Using correlation analysis to assess the reliability of evoked potential components identified by signal averaging

Jackie Campbell[a] **and Massimo Leandri**[b]

[a] Faculty of Health, Education and Society, University of Northampton, University Drive, Northampton, NN1 5PH, UK. jackie.campbell@northampton.ac.uk (corresponding author)
[b] Dipartimento di Neuroscienze (DINOGMI), Università degli Studi di Genova, Largo P. Daneo, 3 I-16132, Italy. massimo.leandri@unige.it

Research article

Highlights

- Evoked potentials are conventionally recorded and analysed using signal averaging
- Averaging can produce misleading results if the conditions are not ideal
- Using pairwise correlation coefficients produces a good measure of signal reliability
- Reliability measures displayed on the averaged trace show the reliable components
- This technique shows the reliability of averaged components and identifies artefacts

1 Introduction

Event related potentials, or evoked potentials (EPs), recorded from over the human cortex have been widely used for many decades as clinical and research tools to investigate the neural pathways associated with different sensory modalities. Their use relies on being able to identify the electrical potentials that are time-locked to the sensory stimuli being presented in the presence of ongoing, unrelated electrical activity. This was first achieved in 1947 using photographic superposition (Dawson, 1947) which became the basis of the signal averaging technique that is still in common use today.

Signal averaging is a simple process and produces outputs which are intuitive and easy to understand. It assumes that samples of both evoked potential (signal) and unrelated activity (noise) are random variables with the expected value of the noise at any given time being zero and the expected value of the signal being the ‘true’ signal value at that point (Collura, 1995). Although each recorded sample, following a single stimulus, will contain the full amount of both signal and noise, the averaged response from successive stimuli will result in decreasing noise with increasing numbers of samples, whilst the signal will remain the same. With an infinite number of samples, the noise component would be zero and recording would be an accurate representation of the signal component. However, it has long been known that the assumptions required for this form of analysis are not met. These are considered in detail in Collura (1995). For averaging to be fully effective the characteristics of both signal and noise should not vary within the recording session, separate components of the signal should be independent of each other, each signal component should have a fixed time relationship to the stimulus, each sample should contain an identical signal and the noise should be completely uncorrelated with the stimulus. In practice, these conditions are not met for a variety of reasons: (i) there can be changes during a recording session due to fatigue, movement, habituation etc, (ii) neurophysiological processes such as facilitation or entrainment can produce non-linear interactions between the signal components and introduce variation in the stimulus response time and, (iii) for stimuli delivered at fixed frequency, there can be noise correlation with extraneous (e.g. power supply frequencies) or intrinsic (e.g. EEG frequencies) sources.

There have been many alternative approaches to the analysis of evoked potential recordings suggested in the light of the potential problems with averaging and, in particular, the desire to reduce the total number of stimuli that need to be administered in order to get a reliable estimate of the underlying signal. Reducing the number of signal samples needed addresses in part the problems of non-stationarity but is also important in the increasing interest in nociceptive evoked potentials where each stimulus is, by definition, unpleasant and where habituation of the response may be an issue. Analysis of evoked potentials has been proposed using wavelets (Quiroga, 2000; Turner et al., 2003), component analysis (Bromm and Scharein, 1982; Aunon et al., 1981), cross-correlation (Patel et al., 2015), variance analysis (Don et al., 1984), pattern recognition (Moser and Aunon, 1986; McGillem et al., 1981) and median averaging (Aunon et al., 1981), amongst others. However, despite a fairly long history of alternative approaches, none of these techniques has found widespread support.

An alternative approach to using a different analytical technique is to be able to assess the reliability of the averaged evoked potential in given situations. This knowledge could then be used to guide research or clinical conclusions by differentiating between components of a recording which are consistent and repeatable and those which are not. Some authors report ways of calculating and reporting internal consistency and suggest novel, more consistent evoked potential parameters (Thigpen et al., 2017) or suggest guidelines for reliability measurement and reporting (Clayson and Miller, 2017). However, these techniques require an entirely different approach to the interpretation of evoked response which is historically on the detection and measurement of individual component latencies and amplitudes. The approach that this paper takes is one of recognising the features of an evoked response that are in common use and giving the researcher or clinician additional information about the reliability of individual components. It recognises the need for individual users to be able to tailor the latency windows of interest for their application.

2 Methods

Somatosensory evoked potentials (SEPs) and electroencephalographic (EEG) recordings were made, with informed consent, using one of the authors (JC) as the subject to obtain genuine biological samples for use in the computer simulations. SEPs were obtained after surface stimulation of the median nerve at the right wrist, with 0.2 ms electric pulses delivered at the rate of 0.83/s and intensity set at the threshold for thumb twitch. Recordings were performed from C3'-Fz scalp derivation with surface electrodes. The same electrodes were used to record equivalent time epochs of EEG spontaneous activity, without any stimulation (called SEP noise later in the text) . Amplification of x 100,000 was used with a bandpass of 0.1-2000 Hz, using 2nd order Butterworth analog filtering (LT amplifiers by Vertigo, Genova, Italy). Signals were then sent to an analog to digital converter (NI PCIe-6320, X Series Multifunction DAQ, 16 Bit, 250 KS/s sampling rate by National Instruments, Austin, Texas). Software was developed using LabView 2017® (National Instruments, Austin, Texas) to acquire 10,000 samples for a period of 1000 ms after each stimulus, thus providing a high definition recording with a dwell time of 0.1 ms. Each response was stored onto hard disc for off-line averaging.

The initial analytical approach was based on that of Coppola et al (1978) who used mean pair-wise correlations between EP recording repetitions to assess signal to noise ratio and response variability in visual and auditory evoked potentials.

An unbiased estimate for the signal to noise ratio, $\hat{\alpha}$, is based on the sample correlation coefficient (Coppola et al., 1978) and is given by:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{A}\left(\frac{\boldsymbol{r}}{\mathbf{1}-\boldsymbol{r}}\right) + \boldsymbol{B}$$ **Equation 1**

where r is the sample correlation coefficient between two repetitions of the recording.

A and B in $\boldsymbol{\alpha} = \boldsymbol{A}\left(\frac{\boldsymbol{r}}{\mathbf{1}-\boldsymbol{r}}\right) + \boldsymbol{B}$ Equation 1 are required in order to make $\hat{\alpha}$ an unbiased estimator. They are defined as:

$$\boldsymbol{A} = \mathbf{exp}\,(-\frac{\mathbf{2}}{(\boldsymbol{N}-\mathbf{3})})$$ **Equation 2**

and:

$$\boldsymbol{B = -\frac{1}{2}(1 - A)}$$ **Equation 3**

Coppola et al (1978) averaged the values of $\hat{\alpha}$ obtained in this way (using the mean) from pairs of repetitions (records) to produce a more stable estimate of signal to noise ratio. However they only used 16 pairs from 32 repetitions which does not fully utilise the available data. The number of combinations of r items from a total of n (where order does not matter) is given by:

$$\boldsymbol{C(n,r) = \frac{n!}{r!(n-r)!}}$$ **Equation 4**

so for 32 repetitions, there are 32!/2!(30)! = 32x31/2=496 different pairs. This combination approach will be used in this paper in order to maximise the stability of the signal to noise estimate.

The correlation analysis was performed with purpose-built virtual instruments using LabView 2017. For each EP comprising N repetitions and voltages at i time points (in ms), an analysis window length, Δt, was specified and the unique pairs of repetitions (based on the combinations of 2 from a total of N, $C(2, N)$) identified. The correlation coefficient, r, for each pair of repetitions was calculated, together with the median and inter-quartile range (IQR) for all rs.

The correlation analysis requires a time window to be specified. It then calculates the median value of all of the r values derived from correlating every possible pair of dissimilar records (repetitions) over the specified time window starting at a given time point.

The characteristics of $\hat{\alpha}$ (signal to noise ratio estimate, which for simplicity will be referred to subsequently as S:N) and r (correlation coefficient) as reliability measures were compared using SEP recordings and simulated evoked potentials.

Simulated single evoked potential components were constructed using NI LabView 2017® based on the generation of a sine wave with variable width and -90° phase shift at a given distance along the time axis. The total length of the simulated signal was 1,000 ms with a dwell time of

0.1 ms (data acquisition rate of 10 kHz) for compatibility with the parameters used in the SEP recordings described above. Amplitude and DC shift were constant within the simulation program but adjustable between simulations. Figure 1 shows examples of the simulated signal and recorded SEP noise used in the development and testing of this method. The simulated single EP component (red) has peak latency at 100 ms and width at base of 20 ms and the blue trace shows a single 200 ms sample of the recorded SEP noise.

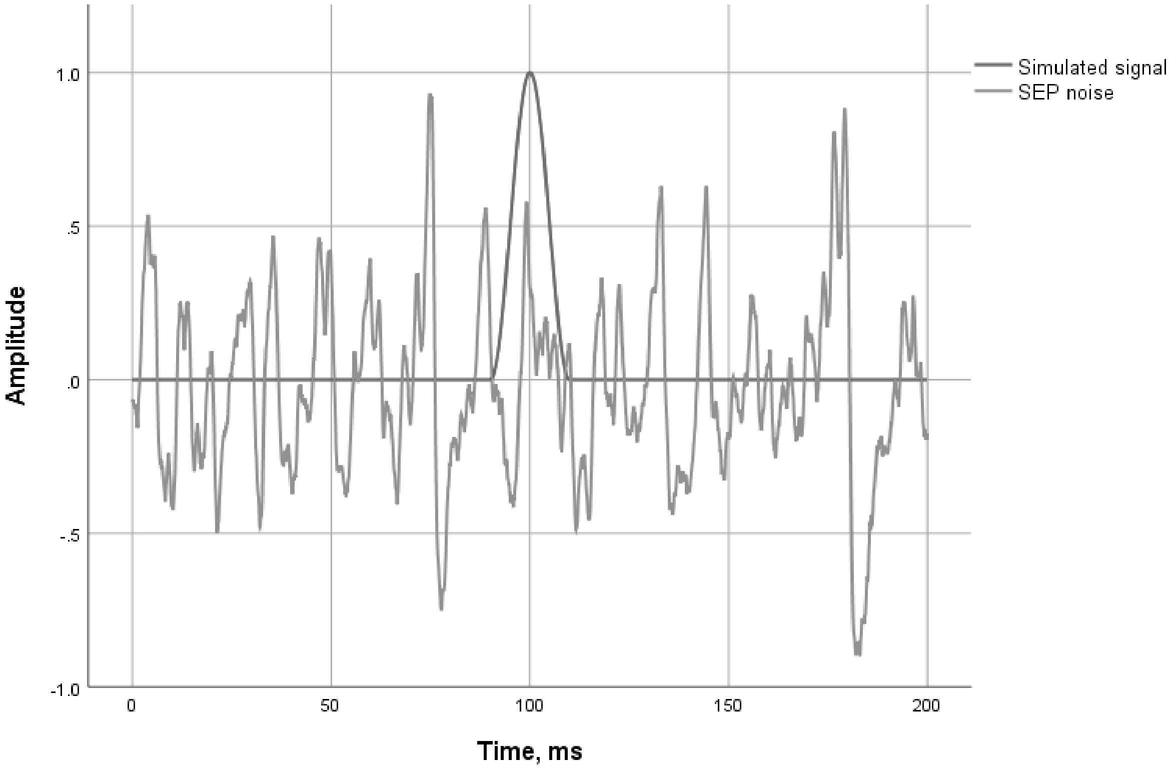


Figure 1: A simulated EP component (peak latency 100 ms, width at base 20 ms) (red trace) superimposed on a single sample of recorded SEP noise (blue trace). Amplitude is shown in arbitrary units.

Complex simulated waveforms were constructed by addition of simulated components with differing amplitude, latency and width. Noise at various amplitudes was added from SEP recordings taken with no stimuli (SEP noise) rather than using simulated noise.

3 Results

The distribution of the correlation coefficients (r) obtained from the set of pairs of repetitions was examined and found to be left (negatively) skewed therefore the average was calculated as the median of each pairwise r, rather than the mean. The median r was recorded for each successive analysis window and displayed as an overlay to the averaged EP in Figures 4-6, below.

3.1 Comparison of signal-to-noise ratio and r as reliability measures

Intuitively, an unbiased estimator of signal to noise ratio (S:N) (see section 2) should be the measure of choice for assessing the reliability of the observed signal in the presence of noise. However , it is based on the quantity r/(1-r) and, as r increases towards 1, it produces increasing large changes for small changes in r. This has the effect of greatly increasing the variance of S:N at higher levels of correlation between the individual evoked potential records.

A comparison between the use of r and S:N was made using the somatosensory evoked potential shown in Figure 2 (see Section 2 for recording details).

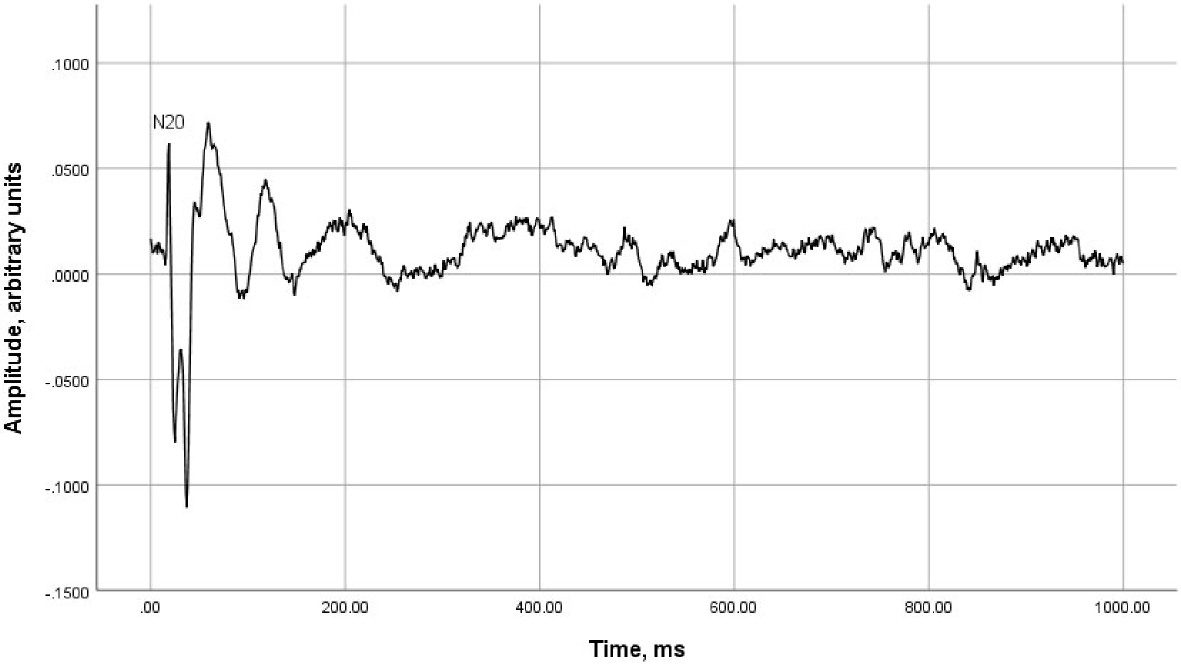


Figure 2: Averaged somatosensory evoked potential (SEP) (right median stimulation, recorded C3-Fz, f=0.83 Hz, 0.2 ms pulse at motor threshold for thenar eminence, recorded at 10 kHz, 256 repetitions). Negative voltages shown as upward deflections in line with neurophysiological recording convention. The first SEP component (N20) is indicated.

Figure 3 shows a comparison of the use of r and S:N for the somatosensory evoked potential illustrated in Figure 2. The median values for each successive 10 ms windows are shown, together with the interquartile ranges (IQR). It can be seen that the variation in the S:N measure is greater than for r where there are known to be stable SEP components (in the first 40 ms post-stimulation). As an example, the median r for the 10-20 ms window is 0.280 with an IQR of 0.553 (IQR:r=1.975) whereas the median S:N for the same window is 0.371 with an IQR of 1.120 (IQR:S:N=3.019). At longer latencies, the uncertainty is larger for r, which reflects the expectation that there is more uncertainty in the later potentials. The general shapes of the median r *v* time and S:N *v* time graphs in Figure 3 are similar, suggesting that r is measuring a similar characteristic to S:N.

As r performs as a good surrogate for the signal to noise ratio and is less variable under stable conditions, the median r will be used for further investigations into a measure for assessing the reliability of evoked potential components.

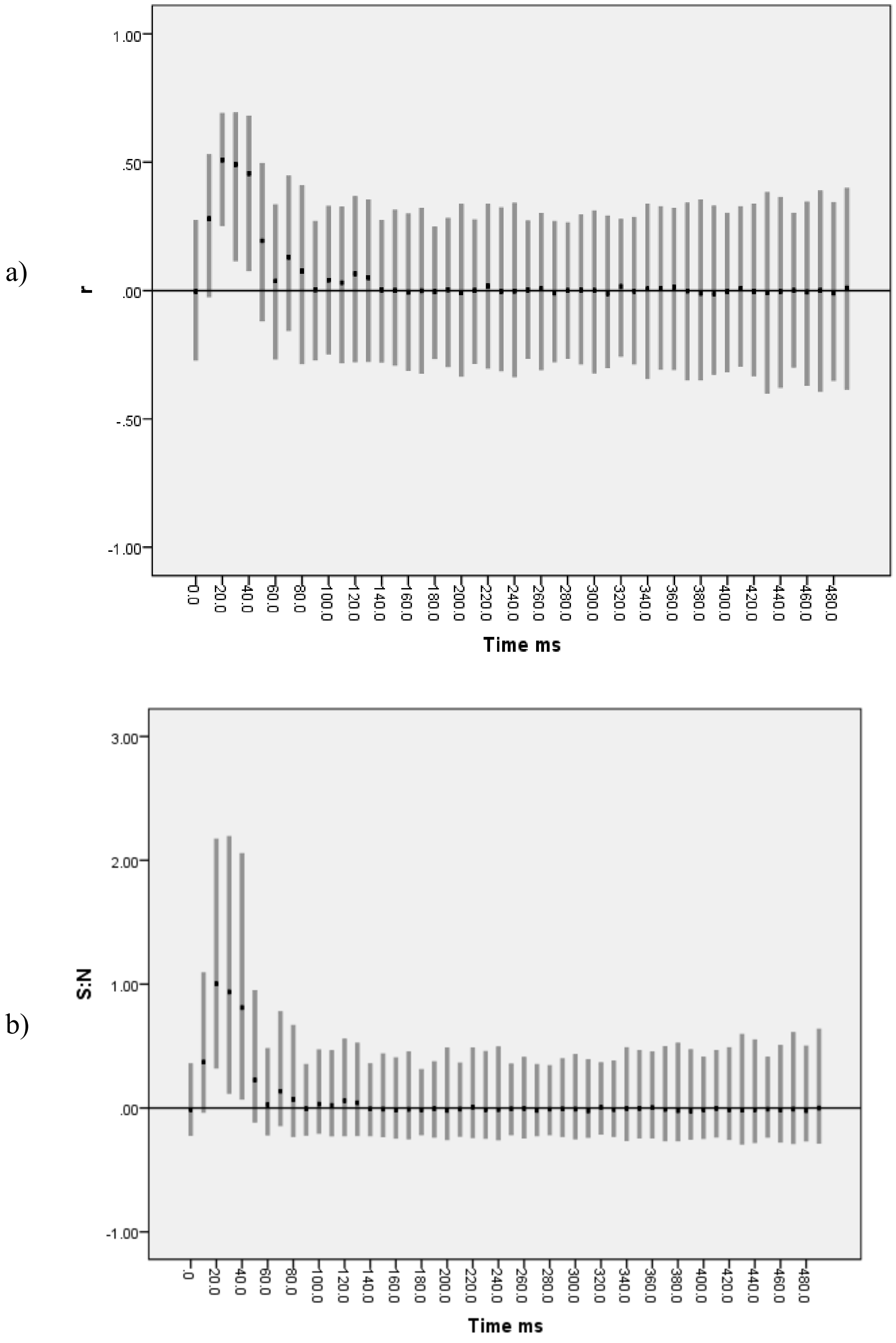


Figure 3: a) Median correlation coefficients (r) and b) signal:noise ($\hat{\alpha}$) calculated for 100 repetitions of the SEP shown in Figure 2, for each window of 10 ms from 0 to 500 ms from stimulation. Median values shown as dots, interquartile ranges as bars.

3.2 Effect of signal amplitude

Correlation coefficients should be a reflection of the similarity of the shape of the signals being compared and should not be affected by differences in amplitude alone. To test this, a single simulated component was created (see Figure 1) with an onset latency of 100 ms, 20 ms width, 0.1 ms dwell time (10 kHz sampling frequency) and total length 1000 ms. Replicates at different amplitudes were created by multiplying this component by 0.1 to 0.9 in 0.1 steps giving a total of 10 components with identical shapes but different amplitudes. The median r for all pairs of replicates were calculated for every 10 ms time window. The median r was 1 for the time windows 100-110 ms and 110-120 ms, which contained the signal centred on 100 ms with a width of 20 ms, and zero elsewhere, confirming that amplitude changes alone do not change the median r.

3.3 Effect of noise amplitude

A single simulated component (width 20 ms, onset latency 90 ms) was constructed and used as the signal element. This was combined with recorded activity from the scalp (C_3-Fz) with no stimulation (see Figure 1). A series of 30 repetitions was constructed with the same signal and 30 successive 1000 ms noise recordings. The amplitude of the simulated component was varied to give differing actual signal to noise ratios. These were then conventionally averaged and analysed to give the median r for 10 ms windows (0-500 ms).

It can be seen from Figure 4 that the median r is high (0.7-0.9) for the width of the simulated signal (90-110 ms) and fluctuates around 0 (±0.08) for the remainder of the record and is an accurate representation of the reliability of the averaged signal for a consistent signal with amplitude less than the ambient noise.

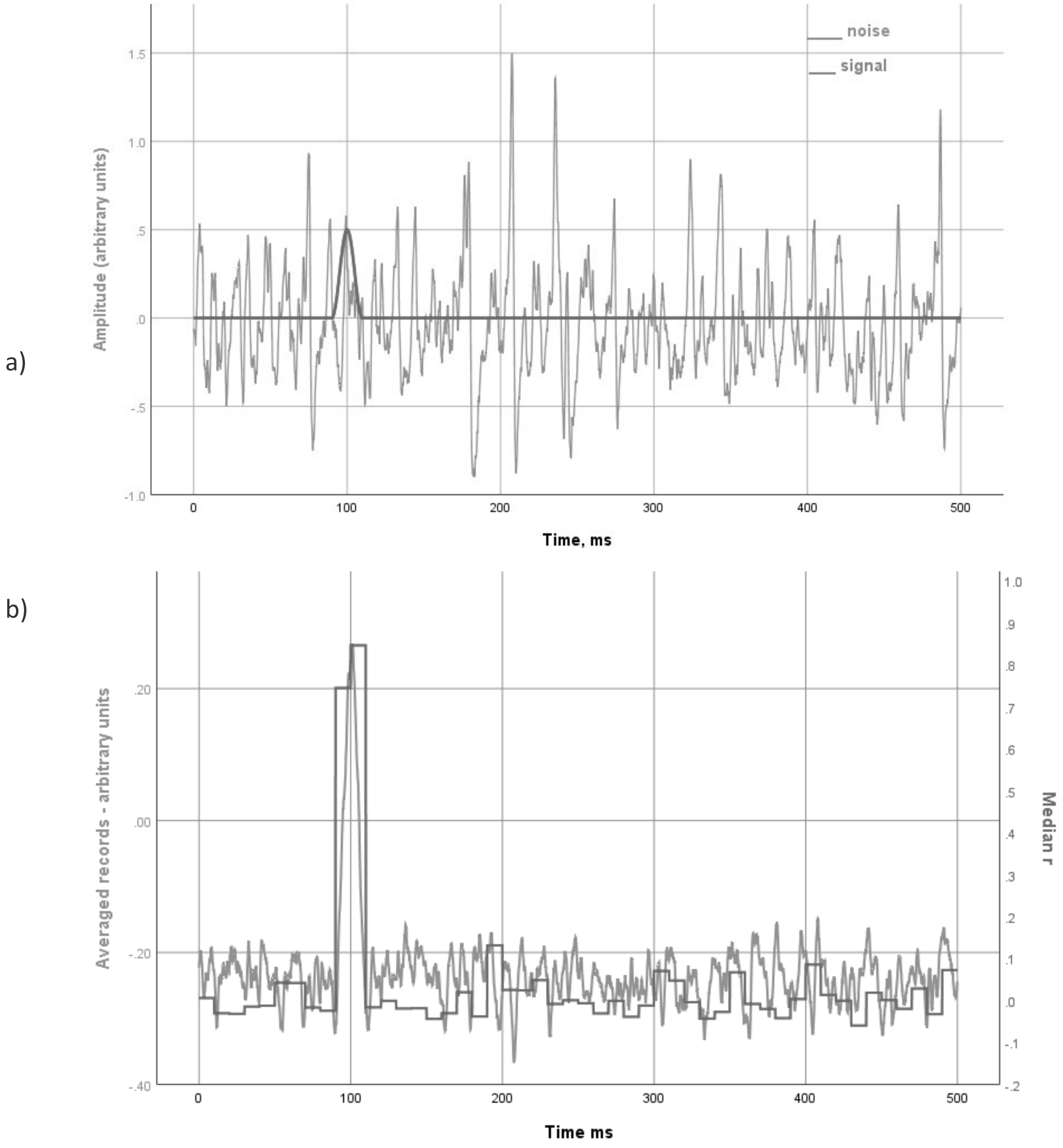


Figure 4. a) Simulated signal (red trace) superimposed on a single record of noise (blue trace) taken with no stimulation during an SEP recording. b) average of 30 records of SEP noise + simulated signal (blue trace) with median r for 10 ms windows (red trace)

3.4 Effect of number of repetitions

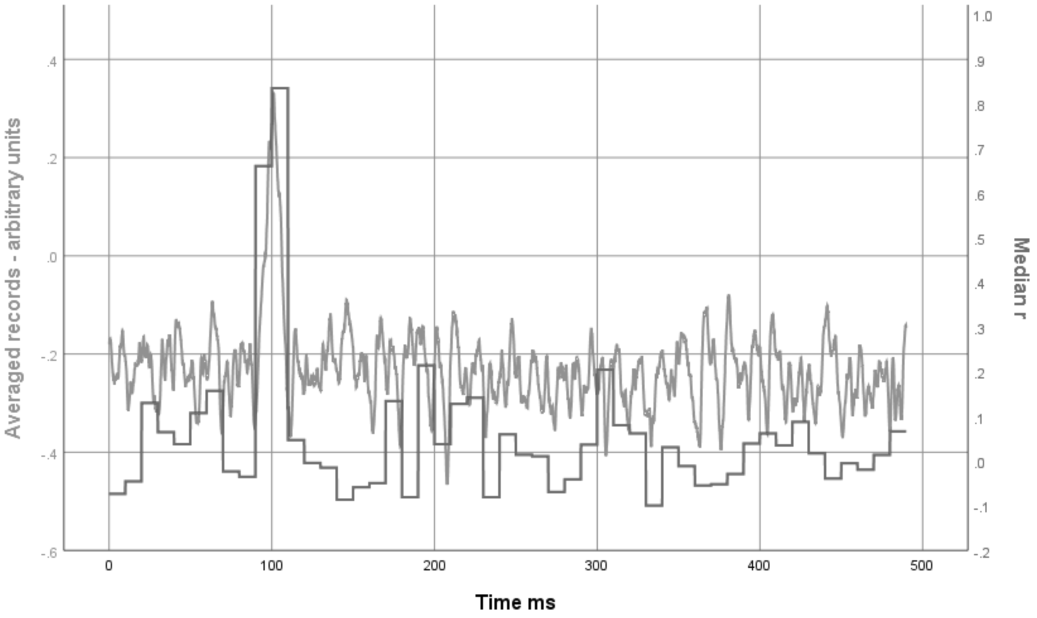
If an additional reliability measure is to be useful in the interpretation of averaged evoked potential components, then it should be able to identify more reliable components when fewer repetitions are available for averaging than would be ideal for the interpretation of averaged signals alone. The simulated signal shown in Figure 4a was combined with SEP noise ($C_{3'}$-F_Z scalp recordings with no stimulation) of similar amplitude. Varying numbers of repetitions were averaged and displayed together with the median r for successive 10 ms windows.

Figure 4b shows that the signal is easily identified when 30 repetitions are averaged with a high reliability (median r=0.8) compared to low reliability elsewhere. Figure 5 shows that the reliability of the signal component remains high (with the same median r) irrespective of the number of repetitions averaged, but the variability of the median r for the noise components is greater as the number of repetitions decreases. With 16 records averaged (Figure 5a), the reliability of the signal component is still clearly higher than the surrounding noise. With only 8 averaged records (Figure 5b), the median r of the observed simulated component is still visibly higher than elsewhere with the noise having a maximum median r of 0.45 and the majority being less than 0.3. It is only when just 4 records are averaged (Figure 5c) that the median correlation coefficients of the noise components become so variable that it becomes difficult to identify the signal component with confidence.

This ability to identify true signal components by locating areas of the averaged potential with visibly higher median r values than their surroundings can be used to minimise the number of repetitions needed. This could be of particular value for situations where the stimuli are unpleasant (for example in nociceptive evoked potentials) or in difficult recording situations. If the median r of all pairs of repetitions could be calculated after each stimulus and displayed in real time, this could be used to identify the point at which the signal can be reliably identified and the recording session ended. This would take into account the individual characteristics of the signal and noise conditions for that recording and minimise the number of stimuli required for a satisfactory recording.

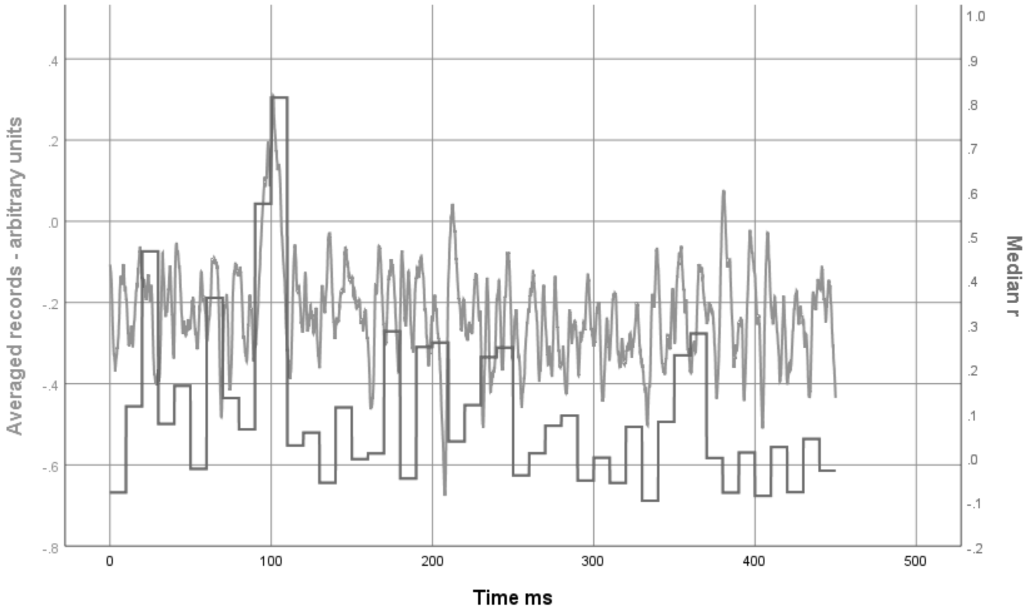
a) 16

records



b) 8

records



c) 4 records

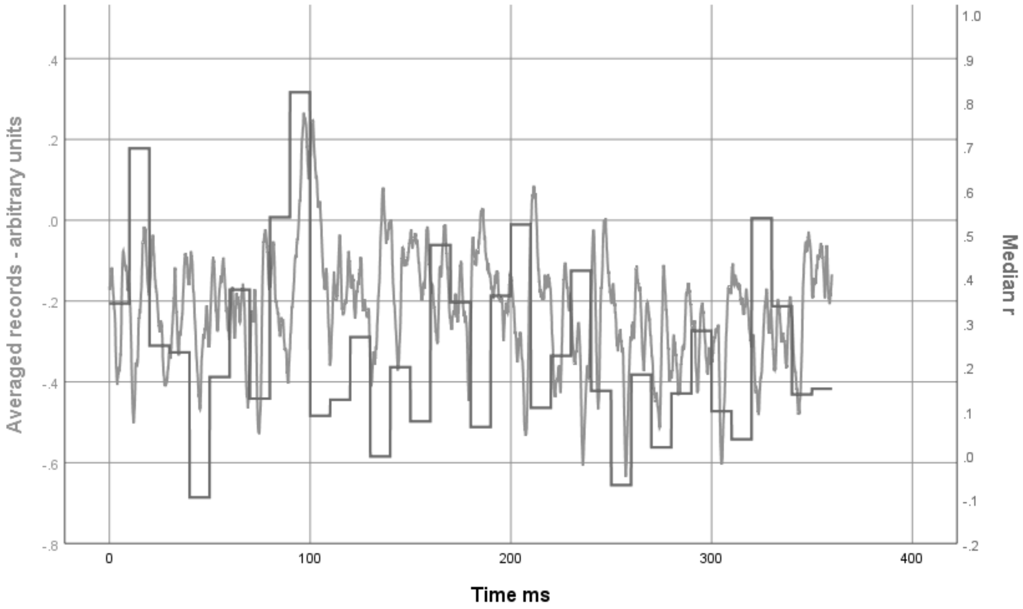


Figure 5: Effect of the number of repetitive records used in averaged recordings using a simulated single component with onset latency 90,s, width 20 ms in recorded SEP noise of similar amplitude. Averaged recordings (blue) are shown together with the median r (red) for successive 10 ms windows.

3.5 Choice of time window width

As the analysis programme is currently configured, a time window is defined by the user prior to analysis and the calculation of the median r values is performed for each successive time window of this width for the whole of the recording. This can be seen in Figure 4b and Figure 5. This approach is adequate and the best performance will be obtained if the window is set at no wider than the narrowest component of interest, so that each element of the signal can be examined individually. For wider components, the values of median r in each of the windows spanning the component can be examined. In future version of the programme, the windows could be tailored to fit each component of interest using varying widths across the time span of the recording.

3.6 Identification of artefacts

Single electrical artefacts which have large amplitude and are unrelated to the events of interest are a problem when interpreting averaged evoked potentials as they can appear to be genuine components. In order to test whether these could be identified and differentiated from true signal components, a simulated EP component was created in the same way as for Figure 1, with a peak latency of 100 ms, width of 20 ms and amplitude of 1 (arbitrary unit). This was replicated 10 times. A single simulated signal of the same shape but with an amplitude of 10 and peak latency of 500 ms was added to the 10th repetition to simulate a large artefact which would look the same as the signal after averaging (but with different latency). Ten successive recordings of 1000 ms of recorded activity from the scalp (C_3-Fz) with no stimulation (EP noise) were then added to the simulated signals, one for each repetition. The peak noise amplitude had a mean amplitude of 1.2 over the 10 repetitions and was therefore about 20% greater than the simulated signal. This combined simulated signal, artefact and noise was then analysed using successive 20 ms windows for the median r calculations.

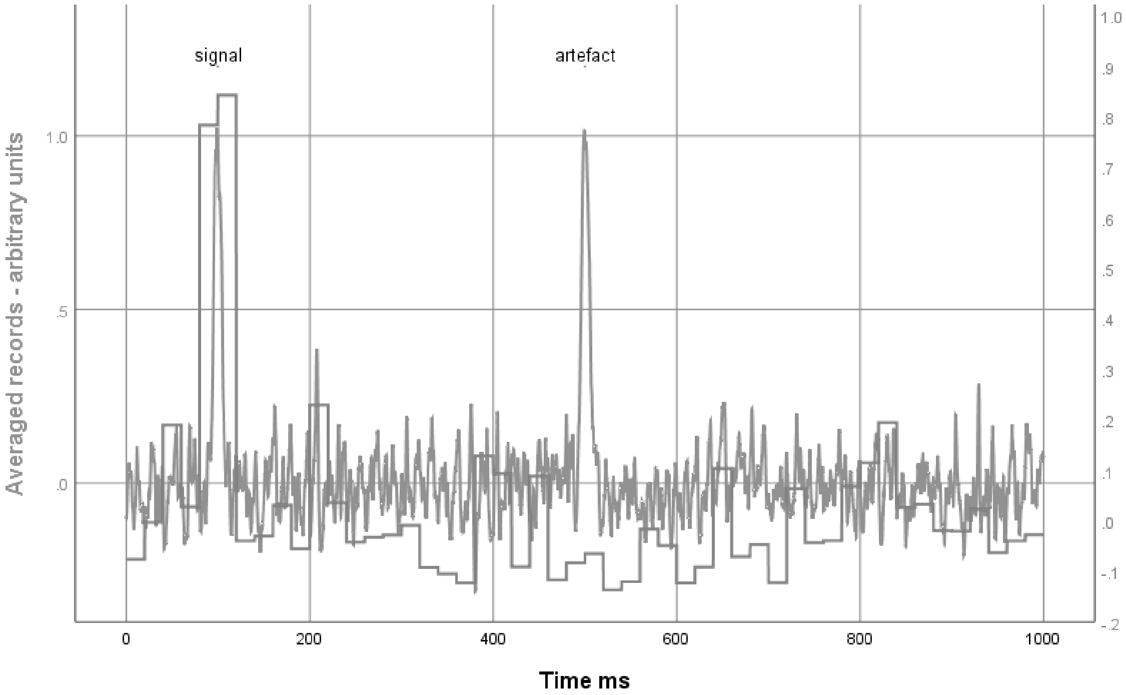


Figure 6: Average (blue) of 10 repeated simulated signals each of amplitude 1 (arbitrary units) with one repetition also having an artefact of amplitude 10. All repetitions have added noise from scalp recordings. Median r for successive 20 ms windows shown in red.

Figure 6 shows the averaged response of the 10 repetitions. It can be seen that the 10 repetitions of the signal and the single occurrence of the simulated artefact (with 10 times the signal amplitude) produce identical components on the average. However, the median r for the two 20 ms windows centred on 100 ms (the signal) is approximately 0.8 which indicates very high reliability, whereas the median r for the windows centred on 500 ms is approximately -0.1 and indistinguishable for the values of median r for those portions of the average which only contains EP noise.

This shows that median r correlation analysis technique can be used effectively to distinguish between transient, high amplitude, and consistent, low amplitude components of an averaged recording, even for averages with few repetitions.

4 Discussion

The practice of evoked potential recording, interpretation and application is firmly based on the use of conventional signal averaging, despite the well-known problems associated with the use of this technique. Innovative analytical methods that have attempted to overcome the drawbacks of signal averaging have not been widely adopted. The correlation method described allows the continued use of the familiar average response but gives additional information relating to the areas of the average which can be considered to be reliable representations of the underlying signal component, and those which are the result of more transient or variable elements.

Conventional signal averaging ideally requires a stationary signal and random noise in relation to the stimulus onset. Correlation analysis also assumes that the signal occurs at fixed positions relative to the stimulus onset. Hence correlation analysis gives a measure of the reliability of signal components where reliability is defined in terms of repeatability. As change in amplitude alone does not alter the median correlation coefficient (see section 3.2), this reliability is therefore expressed as components of similar shape occurring at the same latency.

The results using both simulated signals and real SEP recordings clearly show that it is possible to use the correlation analysis described to identify the reliability of recorded components. Perhaps the main benefit of the proposed methods is its straightforward applicability to scalp responses recorded with the traditional method of averaging.

Early components of SEPs have always been known as the most repeatable and reliable evoked responses that can be recorded from the scalp. They are generated by cortical and subcortical sources after the afferent volley has travelled through fast conducting fibres and just a few synapses, after the electric stimulus delivered at the peripheral nerve trunk produced a simultaneous depolarization of all the fibres. Such characteristics made these responses very popular in diagnosis of the afferent lemniscal pathway with little need to check for reliability. However, there is little clinical utility in the middle and late components of the SEPs because these showed a fairly large amount of variability and they could not be relied upon for diagnosis.

There are other responses obtained with non-electrical stimuli, like the visual and auditory evoked potentials, which, because of insufficient synchronization of the afferent volley also result in later components with a large degree of variability (Sarnthein et al., 2009; Michalewski et al., 1986). The comparatively recent use of nociceptive evoked potentials is particularly important. These are comprised exclusively of late components belonging to the family of endogenous event related potentials. Such responses are heavily processed by several cortical areas, are very prone to habituation and are dependent upon the level of attention. They may be of large amplitude, but are extremely variable in both amplitude and latency (for a review see Lefaucheur, 2019). In addition, most of these late potentials have similar waveshapes to artefacts caused by muscular potentials which are often not automatically rejected because they have similar amplitudes as the genuine signal components. The number of repeated stimulations used for averaged nociceptive evoked potentials is limited (typically 20-50) because of the risk of habituation with all the late potentials and so one single large artefact might be incorporated into the averaged signal and cause false interpretation.

5 Conclusions

The cross-correlation analysis that we propose can add reliability information to evoked potential components identified using conventional signal averaging. A comparison between the averaged components and the median correlation coefficient values of suitable time windows will provide visual information about reliability of each component of interest in the recording. It can also be used to identify contamination of the averaged response by irregular, large amplitude artefacts. It can easily and quickly be applied to recorded responses and the post-processing necessary for this analysis would result in only small delays. For applications that need to minimise the number of stimuli used to produce the evoked potential, it should also be possible to incorporate real-time generation and display of the median correlation coefficient superimposed on the average as it builds with each stimulus. This would enable the recording to be stopped as soon as the components of interest are identified with suitable reliability.

References

Aunon, J.I., McGillem, C.D., Childers, D.G., 1981. Signal processing in evoked potential research: averaging and modelling. Crit. Rev. Bioeng. 5, 323–367.

Bromm, B., Scharein, E., 1982. Principal component analysis of pain-related cerebral potentials to mechanical and electrical stimulation in man. Electroencephalogr. Clin. Neurophysiol. 53, 94–103.

Clayson, P.E., Miller, G.A., 2017. ERP Reliability Analysis (ERA) Toolbox: An open-source toolbox for analyzing the reliability of event-related brain potentials. Int. J. Psychophysiol. 111, 68–79. https://doi.org/10.1016/j.ijpsycho.2016.10.012

Collura, T.F., 1995. Averaging, Noise, and Statistics, in: Comprehensive Clinical Neurophysiology. Elsevier, pp. 11–18.

Coppola, R., Tabor, R., Buchsbaum, M.S., 1978. Signal to noise ratio and response variability measurements in single trial evoked potentials. Electroencephalogr. Clin. Neurophysiol. 44, 214–222. https://doi.org/10.1016/0013-4694(78)90267-5

Dawson, G., 1947. Cerebral responses to electrical stimulation of peripheral nerve in man. J. Neurol. Neurosurg. Psychiatry 10, 134–140.

Don, M., Elberling, C., Waring, M., 1984. Objective detection of averaged auditory brainestem responses. Scand Audiol 13, 219–228.

Lefaucheur, J.P., 2019. Clinical neurophysiology of pain, 1st ed, Handbook of Clinical Neurology. Elsevier B.V. https://doi.org/10.1016/B978-0-444-64142-7.00045-X

McGillem, C.D., Aunon, J.I., Childers, D.G., 1981. Signal processing in evoked potential research: applications of filtering and pattern recognition. Crit. Rev. Bioeng. 6, 225–265.

Michalewski, H., Prasher, D., Starr, A., 1986. Latency variability and temporal interrelationships of the auditory event-related potentials (N1, P2, N2 and P3) in normal subjects. Electroencephalogr. Clin. Neurophysiol. 65, 59–71.

Moser, J.M., Aunon, J.I., 1986. Classification and Detection of Single Evoked Brain Potentials Using Time-Frequency Amplitude Features. IEEE Trans. Biomed. Eng. BME-33, 1096–1106. https://doi.org/10.1109/TBME.1986.325686

Patel, R., Janawadkar, M.P., Sengottuvel, S., Gireesan, K., Radhakrishnan, T.S., 2015. Effective extraction of evoked potentials using template cross correlation. 2015 Int. Conf. Commun. Signal Process. ICCSP 2015 35–39. https://doi.org/10.1109/ICCSP.2015.7322909

Quiroga, R.Q., 2000. Obtaining single stimulus evoked potentials with wavelet denoising. Phys. D Nonlinear Phenom. 145, 278–292. https://doi.org/10.1016/S0167-2789(00)00116-0

Sarnthein, J., Andersson, M., Zimmermann, M.B., Zumsteg, D., 2009. High test-retest reliability of checkerboard reversal visual evoked potentials (VEP) over 8 months. Clin. Neurophysiol.

120, 1835–1840. https://doi.org/10.1016/j.clinph.2009.08.014

Thigpen, N.N., Kappenman, E.S., Keil, A., 2017. Assessing the internal consistency of the event-related potential: An example analysis. Psychophysiology 54, 123–138. https://doi.org/10.1111/psyp.12629

Turner, S., Picton, P., Campbell, J., 2003. Extraction of short-latency evoked potentials using a combination of wavelets and evolutionary algorithms. Med. Eng. Phys. 25, 407–412. https://doi.org/10.1016/S1350-4533(03)00021-3

CRediT author statement

Jackie Campbell: conceptualisation, methodology, software, validation, formal analysis, data curation, writing - original draft, review and editing, visualization

Massimo Leandri: conceptualisation, methodology, software, investigation, data curation, writing - original draft, review and editing.