenase deficiency in 275,000 babies. *Arch Dis Child Fetal Neonatal Ed* 2001; 85: F105-9

Chatfield S, Owen G, Ryley HC, Williams J, Alfaham M, Goodchild MC, Weller P. Neonatal screening for cystic fibrosis in Wales and West Midlands: clinical assessment after 5 years of screening. *Arch Dis Child* 1991; 66: 29-33

Farrell PM, Korosok MR, Rock MJ, Laxova A, Zeng L, Lai HC, Hoffman G, Laessig RH, Splaingard ML. Early diagnosis of cystic fibrosis through neonatal screening prevents severe malnutrition and improves long-term growth. *Pediatrics* 2000; 107: 1-13

Guerina NG, Hsu HW, Meissner HC, *et al.* Neonatal screening and early treatment for congenital Toxoplasma gondii infection. *N Engl J Med* 1994; 330: 1858-63

Helton JL, Harmos RR, Robinson N, Accurso FJ. Parental attitudes towards newborn screening for cystic fibrosis. *Pediatr Pulmonol Suppl* 1991; 7: 23-8

Joseph R, Ho LY, Gomez JM, Rajdurai VS, Sivasankaran S, Yip YY. Mass newborn screening for glucose-6-phosphate dehydrogenase deficiency in Singapore. *Southeast Asian J Trop Med Public Health* 1999; 30 (suppl 2): 70-1

Kwon C, Farrell PM. The magnitude and challenge of false-positive newborn screening test results. *Arch Pediatr Adolesc Med* 2000; 154: 714-8

Lees CM, Davies S, Dezateux C. Neonatal screening for sickle cell disease (Cochrane review). Oxford: updated software. In: The Cochrane Library, Issue 1, 2001.

Massie RJ, Wilcken B, Van Asperen P, *et al.* Pancreatic function and extended mutation analysis in deltaF508 heterozygous infants with elevated IRT but normal sweat electrolyte levels. *J Pediatr* 2000; 137: 214-20

Neonatal screening for cystic fibrosis: position paper. *Pediatrics* 1983; 72: 741-5

Pass KA, Lane PA, Fernhoff PM, *et al.* US newborn screening system guidelines II. Follow-up of children; diagnosis, management and evaluation. *J Pediatr* 2000; 137: S1-S46.

Pollitt RJ, Green A, McCabe CJ, *et al.* Neonatal screening for inborn errors of metabolism: cost, yield, and outcome. *Health Technol Assess* 1997; 1(7):i-iv, 1-202

Pollitt RJ. Priciples and performance: assessing the evidence. *Acta Paediatr Suppl* 1999; 432: 110-4

Scriver CR, Kaufman S. Hyperphenylalaninaemia: phenylalanine hydroxylase deficiency. In: Scriver CR, Beaudet AL, Sly WS, Valle D, eds. The metabolic and molecular bases of inherited disease. 8th ed. New York; McGraw-Hill 2001: 1667

Seymour CA, Thomason MJ, Chalmers RA, Addison GM, Bain MD, Cockburn F, Littlejohn P, Lord J, Wilcox AH. Newborn screening for inborn errors of metabolism: a systematic review. *Health Technol*

Assess 1997; 1 (11): i-iv, 1-95

Smith I, Beasley MG, Ades AE. Intelligence and quality of dietary treatment in phenylketonuria. *Arch Dis Child* 1990; 65: 472-8

Sorensen JR, Levy HL, Mangione TW, Sepe SJ. Parental response to repeat testing of infants with "false positive" results in a newborn screening program. *Pediatrics* 1984; 73: 183-7

Sveger T, Thelin T. Four-year-old children with alpha 1-antitrypsin deficiency. Clinical follow-up and parental attitudes towards neonatal screening. *Acta Paediatr Scand* 1981; 70: 171-7

Thil'en A, Nordenstrom A, Hagenfeldt L, von Dobeln U, Guthenberg C, Larsson A. Benefits of neonatal screening for congenital adrenal hyperplasia (21-hydroxylase deficiency) in Sweden. *Pediatrics* 1988; 101: E11

Tillotson SL, Fuggle PW, Smith I, Ades AE, Grant DB. Relation between biochemical severity and intelligence in early treated congenital hypothyroidism: a threshold effect. *BMJ* 1994; 309: 440-5

Waggoner DD, Buist NR, Donnell GN. Long-term prognosis in galactosaemia: results of a survey of 350 cases. *J Inher Metab Dis* 1990; 13: 802-18

Wilcken B. Rare diseases and assessment of intervention: what sort of clinical trials can we use? *J Inher Metab Dis* 2001; 24: 291-8

Wilcken B, Wiley V, Sherry G, Bayliss U. Neonatal screening for cystic fibrosis: a comparison of two strategies in for case detection in 1.2 million babies. *J Pediatr* 1995; 127: 965-70

Woods WG, Tuchman M, Robison LL, Bernstein M, Leclerc J-M, Brisson LC, et al A population-based study of the usefulness of screening for neuroblastoma. *Lancet* 1996; 348: 1682-5

Zitkovik TH, Fitzgerald EF, Marsden D, *et al.* Tandem mass spectrometric analysis for amino, organic, and fatty acid disorders in newborn dried blood spots: a two-year summary from the New England Newborn Screening Program. *Clin Chem* 2001; 47: 1945-55