International Journal of Pediatric Otorhinolaryngology (2009) **73**, 21—42

ELSEVIER

International Journal of Pediatric Otorhinolaryngology

www.elsevier.com/locate/ijporl

Air-conduction estimated from tympanometry (ACET) 1: Relationship to measured hearing in OME☆

MRC Multi-centre Otitis Media Study Group*[,1]

Received 17 July 2008; received in revised form 5 September 2008; accepted 9 September 2008
Available online 28 October 2008

KEYWORDS
Otitis media with effusion (OME);
Tympanometry;
Hearing thresholds;
Prediction;
Surrogate estimates;
Imputation

Summary

Objective: In otitis media with effusion (OME), the accuracy of predicting air-conduction hearing-level (HLs) from tympanometry has generally been seen as too poor for use in clinical practice. Previous studies of the relationship have mostly concerned single ears, many using samples with predominantly mild cases of OM and weak statistical approaches. A better understanding of the interrelations between these tests might improve efficiency in testing and decision-making for individuals.
Method: Binaural average HL was adopted as the measure to be predicted most relevant to auditory disability. Multiple regression from modified Jerger tympanogram categories B, C2, C1 and A tympanogram types on 3085 children aged 3(1/4)—6(3/4) years gave formulae which we tested for replication, stability and generalization across distributions differing in severity.
Results: Age-adjusted formulae explained up to 49% of the variance in binaural HL (i.e. a multiple correlation of 0.70), and were robust across phase of disease. Best predictions were seen in a severe sample permitting exploitation of the strong conditioning effect by a B tympanogram in one ear upon the tympanometry/HL relationship in the other. This permits a trichotomous approximation (0, 1, or 2 B-tympanograms) to also perform well.
Conclusions: We name the HL prediction formula "ACET" — Air Conduction Estimated from Tympanometry. We do *not* recommend replacing audiometry with tympanometry, particularly not at first assessment. However, *where the diagnosis is, or likely from history to be, OME* (*even if fluid is absent on test day*), the informativeness of further air-conduction audiometry on the same or later occasion may not always be worth the further effort or cost. It is therefore clinically useful to have a dB measure,

Abbreviations: OME, otitis media with effusion; HL, hearing level; ABG, air-bone gap; MEP, middle ear pressure; MEC, middle ear compliance.

☆ Guarantor: Professor Mark Haggard, Writing and analysis team, M.P. Haggard, J.M. Higson, H. Spencer.

* Correspondence to: Mark Haggard, Department of Experimental Psychology, Downing St, Cambridge CB2 3EB, UK.
Tel.: +44 1223 363609; fax: +44 1223 335654.

[1] Full list of authors and contributors in Appendix 2.

0165-5876/$ — see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijporl.2008.09.014

from an evidence-based formula justifying a principled estimate. Non-clinical uses include imputation when research data are missing, and non-intensive applications where audiometry is impracticable, e.g. field clinics and large scale or longitudinal research. A companion paper shows how the part of the air-conduction HL variance that is *not* explicable by ACET, also offers a surrogate, but for bone-conduction HL (BC), where BC testing may be problematic, as in the very young. This surrogate can also define cases needing true BC testing.

1. Introduction

In detecting or confirming otitis media, both acute (AOM) and with effusion (OME), tympanometry has long played an important role, by conveying presence/absence of middle ear fluid [1]. It contributes objectivity and validity to diagnosis and monitoring of middle ear status as the condition evolves. However, presence of ear fluid is very common and can be transitory, so the emphasis of assessment in OME must be on economically establishing persistence over time, degree of impact and any special susceptibilities or comorbidities, as well as degree of hearing loss (e.g. [2]). There are not yet standard accepted measures for impact, so the degraded auditory sensitivity (i.e. raised hearing threshold level, HL) is usually taken as a surrogate measure for it. In other words, hearing loss is assumed to be the main if not unique path to developmental impact. This is despite HLs not having been extensively validated as a surrogate impact measure in OME, and the known fluctuation in HL making it likely that prediction from HL on a single occasion will be poor. Nevertheless, HL has the advantage of a clearly defined scale, with standardized and widely available equipment and procedures, and it is an accepted correlate of many attributes of clinical concern. Thus, along with tympanometry, it constitutes the chief starting point for assessment in OME. Having two such standard tests requires that the relationship between them be understood, so that the respects both in which they are similar and in which they differ can best be exploited. The relation of tympanometry to HL in OME has not been approached in this context of optimum use of information.

Broad categories (e.g. the modified Jerger system — [3,4]— see Table 1) are the most common summary of the range of tympanometric states. The normal to mild A-C2 categories, have a definable admittance peak but are distinguished by the raw middle ear pressure (MEP) at which it occurs. In this region, Lildholdt et al. [5] demonstrated a near-linear relationship of pressure with hearing level, giving a monaural 0.54 correlation (29% variance explained). The percent variance explained for the binaural mean HL (quoted at 93%) was inflated by inappropriately including repeated measures, a common error of method. Unfortunately, Lildholdt et al.'s particular sample and class of formula were of limited clinical relevance, as pressure is only clearly defined for non-flat (high admittance) traces that are at test time mostly physiologically near-normal. Margolis et al. [6], found correlations with HL of 0.45 and 0.53 for static admittance and tympanogram width, respectively, essentially confirming Lildholdt et al., but again on mostly non-flat traces (94 ears). Although they advocated wider use of tympanometry, again under 30% of the variance was explained.

Sichel et al. [7] attempted to define MEP on type B traces, using planimetry (deriving the area-under-curve, AUC) to handle the differing shapes of curve that sometimes occur. They expected to improve the discrimination among type B traces that are not completely flat, despite peaks at low compliance levels being somewhat poorly defined [8]. Sichel

Table 1 Definitions within modified Jerger classification[a] of tympanograms as described by Zielhuis et al. [4].

Tympanogram type	MRC (ml)	MEP (da Pa)
Type A	≥0.2	Above −100
Type C1	≥0.2	Below or equal to −100 and above −200
Type C2	≥0.2	Below or equal to −200 and above −400
Type B	<0.2	Below or equal to −400

MEC, Middle ear compliance (or admittance); MEP, middle ear pressure. Often clinicians combine categories A and C1 to a single, essentially normal category, and some of the analyses reported here directly justify this practice.

[a] Appropriate for use only on non-ventilated ears [45].

et al. found correlations between MEP (where definable) and air-bone gap (ABG), and between AUC and ABG, of magnitude 0.43 and 0.47, respectively (18% rising to 22% variance explained), i.e. a modest but worthwhile gain for planimetry. They acknowledged that even the AUC relationship to ABG is too poor for quantitative (continuous scale) prediction, but hinted at greater usefulness for the categorical distinction of ABG above/below 20 dB. However, their preliminary study provided only significance, not magnitude, information on the latter, another common methodological deficiency. None of these studies quantified the error of prediction. From data-points in Sichel et al.'s Fig. 3, we have calculated an odds-ratio (OR) for what appears to be the optimum planimetric criterion of 10 cm^2. Giving benefit of doubt to 3 readings exactly on this cut-off, of which two have ABG $>$ 20 dB, the OR is in the region of 10. This is a fairly strong relationship, corresponding to a high sensitivity of approximately 95% but only modest specificity of about 64%. The strength of relationship may be limited by the poorer reliability of ABG despite its closer physiological link to tympanometry: the differencing of the air- and bone-conduction measures results in addition of the error variances. Furthermore, within the framework of impact on children and their need for treatment, it is the total HL not the ABG that is relevant to auditory disability [9]. Hence, other things being equal, there is greater justification for surgical intervention in the OME child also having a non-conductive loss [2,10].

Appropriate use of the tympanometry-HL relationship requires an appreciation of the caseload to which it may be applied. In ears with OME the effective range of HL is about 10—45 dB HL [11—13]. Other things equal, the strongest relationship must be found in samples having about half their cases in the upper half of this range. Regrettably, the basic physiological relationship between HL and tympanometric measures is only tight and near-linear in the mild or resolving cases whose tympanograms are not B, and whose HLs are largely below 20—25 dB [5]. That part of the range showing a neat psycho-physical relationship is not the clinically important region. Any sensible and general approach to the relationship therefore has at very least to define, and preferably to justify, the severity distribution of the sample(s) used. Similarly, any procedure offered for application has to state whether it applies to sub-clinical cases or to the entire range, or whether it includes part-solutions for each.

In the absence of comprehensive studies to date, the usefulness of the tympanometry/HL relationship has not been properly determined. Brookhouser [8] thought "the relationship between immittance results and auditory acuity is not consistent enough to permit *accurate* prediction of hearing thresholds by immittance testing" (italics ours). In contrast, Sichel et al. thought that the: "...tympanogram can supply clinicians with ...quantitative information...useful for the objective evaluation of hearing in very young children with middle ear effusion". Work in this area to date has not addressed the potential clinical usefulness of approximate but optimized predictors, and the aspects of method (large sample size, serious cases meriting consideration for treatment, etc.) necessary to tackle the issue properly have not attracted critical comment.

The HL/tympanometry relationship has in the past been largely viewed on a unilateral basis (e.g. [7]), as if the physical determinants and isolated severe ear pathology were of chief concern. Some clinical decisions are made on unilateral evidence of OME [14,15] but this is incompatible with contemporary thinking on what intervention hopes to achieve. The vast majority of management decisions in OME are not about pathology in a single ear, but are rightly made on a binaural basis, through concern about developmental impact and auditory disability in the whole child. Thus, it is the disease effect on the whole child that determines need for management [16,17]. Binaural hearing levels better predict broad outcomes such as speech-in-noise performance or perception of benefit from surgery than single ear HLs do [18,19]. When averaged, the two ears' data also provide a more reliable summary than a single ear measure does, raising the ceiling for predictability. The general failure to appreciate that the most clinically relevant relationship is the one between the *binaural* HL and the *binaural* tympanogram is a possible further contributor to unnecessary pessimism about useful prediction of HL from tympanometry and to the absence of evidence that tympanometry is widely and effectively used in this way. The present work therefore takes a binaural approach.

To avoid mis-attribution to us of claims that we are in fact not making, we state in advance: (1) The procedure only applies to the (many) cases where there is a history, suspicion or presumptive diagnosis of middle ear fluid; (2) HL prediction from tympanometry is for optimizing the information value from *further* audiometry or other tests relative to cost and effort, but ideally after previous acquisition of at least one set of air-conduction thresholds. We offer a substitute for when this ratio of extra information to extra cost or effort may be unfavourable, and do not advocate replacing pure-tone audiometry in any general way.

2. Method

2.1. Study design

Participants were children giving baseline data in the recruitment phase of the UK Trial of Alternative Regimens for Glue Ear Treatment (TARGET) evaluating the effectiveness of the placement of ventilation tubes, with or without adjuvant adenoidectomy [20—23]. Children were aged 3(1/4)—6(3/4) years on first visit. The protocol shaped the formation of the samples available, so is summarized here. To obtain sufficient children re-qualifying with 20 dB HL (better ear) after a 3-month watchful wait, a preliminary stage (''Visit 1'') screened over 4000 within-age referrals to hospital ENT from general practitioners or community pediatricians, with suspected OM(E) diagnosis. In the process, nearly all gave tympanometry data and most, but not all, supplied HL data.[2] Of the 1199 meeting the criterion of at least (B, C2) *plus* at least 20 dB HL in the better-hearing ear, over 500 met it again at Visit 2, 3 months later, i.e. after formal ''watchful waiting'', of whom 376 accepted randomization to treatment. For qualifying children, audiometry, tympanometry, otoscopy and a range of questionnaires were administered at each visit. The present paper uses only minimal demographics, pre-randomization audiometry and tympanometry data from the first two visits.

2.2. Audiometry and tympanometry

Details of audiometric and tympanometric instruments and their calibration are given elsewhere [21,45]. Audiometry was performed at 0.5, 1, 2 and 4 kHz using warble tones (±5 Hz of the center frequency) through headphones. To preclude any important cochlear hearing loss, unmasked bone conduction threshold at 1 kHz was also obtained with the bone vibrator placed on the mastoid process of the ear having the better average air-conduction threshold (averaged across 0.5, 1, 2 and 4 kHz); 2701 cases gave such data for the 2nd paper of this pair. Method A of the British Society of Audiology's Recommended Procedure [24] was employed, with the option (as with [25]) of using play techniques for younger children.

Table 2a Binaural average HL (0.5—4.0 kHz) by tympanogram combination at Visit 1.

Tympanogram combination	Mean HL (dB)	SD	*N*	%
A/C1 with A/C1	14.4	7.6	470	15.2
C2 with A/C1	16.3	5.7	210	6.8
C2 with C2	18.3	6.5	120	3.9
B with A/C1	20.0	7.2	243	7.9
B with C2	23.5	6.9	354	11.5
B with B	32.1	7.9	1688	54.7
Total	25.8	10.4	3085	100

Middle ear pressure (MEP), maximum compliance (MEC, or admittance), volume (Veq) and tympanogram gradient were recorded from a sweep of the ear canal pressure through the range +200 to −400 da Pa. The first two parameters were used to classify each tympanogram according to the Modified Jerger classification (Table 1).

2.3. Sample characteristics

Tables 2a and 2b gives a breakdown of mean HL by a binaural combination of the 3-level tympanogram classification, for the first two TARGET visits, 3 months apart. The 1199 Visit 2 cases are all those who had a (B, B) or (B, C2) tympanogram combination 3 months earlier (i.e. the last 2 rows in Table 1a), apart from 40 cases fast-tracking to treatment on grounds of severity, thereby bypassing Visit 2 and not being randomized, and 2 other exclusions (not related to ear status). Comparison of HL for these 1199 children showed 5.9 dB of resolution between visits. In contrast, comparison between the *total* samples available for the two visits shows them to be very similar. The proportion of (B, B) or (B, C2) tympanograms was similar across total samples for each visit: approximately two thirds (67%) at Visit 1 versus only slightly more (71%) at Visit 2. HLs within the *same tympanogram combinations* across visits were better at Visit 2 than

Table 2b Binaural average HL (0.5—4.0 kHz) by tympanogram combination at Visit 2.

Tympanogram combination	Mean HL (dB)	SD	*N*	%
A/C1 with A/C1	13.0	5.9	132	11.0
C2 with A/C1	14.9	4.5	75	6.3
C2 with C2	16.1	5.3	61	5.1
B with A/C1	19.6	7.5	80	6.7
B with C2	20.7	5.7	161	13.4
B with B	30.3	8.6	690	57.5
Total	24.7	10.2	1199	100

[2] Based on the good specificity of tympanometry as predictor of a hearing loss [13], the protocol only required acquisition of HL data given one of the binaural tympanogram combinations (B, B) or (B, C2). However, it was the usual practice of several centers to perform audiometry prior to tympanometry to avoid possible loss of cooperation with the slightly more invasive technique. Thus we have sufficient non-Bs with audiometry data for various analyses including the present ones.

at Visit 1, but only by 1—2 dB. The wider variance better spanning the full range of tympanogram categories and the larger *N* make Visit 1 the distribution of choice for the main ACET derivation. However, the overall severity does not differ greatly between cases at the two visits, because resolution offsets the pre-selection of the second visit sample which is thereby suitable for generalization testing.

2.4. Analysis steps

2.4.1. Step 1—Derivations of formulae

Multiple regression models, all predicting the binaural average HL, were used to compare several tympanometric coding schemes (listed as (i)—(vi) below). These schemes contrast differing degrees of simplicity in tympanometric data reduction, within otherwise similar regression models. We do not present detailed results for all codings, distributions and models run and compared, but distinguish six contrasting coding schemes for the binaural tympanometric data used to predict binaural HL. Every model had 1 extra degree of freedom (df) with the adjustment for age, beyond the df figure given for the tympanometry coding. The six coding schemes were:

(i) The four-category modified Jerger classification (A, C1, C2, B) for each ear, giving a model with 3 + 3 = 6 degrees of freedom (df).

(ii) A 3-category version with A and C1 combined to form a single group, versus C2 and B (4 df).

(iii) and (iv) As for (i) and (ii) but additionally including the interaction term between categories on the two ears, giving non-independence models with [6 + (3 × 3)] = 15, and [4 + (2 × 2)] = 8 df, respectively.

(v) A collapsed binaural tympanometric trichotomy, coding for number of B-tympanograms: 0, 1 and 2 (2 df).

(vi) Two further collapsed dichotomous codings: (B, B) versus the rest and (B, B) or (B, C2) versus the rest (1 df each).

The nearness of the 6 df in (i) to the 8 df in (iv) provides a type of control in later comparisons.

2.4.2. Step 2—Replicability, generalizability and applicability

With a higher number of variables significantly entering a model, there is usually enhanced prediction but always a statistical penalty of the number of degrees of freedom (lack of parsimony), at which this is purchased. For clinical application, it is necessary to balance the trade-offs between three considerations: the effort and overhead costs of being complex rather than parsimonious, the usual opportunity costs (and possibly even risks) from not being able to do something else with the time, and the clinical benefits from greater certainty.

We adopted two bases of evaluation and recommendation: (a) the predictive accuracy of the formulae in derivation on Visit 1 data, and (b) the replication, stability and generalization of the formulae to data other than those on which they were derived. We also tested generalization of formulae to a later disease phase (Visit 2) and application to a milder distribution, formed by discarding cases with bilateral B tympanograms.

2.5. Statistical concepts

The nature, advantages and limitations of multiple regression are briefly explained in Appendix 1. In this work, the existence of some association between HL and tympanometry is already known to be genuine, so null hypothesis testing is not appropriate. The overall multiple correlations giving values of percent variance explained simply quantify the relative strength of capture of the relationship against the background of some inevitable error in it, according to different ways of coding the tympanometric data. These correlations were mostly of extremely secure levels of significance, given the large sample. Thus, *p*-values are mostly not appropriate or useful here and hence only given sparingly. A more appropriate use of *p*-values comes in judging whether a possible extra model parameter is really worth including in the model (by whether the apparent increase to variance explained is secure): in other words how complex a model is adequately justified by the data. An example indicates what advantage of one model over another is worth talking about. Within the general range of model fits here explaining 41—49% of the variance, an increment of 0.5% (one half of 1%) difference in the explained variance for the sample size used is worth discussing, interpreting or using to improve a formula. As benchmark, such a 0.5% increment in a sample size of over 3000 corresponds to a *p*-value of 0.025, so is worthy of scientific discussion. However, larger increments, of at least 1% (or even 2%, depending upon the additional complexity of procedure or coding) would be needed for worthwhile gain in practice. Most values in tables are to 3 decimal places, to avoid small coefficients being rounded to zero.

Table 3 Illustration of contralateral conditioning: mean left ear hearing level at Visit 1, with standard deviation and N for each combination of tympanogram types at Visit 1.

Tympanogram type		Mean left ear HL (dB)	SD (dB)	*N*
Right ear	Left ear			
A	A	13.10	7.77	273
	C1	16.66	7.31	65
	C2	16.62	4.43	41
	B	*25.13*	10.45	55
	All groups	15.49	8.78	434
C1	A	13.44	7.08	53
	C1	16.60	8.80	79
	C2	18.72	6.46	70
	B	25.70	9.92	55
	All groups	18.48	9.13	257
C2	A	13.79	9.04	34
	C1	14.98	5.15	65
	C2	18.24	7.19	120
	B	26.75	9.05	165
	All groups	20.95	9.49	384
B	A	14.84	8.99	58
	C1	16.48	6.75	75
	C2	19.45	7.64	189
	B	*32.02*	9.14	1688
	All groups	29.77	10.34	2010
All groups	A	13.44	7.97	418
	C1	16.21	7.20	284
	C2	18.71	7.09	420
	B	31.21	9.41	1963
	All groups	25.72	11.45	3085

The contralateral conditioning is seen most readily by contrasting the two italicized and underlined entries and then noting that the corresponding cells in intervening fields show it is the contralateral B rather than any gradation A, C1, C2 that makes the difference.

3. Results

3.1. Illustration of the basis of improved HL prediction with B tympanograms

Although the overall emphasis in this work is binaural, the simplest illustration of the influence of the "other" ear, hence the need to consider influences from *both* ears' tympanogram states, can be seen in data where the variable influenced comes from a single ear. Table 3 gives left ear hearing data, stratified by tympanogram status on each ear, for the 3085 cases at Visit 1. The effect of the tympanogram state of the "other" ear is shown via the 6.9 dB range of means (32.0—25.1 dB) for left ear hearing level for a fixed category B tympanogram on the left, as conditioned by the tympanogram on the right ear. The example chosen is the least favorable: corresponding data for the right ear HL, and for each ear at Visit 2 show even stronger contralateral conditioning (full data not shown; Visit 1: right ear range 10.0 dB, and for visit 2: 8.4 dB, left and 10.6 dB, right).

This 7—10 dB of conditioning from a contralateral B is large and offers some escape from the lack of discriminable relationship between HL and B tympanograms on a single ear. It explains why a simple additive linear model for the two ears' states cannot be optimal when predicting HL. It justifies fitting, and at the margins preferring, a statistical interaction term in subsequent binaural formulae. The R → L and the L → R conditioning processes for the monaural mean HLs both contribute to the high HLs for bilateral Bs. Indeed, the corresponding binaural table to Table 3, but combining effects of conditioning with averaging (see Tables 2a and 2b) gives a difference of 12 dB between (B, B) and (B, A) tympanogram combinations, when averaged for the two ears. Hence in caseloads where B tympanograms are at all frequent, a supra-additive model (i.e. a conditioning, interactive or synergistic—these words are all equivalent here) must perform better than a merely additive one.

3.2. Binaural hearing level from the 4-category Jerger classification: non-interaction model

The non-interaction (additive linear) model to predict binaural average hearing level from the 4-category tympanogram codings on each ear and age explained 48.4% of the variance on Visit 1 data (Table 4). This corresponds to an equivalent $r = 0.696$; [95% CI: 0.677—0.713]. For the chief models discussed, confidence intervals for equivalent *r* (i.e. square root of percent variance explained) are given in Fig. 1. Other covariates to adjust for per-

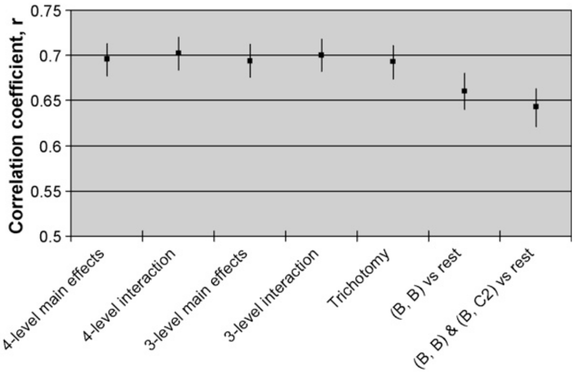

Fig. 1 Correlation coefficients (square root of percent variance explained) and 95% confidence intervals for chief derivation models predicting HL from tympanometric data.

Table 4 Regression model (without interaction term) for predicting binaural average hearing level at Visit 1 using the 4-category Jerger tympanogram variable on each ear ($N = 3085$). This model explains 48.4% of the variance.

Parameter	*B*	SE	*t*	*p*-Value
Intercept	35.685	0.801	44.561	<0.001
Type A (L)	−10.280	0.534	−19.269	<0.001
Type C1 (L)	−8.965	0.526	−17.030	<0.001
Type C2 (L)	−7.399	0.429	−17.233	<0.001
Type A (R)	−8.489	0.525	−16.156	<0.001
Type C1 (R)	−7.093	0.548	−12.954	<0.001
Type C2 (R)	−7.001	0.442	−15.834	<0.001
Age (months)	−0.065	0.013	−5.026	<0.001

The reference category is type B, so the intercept represents the predicted HL when both tympanograms are type B plus the effect of adjustment for age. Each parameter estimate gives a value to be subtracted from the intercept, giving lower ACETscores required for milder tympanogram categories to give a good fit to the actual HL data.

formance aspects of audiometry (e.g. gender, socioeconomic status, audiologist's rating of child concentration) were examined, but these were not significant once age was already in the model. The left ear contributes slightly more strongly than the right here, thus compressing the contribution from tympanogram categories on the right ear, once the left is taken into account. (Such apparent left-right asymmetries should not be over-interpreted; the optimizing algorithm in multiple regression leads stronger variables to suppress weaker ones when the independent variables are highly intercorrelated, thus magnifying such ear-asymmetries, which are usually slight.)

3.3. Binaural hearing level from tympanometry: interaction models

The merely additive model in Table 4 does not capture the strong contralateral conditioning seen in Table 3, requiring an additional (binaural interaction) term. With 4-level tympanometric coding, variance explained by the model including main-effects and interaction (not shown) rose to 49.3%. The 4-level model is not favored because the total df required to express the interaction model with two ear main effects each having 4 levels rises to 16 [1 + 3 + 3 + (3 × 3)], and this may require caution about instabilities (i.e. overfitting). For the interaction term itself, *p*-values were highly significant for all but one of the 9 coefficients that compose the pattern of interaction, not suggesting that the interaction itself becomes unreliable. In both the main-effects and the interaction terms, the differences between A and C1 categories shrank to about 1 dB in this model, i.e. only about 1 standard error, and so not statistically reliable. This fact, plus a corresponding narrow spacing of component estimates within the interaction term involving A and C1, justifies our collapsing A and C1 categories, as often

Table 5 Interaction model for binaural hearing at Visit 1 from 3-category tympanograms, age and the 4 component terms of the left*right interaction.

Parameter	*B*	SE	*t*	*p*-Value
Intercept	36.094	0.797	45.282	<0.001
Type A/C1 (R)	−11.188	0.733	−15.260	<0.001
Type C2 (R)	−8.465	0.608	−13.921	<0.001
Type A/C1 (L)	−12.854	0.671	−19.158	<0.001
Type C2 (L)	−8.399	0.572	−14.674	<0.001
Type A/C1 (R)* Type A/C1 (L)	6.471	1.036	6.247	<0.001
Type A/C1 (R)* Type C2 (L)	4.300	1.154	3.725	<0.001
Type C2 (R)* Type A/C1 (L)	5.230	1.161	4.506	<0.001
Type C2 (R)* Type C2 (L)	3.212	1.061	3.026	0.002
Age (months)	−0.067	0.013	−5.175	<0.001

Type B is reference value for each ear. Adjusted Rsq = 49.0%. This preferred model has an interaction term with 4 degrees of freedom and is expressed by the set of coefficients linked with asterisk. The formula provides an ACET value for clinical use, using the intercept and *B* coefficients from the table that correspond to the combination of tympanograms and age, as given in the following examples. Example 1: For a child aged 5 years (60 months) with a C1 tympanogram on the right and C2 tympanogram on the left: ACET = 36.094 − 11.188 − 8.399 + 4.3 − 0.067 × 60 = 16.787 dB or 17 dB to the nearest dB. Example 2: For a child aged 5 years 6 months (66 months) with a C2 tympanogram on the right and a type B on the left: ACET = 36.094 − 8.465− 0 + 0 − 0.067 × 66 = 23.207 dB or 23 dB to the nearest dB.

done clinically, to a 3-level coding for each ear with 9 df (2 + 2 + 1 + [2 × 2]). This demonstrates one justification for considering C1 tympanograms as within the normal range.

Table 5 gives the parameter estimates for the consequent interaction model using the 3-level tympanogram categories, plus examples of its use in calculating ACET values. At 49.0%, the variance explained here is close to that from the interaction model using 4 category levels, but the parsimony of 8 + 1 = 9 df makes it more robust. The corresponding 3-level non-interaction model (shown later—Table 8i) gave 48.2%. This 0.8% gain for including an interaction term expresses the strength of the interaction effect within the particular model (all four interaction components highly significant here, three of them $p < 0.001$). The general similarity of all these net variance-explained figures, at around 49%, makes empirical predictive accuracy alone an inadequate basis for clear choice of optimum model (see later). However, the agreement confirms that this explicability of nearly 50% of the variance is generalizable, not an isolated and opportunistically selected finding.

3.4. A trichotomous model for manual application

The preceding models are preferred as making most precise use of the data, and they are explainable in physiological terms, but require decimal arithmetic, making them impractical for real-time application in clinics. We therefore sought a simpler model that captured the essence of the contralateral conditioning, but which could be applied by a simple rule or look-up table, not requiring decimal arithmetic. We scored tympanogram types according to the trichotomy: *binaural B, unilateral B, no Bs at all*, requiring only 2 df (compare 8 df for the 3-category reduction of the modified Jerger scoring used in the ACET derivation of Table 5). This binaural trichotomy explains a surprisingly high percentage of the variance in binaural HL (48.0%, Table 6). The contralateral conditioning (i.e. binaural interaction) is present but implicit rather than explicit in this coding of the variables. This is because the poorer HL for binaural B-type compared with unilateral B-type tympanograms, combines the general effect of having a B on the other ear (cf additive term for second ear) with its conditioning effect upon the first B (interaction term). The higher predicted HL for (B, B) (e.g. 32 dB for a 5 years old) versus the other two codings (22 dB for unilateral B and 16 dB for non-B combinations) is due to the strongest contralateral conditioning occurring from a type B to a type B. The respective separations between the category estimates, i.e. 10 dB (the extra effect of a double-B compared to single-B) and 6 dB (effect of a single-B compared to others without one), summarize the trends seen in the more complex models, but the formula has the additional virtue of easy memorability for practice. *Simply, for 5 years old, ACET = 16 dB, 22 dB and 32 dB for the 3 tympanometric combinations having 0, 1 or 2 type Bs. These values change by + 1 dB for 3(1/2) years old and −1 dB for 7 years old.*

3.5. Advantages over cruder models

In terms of variance explained, the simple binaural trichotomy comes an impressive close second to the more comprehensive models with 3 and 4 categories per ear. To understand the crucial features in this success demands comparison with some cruder models that fail. A simpler dichotomous collapsing of the binaural tympanogram combinations would have

Table 6 Prediction model for binaural hearing at Visit 1 from trichotomous tympanogram coding (bilateral B, unilateral B and non-B) with age.

Parameter	*B*	SE	*t*	*p*-Value
Intercept	35.853	0.804	44.590	<0.001
Non-B tympanogram combinations	−16.489	0.324	−50.964	<0.001
Unilateral B tympanogram combinations	−9.929	0.359	−27.677	<0.001
Age (months)	−0.063	0.013	−4.823	<0.001

Binaural B-tympanograms are the reference category. $N = 3085$; adjusted Rsq = 48.0%. This alternative simpler model gives an ACET value more readily calculated by hand while offering only slightly reduced precision. We use the same examples as in Table 4, and show predicted values within 1 dB of those predicted by the 3-category interaction model formula. In general, by placing cases into one of the three clusters in Fig. 1c, the trichotomous model will provide values to within 3 dB of the interaction model and to within 5 dB of the true HL. Example 1: For a child aged 5 years (60 months) with a C1 tympanogram on the right and C2 tympanogram on the left: ACET = 35.853 − 16.489− 0.063 × 60 ≈ 36− 16− 4 (calculating to nearest dB) = 16 dB. Example 2: For a child aged 5 years 6 months (66 months) with a C2 tympanogram on the right and a type B on the left: ACET = 35.853 − 9.929 − 0.063 × 66 ≈ 36 − 10− 4 (calculating to nearest dB) = 22 dB. These coefficients are also used to give the easily remembered 32, 22 and 16 dB ACET values for children with 2, 1 or 0 B tympanograms. These values are increased by 1 dB for children 3.5 years and decreased by 1 dB for 7 years old.

only 1 df; this can be done in one of two ways: (B, B) tympanograms versus the rest, or (B, B) plus (B, C2) versus the rest. These two dichotomies gave significantly poorer results, with 43.6% and 41.3% of the variance explained, respectively. The slightly better one of these two cruder and less predictive dichotomies, (B, B) versus the rest, aligns the distinction it makes with occurrence of contralateral conditioning. Unilateral prediction gave even poorer results: neither left nor right ear tympanometry alone (3-levels—data not shown) reached 41% variance explained, further supporting the necessity of a binaural approach.

3.6. Graphic distributions of predicted values for the 3 main models

To give a more intuitive understanding of the differences among the models discussed, Fig. 2 shows the residuals (deviations of true value from predicted value) as a function of predicted value from fitting three models: 3-category main effects only (2a), 3-category interaction (2b) and the *a priori* trichotomy (2c). The ideal would be a continuous horizontal but narrow band. However, the originating tympanometric information is categorical, so the predicted HL values can never be truly continuous. The fairly distinct clusters center on the mean HLs (dB) for each tympanogram combination used in the models, e.g. for Fig. 2c, three of these. The homogeneous cluster width in 2c is due to the same age-range adjustment [(12 × 0.063) dB per year × 3.5 years cohort width = 2.65 dB] being applied to each category; this age-related spread differs only minimally across models. In all the models, the most severe cases separate out to a cluster at approximately 30—33 dB, corresponding to binaural B cases. The non-interaction model (2a) also has two lower clusters, but these are wider (12—18 and 20—26 dB) compared with those of 2c (14—17 and 20—23 dB). This is because each value in the 3-category main-effects model is predicted from separately entered tympanogram types and age; so (A, B) gives a slightly different prediction from (C2, B) for the same age in this model. Thus for models *not* using an interaction term, the extra information in the 3- or 4-level coding (relative to simpler coding) pays off mostly in the better resolution among milder cases of lesser clinical concern. The interaction model (2b) illustrates two consequences of "spending" further degrees of freedom: (i) the category estimates in the lower clusters are fused towards becoming a single continuous cluster, the small conditioning effects seen here combining with age-adjustment to resolve real individual differences. However, (ii) this combined broader cluster is further separated (less continuity overall) from the binaural B by the contralateral conditioning phenomenon applying mostly to the double-B combinations. Overall percentage variance explained is the single most valid and generic metric for predictive accuracy; however, comparisons of models using it may need to consider the resolution achieved in different parts of the distribution. The weighting of the resolution, achieved in each part by the actual numbers there, is one reason why distribution influences correlation magnitude.

3.7. Replicability, stability and generalizability tests

3.7.1. Replicability of formula between equivalent individuals

We tested replicability between individuals using only the 3-category main-effects and interaction ACET formulae. For this, split-half sampling is

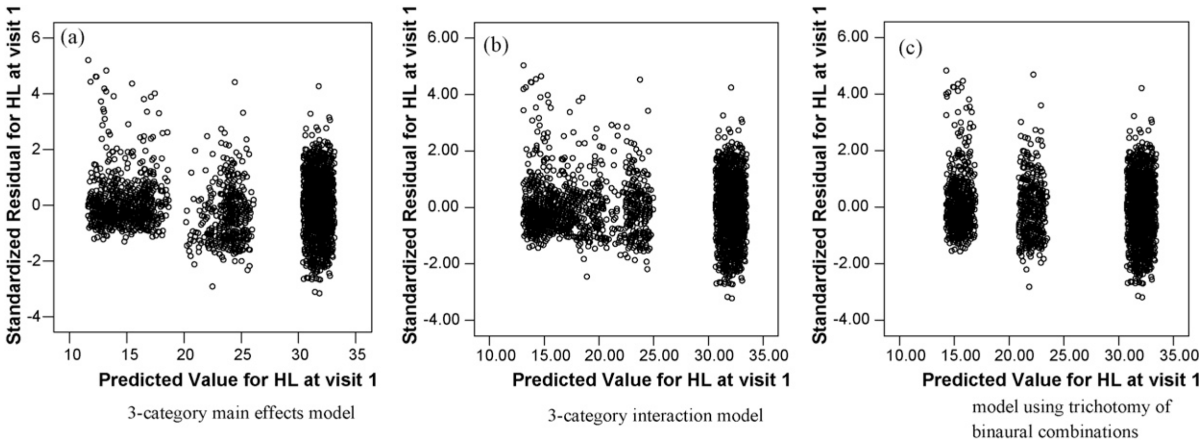

Fig. 2 Plot of residuals against predicted value for 3 models predicting HL at Visit 1 from tympanometry and age: (a) 3-category main-effects model, (b) 3-category interaction model and (c) model using crude trichotomy of binaural combinations.

preferable to test—re-test, which can be confounded by time-order effects or specific time effects such as development. The two 3-category models were first re-derived separately on two randomly generated (without replacement) subsets of Visit 1 data, the two halves of the 1827 cases not meeting TARGET RCT entry criterion at Visit 1 (N = 914—set i; N = 913—set ii). New ACET formulae were derived on one half of the data and then applied to the other half. For the main-effect and interaction models, respectively, the variances explained by the re-derivation versions were 45.7% and 46.0% (set i) and 42.0% and 42.9% (set ii), only slightly lower (as expected) than those derived on the full data. We denote an ACET value calculated on the second half of the data by a main-effects model formula derived on the first half as ACET(M)$_{\mathrm{i,ii}}$. Similarly the other possible ACETs are ACET(M)$_{\mathrm{ii,i}}$, ACET(I)$_{\mathrm{i,ii}}$ and ACET(I)$_{\mathrm{ii,i}}$. These predicted values were then subjected to correlation with true HL on the application sample (HL$_{\mathrm{i}}$ or HL$_{\mathrm{ii}}$ depending on which half is taken as application), giving Pearson correlations of: 0.650 (ACET(M)$_{\mathrm{i,ii}}$, HL$_{\mathrm{ii}}$,), 0.678 (ACET(M)$_{\mathrm{ii,i}}$, HL$_{\mathrm{i}}$), 0.657 (ACET(I)$_{\mathrm{i,ii}}$, HL$_{\mathrm{ii}}$) and 0.679 (ACET(I)$_{\mathrm{ii,i}}$, HL$_{\mathrm{i}}$). The average r of these four is 0.666, for which the corresponding R^2 is 0.444. The square root of the R^2-values from the respective models on their own training sets, the correlations would have been (0.682, 0.678, 0.660, 0.650; average 0.668). (Here the R^2 values used were not adjusted for df, so as to retain comparability with Pearson's r.) The drop in average R therefore when a formula is run on a comparable but different subset of cases is trivial, at 0.002, or one fifth of one per cent. This is only slightly greater than the decimal rounding error when working to 3 decimal places. This very good replication shows the virtue of a very large N and unbiased (random) subset generation, contrasting with the poor replicability of multivariable results on small samples.

3.7.2. Longitudinal stability

In a changeable condition like OME, correlation between measures as far apart as 3 months gives not pure test—retest reliability, but an estimate of the "longitudinal stability" of a condition, or persistence, if all cases start with the condition. Persistence is raised in OME caseloads like the present one that are already somewhat selected for persistence of their OME in the past [26]. The serial correlations must therefore be moderately positive, although far from perfect, and limited by the narrowed distribution at Visit 1 qualification for cases going on and able to give data also at Visit 2. The issue is: how well do the various tympanometric codings reveal the moderate stability expected? For the 1171 cases with complete HL and tympanogram data at both visits, the serial correlations are shown in Table 7, comparing HL, ACET 3-level interaction model, and two crude dichotomies. If anything, the ACET formula reflects longitudinal stability more strongly than HL itself does (perhaps due to less room for performance effects or for changes in them, due to maturation or prior experience with tympanometry compared to HL). Crude dichotomies are inferior here, as in the other comparisons.

3.7.3. Generalization of formulae to milder distributions

We performed a very large number of analyses on different sub-distributions shaped from our master Visit 1 distribution, but briefly illustrate only the two most informative. These confirm the caution that in the extreme, samples differing radically in composition or distribution from ours at Visit 1 could

Table 7 Longitudinal stability as measured by serial correlations between corresponding measures at Visits 1 and 2 (3 months later).

Correlation type	Hearing level	ACET		Dichotomy	
		Visit 1 3-level interaction formula applied to data at both visits	Visit 2 3-level main effects only formula applied to data at both visits	B + B versus all other combinations	B + B and B + C2 versus all other combinations
Pearson r	0.463	0.477	0.472	0.412	0.363
Spearman rho	0.455	0.556	0.553	0.412	0.363

Correlations are on 1171 cases with complete HL and tympanometric data on 2 occasions. They are of moderate magnitude, showing some longitudinal stability (all are highly significant on this N). The values in the two columns giving the serial correlations for two different ACET formulae are very close because their predicted values correlate highly (e.g. 0.99 at Visit 2); the small correlation difference of 0.003—0.005 is due to the identity of the derivation used, not any slight optimality of the local derivation for Visit 2. In this it is not possible to dissociate the subset of data from the class of model, each being the one preferred, but with such a small difference it is not necessary to do so. For the non-scaled dichotomies, the two types of correlation are necessarily the same. The higher value for the Spearman rank correlation with ACET suggests a nonlinear relationship between the ACET values on the two occasions, but still a quasi-continuous scale that the approach preserves.

Table 8i ACET derivations contrasting total sample with a milder sub-sample in which cases with a (B, B) tympanogram combination have been removed. Unselected sample; 3-level main-effects only model ($N = 3085$); Rsq 0.482.

Parameter	*B*	SE	*t*	*p*-Value
Intercept	35.805	0.802	44.668	<0.001
Type A/C1 (L)	−9.800	0.428	−22.886	<0.001
Type C2 (L)	−7.308	0.429	−17.028	<0.001
Type A/C1 (R)	−8.044	0.430	−18.717	<0.001
Type C2 (R)	−6.936	0.442	−15.678	<0.001
Age (months)	−0.067	0.013	−5.163	<0.001

require local variants of the ACET principle to be rederived if the sample is large, or an appropriate milder variant formula to be obtained from ourselves. As shown earlier (in *Sample characteristics* and Tables 2a and 2b), the similar distributions of severity for the two qualifying visits with similar proportions of cases within each tympanogram combination, result from opposed effects of selection and resolution so some differences in covariance structure would be expected, producing some loss of generalization. For the 3-category interaction model there was a 4% drop from percent variance (as explained at Visit 1) when the same Visit 1 formula was imposed on Visit 2 data. This is mainly due to differences between the samples, even though those differences appear small because in opposing directions. Indeed, comparing the two formulae both applied to the Visit 2 data gave only a 0.5% difference between the two models (45.0% and 44.5% for the Visit 1 and Visit 2 formulae, respectively). Hence there is a lowered ceiling at Visit 2 due to its particular covariance structure but the two formula derivations perform very similarly when applied to the same data. Thus there is minimal loss of generalizability between samples despite the differences in covariance structure.

Bilateral B tympanograms can be considered an extreme group in a clinical decision, e.g. always justifying a further test, whereas milder cases might not. We therefore also tested generalization to a subsample of cases ($N = 1397$) containing no bilateral B tympanograms (only unilateral Bs). In Fig. 2, continuity (for the 3-level main-effects and interaction models—2a and 2b) is particularly manifest at lower HL, and so we expected this sub-sample to come closer to achieving continuity. When bilateral B-tympanogram cases are excluded, the interaction no longer offers a justifiable addition to the 3-level model (see Table 3). Table 8i shows the 3-level *main-effects only* models for the whole sample and for this reduced sample, respectively. Fig. 3 shows the largely continuous distribution achieved when the discriminating power of the model is focused in the milder range. On this distribution deprived of double-Bs, the trichotomous model collapses to a dichotomy, 0 versus 1 B. The corresponding data are given in Table 8iii. The truncation reduces the total variance explicable by each formula from 48.2% (Table 8i) for the full dataset, down to 21.3% (Table 8ii) and 15.6% (Table 8iii) on the reduced dataset for the 3-level model and the remaining dichotomy, respectively. The drop in variance explained (and in the beta coefficients) occurs partly because range is compressed relative to error, but also because there remains no important degree of contralateral conditioning (through absence of the (B, B) category) for either an interaction term or the bilateral B category in the trichotomy to take advantage of.

4. General discussion

4.1. Effective accuracy of prediction

Despite the acceptable percentage of variance explained, multivariable prediction for the individual

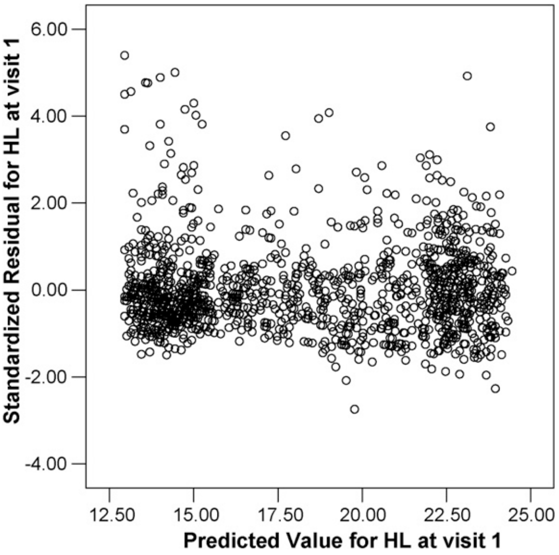

Fig. 3 Plots of residuals against predicted value for the main-effects model predicting HL at Visit 1 on a sample excluding serious cases ($N = 1397$).

Table 8ii ACET derivations contrasting total sample with a milder sub-sample in which cases with a (B, B) tympanogram combination have been removed. All double-B cases removed ($N = 1397$); 3-level main-effects only model; Rsq 0.213; B tympanograms remaining in the unilateral cases provide the reference category for the coefficients.

Parameter	*B*	SE	*t*	*p*-Value
Intercept	31.042	1.272	24.404	<0.001
Type A/C1 (L)	−7.224	0.539	−13.408	<0.001
Type C2 (L)	−4.249	0.588	−7.230	<0.001
Type A/C1 (R)	−5.830	0.507	−11.500	<0.001
Type C2 (R)	−4.102	0.574	−7.151	<0.001
Age (months)	−0.062	0.018	−3.468	0.001

does not look precise by the standards of careful measurement. However, the graphics show that for the most severe tympanogram combination, the effective range of possible HLs has been considerably narrowed from that usually associated with B tympanograms. The error of prediction for a total large sample is concerned with magnitude not confidence intervals on estimates from sub-samples, so is not proportional to sample size; it is given as a standard deviation (SD) not a standard error (SE). Put more precisely, the distribution of prior expectations for the present sample (Visit 1) has a mean of 25.8 dB and standard deviation (SD) of 10.43 dB, specifying the uncertainty before any actual measurement is tagged to a particular individual. Acquiring only the individual's tympanometric information narrows this to a mean estimate for the particular tympanogram category combination with an SD of 7.45 dB for the 3-level interaction model and 7.52 for the trichotomy. Whilst this is still considerably greater than the test/re-test error on particular measurements, the reduction of the uncertainty by a quarter has to be called "modest but worthwhile". Expressed in this continuous metric, the conclusion does not quite do justice to the fact that ACET discriminates particularly well those below from those above 25 dB, the region where a clinical decision is required. The imperfect precision in the ACET procedure reflects the impossibility of escape from the fundamental lack of HL discrimination among B tympanograms, at least in the absence of more sophisticated measures and a more sophisticated model of the middle-ear mechanics and hydrodynamics. However, it offers (a) some improvement on schemes failing to incorporate contralateral conditioning; (b) relevance to clinically important cases; (c) specified performance and accuracy data; (d) a rough but unbiased and justified way of narrowing the prior uncertainty for individuals. This permits interpretation and equivalencing in the familiar dB HL scale and (in the absence of HL) potentially permits showing relationships that HL should show.

Our 3-level interaction model explained about half (49%) of variance in true HL (i.e. correlation of ACET with true HL at 0.7 (95% CI: 0.682—0.718)). This justifies claiming, for a sample with the relevant presumptive diagnosis, a moderate rather than a low association between HL and tympanometry. In turn this encourages more creative thinking about applications. Previous work on tympanometry as surrogate for HL (e.g. [5,7]) was not comprehensive enough to secure wide application, possibly due to shortcomings in the conceptualization, and insufficient severity of the cases giving the data. Our data align with and include replications of others' findings, so our more encouraging message is not based on conflicts between data of the same type, but on more appropriate definition of the appropriate data and clearer conceptualization. The next-but-one section summarizes similarities for contralateral conditioning. Where using continuous measures (middle ear pressure and compliance) and addressing only the same clinically unimportant mild range, we obtained similar accuracy of prediction to Lildholdt et al. We also found that the apparently good relationship there is not due to intrinsic continuity of the measures used. As suggested by the continuity achieved in Fig. 3 (but not shown), the

Table 8iii ACET derivations contrasting total sample with a milder sub-sample in which cases with a (B, B) tympanogram combination have been removed. All double-B cases removed ($N = 1397$); tympanometry dichotomized at (B, C2) versus the rest: Rsq 0.156.

Parameter	*B*	SE	*t*	*p*-Value
Intercept	27.870	1.219	22.863	<0.001
Tymp combinations other than (B, C2)	−7.035	0.445	−15.808	<0.001
Age (months)	−0.070	0.019	−3.780	<0.001

categorical equivalent (A, C1, C2), performed approximately as well as the best transformation and combination of continuous pressure and compliance measures.

Comparisons of detailed results between models and differently shaped distributions suggest that the satisfactory relationship obtained here is due to a conjunction of three factors: (1) the more reliable and valid data available from two ears and predicting the more appropriate and performance-relevant binaural HL; (2) healthcare setting—application to a distribution appropriate for secondary care with many double-B tympanogram combinations, hence many HLs over 25 dB; and (3) an innovation—introduction of an interaction term reflecting strong contralateral conditioning, which increases resolution around the pivot of distributions that contain severe cases. Conditions (1) and (2) should not be seen as limitations to application, as these are precisely the circumstances of most clinical relevance. For example, where the context is disability, it is already widely appreciated that binaural hearing (1) is needed for the common task of understanding speech in noise [18] and for this reason binaural hearing aids and binaural cochlear implants are widely advocated [27—29]. As to severity (2), a recent treatment guideline [17] recommends a general criterion for intervention with ventilation tubes in the HL region of 25 dB. Hence attention concentrates mainly on (B, B) and (B, C2), and their distinction from other tympanometric combinations. All versions of ACET that have been derived on a sample including (B, B) cases are consistent with making the major distinction in this region.

4.2. Extrapolation outside age range

Cognitive/motivational factors in task performance have few convenient markers beyond age and a tester's rating of concentration [21,25]. In preliminary work we found that after adjusting for age (a precise and objective variable) concentration did not enter models of the present type. Therefore age, albeit a minor parameter of the models, captures some of the cognitive and motivational factors influencing behavioral thresholds. Had our sample included children younger than 3 years, concentration might also have entered, or the age coefficient in the regression might have been stronger, for both cognitive reasons and reasons in middle-ear mechanics. Without further derivation work on a younger sample, generalization of the present derivation to under-3 s could be expected to be less good than seen here. Small systematic errors would not matter if the purpose was correlation, but they could if the purpose required absolute reference. The companion paper [30], extending the present work to surrogate bone conduction measurement, offers particular promise in children under about 3—4 years who may not well tolerate headphones and especially not a bone-conduction receiver, so rederiving ACET on a large sample of such young children with well determined thresholds would be valuable. Further derivation is not necessary for applications to an older sample, where OME cases become rare. Extrapolation could proceed by the adoption of small adjuster constants for age, based on the small yearly improvement in thresholds in published developmental norms.

4.3. Generality and interpretation of contralateral conditioning

The "other ear" tympanogram status reflects an underlying severity parameter, so influences the mean HL especially where both are Type B. Among the codings of the tympanograms with fewer df, those that do relatively well (trichotomy and (B, B) versus all other types of tympanogram combination) reflect this conditioning to some extent. Considered for one ear at a time, the present mean HLs for a B tympanogram in isolation are in the upper 20 s, but they are over 30 dB HL when the non-test ear is also B (Tables 2a and 2b). This contralateral conditioning is not restricted to our data. Sabo et al.'s study [25] (2003) of audiometric methods in OME was stratified by presence of effusion on the basis of pneumatic otoscopy rather than by B tympanograms as here, but these markers are closely related. Their data show a very similar contralateral conditioning effect (their Tables 2a and 2b) to ours, but the very small numbers of the more severe cases would not have permitted showing the conditioning to be statistically significant. Fria et al. [11], in a comparison of worse ear (*N* = 155) versus binaural (*N* = 385) average thresholds on a more comparable sample, also show 8 dB of contralateral conditioning in children aged 2—12 years. We do not claim to have "discovered" the contralateral conditioning effect, but claim to have seen its importance and developed an appropriate method to use it in healthcare systems that are appropriately selective for severity in the cases seen at secondary care.

Obviously the contralateral tympanogram is not itself a fundamental "cause" of the tympanogram/HL relation on the first (index) ear, only a marker of an underlying dimension of case severity. However, it summarizes various causal influences underlying the measure in either ear, in a way consistent with a bilateral case being a more serious case in an ultimately causal sense. It goes beyond the limitations of the tympanogram category system to tap within-category continuous influences. It also illustrates that measurement is more than switching on instruments; it embraces the appropriate data-processing. Improvements in instrumentation for mechanical description of the middle ear (e.g. [31—33]) may set a higher ceiling for prediction of the present type but do not evade the requirement for appropriate data processing to mapping ear states and the underlying pathological state that influences the hearing levels.

4.4. Basis of success of trichotomy

The models with 3 or 4 category-levels per ear and even the trichotomy are sufficiently close in predictive accuracy terms that there is no clear overall ''best model''. Recommendation thus has to be based on purpose and on considerations other than the (small) differences in variance explained by the derivation formula. One of these, the pervasiveness of the contralateral conditioning effect for distributions with any material proportion of serious cases has been fully discussed. Another consideration, general parsimony (having few degrees of freedom in the model), is an index of theoretical power, simplicity and elegance. Parsimony favors the trichotomy, although the scientific obligation to seek the best level of measurement feasible (see statistical appendix) does not favour it. Tables 4—6 all suggest how degrees of freedom can be discarded without compromising prediction, provided that the remaining ones are ''spent'' in the most discriminating way. The tympanogram trichotomy has the advantage of practicality, e.g. where computation has to be done by hand. Its material advantage over the simple dichotomy (by 4.4% variance explained) confirmed the advantage of coding that distinguishes bilaterality from unilaterality. Application involves remembering only 3 values: 16, 22 and 32 dB for tympanogram combinations having 0, 1 or 2 type Bs, ±1 dB (according to age), giving an ACET with 7.52 dB standard deviation of prediction error.

4.5. Use of ACET in research

The differing objectives and differing requirements of research and practice entail that a single version of a technique may be good for one but not for the other—hence our developing the trichotomy separately from the full 3-level ACET with interaction term. The drop in correlation reflecting imperfect generalizability between occasions is more a difference of process than of outcome. As noted in the legend to Table 7, there is a correlation over 0.99 in the resulting values. This implies that requirement for a new derivation on new data is minimal, certainly in relation to the stricture about sample size required. In new datasets insufficiently large for a new local derivation, application of the present formula (Table 5) can be encouraged. It is necessary in such application to understand the principles affecting predictive accuracy and to be aware of possibly reduced generalizability to milder samples or narrow distributions. Any more general recommendations here on what to do with a small new subclinical sample (with fewer B tympanograms) would be too open to misinterpretation. On the other hand, if a new study is very large (e.g. $N > 500$), a new derivation can be produced along present lines and retested for replicability as methodological research as well as for use in the particular application. There may be demonstrable advantages in accuracy to be had from optimizing the derivation to the particular distribution or type of case in an epidemiologically differing context, and the hypothesis from the present work to be tested is that these will be minor. We have incorporated the ACET formula in a spreadsheet that can be made available and we can advise further.

We can recommend four types of use of ACET in research: (i) planned use, including interpretation as HL (e.g. to avoid cash cost or opportunity cost in the overheads required to test very young children). Here the research design should consider the nature of the interpretations required and the acceptability of an approximation explaining only 50% of the variance. Such use would overcome practical obstacles to making frequent, short term, measures of HL (e.g. in homes); here tympanometry would be more feasible, in terms of both its simplicity and freedom of need for sound isolation. Regular ACET measures in such settings could, over time, provide a measure of long term auditory deprivation. Their accumulation over time would also be expected to reduce the gap in the distribution of estimated HL values. Use (ii) is similar to (i) but concerned with the cost-effective use of pre-existing datasets to enable aggregation of studies for meta-analysis of trials on individual patients (e.g. [34]) or corroborations in secondary analysis of large epidemiological datasets of good quality but not optimal for the specific purpose in hand (e.g. not containing HL). Obviously re-derivation for the particular data is impossible here. (iii) Use to impute for missing HL values. This raises no issues beyond those already attending the issue of imputation. In the evaluation of treatments but also in epidemiology more generally, authors with much missing data are increasingly required to impute for them. The principle increasingly appreciated is that it is generally better to minimize selection biases by imputing missing values and so to retain cases with incomplete data in an analysis, than to risk biases by having to exclude cases [35]. The fourth use (iv) would be simply as a more powerful summary index for tympanometry, without explicit interpretation as substitute for HL, or even as a control check for confounding possible with HL. As an example, we have recently shown wide variation in mean severity (HL) of child caseloads referred to ENT departments around the UK. The hospital differences in ACET were very similar to those in HL, and this demonstration rules out explanations in terms of possible differences in noise levels or in audiological skills for coaxing reliable thresholds out of young children. The issues in using the greater power of ACET as a quasi-continuous measure are addressed in the Statistical Appendix.

4.6. Use of the ACET formula—clinical practice

ACET provides an explicit estimate more precise than an informal grading of problem/no-problem, for what the binaural HLs might be in children on whom they have not yet been obtained, by acknowledging the severity implications of bilateral B tympanograms. ACET is free from confounding by performance factors (age, confidence, distraction, fatigue, etc.), which true HL is not. Furthermore, it makes distinctions in the clinically important region and can be obtained without the need for high-grade acoustical isolation. In the contexts where these advantages can most generally be realised, the trichotomous approximation, without decimal calculations, will for the foreseeable future be acceptable. Audiological capacity is stretched in many clinics, and in many countries of the world it barely exists. This requires that effort be efficiently distributed to the clinically vital distinctions with regard for the more technically demanding or more informative tests. Many clinicians have always done this informally, but ACET offers precision and efficiency for universal application. The knowledge about the relationship between ACET and HL can underpin decisions about what information it is most important to acquire next.

It is for imaginative clinicians to now generate further potential clinical applications that may be more contentious than those above, in the light of the cautions we have expressed. The benefits should be subjected to formal designed evaluation research, but seem to lie in 2 main areas.

4.6.1. At repeat assessments for monitoring and/or counseling

Examples of calculations are given in the footnote to Tables 5 and 6. Such calculated figures for ACET or the simpler trichotomous equivalent can be used for counseling in terms of a single severity metric. This maps into binaural HL, for relevance of interpretation and comparison with true measurements. For example, in monitoring change it is meaningful to take the difference between true binaural HL with an ACET and the lower precision of the ACET does not render useless the greater precision of the HL; acquiring true HL would become justified if this difference exceeded some criterion. The metric of ACET justifies interpretation to parents as "hearing ability" in some circumstances and especially at later stages in an episode of consultations when some HLs have been obtained. The issue of best wording to confer meaning of an ACET measure to parents (e.g. "the expected hearing ability, given the ear's (lack of) mobility") does not differ much from the issue already met in explaining the reasons for doing tympanometry.

4.6.2. Routing patients to assessment and decisions

ACET provides a flexible metric for using information from both ears to make interim clinical decisions. Tympanometry is sensitive though not specific for detecting HLs at or above 25 dB. Dempster and MacKenzie [13] used this fact to propose efficient tympanometric determination of whether even air-conduction audiometry was required. A subset of the present data have already shown with some precision that their sensitivity and specificity findings are broadly repeatable [20], but their suggestion had three deficiencies, which may explain why it was not widely adopted. Firstly, its emphasis on efficiency was too drastic for an era in which over-intervention was only just being acknowledged. Even getting accurate air-conduction hearing levels on most cases was a recent achievement and a disincentive to over-aggressive tube placement on mere diagnosis of ear fluid, so pointers to limiting the demand for this test went unappreciated. Secondly, having only one suggested cut-off, the procedure was not adapted flexibly to the actual restriction of audiological capacity or to the uncertainty in the individual case. Thirdly, the proposal did not address (either from a clinical or a medico-legal standpoint) the issue of ruling out an underlying cochlear or neural component. ACETaddresses the first two problems. The companion paper [30] shows that a sensorineural component is quite likely in referred cases with a minor conductive component and that the degree by which ACET differs from true HL indicates the likelihood of a sensorineural hearing loss. The companion paper offers a simple decision rule to compress the current wide between-hospital variation in use of bone-conduction testing

[36]. Thus we reconcile the economising spirit of Dempster and McKenzie with the need to rule out a sensorineural component, but some air-conduction HL measurement on each case is necessary.

4.6.3. Clinical framework for combining information from audiometry and tympanometry

The breakdown in Fig. 4 assists more systematic thinking about using both HL and tympanometry in treatment decisions, as well as showing how the number of B tympanograms discriminates as to HL. For efficiency, effort needs to be concentrated on cases where there is genuine uncertainty. For example, let us assume that among the many cases with HLs below 20 dB binaural average (of which only 13% have (B, B) tympanograms), only a very few specially and explicitly justified cases would be offered ventilation tubes; likewise, that most cases above 30 dB (of which 90% have (B, B) tympanograms) would be offered them. In those extreme regions, neither tympanometry nor the usual supplementary considerations (comorbidities, susceptibilities, cultural ability and personnel capacity of the family and school to handle a hearing loss) are crucial because the recommendation can be made largely on HL alone.

However decisions for children with HLs between 20 and 30 dB (in whom 60% have (B, B) tympanograms) are less clear-cut. The child with binaural average HL of 26 dB is not radically different from one with 24 dB and does not reliably obtain more benefit, even though adoption of a 25 dB cut-off in a guideline [17] may seem to imply this. We now illustrate that supplementary considerations are not merely appropriate because clinically wise; they are necessary to resolve uncertainty. Difficulties and inconsistencies can arise in cases close to the cut-offs in guidelines [37] due to failure to consider supplementary information. These can be partly handled by a tie-break approach: spending more effort on assessment and questioning of parents of those children close to cut-off. This sits poorly with the idea of a standard package of testing and explanation for a standard fee. However it is more important for cost-effective treatment policy, and the time resource for more comprehensive yet precise assessment at the margin can be created by economising on the clearer extreme cases as set out above. In cases marginal in this way for a treatment decision, the full audiogram is vital. However the fact that tympanometry will have been considered at the earlier stage of diagnostic confirmation does not stop it also acting as a supplementary consideration. In Fig. 4, the percentage with bilateral Bs in the 20—30 dB band is 60%, centering tympanometry's discrimination on this band. This suggests that discrimination may be achieved in this region by aligning each case to one of the adjacent more certain bands according to its tympanogram status. Indeed, when only a truncated 20—30 dB case range is selected ($N = 400$), the actual HL value 3 months earlier does not significantly add to prediction of later HL dichotomised at 25 dB HL ($p = 0.223$ when ACET is already in the logistic regression). However tympanometry, coded as ACET, does remain useful here; it predicts later HL $\geq$ 25 dB (OR = 1.068 per dB in ACET; CI 1.020—1.228; $p = 0.005$) and this works also for the trichotomy (OR = 10.3 for two B tympanograms versus none; $p = 0.025$, contrast marginal for only one B). This finding is relevant because continued hearing loss is a surrogate for ability to benefit from treatment generally and now a recommended powerful determinant of treatment. We interpret the ability of the tympanogram state (especially when coded as ACET) to predict persistence of hearing loss over time similarly to its ability to condition hearing loss across the two ears, as due to its marking an underlying pathological state; this offsets unpredictive sources of variability in the first HL, e.g. short-term fluctuations. The stronger prediction from a measure whose data come from a different domain long thought non-predictive of HL would appear astonishing, apart from the context of a narrowed HL range exhausting HL's own predictive value *relative to these sources*. An uncertainty region requiring a tie-break is a very relevant context, characteristic of clinical algorithms, and this demands a comprehensive approach to the use of available information. The specific ways to optimize treatment decisions by comprehensive use of information lie beyond the present scope.

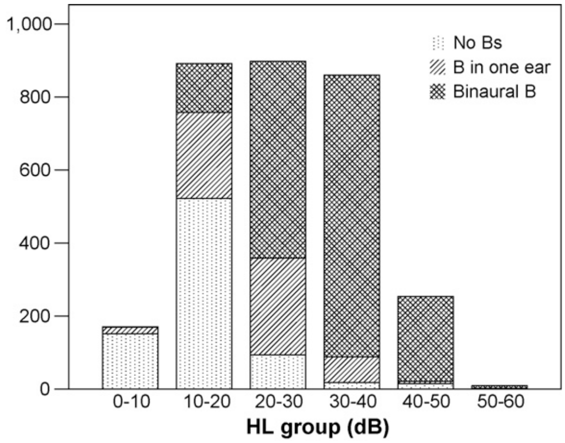

Fig. 4 Bar chart of binaural HL (average 0.5—4.0 kHz) in 10 dB groups by binaural tympanogram combination ($N = 3085$).

5. Conclusions

We have obtained overwhelming and scientifically coherent evidence for moderate validity and high reliability and generalizability of the relationship between tympanometry and hearing level when the tympanometric information is processed appropriately. Economical formulae developed on a large sample of children, aged 3(1/4)—6(3/4) years, predict binaural HL from 2-ear tympanometric data, with a small age adjustment. The formulae take the standard modified Jerger system as input, so are applicable to all data coded in this way without doing new derivations (e.g. on small samples where derivation would be unreliable). The outputs are individual estimates of binaural average HL (0.5—4.0 kHz).

We have named the technique according to what it is: air-conduction threshold estimated from tympanometry (ACET). It distinguishes children likely to have material hearing loss from those not so affected. More generally, the estimates compare usefully with true HL, the optimum formula explaining 49% of the variance in HL-values (equivalent $r = 0.70$) on an adequately severe sample or caseload. A 3-level approximation (0, 1 or 2 B-tympanograms) performs very well in many circumstances. Prediction from ACET is imperfect, but this very fact permits the deviations from prediction (residuals, which are not random error), to convey other information not related to OME.

One main basis for the successful prediction of ACET is a contralateral conditioning phenomenon, whereby an ear with a B tympanogram has HLs approximately 9 dB higher when the tympanogram on the contralateral ear is also a B. Contralateral conditioning can be seen in other data but has not been interpreted or used satisfactorily before. Considering B tympanograms binaurally offers an extra element of prediction that is absent monaurally. When the HL range is truncated to 20—30 dB, the main clinical uncertainty region, ACET better predicts persistence (later HL) than HL does itself.

In research, the binaural ACET formula provides a specified unidimensional *binaural reduction of tympanometry data* for various applications irrespective of whether or not a strong interpretation of equivalence to HL is made, and irrespective of whether the true HL measure is also available. Although its distribution is not unimodal, its scale properties are good.

Clinically, estimated hearing levels can be applied in three main ways given an OM(E) diagnosis: (a) in *counseling* parents in terms of a single severity metric, without necessarily (re-) acquiring HL on every clinical occasion; (b) in making clinically efficient *routing decisions*, e.g. to the use of (further) audiological capacity or otherwise; (c) as a supplementary consideration in treatment decisions when HL is close to 25 dB.

Acknowledgements

We thank Dr. Paola Marchisio for introducing us to the practical research problem of low availability of HL data which developing ACET helped to solve, and Dr. Judith Gravel for constructive comments on earlier drafts

Appendix A. Statistical notes

The statistical techniques used in the derivation work are not familiar to most otolaryngologists. These notes are general, to provide an intuitive introduction to their power and the reasons why they form the appropriate approach to the research question, but also serve to keep brief the reporting of Methods. There are many texts that provide a more complete and formal account (e.g. [41,42]).

A.1. Multivariable regression

A.1.1. Modelling strategy

Multivariable (or multiple for short) linear regression is a powerful technique whereby several sources of variance can be specified to provide a good predictive model. It heeds the scientific imperative for a good general account of how variables relate to each other, rather than being stopped at the elementary or preliminary level with a list of relevant variables, between which some unspecified relationship may exist. Its chief use in medical research is to adjust for known or suspected influences or confounders that are not of prime interest, to obtain a more precise, a less confounded and/or a more conservative account of the chief relationship(s) of interest. Demographics provide perhaps the most frequent examples, as in statistical control for age or socioeconomic effects *not of focal interest* and probably not fundamentally causal; these often mark an unspecified package of probably causal variables. What is of focal interest is defined by the user's main scientific question, and given form by the range of models tested. The analysis part of the methods has to be expressed as an explicit modelling strategy. The statistical control achieved does not guarantee causal relationships, but appropriate contrasts between regression models within a set, according to a planned strategy can allow a stronger causal inference than any set of univariate correlations can. Multiple regression serves to reduce the problem of multiple testing, whereby a large matrix of intercorrelations may not be securely interpretable because several correlations in it can be "significant" under chance alone.

To be stable and replicable, multiple regression models require a number of cases safely larger than the number of degrees of freedom used in the final model. A factor of 10 times that number is often used as a guide for when possible instabilities do not need to be addressed in relation to determining presence of important effects without high Type 1 and Type 2 error. For replicability of the precise values included in the formulae, in the paper we work to a factor of 20. Below 5 is so risky as to be inadvisable under most circumstances. The predictor variables (independent variables) do not need to be normally distributed: analyses tend to be more powerful if they are symmetrically distributed, and this makes it worth considering transformation, except where their distribution already mimics that of the dependent variable (see section on Skew below). The bimodal distribution of ACET (e.g. Fig. 2) is not a bar against its predicting other measures (outcomes) of interest.

Unordered or ordinal category variables (such as the tympanogram categories) can also be entered as predictors but involve a number of degrees of freedom that is 1 less than the number of levels, contrasting with only 1 overall for a dichotomy or for a continuous (linear) dimension. Regression programs make available an estimate with standard error available for each category level (Tables 4—6); the estimate reflects the difference between the raw mean of cases in this category level overall and the reference category, but adjusted for other effects in the model. Worked examples of categorical regression in the present context are given in the legends to Tables 5 and 6.

Mentally grasping all parts of a complex model at once may be hard, which is why the computational procedure is needed. Any one part (e.g. one predictor's coefficient) has been optimized for the context of the other parts of the model, so it may not always make sense if considered in isolation. As one example, between the interaction and non-interaction models reported here, the estimates for the overall (main effect) components differ radically according to whether a more detailed part of their work is being done by an interaction term or not.

Where there are many variables, hence many possible models based on inclusion or otherwise, some terms are likely to be pre-empted by others that better mark the underlying construct. With many variables available, their many combinations (as to significant entry into the model or not) generate many possible models. Here interpretation is based on the features of the better models relative to the poorer ones, requiring some experimentation with an evaluative classification of many models. Thus the most powerful application of multiple regression is not a passive one-pass process, but requires comparison of models, possibly after some iterative convergence on the best reference model. Such interactive guiding requires human judgement, from which it is hard to totally preclude bias, although bias can be minimised by explicit strategy, adherence to sound general principles of method and deliberate avoidance of bias. Any remaining distortions from human judgement after intensive interaction with the data are usually much smaller than the general dangers of error from hurried, selective or uncomprehending one-pass analysis without inspecting the fuller properties of the data, for example, the over-interpretation of a matrix of simple correlations.

A.1.2. Symmetry of distribution (skew)

Linear regression requires a continuous distribution of the outcome (dependent variable) on an equal-interval scale that is transformable to be close to a normal (Gaussian) distribution. This is best done once the model has been roughly developed, as technically the requirement applies to the errors after fitting the model ("residuals"), not necessarily to the variable in its raw form. It is also important that residuals be homogeneous, so plots of the type seen in Figs. 2 and 3 are useful for simple visual inspection: where the points do not form a horizontal band, further transformation may be necessary. The separate issue of discontinuity is addressed in the text. In the modelling reported, we attended carefully to distributions of residuals. The problem was reduced by the good metric properties of HL and the predominance of serious cases appearing both in the tympanograms and the HLs, so having little influence on the distribution of residuals, in effect a type of difference between the two. The skew remaining in the HL residuals for analyses predicting the binaural average HL when not transformed was highly significant because of the very large *N*, but it was not great. Comparisons of transforms showed no material differences in the form of the model or in variance explained, and we did not require literal interpretation of *p*-values as these were not marginal. Therefore, to avoid complexity and differences in scale-values for coefficients due to differing transforms, and for simplicity of interpretation, the natural HL values were not pre-transformed. Hence all estimates here can be interpreted in terms of decibels (dB) hearing level, and so compared. In particular datasets, such as a highly selected subset at either end of the general distribution, a transformation (e.g. logarithmic) of HL may be required in derivations to obtain ACET values. Transformations between either HL or ACET and another variable may clarify a relationship by linearizing it. It then has to be remembered that the scale is no longer metrically equivalent to HL although it remains conceptually equivalent. In applying ACET values as independent variables or in univariate correlations it may be possible to avoid transformation and so keep the HL (or ACET) with its familiar values, aiding comprehension.

Where there is no good prior reason for a particular transformation, one is best sought only after the significant predictors in the model have been largely identified, as generally the problem will have been reduced, by comparison with seeking normality of raw variables. A marginally significant predictor may then be favored or otherwise by the transformation adopted at this point, so the strategy can rarely be completed in a single iteration. Space given to interpretation of marginal effects should be small, but comment on the circumstances that allow them to enter is in order.

A.2. Using ACET values and their distributions

A real problem can arise when ACET is used in a particular application as the outcome (dependent variable), i.e. something which other variables such as age, history or treatment might predict. Fig. 2 shows the marked central dip (bimodality) in the distribution. This section lists a range of practical ways of avoiding the statistical errors to which this could lead. A similar modelling strategy to that in the section above for continuous measures needs to be followed, with fitted variables tending to reduce the seriousness of the distribution problem, but the bimodality is likely to remain with ACET. The difficulty is a by-product of the inherent categoricality of B tympanograms. ACET is a victim of its own success in using the contralateral conditioning phenomenon to best discriminate HLs across the distribution as a whole. Using parametric statistics such as *t*-tests and linear regression with ACET as dependent variable, is not necessarily misleading, but is inappropriate and does not allow literal interpretation of *p*-values. For most purposes the following 8 rules avoid adverse consequences of this bimodality.

(a) Do use the ACET values for summary descriptive statistics (medians, percentiles such as quartiles) and for graphs; transform for skew as generally indicated.

(b) If the sample is sub-clinical, with only a low proportion of double-Bs, perform conventional tests for significance of deviation from normality and it may be acceptable to proceed with a parametric test if kurtosis is not significant.

(c) Where kurtosis is significant, provisionally run the parametric statistics as you would for HL to see roughly what results the analysis strategy is going to lead to, and note as background the (incorrect) *p*-values for the small number of effects that should be examined as of chief interest.

(d) Where the *p*-value for the effects of chief interest are highly significant ($p < 0.01$) or null ($p > 0.10$) you do not need to maximise power, so you can complement the descriptives by dichotomising the distribution in the main dip and running the appropriate non-parametric test for the design adopted, such as Fisher exact or multivariable logistic regression. Chi-squared usually has lower power-efficiency and is inappropriate for small expected cell frequencies.

(e) It is best when performing (d) to have a fixed *a priori* cut-off, but you may have reason to investigate more than one dichotomy point. However, if you are testing a general hypothesis about a relationship and it is not very strong, be aware that you must adjust *p*-values for the number of tests you might have used to support the hypothesis. If the relationship is weak, it is unethical to do more than one test using a range of cut-off points not set a priori and to then quote the one most favourable to the hypothesis. If the relationship is strong overall, then summarizing several analyses from differing cut-offs can be supportive to its generality, provided that the results broadly agree.

(f) If the *p*-value is marginal ($0.10 > p > 0.01$) then the general scientific obligation to avoid both false-positive and false-negative results bears specifically upon you. This requires that you use a non-parametric method but one which maximizes power. The sample size then determines what you do next. If the sample size is small (below about 20), the necessary increased power can be achieved by the relatively powerful but non-parametric tests such as the Wilcoxon T (for a single sample or paired data such as difference scores) and the Mann—Whitney *U* (for two groups). (For notes on use and comments on power-efficiency relative to the *t*-test, see [43]). These tests require an equal-interval or ordered metric scale and so are able to make use of the extra metric sophistication of ACET compared to crude category counts. Thus there is still value in achieving a continuously scaled *metric* via an ACET formula, even for dependent variables where the initial *distribution* and the model residuals will never be continuous or normal.

(g) If you face the obligation in (f) but the sample is large, you need specialist assistance. For any type of analysis, some variant of "bootstrapping" can obtain empirical confidence intervals [44], in effect providing a significance test. These computationally intensive methods grew up because it is obligatory to use the greater metric power of interval measurement where available. However for non-normal distributions, as N increases, the ranking involved in non-parametric tests like the Mann—Whitney becomes astronomically difficult, and slow even for a computer program. The general idea is that the software discards or re-uses cases to generate 1000 or more different versions of what the results might have been like (i.e. if cases had not been encountered or there had been more similar cases). But because only the real data are used, this quantifies in a generalised way whether two distributions overlap substantially or whether they are distinguishable (significant difference).

(h) It is inconvenient though not impossible, to obtain bootstrapped confidence intervals and hence p-values for a whole analysis such as multiple linear regression, and so to keep its greater power than that of logistic regression, or even to distinguish their slightly different meanings. This might be desired, for example, when wishing to test interactions affecting scaled values with multiplicative formulae rather than just disproportionate probabilities in a set of at least 2 × 2 categories as permitted by logistic regression. There exist also fairly powerful methods of ordinal regression that can handle a trichotomy as a dependent variable but these are not standard in all packages and also need expert advice. If an interaction test is essential, it may be simplest to achieve the equivalent by performing in advance a differencing (e.g. between baseline and post-intervention) on which a simple test subsequently captures the idea of the interaction. This often eliminates the non-normality of distribution. You then proceed to a powerful univariate parametric or non-parametric test as given by (f) or (g).

Appendix B. MRC Multi-centre Otitis Media Study Group

The group does not have a formal constitution; regular meetings of the core staff and collaborating consultants were only held during the conduct of the trial in the mid-late 1990s. This list satisfies editorial requirements for authorship and governance and human resource requirements for acknowledged contributorship.

MRC Scientific Staff: Project Leader, Haggard MP[3]; Health Services Researcher/Trial Co-ordinator, Gannon MM; Co-ordinator, Birkin JA; Statisticians, Bennett KE, Nicholls EE, Spencer H[4]; Otolaryngology research fellows, Georgalas C, Daniel M; Audiological Scientist, Higson JM[5]; Psychologists, Smith SC, Hind SE; Epidemiologist, Rovers MM.

Academic Medical Staff: Lead Academic Clinician, Browning GG; Attached otolaryngology research fellows, Georgalas C, Daniel M.

MRC Support Staff: Data Manager, Egner EM; Research Assistants, Hayman T, Greenwood DC, Carroll RA, Jones H, Richmond TB, Wade AR, Braham L, Moorjani P, Pearson DAS, Kirk G; Audiologist, Baskill JL.

RCT centres: Royal Victoria Hospital, Belfast; Ulster Hospital, Dundonald; University Hospital, Nottingham; Leicester Royal Infirmary, Leicester; Royal Hospital for Sick Children, Bristol; Freeman Hospital, Newcastle; Royal Hospital for Sick Children, Edinburgh; Queen Alexandra Hospital, Portsmouth; Sheffield Children's Hospital, Sheffield; Coventry and Warwickshire Hospital, Coventry; University Hospital of Wales, Cardiff.

Other contributing centres: Royal Hospital for Sick Children, Glasgow; Manchester Children's Hospitals; Diana, Princess of Wales and Heartlands Hospitals, Birmingham; Epsom General Hospital, Epsom; Sunderland Royal Hospital, Sunderland; Tyrone County Hospital, Omagh.

We acknowledge the contribution of all the hospital chief executives, medical directors, clinical managers and pharmacists in facilitating the trial.

Collaborating Consultants: Adams DA; Toner JG; Gibbin KP, Jones NS, Birchall JP; Thomas RSA, Narula AA, Murty GE; Griffiths MV; Wilson JA, Meikle D, Johnson IJM, Date S; Kerr AIG; Robinson D, Madden G; Bull PD, Chapman DF, Parker AJ; Curry AR; Williams RG; Morrissey MSC, Geddes NK; Zarod AP, Willatt DJ; Proops D, Pearman K; Robb PJ; MacKee I; Law K.

[3] Guarantor of present paper and co-author.

[4] Statistician for analyses used in this paper.

[5] Co-author.

Nurses/Research Assistants: Casey M; Megarry A; Smith L, Smethurst H; Pearce G, Hook P, Phipps J; Gibbs S, Dunster S; Brown E, Birch EM, Page V; Sim A; Sharples G, Crawford J; Davies C; Bartolo J; Edwards C; Wilson J, Johnson N; Baxendale C, Welsby W, Pickavance N; Cox V, Markham K; Robb JA.

Audiological Scientists/Audiologists: Flanagan R, Housten G, Jordan G, Bingham M, Cassidy J; Burns C, Hamill C, Glover L, Eccles L; Armstrong-Bednall G, Towle B, Dyer K, Carr J, Herrod J, Levick A; Dunmore P, Parker-Emery C, Sutton R, Gordon C; Midgley L, Barnett N, Foley R, Machlin R; Elliott C, Davidson T, Walker L, Keith J; Brady M, Barr-Hamilton R, Wilson S, McKenzie J, Edgar R, Noble H; Thomas H, Paul B, Gilbert M, Hoblyn R, Walter S, Griffiths E, Inge F; Hodges M, Lawrence C, Walsh A, Loxley T, Byrom P; Walker D, Najaran R, Blyth G, Mace L; Rabaiotti W; Trearty C, Carmichael A; Ronson R, France K, Pilkington H; Hilton F, Parsons J; Whitby G, Robinson S; Brown E, James P; Allen R.

References

[1] A.S. Feldman, Tympanometry: application and interpretation, Ann. Otol. Rhinol. Laryngol. 85 (1976) 202—208.

[2] R. Rosenfeld, A practical classification of otitis media subgroups, Int. J. Pediatr. Otorhinolaryngol. 69 (2005) 1027—1029.

[3] J. Jerger, Clinical experience with impedance audiometry, Arch. Otolaryngol. Head Neck Surg. 92 (1970) 311.

[4] G.A. Zielhuis, G.H. Rach, P. Van den Broek, The occurrence of Otitis media with effusion in Dutch preschool children, Clin. Otol. 15 (1990) 147—153.

[5] T. Lildholdt, J. Courtois, B. Kortholm, J.W. Schou, H. Warrer, The correlation between negative middle ear pressure and the corresponding conductive hearing loss in children. A 12-month study of 352 unselected 7-year-old children, Scand. Audiol. 8 (1979) 117—120.

[6] R.H. Margolis, L.L. Hunter, G.S. Giebink, Tympanometric evaluation of middle ear function in children with otitis media, Ann. Otol. Rhinol. Laryngol. Suppl. 163 (1994) 34—38.

[7] J.Y. Sichel, Y. Priner, S. Weiss, H. Levi, G. Barshtein, R. Eliashar, et al., Characteristics of the type B tympanogram can predict the magnitude of the air-bone gap in otitis media with effusion, Ann. Otol. Rhinol. Laryngol. 112 (2003) 450—454.

[8] P.E. Brookhouser, Use of tympanometry in office practice for diagnosis of otitis media, Pediatric Infect. Dis. J. 17 (1998) 544—551.

[9] G.G. Browning, S. Gatehouse, I.R. Swan, The Glasgow Benefit Plot: a new method for reporting benefits from middle ear surgery, Laryngoscope 101 (1991) 180—185.

[10] R.J. Ruben, Host susceptibility to sequelae, in: R.M. Rosenfeld, C.D. Bluestone (Eds.), Evidence-Based Otitis Media, BC Decker, Hamilton, 2003, pp. 505—514.

[11] T.J. Fria, E.I. Cantekin, J.A. Eichier, Hearing acuity of children with otitis media with effusion, Arch. Otolaryngol. 111 (1985) 10—16.

[12] M. Sente, R. Sente, Evaluation of the level of hearing loss based on the results of tympanometry in children under 5 years of age with eustachian tube dysfunction, Med. Pregl. 53 (2000) 559—563.

[13] J.H. Dempster, K. MacKenzie, Tympanometry in the detection of hearing impairments associated with otitis edia with effusion, Clin. Otolaryngol. Allied Sci. 16 (1991) 157—159.

[14] N.A. Black, C.F. Sanderson, A.P. Freeland, M.P. Vessey, A randomised controlled trial of surgery for glue ear, BMJ 300 (1990) 1551—1556.

[15] R. Maw, R. Bawden, Spontaneous resolution of severe chronic glue ear in children and the effect of adenoidectomy, tonsillectomy, and insertion of ventilation tubes (grommets), BMJ 306 (1993) 756—760.

[16] M. Richards, C. Giannoni, Quality-of-life outcomes after surgical intervention for otitis media, Arch. Otolaryngol. Head Neck Surg. 128 (2002) 776—782.

[17] CG60 Surgical management of otitis media with effusion in children, National Institute for Clinical Excellence, UK, 2008, http://www.nice.org.uk/nicemedia/pdf/cg60fullguideline.pdf.

[18] P. Persson, H. Harder, S. Arlinger, B. Magnuson, Speech recognition in background noise: monaural versus binaural listening conditions in normal-hearing patients, Otol. Neurotol. 22 (2001) 625—630.

[19] G.G. Browning, Reporting the benefits from middle ear surgery using the Glasgow Benefit Plot, Am. J. Otol. 14 (1993) 135—140.

[20] MRC Multi-Centre Otitis Media Study Group, Sensitivity, specificity and predictive value of tympanometry in predicting a hearing impairment in otitis media with effusion, Clin. Otolaryngol. Allied Sci. 24 (1999) 294—300.

[21] MRC Multi-centre Otitis Media Study Group, Influence of age, type of audiometry and child's concentration on hearing thresholds, Br. J. Audiol. 34 (2000) 231—240.

[22] MRC Multi-centre Otitis Media Study Group, Surgery for persistent otitis media with effusion: generalizability of results from the UK trial (TARGET). Trial of Alternative Regimens in Glue Ear Treatment, Clin. Otolaryngol. 26 (2001) 417—424.

[23] MRC Multi-centre Otitis Media Study Group, Risk factors for persistence of bilateral otitis media with effusion, Clin. Otolaryngol. Allied Sci. 26 (2001) 147—156.

[24] Anon, Recommended procedures for pure-tone audiometry using a manually operated instrument, Br. J. Audiol. 15 (1981) 213—216.

[25] D.L. Sabo, J.L. Paradise, M. Kurs-Lasky, C.G. Smith, Hearing levels in infants and young children in relation to testing technique, age group, and the presence or absence of middle-ear effusion, Ear Hear. 24 (2003) 38—47.

[26] R.M. Rosenfeld, D. Kay, Natural history of untreated otitis media, Laryngoscope 113 (2003) 1645—1657.

[27] G.A. Day, G.G. Browning, S. Gatehouse, Benefit from binaural hearing aids in individuals with a severe hearing impairment, Br. J. Audiol. 22 (1988) 273—277.

[28] R. Van Hoesel, R. Ramsden, M. Odriscoll, Sound-direction identification, interaural time delay discrimination, and speech intelligibility advantages in noise for a bilateral cochlear implant user, Ear Hear. 23 (2002) 137—149.

[29] J. Murphy, G. O'Donoghue, Bilateral cochlear implantation: an evidence-based medicine evaluation, Laryngoscope 117 (2007) 1412—1418.

[30] MRC Multi-centre Otitis Media Study Group, Air conduction estimated from tympanometry (ACET): the use of hearing level-ACET discrepancy (HAD) to determine appropriate use of bone-conduction tests in identifying permanent and mixed impairments, IJPORL 2 (2008).

[31] D.H. Keefe, J.L. Simmons, Energy transmittance predicts conductive hearing loss in older children and adults, J. Acoust. Soc. Am. 114 (2003) 3217—3238.

[32] M.P. Feeney, I.L. Grant, L.P. Marryott, Wideband energy reflectance measurements in adults with middle-ear disorders, J. Speech Lang. Hear. Res. 46 (2003) 901—911.

[33] C. Dai, M.W. Wood, R.Z. Gan, Tympanometry and laser Doppler interferometry measurements on otitis media with effusions model in human temporal bones, Otol. Neurotol. 28 (2007) 551—558.

[34] M.M. Rovers, N. Black, G.G. Browning, R. Maw, G.A. Zielhuis, M.P. Haggard, Grommets in otitis media with effusion: an individual patient data meta-analysis, Arch. Dis. Child 90 (2005) 480—485.

[35] G.J.M.G. van der Heijden, A.R.T. Donders, T. Stijnen, K.G.M. Moons, Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example, J. Clin. Epidemiol. 59 (2006) 1102—1109.

[36] M. Haggard, J. Higson, H. Spencer, Bone-conduction measurements in children with OME—an issue for professional debate, BSA News 51 (2007) 33—37.

[37] M. Haggard, The relationship between evidence and guidelines, Otolaryngol. Head Neck Surg. 137 (2007) S72—S77.

[41] A. Rutherford, Introducing ANOVA and ANCOVA; A GLM Approach, Sage Publications, London and Thousand Oaks, CA, 2001.

[42] L.S. Aitken, S.G. West, Multiple Regression: Testing and Interpreting Interactions, Sage Publications, London and Thousand Oaks, CA, 1991.

[43] S. Siegel, N.J. Castellan, Nonparametric Statistics for the Behavioral Sciences, McGraw Hill, New York, 1988.

[44] B. Efron, R. Tibshirani, Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, Stat. Sci. 1 (1986) 54—75.

[45] MRC Multi-centre Otitis Media Study Group, An extension of the Jerger classification of tympanograms for ventilation tube patency: specification and evaluation of equivalent ear-canal volume criteria, Ear Hear. 29 (2008) 894—906.

Available online at www.sciencedirect.com
ScienceDirect