doi:10.1093/brain/awn091 Brain 2009: 132; 1–3 | e102

BRAIN
A JOURNAL OF NEUROLOGY

LETTER TO THE EDITOR

A plea for confidence intervals and consideration of generalizability in diagnostic studies

Stefan Klöppel,[1,2,*] Cynthia M. Stonnington,[2,3,*] Carlton Chu,[2] Bogdan Draganski,[2] Rachael I. Scahill,[4] Jonathan D. Rohrer,[4] Nick C. Fox,[4] John Ashburner[2] and Richard S.J. Frackowiak[2,5,6]

1 Department of Psychiatry, University Clinic Freiburg, Freiburg, Germany
2 Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, United Kingdom
3 Department of Psychiatry and Psychology, Mayo Clinic, Scottsdale, AZ , USA (CMS) and Department of Radiology, Mayo Clinic, Rochester, MN, USA (CRJ)
4 Dementia Research Centre, Department of Neurodegenerative Disease, Institute of Neurology, University College London, London, United Kingdom
5 Département d'études cognitives, Ecole Normale Supérieure, Paris, France
6 Laboratory of Neuroimaging, IRCCS Santa Lucia, Roma, Italy

*These authors contributed equally to this work.

Correspondence to: Stefan Klöppel, MD,
Department of Psychiatry, Hauptstr. 5,
79104 Freiburg, Germany
E-mail: stefan.kloeppel@uniklinik-freiburg.de

Sir, Many thanks for letting us respond to the interesting letter concerning our recent paper. We are grateful for the chance to clarify the points raised, which suggest our conclusions were too optimistic. In our paper (Kloppel *et al*., 2008), we used MRI scans from pathologically proven cases of Alzheimer's disease and frontotemporal lobar degeneration (FTLD) to validate trained sets for a machine learning-based support vector machine (SVM) approach to the categorization of structural scans from normal and each other.

This rigorous approach substantially limited the number of available subjects, which we made perfectly clear in our article, but which was unavoidable given our novel approach. Frost and colleagues are right to point out that such low numbers result in larger confidence intervals than if we were able to include more scans. This is an object of our further empirical studies—what is the improvement in classification gained using this technique with greater numbers of scans in the trained set? The graph below (Fig. 1) illustrates diagnostic accuracy when the whole brain grey matter segment is used to separate probable Alzheimer's disease patients from all clinical stages (MMSE range of 3 to 30; defined clinically in the same way as group III, in our original paper) from controls. Classification is performed repeatedly and after removing one Alzheimer's disease patient and one

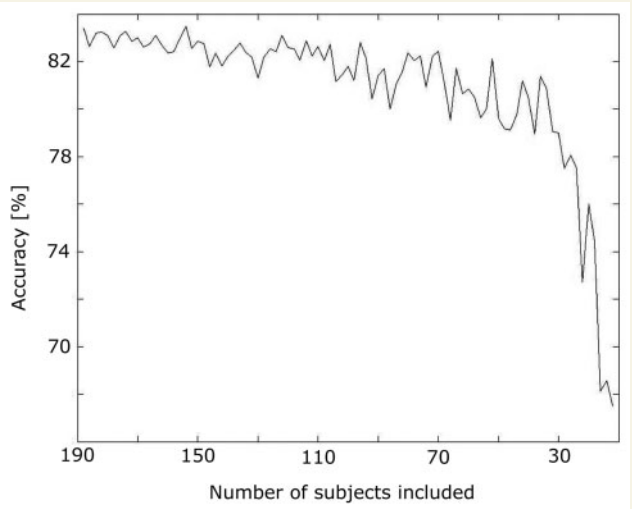

Figure 1 Alzheimer's disease patients from all clinical stages (including all subjects from group III) and an equal number of age and sex matched cognitively normal controls are separated repeatedly. Before each classification, one patient and one control are removed to illustrate the robustness of classification with shrinking group size.

Received and Accepted April 18, 2008. Advance Access publication May 21, 2008
© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Table 1 Demographic information on groups I, II and IV with post-mortem confirmation of Alzheimer's disease obtained at different centres

Group (*n*)	Group I		Group II		Group III		Group IV	
	Alzheimer's disease (20)	Controls (20)	Alzheimer's disease (14)	Controls (14)	Alzheimer's disease (33)	Controls (57)	Alzheimer's disease (18)	FTLD (19)
Sex (F/M)	11/9	10/10	5/9	5/9	10/23	16/41	6/12	8/11
Age (mean, range) at MRI-scan	81.0 (51–102)	79.5 (55–91)	65.0 (53–85)	63.0 (51–81)	73.1 (61–80)	71.9 (61–80)	66.0** (53–85)	61.7** (46–73)
MMSE-score (mean, range)	16.7 (7–29)	29.0 (27–30)	16.1* (10–20)	29.2 (28–30)	23.5 (20–28)	29.1 (27–30)	16.2* (5–29)	18.0 (0–26)
Years from MRI-scan to death (mean, range)	1.7 (0.2–3.4)	NA	3.6 (0.3–7.2)	NA	NA	NA	3.5 (0.3–7.2)	5.8 (1.3–11.0)

*MMSE scores obtained around the time of scanning only available from 12 subjects; **groups are age matched ($P = 0.1$). The first and third image sets are from a largely community-based sample, whereas subjects from sample II tended to be younger. No strong family history was present in any of the subjects.
FTLD = frontotemporal lobar degeneration; MMSE = Folstein Mini Mental State Examination.
(*Source*: Kloppel *et al*., 2008).

Table 2 Results of SVM classification using grey matter from the whole brain for image analysis

Group	Correct (%) (95% CI)	Sensitivity (%)[a] (95% CI)	Specificity (%)[a] (95% CI)
Alzheimer's disease and controls group I	95.0 (81.8–99.1)	95.0 (73.1–99.7)	95.0 (73.1–99.7)
Alzheimer's disease and controls group II	92.9 (75.1–98.8)	100 (73.2–100)	85.7 (56.2–97.5)
Alzheimer's disease and controls group III	81.1 (71.2–88.3)	60.6 (42.2–76.6)	93.0 (82.2–97.7)
Dataset I for training, set II for testing	96.4 (79.8–99.8)	100 (73.2–100)	92.9 (64.2–99.6)
Dataset II for training, set I for testing	87.5 (72.4–95.3)	95.0 (73.1–99.7)	80.0 (55.7–93.3)
Group I + II	95.6 (86.8–98.9)	97.1 (82.9–99.8)	94.1 (78.9–99.0)
Alzheimer's disease from Dataset II and FTLD group IV	89.2 (73.6–96.5)	83.3 (57.7–95.6)	94.7 (71.9–99.7)

[a]Considering a correctly identified Alzheimer's disease case as a true positive.
95% CIs are calculated according to the efficient-score method (Newcombe, 1998; http://faculty.vassar.edu/lowry/clin1.html).

Table 3 Results of SVM classification using only grey matter of antero-medial temporal lobe structures for analysis

Group	Correct (%) (95% CI)	Sensitivity (%)[a] (95% CI)	Specificity (%)[a] (95% CI)
Alzheimer's disease and controls group I	90.0 (81.8–99.1)	85.0 (61.1–96.0)	95.0 (73.1–99.7)
Alzheimer's disease and controls group II	92.9 (75.1–98.8)	92.9 (64.2–99.6)	92.9 (64.2–99.6)
Alzheimer's disease and controls group III	85.6 (76.2–91.8)	75.8 (57.4–88.3)	91.2 (80.0–96.7)
Dataset I for training, set II for testing	71.4 (51.1–86.1)	50 (24.0–76.9)	92.9 (64.2–99.6)
Dataset II for training, set I for testing	70.0 (53.3–82.9)	95.0 (73.1–99.7)	45.0 (23.8–68.0)
Group I + II	94.1 (80.5–99.1)	97.1 (82.9–99.8)	91.2 (75.2–97.7)

[a]Considering a correctly identified Alzheimer's disease case as a true positive.
95% CIs computed as above.

control each time. Results are fairly stable but accuracy becomes more variable until a steep decline occurs when less than around 20 subjects per group are included. Suffice it to say we were surprised how well Alzheimer's disease was distinguished from FTLD given the even smaller numbers of validated scans we had available for that classification. To clarify these issues, we provide a table that supplements our data with CIs. Further, we found very similar results using two completely independent datasets and the CIs become relatively small when data from the first two datasets are combined. So, although we agree with the question posed theoretically, practically the results stand as proof of principle.

We also agree that some statements found on the BBC's website (BBC, 2008) are misleading. Specifically, we show that SVMs provide a much faster classification than full clinical workup. Where the website misleads is in implying that they detect early degeneration faster, which is clearly beyond the scope of the current article and again a subject of ongoing study.

It is important to emphasize that such multivariate methods generalize to new data. Figure 1 in our original paper illustrates that during training, samples from those individual subjects (i.e. normalized grey matter segments from either the whole brain or the hippocampus area), which best separate the two groups define the decision boundary. The figure is an example with two dimensions but in reality, the number of dimensions equals the number of voxels used. If a classifier generalizes well, a new scan will be assigned to the same side of the decision boundary as the rest of a diagnostic group. It is a critical part of our results that

the decision boundary defined by data from one imaging centre using different hardware and sequences is sufficiently general to separate data accurately from other imaging centres. This ability is of great practical relevance as a library of well-defined cases can be made available to referral centres as a general trained set to diagnose scans collected there. While our results are promising, as we pointed out in our article, 'a formal comparison with modern conventional clinical assessment is required'. It should be kept in mind that we used very strict inclusion criteria and the extension to relatively poorly defined data from primary referral centres needs to be addressed in a separate study. It is likely that libraries from very early stages of the disease need to be produced, which are then validated longitudinally or pathologically. The issue now is to optimize the variables to maximize sensitivity and accuracy. One lesson we learned is that proper validation of scans included in the trained set is likely to be critical.

Acknowledgements

Funding to pay the Open Access publication charges for this article was provided by The Wellcome Trust.

References

Kloppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, et al. Automatic classification of MR scans in Alzheimer's disease. Brain 2008; 131: 681–9.

BBC. Computers 'spot Alzheimer's fast', 2008. http://news.bbc.co.uk/2/hi/health/7258379.stm

Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 1998; 17: 857–72.