Evidence-Based Practice, Ethics and EHDI Program Quality

Martyn Hyde

The Evidence Revolution

Evidence-based practice (EBP) has evolved over the last decade into what might be considered a mantra of the new millennium. Early on in this evolution, Yale University’s Alvan Feinstein, who was a founding father of modern clinical epidemiology, noted that the evidence-based approach to clinical care ‘... has acquired the kind of sanctity often accorded to motherhood, home and the flag ...’ (Feinstein and Horwitz 1997). The pervasiveness of the evidence-based approach, and the reverence accorded to it by many, have increased in recent years. Mantras and their related totems often evoke a broad range of response, from uncritical enthusiasm through to uncritical aversion. Evidence-based practice is no exception, with its share of devotees and detractors. As history often teaches us, there may prevail eventually a balanced viewpoint that is less doctrinaire, which recognizes the strengths and limitations of the new approach, and which places them more or less in proper perspective. One purpose of this article is to acquaint the reader with some tools, strengths and limitations of the evidence-based approach, at least from my own perspective. This will be done with special reference to the context of EHDI programs. This article is intended to complement, and in some areas to amplify, that by Judy Gravel in this volume. Unfortunately, this subject requires a discussion of ideas, not data, so I apologize to the reader for the wad of text and lack of nice figures.

There have been many attempts to define evidence-based practice or, more specifically, evidence-based medicine (EBM). The University of Oxford’s David Sackett, a leader in the field, described it as ‘... conscientious, explicit and judicious use of current best evidence ...’ in making decisions about clinical care (Sackett, Rosenberg, Gray, Haynes and Richardson 1996). Brian Haynes, another leader in the field, defined it as ‘... a set of tools and resources for finding and applying current best evidence from research for the care of individual patients’ (Haynes 2002). An interesting definition given by Timio and Antiseri (2000) was: ‘... an emerging paradigm of scientifically based clinical care. It de-emphasizes intuition and unsystematic clinical experience ...’ As the reader will see, this definition touches on both key strengths and key limitations of current EBP.

One measure of the progression of interest in EBP is citation rates in databases such as Medline. In fact, the number of ‘evidence based’ methods citations in Medline has grown with astonishing linearity from about zero in 1995 and previously to about 2,100 in 2003.

The enthusiasm for an evidence-based approach extends beyond healthcare. In the realm of education, Mayer (2000) argued strongly, against considerable odds, for a more evidence-oriented approach, as an escape from ‘the abyss of relativism, in which all opinions are equally valid.’ The abyss of opinion-driven action to which he refers is a concern in many areas of public services, including healthcare.

Why Evidence-Based Practice?

Several forces motivated the development of a more evidence-based approach to clinical services specifically and to public services more generally. Perhaps the most powerful was (and still is) the
inexorable pressure of exploding service costs and the need to contain costs and maximize ‘bang for the buck.’ There is also an ethical imperative to continuously strive to improve the quality of services.

The evidence movement was a seed that fell on fertile soil amidst gentle rain in precisely the right season for luxuriant growth. One piece of ground in this case was the emerging discipline of Health Services Research, closely coupled with rapid growth in the areas of Evaluation and Methodology. Those disciplines elaborated useful conceptual models and methods, and refined important constructs such as efficacy, effectiveness, equity and efficiency of procedures and services. A closely linked area is Health Economics, with its array of models and techniques, especially cost-effectiveness, cost-utility and cost-benefit analysis.

A key nutrient was explosive growth of the volume of scientific reports and an emerging, obvious need to provide practitioners with methods of accessing, evaluating, integrating and applying the mass of available evidence in a practicable and useful way. Furthermore, there was a growing disenchantment among the clinical epidemiologic and health services communities with the poor methodologic quality of much published research. A third factor was an abundance of examples of so-called ‘area variations’ in clinical services. What this jargon refers to is observation of marked inconsistency in methods, effectiveness or efficiency, for a given clinical problem, among units of care delivery ‘area’ that may be individual clinicians, professional sectors, institutions, cities, states or countries.

The entire process of EBP is geared towards making available relevant, valid and concise information to support effective clinical decision-making. EBP tools and methods are oriented towards sifting germs of probable truth from a morass of irrelevant or potentially misleading material. Many published reports contribute little but noise to the processes of scientific and clinical advance. The depth and continuing extent of this problem is both startling and puzzling. A key issue for the working clinician who wishes to keep up with the science is not so much how to access the vast amount of published material, but how to decide what NOT to bother reading. EBP provides some useful principles and techniques by which this can be accomplished.

The Mechanics of EBP

1. Primary Scientific and Clinical Reports

The raw fodder upon which the machinery of EBP usually chews is the primary scientific or clinical report. The main ingredient is the published article in a peer-reviewed journal, but other sources of information such as articles that are not peer-reviewed, internal, unpublished reports, and a range of web-based sources, may yield useful information. A particular concern with peer-reviewed articles is publication bias, such as a tendency to publish preferentially articles that describe positive results of ‘interventions,’ whether the intervention of interest is an act of screening, of diagnostic assessment, or of therapy or ‘treatment,’ to use the jargon of the medical model. Another, longstanding concern that dates back well before the time of Galileo is the bias of the scientific and clinical establishment toward articles that are consistent with, or at least do not substantially threaten, widely held hypotheses, beliefs and models. A third concern is that, as just mentioned, much of the available peer-reviewed fodder is not wheat, but chaff.

2. Critical Appraisal of Primary Reports

Critical appraisal is, at a grass roots level, the business of evaluating the quality of published reports and their relevance to the clinical or other question at hand. For example, one may wish to decide on the best course of action for a specific patient or client with a specific clinical presentation and context of action. Alternatively, the challenge may be to design a clinical service program that is grounded in scientific evidence to the greatest extent possible, or to evaluate and improve an existing program of care or pattern of clinical practices.

To achieve these ends, the process may start with a literature search and evaluation. Assuming that one is not naive enough to believe all that is written, even by gurus, then an explicit, careful and consistent process is necessary to evaluate the quality and relevance of the published material.

Typical, key questions that form the basis of critical appraisal are presented below. For each area of concern, the published material must be examined with respect to the completeness and depth with which the issue was described in the report, as well as
with respect to the actual substance of the information given:

i) is the report clearly relevant to the question at hand, especially with respect to the choice of target population, the act of measurement or intervention being studied, and the outcome measures used?

ii) was the study sample appropriate and representative of the intended target population, and were there any obvious sources of bias in the sampling and recruitment methods used?

iii) was the procedure under investigation performed appropriately?

iv) was the outcome measure used valid and reliable?

v) was the sample size sufficient to achieve the study objectives?

vi) were the data analyzed appropriately and reported in sufficient detail?

vii) were the author’s conclusions clearly justified by the data, etc?

Some more specific information on some of these issues is given shortly, but a comprehensive discussion is beyond the scope of this article. For a wealth of introductory information on critical appraisal, the reader might try Google searches on ‘MS1903 Spittlehouse’ and ‘OMNI Greenhalgh.’ The latter will access an excellent and very readable set of articles in the British Medical Journal (BMJ) on the evidence-based approach, including evidence reviews, database methods, statistical methods and more (see, for example, Greenhalgh, 1997 a,b,c).

Every clinician should be aware of at least the basic elements of critical appraisal, for at least two reasons. First, conscientious front-line clinicians seek to continuously evaluate and improve their practices. Second, there is a need for efficiency in extracting relevant information from the avalanche of published material on almost any topic, be it obtained through conventional database searches or, increasingly, from web-based sources. Clinicians with an interest or involvement in program development or evaluation, as well as those with a research interest, should go further and be knowledgeable and skilled in critical appraisal.

3. Reviews, Systematic Reviews and Meta-Analyses

Traditionally, reviews are attempts to survey what is known in a given area and, usually, to boil the material down to a few key findings and/or recommendations. The literature is full of ‘review’ articles, and the word ‘review’ covers a wide range of activities, with radically different validity and value. Traditionally, reviews are often commissioned from ‘experts.’ People perceived to be experts may or may not have profound knowledge or insight in any particular area. It is not easy to distinguish genuine expertise from other attributes such as the mass of publication, professional prominence, or verbal and literal sophistication. There is a natural tendency for experts to look favorably upon published articles that support their current point of view, and the traditional ‘review’ method, which is largely informal, encourages this potential for bias, be it unconscious or conscious.

An experience that many experts have is surprise at how their own work is reported, summarized or misinterpreted by others, especially in reviews. This often reflects a tendency on the part of the reviewer to read selectively and see only what is desired. It also reflects the fact that the scientific and clinical communities inevitably engage in something akin to the whisper game. At the end of a chain of reinterpretation by different individuals, what was originally said is frequently unrecognizable. A related phenomenon that obscures the truth is derivative reporting. Author A may say X, for example. Authors B through E may then report X in various ways, each referencing author A. Author F may then state that X is ‘well-established,’ or even that there is a ‘general consensus’ on X, yet X derived from one finding by one author, which might even have been not a solid finding but only a speculation. This type of distortion is not uncommon in traditional review, but it, like many other sources of untruth, will be detected by a competent systematic review.

These many flaws in traditional evidence reviews have certainly limited the rate of progress of clinical science. They are also responsible in part for conceptual inertia, the promulgation and persistence of false concepts, beliefs and practices, and the generally erratic pattern of development in many fields. Systematic review, as the cornerstone of EBP, represents a more structured and critical approach to the assembly and evaluation of available evidence, and the derivation of valid inferences on which to base
The whole point of systematic review is to minimize the predispositions for bias and error that are inherent in such practices as the traditional 'expert' review. The key problems that systematic review seeks to avoid are: incomplete gathering of evidence; aggregation of evidence of widely differing quality; biased selection of key evidence; inappropriate inference that is not justified by the data; and imprudent generalization of study findings to other populations and contexts.

The essence of a 'systematic' review lies in its comprehensive and explicit methodology. The process starts with the clearest possible definitions of the key clinical questions, the precise target population and the context of clinical operation to be considered. Special techniques are used to minimize search bias and maximize completeness of source identification. Objective and predefined criteria of relevance and quality are then applied as 'filters' to the body of literature identified relating to a given issue, in order to identify in an unbiased way 'key' evidence of high relevance and quality. Special techniques are then applied to the abstraction, integration and reporting of the overall findings and to evaluation of their generalizability and limitations.

There is much more to a good systematic review than just running a Medline search on a couple of obvious keywords. Even searching is in itself a non-trivial art. There are multiple databases with different coverage of the possibly relevant literature. In particular, it can be important to go beyond Medline. Date range restrictions often exclude useful information; there is a mistaken tendency to believe that only recent articles can be relevant, yet it is remarkable how slow is the pace of scientific advance in many areas, and how many things are constantly being re-invented. It is also risky to exclude non-English sources – some very smart people in this world do not necessarily publish their seminal works in English.

Searches often result in highly incomplete retrieval, because of poor or incomplete search term selection. A trivial example is a search on 'hearing aid' without covering the possibilities of 'hearing instrument' or 'amplification'; another example is not using wildcards to cover all the possible endings of a string such as 'audi' in a context such as 'audi (tory) (ologic) (ological) rehabilitation.' Also, one cannot rely fully on the accuracy of classification and search term abstraction by database providers. For further insight into these and related matters of database searching, see, for example, the article by Greenhalgh (1997) on database methods.

While there are standard techniques for systematic review, it is certainly possible to conduct useful reviews that do not involve the full array of methods for formal, systematic review. These are sometimes called 'overviews' or 'semi-systematic' reviews; they typically have explicit search criteria and quantitative study inclusion criteria. A good example of this is the overview of prevalence of permanent childhood hearing impairment (PCHI) by Fortnum (2003). However, not all non-systematic reviews are of such a high quality, and there is a continuum of quality between a good, systematic review, a good but not fully systematic overview, and a poor, traditional, 'expert' or 'ad hoc' review.

Meta-analysis is a process of quantitative integration of results from a set of studies. It is typically preceded by a systematic review, to enhance completeness of source data and to identify key reports of high quality that are sufficiently homogeneous in their approach and methods that they can be combined with validity. Some important issues in meta-analysis relate to homogeneity of target populations and subject groups, appropriateness and consistency of outcome measures, completeness of statistical reporting from source studies, exploration of potential bias especially in relation to sample size, and statistical methods of aggregation and summarization of data. It is not unusual for meta-analyses to be limited in scope or even impossible, because of inadequate description or lack of homogeneity of subjects, methods or outcomes in primary studies. For example, see the comments by Fortnum (2003). A definitive text on meta-analysis methods is due to Hunter and Schmidt (2004). Examples of meta-analyses in audiology are Stapells (2000) and Amlani (2001).

4. Rating Scales for the Quality of Evidence

In order to give appropriate weighting to source studies of various types and to reports with varying scientific standards, it is necessary to apply a hierarchical scale of the quality of evidence. Several such scales exist, and most are based on similar fundamental precepts that emerge from classical domains of experimental design and statistical inference. The key underlying concepts are validity, minimization of bias (systematic error), maximization of precision (minimization of random error), and generalizability,
which relates to the extent to which study findings can be expected to hold true in other samples, populations or circumstances than the exact ones studied. Validity problems often arise due to use of outcome measures that do not, in fact, measure what they are intended to measure, or that may be valid in one population of subjects but not in another. The second major source of invalidity is confounding variables, namely variables other than the intervention of interest that affect the outcome measure. Bias can arise in many ways, commonly by inappropriate (non-representative) selection of the study sample and by inadequate control of confounding variables. By far the most common source of inadequate precision is insufficient study sample size. Generalizability problems arise commonly because of inappropriately restricted sampling of subjects, by the use of inappropriately idealized intervention methods that may not be representative of real-world clinical application, and also by insufficient sample size.

Studies can be classified as experimental or observational, and the opportunities to avoid or control sources of error are much greater in experimental studies. In these studies the allocation of the test ‘intervention,’ be it an act of measurement of some attribute of the subject or an actual ‘treatment’ to be delivered, is under the control of the investigator. In observational studies, the intervention of interest occurs outside the control of the investigator. Experimental studies are necessarily prospective, in that the outcome follows the intervention delivery. Observational studies may be prospective or retrospective, depending on whether the outcome has already occurred at the time of initiation of the study observations.

The randomized, controlled trial (RCT) is widely considered to be the ‘gold standard’ study type. It is experimental in nature, it includes concurrent test and control conditions, such as a novel intervention versus a conventional intervention or no intervention. The key ingredient is random allocation of subjects to interventions which, if done properly, virtually guarantees that sources of bias will operate according to known statistical laws, in which case statistical tests of hypotheses are likely to be valid.

Cohort studies involve the determination of outcomes for ALL members of a defined group of interest (the inception cohort). They may or may not involve control groups and they do NOT involve random allocation of interventions. The key limitation, therefore, is that those subjects who have received the intervention may be different from the population of interest at large, in some manner that affects the outcome of interest. That is, cohort studies are intrinsically more prone to bias than RCTs, because of the inability to allocate intervention randomly. Controls may be concurrent or historical, but there is always the concern that the intervention and control groups differ in some way that is not known or quantifiable, so the outcomes of interest may be confounded by these variables.

Case-control studies are like cohort studies in reverse, in which the occurrence or lack of occurrence of the outcome of interest defines the study groups, which are then examined retrospectively to determine the association between the outcome event and the ‘intervention’ event of interest.

Note that studies of the yield of cases with PCHI from EHDI programs are NOT cohort studies because the essence of a good cohort study is determination of outcome in the entire group of subjects who have the test ‘intervention,’ in this case a hearing screening test. In EHDI programs, typically only the infants who screen positive are followed to determine the presence of PCHI. In contrast, the well-known NIH study of screening methods (Norton et al. 2000) was a cohort study, because an attempt was made to determine the outcome in all subjects screened.

Studies involving case series or individual cases may be case-control or merely case reports, in which there is no means to assess the representativeness of the observed effects or the presence or absence of biasing factors. Such studies typically serve to generate new hypotheses or to call attention to novel phenomena that may not fit with established doctrines or models.

Finally, we have ‘studies’ lacking specific observations but which relate to review of published reports of varying quality and unknown bias. One variant of this is the opinion piece, which may or may not be based on rational analysis of actual data.

These examples of types of evidence reflect a hierarchy of potential validity that seems reasonable. An example of so-called Levels of Evidence Scaling with respect to intervention studies is as follows:

1a Systematic review with homogeneous RCTs
1b Individual RCT with narrow Confidence Interval.
2a Systematic review with homogeneous cohort studies
2b Individual cohort study, or RCT with < 80% followup.
3a Systematic review with homogeneous case-control studies
3b Individual case-control study
4 Case series and poor-quality cohort/case-control studies
5 Expert opinion without explicit critical appraisal

This scale is extracted and simplified from that available from the Centre for Evidence-Based Medicine in Oxford, England. A much more comprehensive listing of levels of evidence, covering a wider range of study types other than therapeutic intervention studies is available at the CEBM website. Note, in particular, that the opinions of experts are at the bottom of the list!

5. Clinical/Professional Practice Guidelines (CPGs/PPGs)

Development of CPGs or PPGs is a common motivation behind many large, systematic evidence reviews and meta-analyses. The purpose may include improvement in quality of care or equity of care access, protection of the public from risk of harm, enhancement of provider accountability, budgetary justification, political or legal defense of individual or program actions, restriction of care options and cost constraint. Many CPGs are directed as facilitating decision-making during the management of individual clients or patients. Typically, the guidance is expressed in terms of clinical algorithms and decision trees.

The translation of findings from systematic reviews into CPGs may be far from straightforward. The more complex and multifaceted the clinical problem, the more difficult is the CPG development. Despite these limitations, CPGs are a growth industry. The US Agency for Healthcare Research and Quality (AHRQ) has been a leader in EBP development, and the reader can Google on ‘AHRQ’ for a wide range of relevant materials. As reported by Judy Gravel in her keynote address to this conference, a relevant initiative is the much-anticipated New York State CPG on Assessment and Intervention for Children 0–3 years with Hearing Loss.

The production of a CPG is not an end in itself. The usual measure of the utility of a CPG is the extent to which it is reflected in the actual processes of care delivered. Do providers follow it? Does it help in determining choice of actions? Most fundamentally, does it improve care processes and/or outcomes, in relation to care effectiveness and/or efficiency? Not surprisingly, the EBP movement has come up with several approaches to assess compliance with, and effectiveness of, CPGs. See, for example, Hutchinson, McIntosh, Cox and Gilbert (2003).

Problems with Primary Research Reports

Earlier in this article, it was alleged that much of the primary clinical research literature has significant limitations of one kind or another and may contribute little to scientific and clinical advance or to clinical guidance in specific situations. There are several indicators that support this assertion. In 1997, Alvan Feinstein noted that ‘in a recent review, more than half of pertinent publications did not fulfill basic methodologic standards for scientific quality’ (Feinstein and Horwitz 1997). In 2001, the U.S. Preventative Services Task Force (USPSTF) reported on their findings arising from a systematic review of the evidence relating to effectiveness of universal newborn hearing screening (UNHS), conducted by the Evidence-Based Practice Center at Oregon Health and Science University. The systematic review was reported by Thompson et al. (2001), and is also accessible by Googling ‘ahrq 02-S001.’ Essentially, the USPSTF reduced the entire body of literature on newborn screening to a mere handful of key evidence papers that formed the basis for their ultimate grading of the evidence. They also remarked on the methodological limitations of the available evidence.

In 2005, the Public Health Agency of Canada’s ‘Canadian Working Group on Childhood Hearing’ has issued a ‘resource manual’ on the evidence related to EHDI (CWGCH 2005). This will include the results of a systematic review of evidence for effectiveness for different approaches to language development in early childhood. Carried out by the prestigious Chalmers Research Group based at the Children’s Hospital of Eastern Ontario Research Institute, the systematic review could draw no conclusions because of methodologic deficiencies in the several hundred primary reports.

Judy Gravel, in her keynote address, noted that in the development of the imminent NY State CPG on assessment and intervention, several thousand primary abstracts were reduced to several hundred.
articles worthy of detailed review. Less than 10% of these articles were selected as constituting evidence of sufficient quality and relevance to provide a basis for the CPG itself.

This is all just the tip of the iceberg of indications that the bulk of published evidence has significant flaws or limitations. While I would be delighted to be proved wrong, I suggest a working hypothesis that maybe one primary paper in ten would make it through any reasonable quality filter targeting methodological and reporting deficiencies. If true, this is bad news for well-motivated editors but good news for clinicians, provided that they can spot the flaws quickly. Audiology is by no means exceptional in this regard. For example, an extensive critique of the flaws and limitations of existing outcomes research in the area of speech-language pathology was given by Pring (2004).

The deficiencies in methodologic quality or presentation of the majority of peer-reviewed clinical research papers in many areas, including audiology, have many causes. Assuming that serious criticism in review would block most papers (but certainly not all), the obvious inference is that the general level of methodologic awareness among reviewers is less than it should be. A long-term solution may be to ensure that the relevant postgraduate courses enhance their program content with respect to critical appraisal, experimental design, clinical epidemiology and statistical methods.

In the meantime, it may be helpful for clinicians to be alerted to a few major flaws that are curiously widespread in the primary literature, despite the fact that they are quite basic from a methodological standpoint. These issues come up time and time again in critical appraisal.

Clarity of Question and A Priori Statistical Hypotheses.

The question addressed by the research must be stated clearly, succinctly, specifically and prospectively. Fuzzy objectives such as ‘to explore the relationship between . . . ’ are typically associated with poor design and analysis. Retrospective formulation of objectives, driven by inspection of data collected with no clear, a priori question is especially counterproductive, because chance manifestations of random sampling errors will tend to be misinterpreted as genuine effects. Many studies lend themselves naturally to very specific statements of statistical hypotheses, and the more specific, the better. Some studies, especially more exploratory studies, are simply descriptive and have no associated hypotheses; however, such studies serve only to generate hypotheses for future study, and they do not themselves yield adequate bases for clinical decision-making. It might be very helpful to the audience if journals clearly identified such reports up front, perhaps as part of a larger scheme of classification. Other studies may have objectives that are formulated clearly, but are questionable in terms of their rationality. For example, is a comparison of click ABR and multi-frequency ASSR in the prediction of frequency-specific thresholds in infants appropriate, given that the click is not a frequency-specific stimulus? Is determination of the ‘accuracy’ of screening ABR using ‘diagnostic’ ABR as a gold standard appropriate, and would the reader be surprised if the screening sensitivity were found to be 100%, given that the screening and reference tests reflect the same underlying phenomena?

Appropriateness of Outcome Measures or Endpoints

Outcome measures must be explicit, precisely defined, relevant and valid. A particular problem is insufficient reported information to fully define the measure, such as a set of ABR thresholds in dBnHL without any SPL reference level or without clear criteria to define ABR presence or absence. Another perennial problem is the use of inadequately validated measures to study intervention effects. If no effect is found, for example, was it the intervention or the measure that was inadequate?

A more fundamental issue is the appropriateness of the outcome measure(s) selected. For example, it seems very popular to focus on long-term language development, as the primary outcome measure by which to evaluate the success of UNHS/EHDI programs. To some, this focus is puzzling and even perverse. Would one evaluate a newborn vision screening program, for example, primarily by measuring reading ability at age 3–5 years? I think not. It can be argued that if poor hearing or poor vision are the primary deficits, then demonstrable improvement in hearing or seeing ability are the most direct, primary outcome measures.
Identification and Management of Confounding Variables

Good studies carefully consider variables other than the procedure of interest that could affect the outcome. Some of those variables can be held constant, though at a risk of losing generalizability of findings. Others may be managed by stratified sampling, such as gender and age decade effects. Yet others may be managed in analysis, such as with analysis of covariance as opposed to simple analysis of variance. For an example of a concern about confounding, consider the finding that infants with PCHI who are identified before six months of age have better language development at age about three years than those identified later. There is some evidence to suggest that the degree of parental involvement in communicative development has a strong effect on outcomes. What if, by some unanticipated mechanism, children with parents who were predisposed to be very involved were over-represented in the early-identified group? Such a bias, whether systematic or a chance occurrence, would tend to produce the observed result, even if there were actually no true effect of early identification. I am not suggesting that such an effect DID occur in existing reports, but only that such a possibility is plausible and must be considered explicitly in any future studies.

Target Population Definition and Accrual of the Study Sample

In a good study, sufficient, detailed information about the intended target population is given such that the study could be replicated meaningfully by others. A cause of confusion in the early literature on ABR screening of high-risk infants, for example, was lack of clarity in distinguishing subgroups such as recent neonatal intensive care unit (NICU) graduates and infants referred for testing due to suspected abnormal development. Of course, the proportion of abnormal ABRs differs according to the weighting of these subgroups. A more subtle version of the same problem is variability between studies because of differing levels of effort to properly characterize the sample. One can easily conclude that a baby is not at risk, for example, simply by not putting much effort into the assessment of risk. This has no doubt contributed not only to the marked variation in estimates of the proportion of infants at risk for PCHI, but also to variation in estimates of hearing loss prevalence in ‘low risk’ and ‘high-risk’ groups.

In its largely negative evaluation of the literature on effectiveness of newborn hearing screening, the USPSTF (2001) commented on ‘methodologic limitations’ that included (in the Task Force’s opinion) the use of ‘convenience samples.’ This refers to the acquisition of study samples by ad hoc or naturally occurring processes, as opposed to sampling according to a standardized scheme, such as random sampling from a defined target population. Convenience samples are subject to a host of potential sources of bias that may confound the outcome measurement.

Determination of Statistical Power and Required Sample Size

This is an enormous problem in the general clinical literature, despite the fact that for most funding agencies, a sample size justification is absolutely mandatory in a research proposal. The issue is crucial in hypothesis-testing studies. The key question is statistical power, that is, the probability of detecting the effect one is seeking, if it were truly present. The conventional, minimum required power is 0.8 or 80%. If the actual power were only 50%, for example, then one could forget about doing the study and simply toss a coin! Yet, an astonishing proportion of studies are seriously underpowered (see, e.g., the excellent and readable text by Lipsey 1990). If power is low, then a negative finding will happen most of the time and is virtually meaningless. Despite this, there are many reports in which an underpowered negative finding is over-interpreted as convincing evidence of no effect.

Any study that involves testing a primary hypothesis should incorporate a power and sample size calculation. If the resources of the study do not permit the desired sample size, then there must be a calculation of the size of effect that would be reliably detected, given the largest practicable sample size. Another strategy is to lower the variability of the outcome measures by methods such as restriction of subject diversity, stratification, covariate adjustment and within-subject replication. Reduced error variation increases power, other things being equal.

Choice and Delivery of the ‘Intervention’

If the measurement act or intervention is poorly defined or not delivered correctly, the study cannot be
replicated and the findings are likely to be irrelevant. To illustrate, in order to study the relative merit of two techniques, it is essential that both be done appropriately and, preferably, in an optimal way. What is the value, for example, of showing that ASSR is a more accurate predictor of true hearing levels than an ABR done with an inappropriate recording bandwidth or averaging strategy? Worse still, what is the value of a claim of superiority of one technique over another when all the salient parameters defining the technique are not described?

Methods to Reduce or Eliminate Bias

The usual methods for bias reduction include random sampling of subjects from a defined pool of potential subjects, random allocation of subjects to procedures, matching of subject groups on specified covariates, and blinding of measurement to group membership or to potential correlates of the primary outcome variable. Blinding is especially important when there is any possibility that the outcome measure could be affected by subjective opinion of the producer of the measurement. This may be an issue not only in obvious areas such as in subjective ratings of intervention effectiveness, but also in areas such as evoked potential or VRA threshold estimation, for which there is a subjective component in response assessment.

Completion of the Outcome Measurement Protocol

Failure to complete the measurements on a significant proportion of subjects is a strong source of possible bias. For example, surveys with return rates of less than 80% are commonplace, yet if those who respond do so because they are special in any way, the generalizability of the findings to all subjects is highly suspect. Similarly, in experimental studies, what if those who comply fully with a lengthy intervention process do so preferentially because they believe a given intervention is working? The potential for bias is obvious, and while there are standard remedies, such as ‘intention-to-treat’ analysis, these are often not employed.

A fundamental problem here is that methodologists often (and perhaps justifiably) set standards for follow-up coverage that may be reasonable in the context of, say, adult drug trials for acute conditions but are very difficult to achieve in the context of prevalence or (re)habilitation studies. The question then is: should all studies with less than 80% response rates or follow-up completion rates be scrapped? The answer is no, but their results certainly cannot be taken at face value, without a careful analysis of potential sources of bias and limitations on generalizability.

 Appropriateness of Statistical Analysis

This area of major deficiency in peer-reviewed publication could fill an encyclopedia, and all that can be done here is to give a few examples. The first is to illustrate an obsession with measures of central tendency, namely mean values, and a willingness to ignore variation.

Descriptive studies often focus on producing a mean value for some outcome, such as a performance level on some auditory test, or a correlation between two measures. This mean value is a ‘point estimate’ of the true, underlying quantity, and is a measure of so-called ‘central tendency’ in the underlying statistical distribution of possible values that might have been obtained from the study. The problem is that the practical or theoretical usefulness of the point estimate can only be assessed if it is accompanied by an ‘interval estimate,’ which is a statement of the range of plausible alternative values that are consistent with the study data. For example, two studies might yield 0.85 as the estimated sensitivity of, say, a screening test for PCHI. Study A might yield an interval estimate of 0.6 to 1.0, and the other study (B) an interval of 0.81 to 0.88. The point is that study A is virtually useless, but study B is highly informative. Thus, if there is no interval estimate actually given in the publication, the credibility and utility of the stated result cannot be determined! Despite this profound limitation, studies still get published without interval estimates on their primary outcomes. These are ‘confidence intervals,’ of course, and the mystery is that while everyone learns about confidence intervals in basic statistics courses, there appears to be such a lack of understanding that the interval is what really matters, not the point estimate itself! See the review by Fortnum (2003) for comment about this issue.

Another side of this coin is the number of publications that do not make available the actual, observed values of the source data. The individual data points are much more informative than summary statistics
alone. Ideally, the reader should be given enough information to be able to re-analyze the data in ways that may differ from those used by the study authors. At the very least, the reader should be put in a position to be able to assess the validity and appropriateness of the author’s analytical methods and inferences. One very important tool is the scatterplot, particularly if the purpose of the study is any kind of correlation or regression analysis. It is astonishing how many published articles either do not give scatterplots or, even if they do, show data patterns that reveal absolutely basic errors of inference that are described in any introductory textbook on correlation or regression analysis!

In hypothesis-testing studies, the most common error is a lack of power, as already discussed. But there are several other traps for the unwary. For example, some studies essentially seek to prove that measure B or intervention B is equivalent to or at least no worse than measure or intervention A, and proceed to test the very familiar null hypothesis that A=B, with a one-tailed (upper tailed) alpha of 0.05, say. The null hypothesis, in fact, is the hypothesis that the study seeks to DISprove, not to prove. In this case of so-called ‘non-inferiority’ study, the proper null hypothesis would typically be of the form B>=A-M, where M is known as the acceptable ‘non-inferiority’ margin. This equation reads: B is not worse (smaller) than A by more than a clinically acceptable decrement of size M.

Another trap that is common relates to the importance of formulating a limited number (preferably ONE or TWO) of a priori hypotheses involving a limited number (again, preferably one or two) of outcome measures. The more hypotheses and the more outcome measures there are, the greater the sample size demands and the greater the likelihood of false detection of statistically significant findings. Plugging in popular procedures for ‘correction’ of significance levels, such as the Bonferroni adjustment, does not always solve the issue of multiple comparisons appropriately. See the excellent text by Stevens (2001) for a discussion of this and many other significant issues in statistical analysis.

Even more subtle than multiple a priori comparisons or ‘contrasts’ (eg among groups receiving different ‘treatments’) are so-called post-hoc contrasts, wherein the comparisons of interest are not established before the study but are suggested by either preliminary or final inspection of the data. Here, the optimal methods for adjustment of significance levels depend quite strongly on the precise nature of the data.

There are many other areas of concern, such as the appropriateness or otherwise of nonparametric versus parametric analyses, any of which could fill a textbook. However, it is clear that some of these illustrated areas of error or omission in published works are quite straightforward, so much so that their frequency is very puzzling, given basic statistical education and the peer review process. Other issues are much more subtle and it would be unreasonable to expect every researcher or reviewer to have the required depth of expertise. It appears that there is a much greater need for consultation with a statistician than is commonly acknowledged. The time to bring in such expertise is when the study is being formulated, not after it is done. One issue is that consulting with statisticians can be painful. Typically, the process requires a clarity of question formulation, a restriction of objectives, and a commitment of resources that are all greater than anticipated.

Linkage of Study Results to Inferences and Conclusions

Finally, the reader of scientific journals should examine carefully whether the author’s conclusions are actually justified by the data presented. Frequently, authors over-generalize their findings and, on occasion, it can be difficult to perceive any obvious relationship between the data and what is concluded. In a recent example of this plucked from the area of ASSR and hearing threshold estimation, the authors of a not unsophisticated study noted a typical standard deviation of predicted hearing level based on ASSR of about 15 dB. If true, this is expected to yield a 95% confidence interval range of about ±30 dB HL for prediction of hearing level given any specific ASSR result. In a remarkable feat of intellectual elasticity, the authors concluded that ASSR is a reliable tool for estimating frequency-specific hearing thresholds! The point here is not whether or not ASSR actually IS a reliable tool, but that a potential prediction range of up to 60 dB could be considered to reflect reliability and clinical utility. Often, the biases inherent in the point of view of the authors of any given study are quite obvious even from the literature review, and this provides clues for the reader in the evaluation of potential bias or over-generalization of the study findings.
Limitations of EBP

Issues with Randomized Controlled Trials

The RCT is the holy grail of the evidence-based movement, and with some good reason. As mentioned earlier, RCTs, when done properly, are free from many of the sources of bias that confound lesser study designs, and they tend to yield more valid statistical significance levels and estimates of effect sizes. There have been many situations in which RCTs have debunked conventional clinical practices based largely on opinion and inadequate studies. As is clear from the earlier scale of evidence quality, it is believed by many that there is nothing more convincing than a good meta-analysis of several good RCTs with concordant results, and even one good RCT is highly rated as evidence. Yet, not everyone is so enamoured with the RCT, and here are some reasons why even RCT evidence should be appraised very critically. See, for example, Julian (2003) for a more detailed analysis.

Before getting into some of these issues, remember that the idea of a ‘treatment’ and a ‘control’ group can be conceived very generally. A ‘treatment’ does not have to be a conventional act like giving a drug or fitting a particular type of hearing aid. A ‘treatment’ is basically any well-defined act of service provision. The ‘control’ condition can include any alternative ‘treatment,’ a conventional standard of care, or no treatment at all.

In the area of EHDI specifically, to my knowledge there have been no large RCTs that would stand up to critical appraisal. The well-known Wessex trial of UNHS in the UK was a good but non-randomized clinical trial (Wessex 1998). A true RCT of UNHS would involve randomization to an ‘unscreened’ group and would be unethical as evidence. Yet, not everyone is so enamoured with the RCT, and here are some reasons why even RCT evidence should be appraised very critically. See, for example, Julian (2003) for a more detailed analysis.

Next, the study must be feasible. RCTs tend to be large and expensive. They also require substantial sample size and this may be a difficulty if the supply of subjects is inherently limited, as it may be when studying relatively rare conditions. The prevalence of infant PCHI, for example, is about 2 per thousand live births and it would be unusual to have a useful non-exploratory study involving less than about 50 children with PCHI. This might imply a need for an initial target population size of at least 25,000 newborns. Thus, even if any ethical concerns can be dealt with, it may not be really practicable to conduct properly sized RCTs of this population except in the context of regional centers of excellence, or of state- or national-level access to subjects. Furthermore, if samples of sufficient size are accrued at a slow rate, there are many additional threats to study validity and relevance. Procedures may change in subtle ways as the field moves forward in terms of concepts, methods or technologies, and a study that takes three or four years to complete may find itself answering a question that is no longer of interest to the clinical community.

Next, the RCT must involve appropriate procedures and outcome measures. The big concern here is that RCTs are often conducted with exquisite care by investigators who have unusual levels of expertise. When the study is done, the new treatment may be found to be an advance and then disseminated into more widespread clinical practice. It is not uncommon for the RCT-based promise of the treatment not to be realized in general use. There is, therefore, a limitation of the generalizability of the RCT results. This may happen because the treatment delivery is less consistent or of poorer quality than in the trial context, and similarly for the outcome measurements. This is one facet of the difference between treatment efficacy and treatment effectiveness. Efficacy relates to whether a treatment CAN work, whereas effectiveness relates to whether it DOES work in the real world. It is, however, possible to conduct RCTs that deliberately attempt to address effectiveness. Typically, this type of trial is undertaken after earlier trials have shown efficacy under more restricted or idealized conditions.
Another problem related to generalizability arises through the accrual of the study sample. RCTs typically have a lengthy list of intake criteria and exclusion criteria. The general idea is to reduce sources of bias and variability by focusing the study on a relatively restricted and 'purified' group of subjects with the target condition and few, if any, contaminating factors that might affect the outcome. The problem with this is that in the real world, patients or clients may present with all manner of subtle variations of the target condition, and with numerous contaminating variables that were systematically eliminated from the RCT sample. It is no surprise, then, that the RCT findings again do not play out in the same way in general practice.

A related aspect is that RCTs typically and deliberately have a strong focus on a limited number of outcome variables (often just one primary outcome) and a limited number (maybe two or three) of covariates or potential confounders. This is commonly held to be good design practice, mainly for reasons relating to the validity of statistical significance levels. However, outcomes may have multiple facets and what is valued by one patient or client may be different from that valued by another. Also, there may be many more than three covariates in play for any given individual client. This is another source of 'impurity' and variability in the real world.

Finally, it is intrinsic to the nature of RCT design and analysis by such methods as Analysis of Variance (ANOVA) that there is a data focus on mean values, namely measures of central tendency for large groups of subjects. Differences from the group mean for individual subjects are considered as random error components. It is only in more complex designs that involve numbers of repeated measures within subjects, that there can be any consideration of treatment effect variation from subject to subject. For the majority of RCTs, the results are limited to group behavior. A deep question that arises, therefore, is: what is the relevance to the individual client or patient of a study result that tends to be true on average for a large group, but which may not be remotely true for the individual?

Note that the field of cochlear implants, or in fact, hearing aid outcome research in general, is unusual in this regard, provided that test materials that involve multiple items that are essentially equivalent are used. Standard word lists or sentence lists, for example, can be considered as repeated measures, and the scores for individual subjects can often be modeled statistically as binomial variables. In such studies, it is possible to explore and describe outcomes both at the group and individual levels, because it is possible to separate 'pure' within-subject error variance from inter-subject differences. It is not uncommon to find subgroups of subjects who differ substantially in response to key parameters of the interventions.

In summary, there are several fundamental limitations of RCTs in terms of their capacity to guide real-world clinical practices. These limitations give pause for thought about the 'gold standard' status of the RCT and are good reasons not to accept even high-quality RCT findings at face value, with respect to practice guidance.

Issues with Systematic Reviews

There is no doubt that systematizing the process of appraising evidence is potentially a major advance on traditional, unsystematic reviews. However, one problem that arises from the move to systematic review as a core element in, for example, the design and evaluation of programs such as EHDI is that more often than not, the evidence base for important decisions about program elements is found to be simply inadequate. This is not a deficiency of the systematic review process but, rather, it stems from the limitations of primary evidence discussed earlier. Thus, proper systematic review is not a garbage-in, garbage out process. It can, however, lead to a stark realization of just how little that has to be done in clinical service programs is based on high-quality evidence.

Even if the primary evidence base is adequate, activities such as systematic review and guideline development are not automatically beneficial. One problem is that, like most other human activities, they can be done well or done poorly. A particular issue here is the patina of credibility afforded by the mere words 'systematic review.' The expertise necessary to conduct a proper systematic review is substantial. Even if the clinician is reasonably familiar with critical appraisal as it applies to primary literature, it is a different matter to distinguish systematic reviews of high quality from those that should be disregarded.

The EBP movement, being grounded in methodologic critique, is keenly aware of this problem and has taken a number of steps to address it. Methodologic and procedural guidelines for the proper conduct of
systematic reviews have been developed, as have useful tools to evaluate some aspects of the quality of new systematic reviews. For further information on these areas, see, for example, the websites for the Cochrane Database of Systematic Reviews (CDSR), Ovid EBM, the Database of Abstracts of Reviews of Effectiveness (DARE) and the American College of Physicians (ACP) Journal Club. Googling on any of these phrases will take the reader to interesting places.

There are, though, some intrinsic limitations of the systematic review process; some of these are partially addressed by the various tools that are available for promoting higher quality of review. But limitations, or at least, areas in need of careful attention, remain.

Perhaps the most profound limitation lies in the choice of the question addressed by the review. There is no well-accepted process to define, validate and refine the question itself. Review questions are typically generated by an ad hoc process that may or may not achieve the best possible clarity, relevance and completeness. An example of this will be given later.

Another issue is the effectiveness of the evidence search, as touched upon earlier. Readers who have much experience with searches will understand how easy it is to miss important evidence because of incomplete search terms, date limits, language limits, database abstraction errors, idiosyncrasies of terminology by primary authors, omission of important databases, and so on.

Next, there is the problem of bias in key evidence filters. The set of criteria that define which pieces of evidence are ‘key’ are partially objective but necessarily involve appeal to deep issues in the primary field itself, and often, consultation with ‘experts’ in that field. This inevitably introduces subjectivity, even as it also permits avoidance of some obvious errors that would not be apparent to a methodologist with no primary content expertise. For example, in a review of ABR screening test performance, should studies based on non-automated ABR screening be included or not? Should studies based on specific screening devices be excluded because of some perceived limitation of their design or response detection algorithms? There are always many such issues and they are usually determined by opinion.

Once the key evidence candidates have been selected, there is the procedure of quality rating, either to eliminate some sources or to assign relative weighting to the selected reports. This process, again, is not fully objective. It is not uncommon for methodologists to differ in their opinion about what is important in determining quality. For example, even statisticians differ widely in their opinions on the relative merits of say parametric versus nonparametric analysis, or about procedures to handle covariates or multiple comparisons or a posteriori group comparisons.

The danger, then, is that by moving from an unsystematic process of review to a systematic one, we have essentially replaced clinical expert opinion (the lowest grade of evidence in the New Order) by methodologic opinion. It is not entirely clear why methodologic opinion should be rated more highly than primary clinical opinion. It is difficult to weigh the impact of that issue against the undoubted desirability of attempts to raise general methodological standards and to sort out the gems in the primary literature. But it is clear that just because a particular review is called ‘systematic,’ its aura of validity should NOT be taken at face value.

In meta-analysis, which operates on either systematic review output or primary evidence appraisal, the two substantial issues are the inability to integrate apples and oranges on the one hand, and the extent to which the mysterious fruit is rendered identifiable, on the other. The decision about whether studies are sufficiently similar to justify aggregation of their findings is almost always multi-parametric and usually has a subjective component. For example, can studies that use different electrophysiologic recording parameters or different hearing aid signal processing algorithms be combined in a valid way? Such decisions require intimate knowledge of the technical issues, and there may be no clear answer.

Issues with CPGs

There is substantial variation in quality across the rash of CPGs that has appeared in recent years. Generally, when appraising a CPG, the first concern is its relevance to the clinical problem at hand. The second concern is the strength of critical appraisal that was applied to the evidence base underlying the component clinical processes within the CPGs. The third concern is the strength of the primary evidence base itself; one cannot build a house on a sand foundation. Fortunately, because the EBP community is so acutely aware of issues of quality, the reader might expect that there would be specific tools and guidelines for evaluating the quality of CPGs, and indeed
there are. There have been several useful efforts to develop guidelines for the development of guidelines. Also, there are several evaluative approaches and tools, the general purpose of which is to guide the CPG development process and to critically appraise CPGs. See, for example, Cluzeau (2003).

The True Complexity of the World

It is human nature to approach natural phenomena in a simplistic way. Two relevant observations on this are by Albert Einstein: ‘Everything should be made as simple as possible, but not simpler.’ and by Oscar Wilde: ‘the pure and simple truth is rarely pure and never simple.’ We like to think about group data and central tendencies, not about variation and idiosyncracy. We like to think that relationships among variables are linear, when mostly they are non-linear. We like to think of effects of different variables as additive, when mostly they are interactive. These are some of the reasons why so much of the primary literature is not useful: we typically ask vague and simplistic questions, seeking to grasp the essentials of a formidable complex world. But, in fact, we usually view the elephant through the keyhole. Here is one of many possible examples: What is the Prevalence of PCHI?

It seems an entirely reasonable question, but in fact it explodes when looked at under a microscope. Those readers who are familiar with chaos theory and the Mandelbrot set will recognize the possible analogue of self-similarity at various levels of magnification when examining the problem. The more you analyze the variables determining prevalence, the more parameters you see as potentially relevant. Here is a list of factors that are known to affect the answer to the question: the laterality, severity and frequency range of impairment to be included; the inclusion of structural conductive impairment and auditory neuropathy; the age at which impairment is determined; the inclusion of known progressive impairment; the audiometric method and levels of clinical skills; the completeness of coverage of the target population; adjustments for SPL changes in the maturing infant ear; errors due to concurrent transient impairment; the racial and genetic characteristics of the target population; the quality of health care in the target population; the socioeconomic properties of the target population, etc, etc.

The point is that in fact, the question in italics above is a bad question to which there is no simple and valid answer. A better question is: what are the most important variables that affect the prevalence of PCHI, what are their effects and how do they interact? The prevalence must be seen not as a number that actually exists somewhere and needs only to be discovered, but as a dependent variable in a multiple regression equation with a host of independent variables. For those who like statistics, the prevalence is NOT a unique number but is actually a surface in a multidimensional variable space. No wonder there is so much confusion and variation in reports about PCHI prevalence. The reader could torture herself further by thinking about how one would best meta-analyze prevalence data . . .

A Spectrum of Evidentiability

Taken as a whole, all of these considerations lead me to a proposition. It is that there is a hierarchical spectrum of the degree to which any clinical question can be answered in valid and practicable way by the typical processes of EBP. The hierarchy is based on the degree of complexity of the input and output variables, and the laws underlying their relationships. Here is one way of looking at the concept.

First, we have the action or intervention itself, which may be an act or process of measurement, such as in screening or diagnostic assessment, or of ‘treatment,’ such as provision of amplification. An example of a simple intervention might be injection of a one-time bolus antibiotic. At the other end of the scale, an example might be an audiologic habilitation program or a language development program lasting several years. One key dimension is the temporal one: is the action a single event, a few discrete events, a series of events, or a lengthy and virtually continuous process? The other key dimension is the complexity of the action, which can be thought of in terms of the number of important underlying parameters or components.

Next, we have the set of outcome variables. This may be simple, even unitary, such as a pass or refer result on screening. At the other extreme, we may have complex, multi-parametric outcomes such as language level at say age five years or, even more complex, the entire temporal profile of language development, or a highly complex construct such as quality of life.

Next, we have provider mediation. This reflects
the extent to which the action is rigidly specified or is modifiable or interpretable by its implementer or provider. The more opportunity for variation there is, the more difficult the area will be to study.

Next, we have recipient mediation. This reflects the extent to which the outcomes are modifiable by the recipient, such as by co-morbidities, individual preferences or values. An audiogram, for example, could be an outcome and is not modifiable, whereas a quality of life measure is highly modifiable by individual values.

Finally, we have provider-recipient interaction, which may range from zero (such as for the antibiotic injection) through to profound, such as for a habilitation program. The greater the amount of interaction, the more complex is the actual intervention delivered and the more potentially diverse the array of recipient responses.

Given this, the idea is that these various dimensions of complexity influence the extent to which there is likely to be a body of useful, high quality evidence and whether the processes of classical EBP are likely to be successful, such as by the development of a CPG that is relevant and seen as useful in the real world. Some observations for the EHDI context follow.

Screening

The action is relatively simple and highly discrete in time. The outcome is simple and objective. Provider mediation is modest, the main issues being coverage or access to the target population and skills at test conduct. Recipient mediation is negligible. Interaction is negligible.

Audiologic Diagnosis

The action is moderately complex but is usually discrete in time (except in cases of auditory neuropathy and progressive impairment). The outcome is moderately complex. Provider mediation is moderate, such as by virtue of protocol and technique. Recipient mediation and interaction are slight for ABR-based testing but greater for behavioral testing, due to responsiveness factors.

Amplification

The action is moderately complex and may be extended over time. The outcome is complex and may be time-variant. Provider and recipient mediation are substantial, such as by virtue of options, skills and preferences. Interaction may be strongly influential on outcomes.

Communication Development

The action is complex and extended in time. The outcomes are complex and are extended in time. Provider and recipient mediation are very strong, as are interactions.

Consider some studies involving these areas. It is reasonable to expect studies of screening per se to be relatively straightforward, studies of diagnostics and amplification to be more difficult, and studies of communication development to be very difficult indeed. It is difficult to conceive of a meaningful RCT in the area of communication development, and if one were to be designed it would require exquisite attention to the definition of the intervention and the outcomes, and to control of provider effects, recipient effects and their interactions.

Note that where studies involve more than one area, the more complex area will determine the overall complexity. Thus, if screening were to be studied in relation to diagnosis, it would be the latter that would govern the complexity, whereas if a compound series of screening and diagnostics was to be studied in relation to communication development outcomes, it would be the communication development that would pose major problems in successfully completing a valid study.

Clues in support of this idea are that, in my view, there are many studies with reasonable methodologic quality of screening methods per se, less than ten such studies of screening in relation to diagnostic assessment (only one controlled trial and perhaps three cohort studies), and two or three such studies involving screening through to communication development (a view supported by the findings of the US Preventive Services Task Force 2001). Similarly, there are very few high quality studies of pediatric hearing aid prescription methods in relation to even short-term outcomes. These studies are very demanding, by virtue of their inherent complexity in the terms just outlined.

EBP for Complex Activities

In problem areas that are complex, in the terms just described, EBP is likely to find limited high
quality evidence, at least to the methodological standards espoused by leading authorities in the EBP movement, including the USPSTF (2001). Consequently, CPGs are unlikely to be evidence-based in the strict sense. Here, the drive to produce CPGs may result in opinion-driven CPGs and/or to a liberal reinterpretation of what constitutes high-quality evidence. Examples of this are not hard to find. The utility of such CPGs remains an open question.

How, then, can one go about discovery of the key ingredients of high-quality care, that is, care that is effective for a broad range of individual clients and contexts, as well as being efficient and accessible? Such discovery is essential, not only for designing and implementing superior services and programs, but also for program evaluation and quality improvement, for training of new health professionals and for continuing education.

One clue to an approach lies in the observation that clinicians vary in their skills and effectiveness. In my experience, much of complex clinical activity is an art that is more than a little mysterious. Some people learn to be effective much faster than others, whether in electro-physiologic audiometry, VRA, hearing aid provision or communication development training. Some people manage to achieve consistently good results, whereas others do not. These differences are observable but, typically, excellent clinicians have difficulty explaining and dissecting what it is that they do that works so well. Some characteristics that tend to be found in such clinicians include a broad grasp of technical and procedural principles, a deep empathy with their clients or patients, a concern to do the best possible work and, curiously, a humility and lack of awareness of just how good they are. Characteristics of clinicians who tend to be less effective include a lack of awareness of deep, technical issues, a rigidity of approach and, curiously, hubris and a sublime belief in their own competence.

These variations and characteristics present a golden opportunity to explore the key determinants of outcome, but in order to see the opportunity, one has to accept that the effectiveness of many clinical procedures is not intrinsic to the procedure itself but is profoundly dependent on the behaviors and other characteristics of the provider. This may come as no surprise in areas such as psychiatry, wherein there is evidence that provider effects may explain much more outcome variation than differences in treatment approaches or doctrines. It may also be no surprise in areas of audiologic habilitation and rehabilitation, which usually involve strong interactions between the provider and the recipient. However, it may also be true even in areas that are commonly believed to be much less influenced by psychological factors, such as the estimation of hearing thresholds from evoked potential records, or in VRA, or in the identification of functional overlay in routine audiometry. It is, in fact, true even for newborn hearing screening; it is not the automated screening device that primarily governs such things as the referral rate from screening. It is the detailed behavior of the screening personnel: when to test, how to handle the baby, how to handle probe insertion or non-patent ear canals, when and how often to retest, and so on. These are the human factors, which truly determine screening test performance.

One difficulty with studies of provider effectiveness variation is the notion that all certified professionals in a given discipline are created equal, and all have equally valid opinions and quality of practice, even though those opinions and practices may differ widely. This is nonsense, of course, but it is a real phenomenon that may be a real obstacle to scientific advance. A more respectable stance would be that any and all sources of knowledge to improve effectiveness will contribute not only to the quality of care for the individual client, which is what really matters, but also ultimately to the prestige and security of the profession as a whole.

For complex problems, there needs to be far more investigation not of the properties of things, but of the behaviors of providers and recipients of care. One approach would be to seek out evidence of outcome variation, identify ‘magic’ providers and systematically deconstruct what they do, by careful observation and experimental manipulation.

A complementary approach is to move away from the group-oriented, brute-force approach of the large trial with big but simplistic questions. If an important limitation of such trials is the squelching of interesting sources of variation into global random error, then perhaps a way out of the resulting mire is to deliberately seek out sources of outcome variation, treatment component interaction and meaningful response differences between subjects. Inevitably, this will involve studying smaller and smaller domains more and more intensively. The idea is to answer small questions well, rather than big questions badly. It will also involve more use of repeated-measures designs that are capable of revealing differences between subjects.
The overall thesis here is that the evidence bases for complex clinical problems have to be addressed by a process of piecewise construct validation, analogous to the method used to develop a new and better outcome measure in a field that lacks an external gold standard measure or criterion. Construct validation is a process, sometimes lengthy, of identifying valid functional relationships between relevant variables in a given problem domain. Piece by piece, the architecture of the overall domain is built up from this set of lawful relationships. At some point, it can be said that there is a valid and useful model of the domain that can be applied in clinical care as well as in further scientific study.

A difficulty that must be overcome in this process involving not blunderbuss trials but piecewise construction is parameter explosion. For example, suppose in some hearing aid study there are five parameters of stimulus control, five parameters of signal processing, five parameters of subject characteristics and three components of outcome measures, then there are 125 parameter combinations affecting each of the outcome triplets. And that is just for one study. A comparable study might have say 50 outcome parameter combinations in common with the first study, but another 50 quite different, and two of the three outcome components different as well. How on earth can the results of even two such studies be compared and integrated, let alone a hundred related studies?

The problem is essentially one of communication, organization and strategy. The solution will lie in mechanisms that promote and facilitate large, collaborative, investigation and reporting structures, as well as much greater standardization of methodology and outcome measures. This will not be easy, given the ego strength (or weakness) of many involved in clinical research and practice. One important step, in the area of hearing aid outcomes, is to encourage widespread use of non-intrusive standard outcome components, such as the IOI-HA (Cox and Alexander 2002). More generally, there is an acute need to develop an acceptable set of minimum methodological, procedural and technical components that would be voluntarily incorporated into many studies of any given area, such as EHDI. In this way, at least some parametric ‘anchors’ would become commonplace and facilitate study linkage and meta-analysis. As a simple example, many difficulties in comparing audiometric studies arise from use of different numerical measures of hearing loss severity.

When the audiogram is to be summarized as a single number, one might use any single frequency, or any one of several averages, or a more complex value such as the lowest threshold within a range. If only the research community could be persuaded to ADD a ‘meta-analysis reference condition’ to their analyses, such as the 2 kHz threshold or some average. It matters much less which parameter is chosen than that SOME parameter IS chosen and used. If such strategies could be agreed, the pace of scientific advance would increase substantially. The key issues are: who cares enough to do something, and how would such an initiative be funded and operationalized effectively?

Evidence and Ethics

So far, we have touched upon the discovery, manipulation and use of scientific evidence. But with respect to delivery of clinical services and programs, these matters of evidence exist not in a vacuum but within a framework of underlying ethical principles. The core principles that form the basis for much ethical analysis and argument are: beneficence, non-maleficence, autonomy and justice. In basic terms, beneficence relates to doing good or to the conferral of benefits, non-maleficence relates to avoidance of doing harm, autonomy relates to the rights of individuals to choose whether or not to be the recipients of specific actions, and justice relates to fairness in the distribution or allocation of benefits among members of society. For an entire journal issue devoted to a wide range of important topics relating EBP and ethics, see the Journal of Medical Ethics 2004, 30(2). See, in particular, the very cogent comments on the history and current status of EBM by Liberati and Vineis (2004).

There are significant concerns about several aspects of the ethics of EBP itself. One concern is that the focus of EBP on study evidence reflecting the properties of populations or large groups of subjects may actually detract from the care provided to individuals, who may not fit the group norm. Particularly important are patient preferences among intervention options, the relevance of which is a cornerstone of client-centered care and informed consent. An issue is the extent to which these highly individual factors can be accommodated within an EBP style of care, respecting the autonomy of the patient (Slowther et al. 2004).
The most obvious way in which ethics have influenced EHDI programs is reflected in the deliberations of the USPSTF (2001) concerning the quality of evidence in support of UNHS. The Task Force explicitly seeks to weigh the evidence for benefits and harms associated with such screening programs. In particular, it addresses the balance of potential benefits from early identification and effective intervention and the potential harms arising from, say, false-positive screens, audiologic misdiagnoses and inaccessible or ineffective interventions.

Much of the Task Force’s approach to evaluation of evidence related to screening has its origins in the World Health Organization’s early treatise on the principles and practice of screening (Wilson and Jungner 1968). The key point is that the quality of evidence required to justify mass screening should be higher than that necessary to justify ordinary clinical investigations. When an individual proactively seeks clinical care, (s)he is usually deemed to have a right to choose among options, having been reasonably informed about the balance of benefits and harms for each option. In mass screening, actions are imposed upon, or at least offered to, large numbers of individuals who have generally not sought such actions. Also, if there is to be benefit, it is incurred by a small number of individuals, whereas many are exposed to the possibility of harms due to screening. It is these concerns that have driven the emphasis on possible harms due to false-positive screens, and to the overall effectiveness of the early detection and intervention process. For a further discussion of these issues, see Ewart (2000).

At a deeper level, it can be argued that the entire evidence-based movement is ethically driven. This appeals to the notion that interventions based on scientific evidence are intrinsically more respectful of the principle of autonomy, because in the absence of such evidence, there can be no demonstrably valid statements of benefits and harms to underlie informed choice. Also, there is an obligation to maximize overall beneficence and non-maleficence, and it is widely believed that scientific evidence is a more valid approach to that end than, say, the exercise of ad hoc practices based on clinical intuition. Furthermore, scientific evidence is likely to enhance distributive justice, because it may indicate preferential candidacy for specific interventions as well as differing balance of benefits and harms for various population subgroups.

Another aspect is family-centered care. This popular approach satisfies the principle of autonomy more fully than do traditional approaches of authoritarian care (the doctor knows best). Whereas authoritarian intervention does not respect autonomy, it may be very effective in acute care, such as giving an antibiotic for an infection. In marked contrast, in areas of habilitation and rehabilitation, there is wide recognition that active endorsement and engagement of the individual or the family in the choice of care options and in their delivery process is likely to increase effectiveness enormously. There is certainly evidence to this effect in relation to early language development (Moeller 2000).

Perhaps the most profound area of interaction between ethics and evidence of immediate interest lies in the rationale for implementing EHDI programs. The usual rationale is that early identification improves language development, that is, an argument of beneficence. Following a systematic review of evidence, the USPSTF (2001) rendered the opinion that the evidence for effectiveness in terms of long-term language development was inconclusive, largely because of study design limitations. Does this mean that EHDI programs should not be implemented, especially given the historical importance attached to the WHO principles of screening? Factors that weigh in the argument include the fact that in the absence of EHDI, the data indicate that many children with impaired hearing will not be detected until age two years or more. Inevitably, such children will have delayed language development, even if they ultimately achieve normal age-level performance after detection and intervention. What is the harm to the individual and to the family of the shortfall in language due to late identification? What is the benefit from earlier identification and intervention? These are not exclusively scientific questions, but also questions of societal values and ethics.

Another important aspect of this issue is the intrinsic value of the ability to hear. Most infants with PCHI cannot hear family speech and other sounds normally. Many such infants, excepting those with profound impairment, can be given improved audibility by, say, personal amplification that is properly fitted. What is the value that should be assigned to, say, a year of improved hearing for an infant, or the negative value assigned to an unnecessary year of sensory deprivation? Ethical questions, again.

Note that the evidence review process can address, for example, whether or not it is clear that UNHS leads to substantively earlier detection and
diagnosis, and this was found to be the case by the USPSTF (2001). But, in itself, evidence review cannot attribute quantitative value to such outcomes unless there has been specific study of the way people assign value or utility to these kinds of outcomes. This type of research is feasible. It is called utility theory, a branch of decision theory, and for more information, try Googling on ‘standard gamble’ and ‘time trade-off.’ But, as yet, it has not been reported in EHDI-land.

Ethical statements tend to have enormous power and to outweigh conventional issues of scientific evidence. The stem cell and cloning debates are examples. In the EHDI context, what would be the impact of a statement of ethical principle such as; ‘every child has an intrinsic right to choose to hear’? The ensuing debate would dwarf conventional issues of evidence. Such a statement could be viewed as assigning a very high benefit weighting to effective amplification following early detection. It would also lead to some interesting discussion about the rights of the child in relation to the exercise of autonomy by the family. This is an area of considerable ethical debate, especially when the choices made by the family may restrict the current and future abilities of the child to exercise personal freedom of choice with respect to hearing and communication development. For some further insight, see Levy (2002) or Google on Joel Feinberg’s concept of an ‘open future’; that important concept relates to the intrinsic right of a child to future autonomy of choices, balanced against the right of a family to exercise autonomy of choice that may restrict the range of future choice by the child.

Another ethical aspect of the rationale for EHDI involves the principle of distributive justice. Consider an adult with normal hearing who acquires sudden, sensorineural, moderate hearing loss. Such an adult would probably be struck with anxiety and would immediately seek a remedy because, of course, the ability to hear is highly valued by most people. Now, consider an infant who acquires exactly the same impairment perinatally, by any of several mechanisms. The infant will not seek any remedy, because it physically cannot. If society does not provide hearing health care services to that child, it is because the child did not seek services. The concept of withholding such services, should infants miraculously become able to seek service by themselves is unthinkable. Yet, the child does not seek services simply because it cannot, not because it chooses not to. Is there an ethical onus on society to act so as to identify children in that predicament? I suggest that the fact that infants do not seek care because they cannot seek it must modify the force of arguments in the WHO considerations about screening and the required quality of evidence. It might be argued on ethical grounds that failure to proactively identify infant hearing impairment violates at least the core principle of distributive justice.

Related, non-trivial arguments of ethical principle can also be developed in relation to the right of the family to the earliest possible knowledge of the hearing status of their child, the right to make timely and informed decisions, whether about further child-bearing or about communication development options for their baby, and the right of the family to avoid inappropriate communication strategies such as singing to their baby in the dark or speaking without facing the child.

Any or all of these ethical issues, if accorded high value by society at large, would essentially overwhelm scientific debates about levels of evidence for this or that amount of benefit or harm, unless harms were common, obvious and substantial.

Issues of Program Quality

Finally, a few remarks about the role of EBP and ethics in the context of EHDI program quality. Up to this point, from an international perspective, questions about the quality of scientific evidence have dominated recent discussions about whether or not to implement EHDI programs. These questions are legitimate and must be addressed. But, I suggest that it is ethical matters such as the right of the child to improved hearing, or the right of the family to make timely choices for themselves and their child that should dominate the discussion and the major program decisions. Note that the matter of EHDI impact on long-term language development is not mentioned here and may not, in fact, be the most important and compelling issue, from a broad, ethical standpoint. In fact, EBP has a limited role in informing such ethical debate, and its real, practical role lies in informing not the decision whether to implement EHDI, but the processes of how best to implement it.

It is interesting also to examine the role of ethics in program delivery. It is, of course, unethical to implement knowingly a program of poor quality. For example, it is certainly unethical to implement UNHS in the absence of adequate follow-up services. However, there are several other ways in which
ethically principles can have a strong impact on program design and execution. The principle of justice, for example, has several important implications. Distributive justice drives the important concept of equity, both equity of access to services and equity of quality of care. Equity implies that family A and baby A are entitled to the same quality of care as are family B and baby B. Among other things, this means that if baby A is screened by screener X and managed by audiologist X, and baby B is screened by screener Y and managed by audiologist Y, the quality of care must be as equivalent as is humanly possible. It is not enough that the two babies are seen by certified audiologists; it is the quality of care that must be equal, not the professional status of the provider. This pursuit of equity has enormous implications for consistency of procedures and compliance with evidence-based protocols to the maximum possible extent. For example, if audiologist X employed click ABR and audiologist Y employed tonepip ABR, the principle of equity goes out the window because click ABR cannot possibly guide amplification prescription to the extent possible with tonepip ABR performed correctly. What is important here is not the right of the professionals to exercise idiosyncratic judgements about what procedures they will perform, but the right of the babies to equitable care. Thus, in pursuit of equity of care, there is an ethical onus on the providers of care to maximize their consistency of practice across the entire program, guided by such evidence as exists.

Another crucial factor is program evaluability. It is difficult enough to establish an EHDI program, but it is a greater challenge to sustain, evaluate and optimize that program over the long term. Maximization of benefit, maximization of equity and minimization of harm are crucial endeavours and a good program will pursue these objectives continuously. No such optimization is possible unless the program can be evaluated quantitatively AND corrective action taken, as required. In order for an EHDI program to be evaluable, its operating procedures must be explicit, must be fully documented, and must be actually delivered in compliance with protocol. It is simply not possible to evaluate a program that is not grounded in defined protocols and consistent practices. EBP is the only rational and ethical approach to the determination of what merits inclusion in an obligatory protocol, versus what can be and should be left to the intuition and unique skills of the individual provider. Given the current, limited state of evidence about many areas of EHDI, and given the complexity of the overall EHDI process, there is ample room for both styles of activity. EBP should be seen not as a threat to freedom of practice, but as a tool to raise the quality of practice bar and to liberate genuine clinical skill and intuition to focus on those aspects of care that are truly complex and subtle.

References


Haynes, R. B. 2002. What kind of evidence is it that Evidence-Based Medicine advocates want health care providers and consumers to pay attention to? *BMC Health Services Research* 2:3.

Hunter, J. E., and Schmidt, F. L. 2004. *Methods of meta analysis: Correcting error and bias in research find-


---

**Appendix**

**Recommended Google search strings:**

OMNI Greenhalgh
MS1903 Spittlehouse
CEBM
AHRQ
AHRQ 02-S001