A group sequential test for ABR detection

M.A.Chesnaye, S.L. Bell, J.M. Harte, & D.M Simpson

Abstract

OBJECTIVE: To detect the auditory brainstem response (ABR) automatically using an innovative sequentially applied Hotelling's T^2 test, with the overall goal of optimising test time whilst controlling the false-positive rate (FPR).

DESIGN: The stage-wise critical decision boundaries for accepting or rejecting the null hypothesis were found using a new approach called the Convolutional Group Sequential Test (CGST). Specificity, sensitivity, and test time were evaluated using simulations and subject recorded data.

STUDY SAMPLE: Data consists of click-evoked ABR threshold series from 12 normal hearing adults, and recordings of EEG background activity from 17 normal hearing adults.

RESULTS: Reductions in *mean* test time of up to 40-45% were observed for the sequential test, relative to a conventional 'single shot' test where the statistical test is applied to the data just once. To obtain these results, it will occasionally be necessary to run the test to a higher number of stimuli, i.e. the *maximum* test time needs to be increased.

CONCLUSIONS: The CGST can be used to control the specificity of a sequentially applied ABR detection method. Doing so can reduce test time, relative to the 'single shot' test, when considered across a cohort of test subjects.

1 Introduction

Auditory brainstem response (ABR) detection can be achieved by repeatedly presenting a brief acoustic stimulus to a subject (e.g. a click, a chirp, or a tone burst), and visually examining the accumulating electroencephalogram (EEG) data. Although potentially quite sensitive (Arnold, 1985), it is well known that visual inspection is dependent on the expertise and mental state of the examiner, and thus introduces a variable and subjective element to the process (Arnold, 1985; Vidler & Parker, 2004). This has led to the development of many different objective measures for detecting the ABR, i.e. methods with a firm foundation in statistics, capable of producing highly sensitive and reliable results.

Usually, the main goal for objective ABR detection methods is to assist the examiner during the visual inspection task, and to improve the specificity, sensitivity, and efficiency of the test. ABR detection methods are therefore typically used in conjunction with visual inspection, and are thus applied repeatedly to the accumulating EEG data until a decision in terms of response present or response absent has been reached. A complication with such sequential test procedures is that most statistical detection methods are designed under the assumption of a 'single shot' application, i.e. it is assumed that they are applied to the data just once. If they are instead applied repeatedly, then the false-positive rate (FPR) will tend to be larger than the nominal α-level of the test (see e.g. Armitage et al., 1960). The latter is also known as an 'inflated FPR', and adjusted critical decision boundaries (for rejecting or accepting the null hypothesis H_0 of 'no ABR present') are required in order to obtain the intended FPR. This adjustment is challenging in the context of sequential testing (Bauer & Köhne, 1994; Proschan & Hunsberger, 1995; Lehmacher & Wassmer, 1999; Muller & Shafer, 2001; Liu & Chi, 2001; Brannath et al, 2002; Hartung & Knapp, 2003; Stürzebecher, Cebulla, and Elberling, 2005; Chang, 2006; Sheng & Qiu, 2007).

Besides having to deal with inflated FPRs, sequential testing also introduces trade-offs between statistical power and test time. In particular, the sequential test allows H_0 to be rejected early for the higher signal-to-noise ratio (SNR) responses (thus reducing test time), which comes at the cost of a reduced statistical power, i.e. analysing N samples using a single test will have a higher statistical power relative to analysing the same N samples using multiple sequentially applied statistical tests (Bauer & Köhne, 1994).

In addition to early rejection of H_0, sequential testing also allows the test to be stopped early in favour of H_0, i.e. when the SNR is very low or a response is absent, additional data collection might be deemed futile. The danger associated with early stopping in favour of H_0 is of course that H_0 is accepted when a response is, in fact, present, resulting in an increased false-negative rate (FNR) and hence a reduced test sensitivity. The critical decision boundaries for accepting H_0 should therefore be chosen conservatively if a reduced test sensitivity is to be prevented.

The aim for this paper is in evaluating the performance of a new sequential test procedure for ABR detection with a view to optimise statistical power and test time whilst controlling the FPR. The test is built around a new method called the Convolutional Group Sequential Test (CGST; Chesnaye et al, in press), which is a flexible and intuitive approach for finding the stage-wise critical decision boundaries and controlling the FPR of sequentially applied statistical tests. In this study, the CGST is used to find the critical decision boundaries (for accepting or rejecting H_0) for a sequentially applied Hotelling's T^2 test for ABR detection. Test performance is then evaluated using simulations and subject recorded data, with the goal to explore and optimise the trade-off between statistical power and test time as a function of the number of sequential stages used for the analysis.

2 Methods

This section describes the data used throughout this study (Section 2.1), after which brief descriptions of the Hotelling's T^2 test and the CGST are provided (Sections 2.2 and 2.3, respectively). Sections 2.4 and 2.5 then describe methods for evaluating the specificity, sensitivity and test time (single ear, and per dB condition) of a sequentially applied Hotelling's T^2 test.

2.1 Data

Data consists of (i) an ABR threshold series obtained from a sample of normal hearing adults, (ii) a relatively large database of recordings of EEG background activity (no stimulus was used), and (iii) simulated coloured noise. The simulated coloured noise and recordings of EEG background activity are first used to evaluate specificity (Section 2.4), after which simulations and the subject ABR data are used to explore sensitivity and test time (Section 2.5). Throughout this work, EEG recordings (either recorded or simulated) are structured into ensembles of 'epochs', where an epoch is defined as a short time interval within the EEG, typically following the onset of an acoustic stimulus.

ABR threshold series

The subject ABR threshold series has previously been described in Lv et al (2007), and was collected from 12 subjects (six female and six males) ranging from 18 to 30 years of age. All subjects had normal hearing levels (<20 dB HL for 0.25, 0.5, 1, 2, 4, and 8 kHz tones), as determined with conventional pure tone audiometry. The mean hearing thresholds (across frequencies) were 1.2 dB HL for the males, and 4.3 dB HL for the females. The acoustic stimulus for evoking the ABR was a rectangular 100 μs click. Click calibration was carried out by playing clicks through ER-2 insert earphones (Etymotic, Elk Grove Village, IL), mounted in an IEC type-4157 occluded ear simulator with a 0.5 inch microphone for measuring the SPL. The output of the ear simulator was then measured using a Bruel and Kjaer spectrometer. The clicks were then delivered to the subjects at a stimulus rate of 33.11 Hz (using the ER-2 insert earphones) at intensity levels 0, 10, 20, 30, 40, and 50 dB SL (sensation level), i.e. relative to the individual's behavioural hearing thresholds (for the click). The behavioural hearing thresholds (for the click) were estimated using a simple 'up-down' approach where the amplitude of the click was reduced in steps of 10 dB for every correct response, and increased in steps of 5 dB for every missed response. ABRs were then recorded between the vertex and the nape of the neck, with a ground electrode placed at mid-forehead. EEG measurements were obtained at a sampling rate of 10,000 kHz using a Cambridge Electronic Design (CED) micro 1401 data acquisition unit along with a CED 1902 amplifier (with a built in 3rd order 30-3000 Hz band-pass filter and a 50 Hz notch filter). The electrode impedances remained below 5 kΩ throughout the recordings. Finally, artefact rejection was applied by throwing away 10% of the noisiest epochs, as determined by their absolute maximum values. There were a minimum of 3000 artefact free epochs available, per subject and per dB SL condition.

EEG background activity

Recordings of no-stimulus EEG background activity were previously collected by Madsen et al. (2017) from 17 subjects (five female and 12 male) under four conditions; (1) *asleep*, where the subjects were asked to try and fall asleep (although sleep was not confirmed), (2) *still*, where the subjects were asked to lie still with their eyes closed (but not to fall asleep), (3) *blink*, where the subjects were asked to blink as a circle appeared on a screen in front of them (once every 1-3 seconds), and (4) *move*, where the subjects were instructed to move according to a random animation, also shown on a screen in front of them. The data were recorded with the aim of emulating EEG recordings of varying quality and morphology. Measurements were obtained at a sampling rate of 20,000 kHz using a Compumedics Neuroscan II EEG amplifier (with a built in 3rd-order 30-3000 Hz band-pass filter) with electrodes placed on the left mastoid, the upper forehead (reference), and the right cheek (ground). The electrode impedances remained below 1 kΩ throughout the recordings. Artefact rejection was applied by throwing away 10% of the noisiest epochs, as determined by their absolute maximum values. After artefact rejection, there was a total of approximately 10 hours of raw EEG available.

Simulated coloured noise

Simulated coloured noise was generated by filtering Gaussian white noise with an all-pole filter, where the poles of the filter were given by the parameters of an autoregressive (AR) model. The AR models were estimated from the recordings of EEG background activity using the Modified Covariance method (Marple, 1987), with a new AR model being fitted to each recording. The order for the AR models was determined by visually comparing the power spectral densities (obtained using the Welch method; Welch,

1967) from the original EEG recordings to those from the simulated EEG recordings. A relatively high model order of 60 was then chosen to ensure a close match between the original and the simulated recordings in terms of spectral content.

2.2 The one-sample Hotelling's T^2 test

The one-sample Hotelling's T^2 test (Hotelling, 1931) is the multivariate extension to Student's t-test, and can be used to evaluate the null hypothesis H_0 that Q feature means are equal to Q hypothesized values, i.e. H_0: $\boldsymbol{x_1} = \boldsymbol{\mu_1}$, $\boldsymbol{x_2} = \boldsymbol{\mu_2}$, ..., $\boldsymbol{x_Q} = \boldsymbol{\mu_Q}$, where $\boldsymbol{x_i}$ is the observed mean value for feature i, and $\boldsymbol{\mu_i}$ the hypothesized value to test against for feature i. To clarify, Q features would be extracted from each epoch, giving an NxQ-dimensional feature matrix $\mathbf{V}$ where N is the number of epochs:

$$\mathbf{V} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1Q} \\ v_{21} & v_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ v_{N1} & \cdots & \cdots & v_{NQ} \end{bmatrix}$$

where v_{ij} is the j_{th} feature extracted from the i_{th} epoch, and where mean feature value $\boldsymbol{x_i}$ is found by taking the mean down the ith column of V. The T^2 test statistic is then given by (Rencher, 2001, p.118):

$$T^2 = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0})\mathbf{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0})^H \tag{1}$$

where $\bar{\boldsymbol{x}}$ is the Q-dimensional vector of feature means, $\boldsymbol{\mu_0}$ is the Q-dimensional vector of hypothesized values to test against, $\boldsymbol{S^{-1}}$ is the inverse of the covariance matrix of the NxQ-dimensional feature matrix $\mathbf{V}$, and ^{H} superscript denotes Hermitian transpose. One can view T^2 as a weighted sum of the mean feature values, with the weighting given by the covariance matrix of the features. This essentially allows correlated features with different scales and variances to be combined optimally, in order to test the H_0. The T^2 statistic can then be transformed into an F statistic using:

$$\mathrm{F} = \frac{N-Q}{Q(N-1)}T^2 \tag{2}$$

which follows an F-distribution with Q and $N - Q$ degrees of freedom (DOF) under H_0. It is worth noting that the number of epochs N should be larger than the number of features Q, else $\boldsymbol{S^{-1}}$ cannot be calculated.

Statistical features

In the current work, the statistical features chosen for the Hotelling's T^2 test are mean voltages, taken across short time intervals of EEG. In particular, the 15 ms windows following stimuli onset are split into 25 segments of equal duration (0.6 ms per segment), and the mean is taken across each segment, giving an Nx25-dimensional feature matrix $\mathbf{V}$. The choice to use 25 'voltage means' as statistical features was based on findings from Chesnaye et al (2018), which show a sensitive and robust performance for the Hotelling's T^2 test when using 25 voltage means for ABR detection.

2.3 The Convolutional Group Sequential Test

The CGST tests for the presence or absence of an evoked response in multiple sequentially applied stages (Chesnaye et al, in press). At each stage of the sequential analysis, a p value is generated by the statistical test (here the Hotelling's T^2 test), and a summary statistic is constructed by combining this p value with all previously generated p values through summation (potentially following some p value transformation; further clarified below). The null hypothesis H_0 of 'no response present' is then evaluated using this summary statistic. That is, if the summary statistic exceeds some upper boundary, data acquisition is stopped and H_0 is rejected, whereas if the summary statistic falls below some lower boundary, H_0 is accepted. The CGST also requires all stage-wise p values to be independent under H_0, which implies that data must be analysed in disjoint blocks of epochs. As an example, when analysing 500 epochs in two stages, a first test might be performed on epochs 1-250, and a second test on epochs 251-500.

For this study, the stage-wise p values are log-transformed and combined using Fisher's method (Fisher, 1932), which has some desirable properties in terms of efficiency when combining p values (Littell & Folks, 1971). When using a total of K stages for the sequential analysis, then the summary statistic is updated K times (once per stage). The summary statistic at stage k, say Σ_k, is then given by:

$$\Sigma_k = \sum_{i=1}^{k} -2\ln(p_i) \tag{3}$$

where p_i is the p value generated by the statistical test at stage i. At each stage of the sequential analysis, the test can be stopped for efficacy when $\Sigma_k > A_k$, or for futility when $\Sigma_k < C_k$. Efficacy implies that there is sufficient evidence for rejecting H_0 at level α. Futility implies that summary statistic Σ_k is sufficiently far from statistical significance, such that additional data collection is deemed to be futile, and H_0 is accepted. It is also worth emphasizing here that the ln-transform gives large values for small p, i.e. the Σ_k summary statistic increases for decreasing p values.

The key challenge is to find suitable values for the A_i and C_i decision boundaries (for $i = 1, 2, ...K$), such that the nominal α-level of the full sequential test is preserved. The latter is achieved by the CGST in a mathematically rigorous and flexible way. For a full explanation of the approach, including graphical illustrations, the reader is referred to Chesnaye et al (in press). Here, a summary will be provided; At the core of the CGST lies the convolution theorem, which states that the sum of two independent random variables is given by the convolution of their individual null distributions (Grinstead & Snell, 1997). Hence, if the null distributions for the (transformed) stage-wise p values p_i are known (the null distributions for $-2\ln(p_i)$), then these can be convolved, two at a time (an additional convolution for each stage), to generate the null distribution for summary statistic Σ_k (say ϕ_{Σ_k}), which can then be used to find A_k and C_k. In this study, the stage-wise p values are assumed to be uniform on the [0,1] interval under H_0, in which case the stage-wise null distributions for the transformed p values (the distributions for $-2lnp_i$) are given by χ^2 distributions with 2 DOF. A final caveat is that distribution ϕ_{Σ_k} changes when proceeding from stage k to stage $k+1$, as it is not possible to enter stage $k+1$ with $\Sigma_k > A_k$ or $\Sigma_k < C_k$ (else the trial would already have been stopped for efficacy or futility, respectively). The stage H_0 rejection and acceptance regions of ϕ_{Σ_k} for stage k should therefore be truncated prior to entering stage $k+1$.

CGST design parameters
The CGST requires various parameters to be specified in advance. This firstly includes the nominal α-level of the test and the number of stages for the sequential analysis, say K. For this study, α is set to either 0.01 or 0.05, whereas K can take values ranging from 1 (giving a single shot test) to 9. After specifying K and α, a choice needs to be made with regards to how α is 'spent' across the K stages, i.e. the stage-wise FPRs, denoted by α_i, need to be specified, such that $\sum_{i=1}^{K} \alpha_i = \alpha$. For this study, α is always spread equally across the K stages, giving α_i values of $\frac{\alpha}{K}$ for all i and K. With respect to early stopping for futility, this can be controlled through the γ_i values (for $i = 1, 2, ..., K$), which are the stage-wise true-negative rates (TNRs; defined as the fraction of tests to be stopped in favour of H_0 when H_0 is indeed true). The danger with early stopping for futility is that the test may be stopped prematurely and H_0 is incorrectly accepted, giving a reduced test sensitivity. The γ_i values should therefore be chosen conservatively if a reduced test sensitivity is to be prevented. Additional simulations (details not presented) suggest that a relatively conservative choice for the γ_i values (little to no loss in test sensitivity) is given by $\gamma_i = \frac{1-\alpha}{K}$ for all i and K, which is the approach adopted throughout this work. The K, α, α_i and γ_i parameters and their underlying trade-offs are further considered in the discussion. All resulting critical decision boundaries $[C_i,\ A_i]$ for $i = 1, 2, ..., K$ (and for $K = 2, 3, ..., 9$) can be found in the Appendix.

The assumptions underlying the CGST
The assumptions underlying the CGST include (1) that the null distributions for the stage-wise p values are known (uniform on the interval [0,1] for this study), and (2) that the stage-wise p values are mutually independent. Note that (1) is only satisfied when the assumptions underlying the statistical detection method (the Hotelling's T^2 test in this case) are also satisfied; in the case of Hotelling T^2, this may be violated for example when the data is not normally distributed (due to e.g. large outliers), or when statistical dependence between successive epochs is not satisfied. With respect to (2), this is satisfied when data analysed in stage i is independent of data analysed in all previous and subsequent stages. The spectral content of EEG measurements is therefore a potential concern for the CGST, as this might introduce long term correlations between epochs of the EEG, potentially resulting in a violation of the independence assumption. The independence assumption is therefore considered in more detail in the specificity assessment below.

2.4 Specificity assessment

This section uses simulations and recordings of EEG background activity to evaluate the specificity of a sequentially applied Hotelling's T^2 test. Simulations are first used to conduct a powerful assessment (a large amount of data is generated) of specificity under well controlled conditions, after which the recordings of EEG background activity are used to verify that specificity is controlled as intended for recorded EEG data.

Independence assessment
The independence assumption takes two forms in this study: (1) independence between epochs (assumed by the Hotelling's T^2 test), and (2) independence between the stage-wise p values (assumed by the CGST). The extent to which independence is violated is dependent on the dominant frequency within the data (the frequency with the largest amplitude). For EEG data, which follows an approximate $\frac{1}{f\alpha}$ spectrum (with $\alpha \approx 1$, Pritchard 1992), the dominant frequency will tend to be the lowest frequency in the recording, and is hence determined by the high-pass cut-off frequency. The distance in time between EEG measurements (determined through the stimulus rate) then determines whether the correlations tend to be positive, negative, or close to zero. In the independence assessment that follows, the FPR is evaluated as a function of the stimulus rate and the high-pass cut-off frequency.

Data consists of 1 000 000 sets of simulated coloured noise, generated as described in Section 2.1. Note that this noise is a Gaussian, stationary, zero-mean noise with similar spectral content as real EEG background activity, i.e. all statistical assumptions (underlying the Hotelling's T^2 test and the CGST) are satisfied for this data, except (potentially) the independence assumptions. The simulated recordings were all band-pass filtered using a 3rd order Butterworth filter from either 30-1500 Hz or from 100-1500 Hz (to cover the lower and upper limits of what is customary for ABR detection), after which they were structured into ensembles of $N = 500$ 15 ms windows (to ensure that the full ABR was analysed). The distance between the 15 ms windows, denoted by τ, was then varied from 0 (contiguous windows) to 25 ms, in steps of 0.4 ms, which corresponds to a (hypothetical) stimulus rate of $\frac{1000}{15+\tau}$, covering stimulus rates of 25.13 Hz up to 66.67 Hz. The 15 ms windows of the ensembles were then analysed in K sequential stages using the Hotelling's T^2 test, where K took values ranging from 1 to 9. The assessment was performed both with and without the option to stop early for futility (H_0 accepted), but always with the option to stop early for efficacy (H_0 rejected).

Recordings of EEG background activity
The recordings of EEG background activity were band-pass filtered from 100 to 1500 Hz, and structured into 30.2 ms epochs, corresponding to a (hypothetical) stimulus rate of 33.11 Hz. A 100 Hz cut-off frequency was chosen based on results from the previously described independence assessment, which show that independence is satisfied when using $f_c = 100$ Hz in combination with a stimulus rate of ~33.11 Hz (see results, section 4). The 33.11 Hz stimulus rate was also chosen to match the stimulus rate of the clicks in the ABR threshold series. Each pre-processed recording was then structured into ensembles of N epochs, where N took values of either 300, 500, or 1000 epochs. There was sufficient data for constructing 3652, 2156, and 1018 ensembles with ensemble sizes of 300, 500, and 1000 epochs, respectively. The initial 15 ms windows of the resulting ensembles were analysed using the Hotelling's T^2 test in K sequential stages, where K was varied from 1 to 9. The analysis was again conducted both with and without the option to stop early for futility. Finally, it is worth emphasizing here that K determines how often (and when) to apply the statistical test to the data, e.g. when using $K = 2$ and $N = 1000$, the Hotelling's T^2 test would first be applied to the data after epochs 1-500 have been collected. If the test is then not stopped for futility or efficacy, it would be applied a second time after epochs 501-1000 have been collected.

2.5 Sensitivity and test time assessment

This section describes simulations and the subject ABR data to explore the trade-off between sensitivity and test time as a function of the number of sequential stages K used for the analysis.

Data for the simulations consists of simulated coloured noise (constructed as described in Section 2.3) for representing the EEG background activity, along with coherently averaged and scaled ABR templates for representing a response. The ABR templates were obtained from the subject ensemble coherent averages from the ABR threshold series, under the condition that the ensemble coherent average contained a clear response. The latter was determined through visual inspection by an experienced audiologist: The audiologist inspected the repeatability of the ABR waveform by comparing two replicates of the coherent average, obtained by taking the coherent average across epochs 1-1500, and again across epochs 1501-3000. The audiologist also used the 3-1 signal to noise criterion (see Sutton et al. 2013) as additional guidance, but was ultimately left free to decide whether a response was present or not. This resulted in 0, 4, 7, 8, 7 and 8 templates with a clear response for the 0, 10, 20, 30, 40 and 50 dB SL conditions, respectively. Note therefore that the 0 dB SL condition could not be simulated, as no clear responses were observed in any of the subjects under this condition. The 10 dB SL condition was also excluded in the subsequent simulations, as there were just 4 (potentially noisy) templates available.

When simulating a response for one of the dB SL conditions, an ABR template was selected at random (from the dB SL condition in question), after which it was rescaled, and added to all epochs within the ensemble in question. The scaling factor was chosen such that a specific SNR was obtained, which was calculated using:

$$SNR = 10\log10(\frac{P_{Template}}{P_{Noise}}) \tag{4}$$

where $P_{Template}$ is the mean square value of the scaled ABR template in question, and P_{Noise} is the mean square value of the ensemble of epochs (prior to adding the template) when treated as a continuous recording. The SNRs for the simulated response were furthermore given by the SNRs estimated from the subject ABR data, which were similarly calculated using Equation (4), where $P_{Template}$ was now the subject ensemble coherent average, and P_{Noise} the ensemble of epochs when treated as a continuous recording.

Simulations I: true-positive rate fixed at 0.99
For the first set of simulations, the true-positive rate (TPR) was fixed at 0.99 for all K and all dB SL conditions. This was achieved by repeatedly generating 10 000 ensembles with increasing or decreasing ensemble sizes N, until a TPR of 0.99±0.005 was obtained. As an example, when using $K = 3$, a total of 3656 epochs (split across the 3 stage) were required for obtaining a 99% detection rate for the 50 dB SL condition, whereas 7210 epochs were required for the 20 dB SL condition. Needless to say, this approach is not feasible in a clinical setting, and was only included here to provide a fair comparison of test time (all tests have equal test sensitivity) per dB SL condition for different choices of K. The simulated recordings were band-pass filtered from 100-1500 Hz using a 3rd-order Butterworth filter, and structured into 30.2 ms epochs (corresponding to a stimulus rate of 33.11 Hz). The initial 15 ms windows of the ensembles were then analysed in K sequential stages using the Hotelling's T^2 test, where K was varied from 1 to 9. The analysis was performed both before and after simulating a response, but always using the required N for a 0.99 TPR. Note that the no-stimulus condition then corresponds to the scenario where a response is absent (due to e.g. hearing loss), but where the test was

still designed for normal hearing subjects (N was still optimised under the assumption that a response was present). The analysis was again performed both with and without the option to stop the test early for futility.

Simulations II: N fixed at 3000
For the second set of simulations, the ensemble size N was fixed at 3000 epochs, for all K and all dB SL conditions. Contrary to Simulations I, the loss in statistical power for increasing K cannot be compensated for by increasing N, i.e. a reduced TPR can be expected for increasing K. It is worth noting here that these test settings are identical to those used when analysing the subject ABR data described below. The goal for these simulations is indeed to provide a more powerful assessment of the trade-off between statistical power and test time when using a fixed N of 3000 epochs. A total of 10 000 recordings were again simulated, which were band-pass filtered from 100-1500 Hz (using a 3rd-order Butterworth filter), and structured into ensembles of $N = 3000$ 30.2 ms epochs. The initial 15 ms windows of the ensembles were analysed in K sequential stages using the Hotelling's T^2 test, where K was varied from 1 to 9. The analysis was performed both before and after simulating response, and both with and without the option to stop the test for futility.

Subject recorded ABR threshold series
The subject data was band-pass filtered from 100-1500 Hz using a 3rd-order Butterworth filter, and structured into ensembles of $N = 3000$ 30.2 ms epochs (time-locked to the stimuli). The initial 1-16 ms windows of the ensembles were analysed in K sequential stages using the Hotelling's T^2 test where K ranged from 1 to 9. The initial 0-1 ms interval of the epochs was excluded from the analysis to avoid potential contaminations from a stimulus artefact. Note that the test settings for this analysis are identical to those used for 'Simulations II' described above. The analysis was performed both with and without the option to stop the test early for futility.

3 Results

3.1 Specificity

Independence assessment
The FPRs (using $\alpha = 0.05$) from the independence assessment are presented in Figure 1 as a function of the (hypothetical) stimulus rate, for high-pass cut-off frequencies of either 30 Hz (plot A) or 100 Hz (plot B). The FPRs for $K = 3, 4, ..., 8$ were all quite similar (they fell between the FPRs from $K = 2$ and $K = 9$), and are excluded from the Figures to avoid cluttering. Results from Figure 1 were furthermore generated *with* the option to stop early for futility. When early stopping for futility was not permitted, results were comparable to those in Figure 1 (not presented to keep the results concise). The two-sided 95% confidence intervals for $\alpha = 0.05$ are furthermore given by [0.0496, 0.0504] (1 000 000 tests performed), and are shown in Figure 1 as red dotted lines. The confidence intervals were found using a binomial distribution constructed from 1 000 000 observations, where the probability of a single successful Bernoulli trial (defined as a false-positive) was set to 0.05 (the theoretical probability of a false-positive). For the single shot test ($K = 1$), results demonstrate significant fluctuations around $\alpha = 0.05$ as a function of the stimulus rate and the high-pass cut-off frequency, which can be attributed to a violation of the independence assumption between epochs (all remaining assumptions were satisfied). For the sequential test ($K > 1$), the FPRs follow a similar but more pronounced trend, which implies that additional assumptions underlying the CGST were violated. Additional simulations (results not presented) demonstrate that independence between the stage-wise p values was satisfied (or that violations were negligible), and that the additional violation originating from the CGST was solely due to the stage-wise p values being no longer uniform on the assumed [0,1] interval. The latter can, in turn, be attributed to the independence violation between epochs (underlying the Hotelling's T^2 test). The underlying statistical assumptions are further considered in the discussion.

EEG background activity
The FPRs (using $\alpha = 0.01$) generated from the EEG background activity (pre-processed using $f_c = 100$ Hz, and using a hypothetical stimulus rate of 33.11 Hz) are presented in Table 1, for different N and K, both with and without the option to stop the test early for futility. The binomial distribution was used to construct two-sided 95% confidence intervals for $\alpha = 0.01$, giving confidence intervals of [0.0068, 0.0140] for $N = 300$ (3652 tests were performed), [0.0060, 0.0153] for $N = 500$ (2156 tests), and [0.0039, 0.0187] for $N = 1000$ (1018 tests). Significant deviations ($\boldsymbol{p < 0.05}$) from $\alpha = 0.01$ are indicated in Table 1 by asterisks. Results show that the observed FPRs tend to fall within the expected boundaries, with the exception of three test conditions, which show a conservative test performance (further considered in the discussion).

3.2 Sensitivity and test time

Simulations I: TPR = 0.99
Results from Simulations I (TPR fixed at 0.99) are presented in Figure 2 (plots A, B, C, and D). Results firstly demonstrate an increased *maximum* test time for increasing K (plot A), i.e. as K is increased, statistical power is decreased, and the ensemble size N needs to be increased in order to maintain the 0.99 TPR. Note that although the ensemble size N was increased with K, the *mean* test time was still decreased (plot B), with reductions in mean test time of 40-45% when using $K = 6$ (relative to $K = 1$). The latter is due to the test being stopped early (and H_0 rejected) for the higher SNR responses, i.e. the final stage of the analysis is typically not reached (and the maximum test time is not used). This is in contrast to

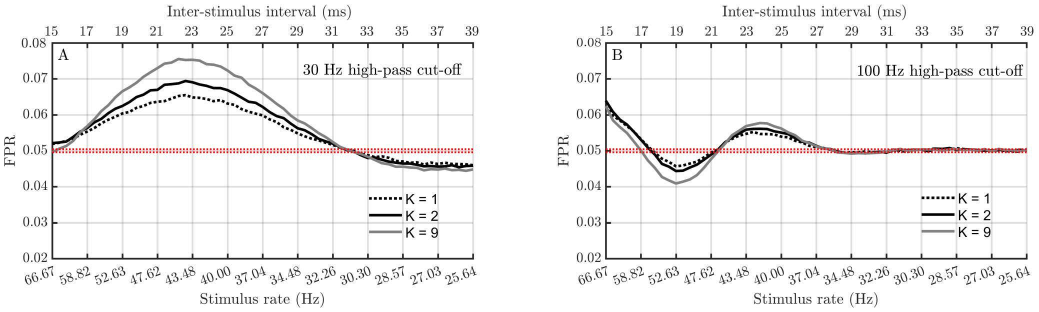

Figure 1: FPRs generated by the Hotelling's T^2 test when applied to simulated coloured noise, as a function of the (hypothetical) stimulus rate, when using band-pass filter settings of either 30-1500 Hz (plot A) or 100-1500 Hz (plot B). Results are presented for $K = 1$ (giving a single shot test), $K = 2$, and $K = 9$. Each FPR was generated from 1 000 000 simulated tests using an ensemble size of $N = 500$, split equally across the K stages.

Table 1: The FPRs (using $\alpha = 0.01$) generated by the Hotelling's T^2 test when applied to the recordings of EEG background activity for different N and K. The recordings were pre-processed using $f_c = 100$ Hz and a (hypothetical) stimulus rate of 33.11 Hz. Significant deviations ($\boldsymbol{p < 0.05}$) from α are indicated by asterisks.

	Futility stopping permitted			**Futility stopping not permitted**		
	N = 300	**N = 500**	**N = 1000**	**N = 300**	**N = 500**	**N = 1000**
K=1	-	-	-	0.0101	0.0102	0.0088
K=2	0.0115	0.0079	0.0088	0.0110	0.0074	0.0088
K=3	0.0090	0.0130	0.0069	0.0090	0.0125	0.0079
K=4	0.0077	0.0130	0.0039 $*$	0.0082	0.0121	0.0059
K=5	0.0079	0.0093	0.0088	0.0085	0.0093	0.0088
K=6	0.0088	0.0111	0.0088	0.0077	0.0116	0.0079
K=7	0.0088	0.0056 $*$	0.0079	0.0079	0.0051 $*$	0.0079
K=8	0.0088	0.0097	0.0079	0.0096	0.0093	0.0079
K=9	0.0090	0.0097	0.0098	0.0082	0.0097	0.0098

the no-stimulus condition; when a response was absent, and early stopping for futility was *not* permitted (plot D), then the *mean* test time is increased with K. In particular, the *mean* test time is close to the *maximum* test time, as the test will proceed to the final stage of the analysis in $(1 - \alpha)$x100% of the cases. When early stopping for futility is permitted, then the increased mean test time for the no-stimulus condition is greatly reduced (plot C). Finally, it is worth noting that early stopping for futility had no noticeable effect on the stimulus condition; results presented in plots A and B were generated *with* the option to stop early for futility.

Simulations II: N = 3000

Results from Simulations II (N fixed at 3000) are also presented in Figure 2 (plots E, F, G, and H). Note again that for these simulations, the reduced statistical power for increasing K cannot be compensated for by increasing N. Coincidentally, a reduced TPR is observed for increasing K (plot G). The decrease in mean test time for increasing K (plots E and F) was now also more pronounced; Reductions in test time of up to 50-60% are observed for $K = 4, 5$, relative to $K = 1$. Early stopping for futility furthermore had no noticeable effect on the TPR (results from plot G were generated *with* the option to stop early for futility), but had a small effect on mean test times, most notably for the 20 dB SL condition (compare plots E and F). With respect to the no-stimulus condition (plot H), mean test time was more or less constant across K when early stopping for futility was *not* permitted, i.e. the test reached the final stage of the analysis (and the full $N = 3000$ epochs were analysed) in $(1 - \alpha)$x100% of the cases, regardless of the choice for K. When early stopping for futility was permitted, mean test time was reduced, up to around 45% for $K = 9$ relative to $K = 1$.

Subject ABR data

Results from the subject ABR data are presented in Figure 3. Similar to Simulations II, N was fixed at 3000 epochs. Consequently, the reduced statistical power for increasing K cannot be compensated for by increasing N, and a reduced test sensitivity for increasing K is observed (plot A). Large reductions in mean test times are also observed for increasing K (plots B and C); up to 45-55% for $K = 4$ (relative to $K = 1$). The option to stop early for futility again had no noticeable effect on test sensitivity (results presented in plot A were generated *with* the option to stop early for futility), but reduced the mean test time, most notably so for the the 0, 10, and 20 dB SL conditions (compare plots B and C). Similar to Figure 2, K values above $\sim$4 do not lead to further reductions in mean test times, but still tend to reduce test sensitivity.

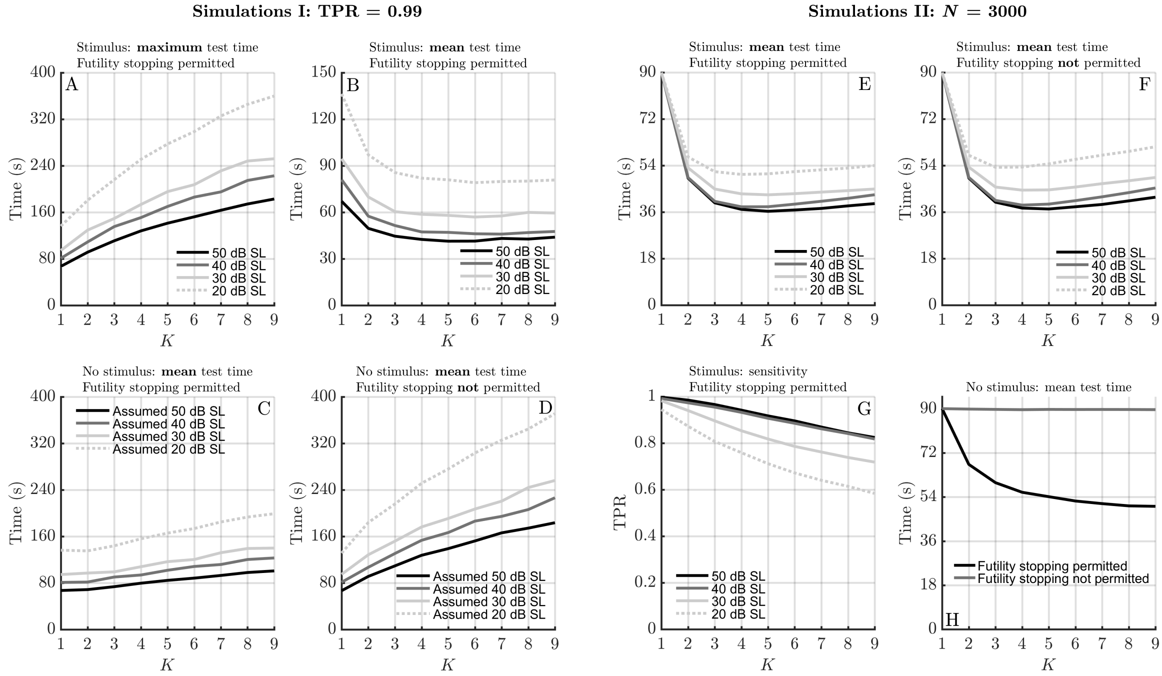

Figure 2: Simulation results for exploring the trade-offs between statistical power and test time, as a function of the number of sequential stages K used for the analysis. Note again that K essentially functions as a surrogate measure for when (and how often) to apply the statistical test to the data. Results from simulations I (the TPR was fixed at 0.99) are shown on the left: the maximum and mean test times for the stimulus condition are shown in plots A and B, respectively, whereas the mean test times for the no-stimulus condition are shown in plots C (early stopping for futility is permitted) and D (early stopping for futility is not permitted). Results from Simulations II (the ensemble size N was fixed at 3000 epochs) are shown on the right: The mean test times for the stimulus condition are shown in plots E (early stopping for futility is permitted) and F (early stopping for futility not permitted). Plot G then shows the TPR for the stimulus condition, and plot H shows the mean test times for the no-stimulus condition. Results demonstrate that when the maximum test time is fixed at $N = 3000$ epochs (plots E, F, G, and H), that the number of stages K should be kept low (<4 stages). On the other hand, if a reduced test sensitivity can be prevented by increasing N with K (plots A, B, C, and D), then the number of stages K should ideally be around 5 or 6. Results are further discussed in the text.

4 Discussion

This study explored the specificity, sensitivity, and test time of a new sequential test procedure for ABR detection. The approach was built around the Convolutional Group Sequential Test (CGST), which was used to find the stage-wise critical decision boundaries (for accepting or rejecting H_0) and control the FPR of a sequentially applied Hotelling's T^2 test. The Hotelling's T^2 test was chosen as the detection method as it has previously shown a good performance for ABR detection (Chesnaye et al, 2018). It is however worth emphasizing that the CGST can be used in combination with a wide range of statistical tests, under the condition that the tests's underlying assumptions are satisfied. The CGST furthermore requires various parameters to be specified *a priori*, which include the number of sequential stages for the analysis K, the nominal α-level of the test, the stage-wise FPRs α_i, and the γ_i values (the true-negative rates). These parameters introduce trade-offs between statistical power and test time, and should be chosen carefully. In the discussion that follows, the specificity of the CGST (and the Hotelling's T^2 test) is considered in more detail, along with the trade-offs between statistical power and test time underlying the aforementioned CGST design parameters.

It is however worth noting that an alternative sequential test procedure has previously been described for Amplitude-Modulation Following Response (AMFR) data by Stürzebecher et al (2005). In Stürzebecher et al (2005), data is continuously pooled in a single ensemble, which is analysed using a statistical test at various pre-determined time intervals. The critical decision boundaries for rejecting H_0 and controlling the FPR are then found *a priori* using Monte-Carlo simulations. The approach was later optimised in terms of statistical power and test time by Stürzebecher & Cebulla (2013) and Cebulla & Stürzebecher (2015). Stürzebecher & Cebulla (2013) optimised the approach in terms of the choice for the stage-wise critical decision boundaries, which essentially determines how statistical power accumulates throughout the sequential analysis. By analogy with the CGST, the latter would correspond to how the available α is spread across the K stages, i.e. the choice for the α_i values.

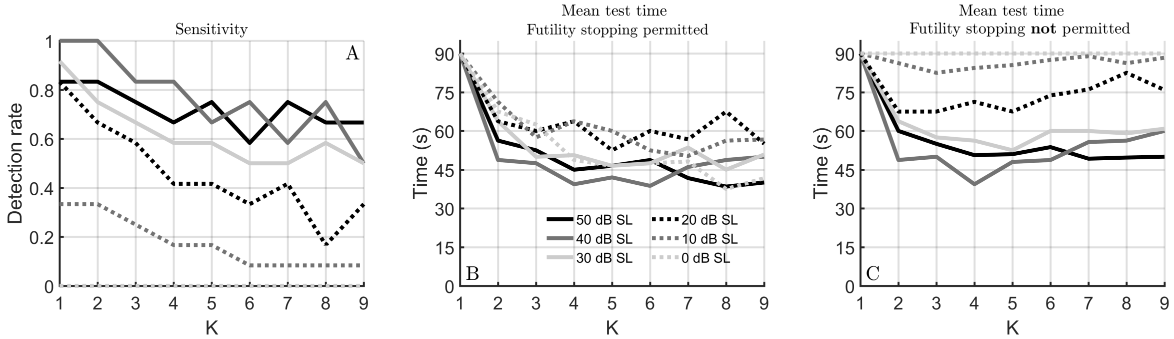

Figure 3: Results from the subject recorded ABR threshold series. Plot A: the detection rate, as a function of K, per dB SL condition. Results were generated *with* the option to stop early for futility. The mean test times (taken across 12 subjects) as a function of K, per dB SL condition, are shown in plots B (early stopping for futility permitted) and C (early stopping for futility not permitted). Note again that K essentially functions as a surrogate measure for when (and how often) to apply the statistical test to the data.

Perhaps the most important difference between the CGST and the sequential test in Stürzebecher et al (2005) is the way in which data is analysed. In the CGST, data is analysed in independent blocks of observations and p values are pooled, whereas in Stürzebecher et al (2005), data is pooled in a single ensemble, which is continuously re-analysed at pre-determined time intervals. A potential advantage of analysing data in blocks is that the stationarity assumption underlying most ABR detection methods is relaxed, i.e. data is assumed to be stationarity within each block, as opposed to across the full recording. Similarly, the evoked response is assumed to be deterministic within each block, as opposed to the full recording. The caveat with analysing data in independent blocks is of course the requirement that the blocks of data are indeed independent. As mentioned in the results section, block-wise dependence was not found to be an issue in the current work, although dependence between epochs was. It is also worth noting here that the problem of dependent epochs also applies to the approach in Stürzebecher et al (2005).

Finally, a convenient property of the CGST is that it permits data-driven adaptations to test parameters following each stage of the sequential analysis. In particular, all previously analysed data can be used to estimate additional parameters of interest (e.g. an effect size or a sample variance), which can then be considered when redesigning the remaining stages of the sequential analysis. Test parameters that can be modified following each stage of the analysis include the ensemble size, the statistical test and features, and pre-processing parameters. Note that data-driven adaptations were not explored in this work, but will be considered in future studies when further optimising sequential test procedures. For more on the CGST and the type of adaptations permitted, the reader is referred to the discussion in Chesnaye et al (in press).

4.1 Specificity

This study demonstrated significant violations to the independence assumption between epochs (underlying the Hotelling's T^2 test), as a function of the high-pass cut-off frequency and stimulus rate (Figure 1). This violation resulted in non-uniform p value null distributions, which resulted in an additional violation (now originating from the CGST) of the assumption that the stage-wise p values are uniform on the [0,1] interval under H_0. Results suggest that certain combinations of the high-pass cut-off frequency and stimulus rate are safe (i.e. the FPR $\approx \alpha$), e.g. a high-pass cut-off frequency of 100 Hz and a stimulus rate of $\sim$33.11 Hz, which were the adopted values for all subsequent analyses in this study.

With respect to the recorded EEG background activity, results demonstrate that the FPRs mostly fell within the 95% confidence intervals for the expected 0.01 FPR, although a conservative test performance was observed for 3 of the 51 test conditions. The latter might be attributed to random variation, as multiple (albeit correlated) tests were performed. Alternatively, the underlying statistical assumptions might have again been violated. Note that if the independence assumptions were satisfied (by using a high-pass cut-off frequency of 100 Hz and a 33.11 Hz stimulus rate), that this leaves just the normality and the stationarity assumptions underlying the Hotelling's T^2 test. Additional simulations (results not presented) indeed demonstrate a minor tendency towards a conservative test performance for the single shot ($K = 1$) Hotelling's T^2 test when stationarity is violated. A conservative test performance is similarly observed for normality violations, under the condition that the violation is due to excessive kurtosis.

To summarise, results from the specificity assessment emphasize that care is required to ensure that the assumptions underlying the chosen ABR detection method are satisfied, else additional violations (originating from the CGST) might be introduced. The choice for the ABR detection method is therefore important, as some methods have a more robust control of specificity relative to others. The Hotelling's T^2 test, for example, is considered to have a good control of specificity relative to methods such as the widely used Fsp (Elberling and Don, 1984) and Fmp (Martin et al., 1994). In particular, the Fsp and the Fmp require the degrees of freedom (DOF) of the data to be assumed *a priori*, which is problematic for EEG data where the DOF of the data are known to vary across recordings (Chesnaye et al., 2018). When the assumed DOF are higher or lower than the true DOF, a conservative or liberal test performance can be expected. Consequently, the stage-wise p value null distributions will no longer be uniform on the [0,1] interval under H_0, as is assumed by the CGST. It can therefore be expected that, when used in combination with the CGST, a sequentially applied Fsp or Fmp will have a relatively poor specificity. A potential solution to the unknown DOF of EEG data is to evaluate the significance of the Fsp or the Fmp with a bootstrap approach (as opposed to using theoretical F-distributions), as this does not require the DOF of the data to be assumed (Lv et al, 2007; Chesnaye et al., 2018). Bootstrapped statistics in general are expected to give a good control of specificity, particularly so as they are robust to normality and stationarity violations.

4.2 Sensitivity, test time, and the CGST design parameters

This study explored the trade-off between statistical power and test time for ABR detection as a function of the number of sequential stages K used for the analysis. Simulation results firstly demonstrate reductions in *mean* test time of up to 45%, with no loss in test sensitivity (Figure 2.B). In order to achieve this, the reduced statistical power for increasing K needs to be compensated for by increasing N, i.e. the *maximum* test time needs to be increased (Figure 2, plot A). If N cannot be increased with K (due to e.g. an upper limit of, say, 3000 epochs), then a reduced TPR can be expected for increasing K. The latter was confirmed with both simulations (Figure 2, plot G) and subject data (Figure 3, plot A), and was found to be most prominent when the single shot test ($K = 1$) was already under-powered (i.e. for the 10 and 20 dB SL conditions of the subject ABR data; Figure 3, plot A). Based on the preceding results, the following rough guidelines might be used for choosing K: if the single shot test ($K = 1$) is expected to be under-powered, and N cannot be increased, then it may be beneficial to keep the number of sequential stages for the analysis low, e.g. 1, 2, or 3 stages might be used. If the single shot test is expected to be over-powered, or if N can be increased with K, then a more efficient approach is to use 4, 5, or 6 stages for the analysis.

CGST design parameters
Additional parameters underlying the trade-off between statistical power and test time that have not yet been discussed include the stage-wise type-I error rates α_i, the stage-wise ensemble sizes, say N_i (for $i = 1, 2, ..., K$), and the γ_i values (i.e. the stage-wise fraction of tests to be rejected in favour of H_0, under H_0). For this study, α and N were always split equally across the K stages, giving α_i values of $\frac{\alpha}{K}$ and N_i values of $\frac{N}{K}$ for all i and K. The latter was based on additional simulations (details not presented), which demonstrate a good test sensitivity when using these settings. Note however that if a larger SNR is expected for some stages of the analysis relative to others, then increasing the α_i values for these stages would be beneficial. It is also worth emphasizing here that although the α_i values are chosen freely (under the condition that $\sum_{i=1}^{K} \alpha_i = \alpha$), they should be specified by the user *a priori*, i.e. prior to looking at the data.

Finally, with respect to the γ_i values, a trade-off is again introduced between statistical power and test time; a more liberal choice for the γ_i values increases the probability of stopping the test early for futility (thus decreasing test time), potentially at the cost of an increased false-negative rate (a reduced test sensitivity). Note that the trade-off between statistical power and test time underlying the γ_i values is also strongly dependent on the SNR of the response, along with the stage-wise ensemble sizes N_i. In particular, for low SNRs or small N_i, the Σ_k test statistic (see Eq. 3) will tend to be relatively small, i.e. it will tend to be closer to stage k critical boundary C_k (for accepting H_0), and the probability of stopping the test early in favour of H_0 is increased. In a nutshell, the expected or estimated distribution of Σ_k under the alternative hypothesis (an ABR is present) should be considered when choosing the γ_i values. Alternatively, a conservative choice can be adopted by using $\frac{1-\alpha}{K}$ for all γ_i. Note that this is a conservative choice only under the condition that the test is not strongly under-powered due to e.g. N being too small, or the SNR being too low.

4.3 Study limitations

An important result from this study was that the *mean* test time for the sequential test was reduced relative to the single shot test, with no loss in test sensitivity (Figure 2.B). These results were, however, not reproduced in the subject ABR data, as the *maximum* test time for the subject ABR data in this study was too low (N was limited to 3000 epochs). The subject ABR data (initially described in Lv et al, 2007) was indeed not collected with a view to explore and optimise the test performance of a sequentially applied statistical test. In future work, the approach should be tested and optimised using a much larger cohort of test subjects, potentially with a range of hearing impairments. When doing so, the *maximum* test time should be sufficiently high.

5 Conclusion

This study explored the specificity, sensitivity, and test time of a sequentially applied Hotelling's T^2 test with critical decision boundaries (for accepting or rejecting H_0) constructed by the CGST. With respect to specificity, results show that the FPR was controlled as intended, under the condition that the underlying statistical assumptions are satisfied. With respect to sensitivity and test time, results show relatively large reductions in *mean* test time for the sequential test (relative to the single shot test) of up to 40-45%, with no loss in test sensitivity. In order to achieve these results, the *maximum* test time needs to be increased, else a reduced test sensitivity can be expected. When used in clinical practice, it can therefore be expected that the sequential test will occasionally prolong test time (relative to the single shot test), but will tend to decrease the mean test time when considered across a cohort of test subjects.

Acknowledgments: The authors would like to thank Sara M. K. Madsen for collecting the recordings of EEG background noise and Deborah A. Cane for collecting the ABR threshold series. The authors would also like to acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. The data used throughout this study is openly available at the University of Southampton repository at http://doi.org/10.5258/SOTON/D0168.

Declaration of interest: The authors report no conflicts of interest. This work was supported by the Oticon Foundation and the Engineering and Physical Sciences Research Council ([EPSRC, grant No. EP/M026728/1)].

References

[1] Armitage P., McPherson C.K. & Rowe B.C. (1960). Repeated Significance Tests on Accumulating Data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2), pp. 235-244.

[2] Arnold S.A. (1985). Objective versus visual detection of the auditory brainstem response. *Ear and Hearing*, 6(3), pp 144-150.

[3] Bauer P. & Köhne K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50(4), pp. 1029-1041.

[4] Brannath W, Posch M., and Bauer P., Recursive combination tests, *J. Am. Stat. Assoc.*, vol. 97, no. 457, pp. 236-244, Mar. 2002.

[5] Cebulla M. & Stürzebecher E. (2015). Automated auditory response detection: Further improvement of the statistical test strategy by using progressive test steps of iteration. *International Journal of Audiology*, 54(8), pp. 568-572.

[6] Chang M., Adaptive design method based on sum of p-values, *Statist. Med.*, vol. 26, no. 14, pp. 27722784, Nov. 2006. DOI: 10.1002/sim.2755

[7] Chesnaye M.A., Bell S.L., Harte J.M. & Simpson D.M. 2018. Objective measures for detecting the Auditory Brainstem Response: comparisons of specificity, sensitivity, and detection time. *Int. J. Audiol.*, 57(6), pp. 468-478. DOI: 10.1080/14992027.2018.1447697

[8] Chesnaye M.A., Bell S.L., Harte J.M. & Simpson D.M. in press. The Convolutional Group Sequential Test; reducing test time for evoked response detection. *IEEE Transactions on Biomedical Engineering.*

[9] Elberling C., and Don M. 1984. Quality Estimation of Averaged Auditory Brainstem Responses. *Scandinavian Audiology* 13(3), pp. 187197. doi:10.3109/01050398409043059

[10] Fisher R. A. 1932. Statistical methods for research workers, 11th ed. Oliver and Boyd, Edinburgh.

[11] Grinstead C. M. and Snell J. L. , Chapter 7, in Introduction to Probability, 2nd ed., American Mathematical Society, the United States of America, Aug. 1997, pp. 285.

[12] Hartung J., and Knapp G., A new class of completely self-designing clinical trials, *Biometrical J.*, vol 45, no. 1, pp. 319, Jan. 2003. DOI: 1002/bimj.200290014

[13] Hotelling H. 1931. The Generalization of Student's Ratio.*Ann. Math. Statist.*, 2(3), pp. 360-378.

[14] Littell R. C. and Folks J. L., Asymptotic Optimality of Fishers Method of Combining Independent Tests, J. Am. Stat. Assoc., vol. 66, no. 336, pp. 802-806, Dec. 1971.

[15] Lehmacher W. and Wassmer G., Adaptive sample size calculations in group sequential trials, *Biometrics*, vol. 55, no. 4, pp. 1286-1290, Dec. 1999

[16] Liu Q. and Chi G. Y. H., On sample size and inference for two-stage adaptive designs, *Biometrics*, vol. 57, no. 1, pp. 172-177, Mar. 2001. DOI: 10.1111/j.0006-341X.2001.00172.x

[17] Lv J., Simpson D.M. & Bell S.L. (2007). Objective detection of evoked potentials using a bootstrap technique. *Medical Engineering & Physics*, 29(2), pp. 191198.

[18] Madsen S.M.K., Harte J.M., Elberling C. & Dau T. (2017). Accuracy of averaged auditory evoked potential amplitude and latency estimates. *International Journal of Audiology*, 57(2), pp. 1-9.

[19] Marple S.L.Jr. *Digital Spectral Analysis with Applications.* Prentice-Hall, Englewood Cliffs, NJ, 1987.

[20] Martin W.H., Schwegler J.W., Gleeson A.L., and Shi Y.B. 1994. New Techniques of Hearing Assessment. *Otolaryngologic Clinics of North America*, 27(3), pp. 487510.

[21] Müller H.H. and Shäfer H., Adaptive Group Sequential Designs for Clinical Trials: Combining the Advantages of Adaptive and of Classical Group Sequential Approaches, *Biometrics.*, vol. 57, no. 3, pp. 886891, Sep. 2001. DOI: 10.1111/j.0006-341X.2001.00886.x

[22] Pritchard, W. 1992. The Brain in Fractal Time: 1/f-like Power Spectrum Scaling of the Human Electroencephalogram. *International Journal Neuroscience*, 66 (12): 119129. doi:10.3109/00207459208999796.

[23] Proschan M.A. and Hunsberger S.A, Designed extension of studies based on conditional power, *Biometrics*, vol. 51, no. 4, pp. 1315-1324, Dec. 1995.

[24] Rencher A.C. Methods of Multivariate Analysis. Second Edition. John Wiley & Sons, Inc., 2001.

[25] Sheng J. and Qiu L., p-Value calculation for multi-stage additive tests, *J. Stat. Comput. Sim.*, vol. 77, no. 12, pp. 10571064, Nov. 2007. DOI: 10.1080/10629360600872707

[26] Stürzebecher E. & Cebulla M. (2013). Automated auditory response detection: Improvement of the statistical test strategy. *International Journal of Audiology*, 52(12), pp. 861-864.

[27] Stürzebecher E., Cebulla M. & Elberling C. (2005). Automated auditory response detection: Statistical problems with repeated testing. *International Journal of Audiology*, 44(2), pp. 110-117.

[28] Sutton G., Lightfoot G., Stevens J., Booth R., Brennan S., Feirn R. & Meredith R. (2013). Guidance for Auditory Brainstem Response Testing in Babies. Version 2.1. Reading, UK: British Society of Audiology.

[29] Vidler M. & Parker D. (2004). Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test. *International Journal of Audiology*, 43, pp. 417-429.

[30] Welch P.D. (1967). The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms (PDF). *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-15(2), pp. 7073. DOI:10.1109/TAU.1967.1161901