



**Pitfalls in statistics: hypothesis tests on clinical data**

Journal:	<i>Cerebral Cortex</i>
Manuscript ID:	draft
Manuscript Type:	Feature Articles
Date Submitted by the Author:	n/a
Complete List of Authors:	Bijma, Fetsje; Vrije Universiteit, Faculty of Science, Department of Mathematics Bazelier, Marloes; Utrecht University, Faculty of Science, Department of Mathematics Meester, Ronald; Vrije Universiteit, Faculty of Science, Department of Mathematics
Keywords:	Alzheimer's Disease, functional brain network, path length, statistical test, synchronization likelihood

Pitfalls in statistics: hypothesis tests on clinical data

Fetsje Bijma<sup>1,\*</sup>, Marloes Bazelier<sup>2</sup>, Ronald Meester<sup>1</sup>

<sup>1</sup> Vrije Universiteit, Faculty of Science, Dept. of Mathematics, Amsterdam, The Netherlands.

<sup>2</sup> Utrecht University, Faculty of Science, Dept. of Mathematics, Utrecht, The Netherlands.

\* corresponding author: Vrije Universiteit, Faculty of Science, Dept. of Mathematics, De Boelelaan 1081, 1081HV Amsterdam, The Netherlands, Email: f.bijma@few.vu.nl  
Telephone: 0031-20-5987835, Fax: 0031-20-5987653.

For Peer Review

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2 *Abstract:* We discuss the statistics in a recent paper of Stam and others 2007,  
3 concerning potential differences in the functional brain network between patients with  
4 Alzheimer's disease and a control group. We find that the conclusions in that paper, to  
5 the effect that one can significantly distinguish between the two samples, are not  
6 warranted by their statistical analysis. Our considerations are of interest for all scientists  
7 using statistics.  
8

9  
10 *Keywords:* Alzheimer's disease, cluster coefficient, functional brain network, path  
11 length, statistical test, synchronization likelihood.  
12

13  
14 In a recent paper by Stam and others 2007 differences in the functional brain network  
15 between patients with Alzheimer's disease (AD) and controllers were investigated,  
16 based on EEG data. The conclusion of this paper was (citing from the paper) that "For a  
17 wide range of thresholds, the characteristic path length  $L$  was significantly longer in the  
18 Alzheimer patients. This pattern was still present when  $L$  was computed as a function of  
19 average degree  $K$ ". We believe that this outspoken conclusion cannot be drawn based  
20 on the presented analysis.  
21

22  
23 Differences were claimed to be found, based on different properties of the graphs that  
24 were constructed using the Synchronization Likelihood (SL). After applying a threshold  
25  $d$  to an SL matrix, the matrix can be interpreted as an adjacency matrix of a graph. Two  
26 characteristics of the graphs obtained this way were calculated in Stam and others 2007:  
27 the average shortest path length between two nodes ( $L$ ) and the average cluster  
28 coefficient ( $C$ ). The two groups, AD and controllers, were compared using the  
29 calculated values for  $C$  and  $L$  for different values of the threshold  $d$ .  
30  
31

32  
33 Collaborating with the first author of Stam and others 2007, we have been searching for  
34 a mathematical foundation of these differences. The SL data in our re-analysis differed  
35 slightly from the data in Stam and others 2007, since the authors recalculated the SL  
36 data for the re-analysis with an updated algorithm. The original SL data had apparently  
37 been lost. However, in repeating the analysis, we encounter three difficulties and cannot  
38 endorse their general finding of significant differences between the two groups. The  
39 first problem is the use of parametric tests. The use of t-tests is not justified, or at least it  
40 has not been verified. The second – and possibly main - obstacle is the fact that no  
41 correction for multiple testing has been carried out. Finally, the hypothesis that is being  
42 tested is based on the data itself; this point is very much related to the second obstacle,  
43 though treating it separately clarifies our criticism. These three themes often occur in  
44 statistical analysis of clinical trials. We do not claim that our problems with the statistics  
45 in Stam and others 2007 are highly original, but since abuse of statistics is a widespread  
46 phenomenon, we feel that our attention is justified. We treat the three points one by one.  
47  
48  
49

50  
51 When normality is not evident, the use of nonparametric tests (e.g., Mann-Whitney tests  
52 or median tests) instead of parametric tests is preferable; p-values resulting from a t-test  
53 are not reliable in this situation. This is especially important if group sizes are small to  
54 medium, as is the case in this analysis where groups consist of 13 and 15 individuals. In  
55 Figure 1, p-values resulting from the Mann-Whitney test, the median test and the t-test  
56 on the values for  $L$  of the two groups are shown. P-values are lower near average degree  
57 3, though not significant for any test. In Stam and others 2007, the significant  
58 differences are claimed to be found in a regime around average degree 3. There is a  
59 slight discrepancy between the p-values of the t-tests in our re-analysis and the p-values  
60 found in Stam and others 2007, which is due to the update of the SL data. Nevertheless,  
Figure 1 shows that the p-values claimed significant in Stam and others 2007 must have

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

been only just below 0.05. Hence, one should be utterly cautious in drawing explicit conclusions based on these tests – and this caution does not even mention the arbitrariness of the choice of the threshold 0.05.

In the graph analysis in Stam and others 2007, many values of the threshold  $d$  are tested, which is the basis for the second obstacle. Many tests are performed, and no correction for multiple comparisons is made. The bottom line is that if many comparisons are made, some of these will show significant difference just by chance, and clearly no conclusions can be drawn from this. Results are shown for 21 truly different  $K$ -values, of which 3 are found significant. (The plots suggest a higher resolution than possible: 19 of the 21 values are plotted twice, resulting in 40 plotted values.) A Bonferroni correction is probably too conservative in this situation, since there is dependence amongst the hypotheses. Nevertheless, applying no correction is clearly too liberal and yields unreliable  $p$ -values. In this case it is not evident how the correction should be done.

The third point of consideration is about choosing the null hypothesis. When data are analyzed in many ways and significant differences are found for one or a few specific settings (one or a few values of  $d$ ) it is not correct to perform hypothesis testing as usual for this specific setting. The null hypothesis has been searched for, based on the data, and, subsequently, the result of testing is not reliable. This becomes apparent from the following example. *Suppose we have 100 coins, which we all flip 10 times. As it happens, the 21<sup>st</sup> coin shows head 9 times and tail only 1 time. Subsequently, we set up the null hypotheses that the 21<sup>st</sup> coin is not fair, and we test this hypothesis with the same data.* This example illustrates that testing hypotheses that are based on the data is quite dangerous; in this case it is obvious that the rejection of the null hypothesis is totally meaningless. Indeed, the data have been used twice: once for choosing the hypothesis and once for testing that hypothesis. These two steps should be performed on different data sets. In the analysis in Stam and others 2007, a few values of the threshold  $d$  suggest differences between the AD patients and the controllers. In order to verify this claim, this hypothesis should be tested on a new data set.

A final word about the reported significances. As mentioned before, significant differences were found in a regime around average degree 3. However, in this regime, 27 of the 28 graphs are unconnected. Hence,  $L$  cannot be interpreted as the average shortest path at all, since some shortest paths have infinite length. In Stam and others 2007 it is stated that “ $L$  is a measure of how well connected a graph is”. However, in their analysis this interpretation of  $L$  is misleading due to the connectivity problem.

Our conclusion is that the statistical analysis carried out in Stam and others 2007 does not lead to their claim that “ $L$  is significantly longer in AD patients for  $2.85 < K < 3.15$ ”. It seems that further investigation is necessary in order to come to any conclusion at all.

## References

Stam CJ, Jones BF, Nolte G, Breakspear M, Scheltens Ph. 2007. Small-world networks and functional connectivity in Alzheimer's disease. *Cereb Cortex* 17: 92-99

## Captions

Figure 1:

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

P-values for the Mann-Whitney test (squares), the median test (open circles) and the t-test (solid circles) for L as function of average degree. The dashed line shows the 0.05 significance level.

For Peer Review

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

p-values for three different tests

