ELSEVIER

Electroencephalography and clinical Neurophysiology 106 (1998) 457–459

Letter to the editor

Sensitivity and selectivity for continuous perception values – a comment

Michael A. Black, Richard D. Jones*

Department of Medical Physics and Bioengineering, Christchurch Hospital, Christchurch, New Zealand

In a study of the detection of epileptiform activity in the EEG by multiple readers, Wilson et al. proposed *continuous-valued* sensitivity and selectivity to allow a probabilistic approach to spike detection to be taken (Wilson et al., 1996). Although we believe that such a probabilistic approach has merit, we are concerned that the use of alternative definitions for such universal measures of performance as sensitivity and selectivity can be both confusing and misleading.

Wilson et al. stated the standard definition of inter-reader sensitivity, and the relationship between sensitivity and selectivity for spike detection, as

$$Sensitivity_{AB} = \frac{Spikes_{AB}}{Spikes_{B}} = Selectivity_{BA} \quad (1)$$

where $spikes_B$ is the number of spikes reported by reader B, and $spikes_{AB}$ is the number of spikes reported by both reader A *and* reader B. That is, the sensitivity of reader A relative to reader B is the proportion of the spikes reported by reader B that were also reported by reader A. Conversely, the selectivity of reader A relative to reader B is the proportion of the spikes reported by reader A that were also reported by reader B.

The standard and universally-accepted definition of sensitivity has a minimum score of 0 and a maximum of 1 – a score of 0 corresponding to reader A failing to detect *any* of the spikes reported by reader B and a score of 1 *if and only if reader A detects all of the spikes detected by reader B*.

Wilson et al. correctly stated that the standard definition of sensitivity is only able to deal with spikes that are reported in a *dichotomous* manner - that is, events classified as spikes or non-spikes only. In their study, however, spikes were treated as being probabilistic in nature and readers were required to assign each event that they regarded as being epileptiform a *perception* value of 0.25, 0.5, 0.75 or 1 to reflect their subjective assessment of the probability that the event detected was a true epileptiform spike. Incorporating perception values into the analysis prevented readers from being heavily penalised for missing low perception spikes reported by other readers.

In order to include the readers' spike perception values in the sensitivity calculations, Wilson et al. defined the following *continuous-valued* sensitivity formula

$$Sensitivity_{AB} = \frac{\sum_{i=1}^{N^{A\cup B}} x_{Ai} \cdot x_{Bi}}{\sum_{i=1}^{N^{A\cup B}} x^{2}_{Bi}} = Selectivity_{BA} \quad (2)$$

where $N^{A\cup B}$ is the number of events reported by either or both of raters A and B, and x_{Ai} is the perception value (i.e. 0, 0.25, 0.5, 0.75 or 1) given to event i by reader A, and x_{Bi} is the perception value given to event i by reader B.

Although the new formula gives the same results as the standard definition of sensitivity when dichotomous data is used, the use of continuous perception values means that the readers' sensitivities no longer have an upper limit of 1. This occurs because (as Wilson et al. noted) reader A is considered to be more sensitive if they detect reader B's events with a higher perception value than if they had detected the events with identical perception values. This can sometimes result – as occurred in Wilson et al.'s paper – in sensitivities greater than 1, which raises a question as to the validity of this formula as a measure of sensitivity. That is, we consider it inappropriate for a measure to be termed sensitivity if it can take values above 1.

At this point it is tempting to consider normalising the measure to constrain its values between 0 and 1. This is certainly possible, as the maximum attainable sensitivity is defined by the range of perception values chosen. However, normalisation would mean that a sensitivity of 1 could only be obtained by a reader who always assigned a perception value of 1. For example, the maximum possible non-normalised sensitivity attainable using Wilson et al.'s perception values was 4, corresponding to reader A detecting events with a perception value of 0.25 and reader B detecting all of the same events with a perception value of 1. Normalisation gives reader B a sensitivity of 1 in this situa-

* Corresponding author. Tel.: +64 3 3640853; fax: +64 3 3640851; e-mail: r.jones@chmeds.ac.nz

PII S0013-4694(98)00012-1

EEG 97082

Table 1

Perception values assigned to 4 events by 3 EEGers, and group scores calculated by taking the mean

	Readers			Group scores		
	A	B	C	AB	AC	BC
Event 1	1.000	1.000	0.750	1.000	0.875	0.875
Event 2	0.000	0.250	0.500	0.125	0.250	0.375
Event 3	0.250	0.000	0.000	0.125	0.125	0.000
Event 4	0.500	0.250	0.000	0.375	0.250	0.125

tion but would also mean that exact agreement between the readers on both events and their perception values would correspond to a sensitivity of less than 1. Such results suggest that normalisation is not an attractive option.

A further problem with the measure is that the nature of the formula means that reader A's sensitivity relative to reader B can be greater than or equal to 1 *even if A does not detect all of B's events*. Such a situation would seem to go against the generally accepted meaning of sensitivity. A reader's sensitivity should not equal 1 (or greater, as is possible here) unless they have detected all of the spikes reported by the other reader.

The following example gives a brief overview of Wilson et al.'s technique and illustrates the problems encountered if the new definition of sensitivity is used. Table 1 contains theoretical data for 3 expert readers' impressions of 4 epileptiform events, with each reader giving the event a perception value if it is considered to be definitely or possibly epileptiform and a value of 0 assigned if the event was not marked. Only one event was detected by all 3 readers, with readers A and B assigning a perception value of 1 and reader C assigning a perception value of 0.75. The remaining 3 events were detected by one or two readers with perception values of 0.25 or 0.5.

The group score for each event is also listed for each combination of two readers. This was used to compare each reader to the other two readers. The method of combining chosen by Wilson et al. for the continuous data was to take the mean of the individual scores. For example, the group score for the first event for readers B and C was the mean of the perception values those two readers gave to that event, that is, 0.875. These scores were then used to calculate reader A's sensitivity and selectivity relative to the combined readers B and C.

Table 2 contains the inter-reader sensitivities and selectivities that were calculated for the data using Eq. (2). Note that the values on the diagonal are automatically equal to 1 since each reader has perfect sensitivity and selectivity relative to themselves. The sensitivity for each reader is presented vertically, and their selectivity horizontally. It can be seen that the entry corresponding to reader B's sensitivity relative to reader C is greater than 1. This occurs because reader B gave the first event a perception value of 1, while reader C only gave a value of 0.75 and, even though the situation was reversed for the second event (reader B marking 0.25 and reader C marking 0.5), the high scores given to the first event outweighed this. Also of concern in Table 2 is the score of 1 for reader A's sensitivity relative to reader B, *despite the fact that A did not detect all of B's events*. Such a score was possible because although A missed the second event, A assigned a higher perception value than B to the third event, which served to balance the equation. Clearly, in this context a sensitivity of 1 is very misleading.

The data in Table 3 shows the readers' individual sensitivities and selectivities to the group scores and the mean sensitivity and selectivity. The latter were what Wilson et al. placed most emphasis on in their paper, with only the means being reported and not the readers' individual scores. The calculations were made using Eq. (2) since the group scores were treated in the same way as those of an individual reader. Table 3 shows that in this case Wilson et al.'s continuous-valued definition of sensitivity and selectivity gave each reader a sensitivity or selectivity of greater than or equal to 1. Furthermore, readers A and B had sensitivities greater than 1 despite both having missed an event that was found by the rest of the group.

Overall, this resulted in a mean sensitivity of 94% despite reader C having a sensitivity of only 69% and despite *none* of the readers detecting 94% of the events reported by the rest of the group. Also, to report only the means (as Wilson et al. did) in such a situation is inappropriate, as this gives no indication of the spread of the readers' scores, some of which are unexpectedly (for a supposed measure of sensitivity) greater than 1. No matter what method is used, some measure of the readers' variability needs to accompany the mean, to help describe the distribution from which the data came. Simply reporting the range of the readers' sensitivities would be enough in this case.

From this example it can be seen that care must be taken if one wishes to use Wilson et al.'s continuous-valued definition of sensitivity/selectivity. Even when used without reference to other studies, these methods can produce results which are difficult to interpret and their non-standard nature makes direct comparison with results from other studies using standard methods impossible and potentially misleading. The main problem is not the method itself but the naming of it as a measure of sensitivity/selectivity, as these are labels which already have clearly defined meanings and properties. It would seem, therefore, that if these measures are to be used, they should not be called sensitivity

Table 2

Inter-reader spike detection sensitivities and selectivities with continuous-valued spike perceptions

Reader	A	B	C	Average selectivity
A	1.000	0.857	0.571	0.714
B	1.000	1.000	0.778	0.889
C	0.923	1.077	1.000	1.000
Average sensitivity	0.962	0.967	0.675	0.868

Table 3

Group score calculations of sensitivity and selectivity for continuous valued spike perceptions

Reader	Sensitivity	Selectivity
A	1.017	0.714
B	1.103	0.889
C	0.693	1.000
Mean	0.938	0.868

and selectivity, as they do not conform to our prior expectations of such measures. To obtain estimates of these which *do* conform to the standard definitions we would suggest the alternative method used by Wilson et al. which involves using dichotomous data from each reader together with dichotomous group scores.

It should be noted that Wilson et al.'s use of continuous-valued specificity has not been considered in this letter. This is because the nature of spike detection requires that the number of true negative events in an EEG must be artificially specified in terms of the number of events which have *not* occurred over a period of time. This procedure effectively makes specificity arbitrary and hence of minimal value, regardless of whether continuous or dichotomous data is used.

Despite the problems met in attempting to define acceptable performance measures for use with continuous perception values, we agree with Wilson et al.'s assumptions that spike detection is probabilistic and that there is no gold standard reference. We suggest that further work be done in this area to develop valid and appropriate performance measures which satisfy these assumptions, as there are considerable benefits to be gained by standardising the measure of inter-reader performance, especially in the area of automated spike detection.

References

Wilson, S.B., Harner, R.N., Duffy, F.H., Tharp, B.R., Nuwer, M.R. and Sperling, M.R. Spike detection. I. Correlation and reliability of human experts. Electroenceph. clin. Neurophysiol., 1996, 98: 186–198.