# Comments on some published Watermaze Data

## Statement about myself

My name is Richard Morris. I am neuroscientist working the UK whose research focus is on the neurobiology of learning. One contribution I made was the development, in 1981, of the open-field watermaze that in some publications (not my own) is described using my name. I now have extensive experience of using the watermaze with rats and mice, and have tried to offer constructive advice (freely) to other laboratories when they run into problems. By and large, the standard version of the watermaze task works well, but success in using the protocol depends on a range of factors (size of pool, extent and distribution of extramaze cues, water temperature, effective and careful handling of the animals to minimise stress, experimenter, sessions run blind and other parameters). Variability is sometimes seen across studies within and across laboratories. Many excellent papers using the watermaze have been published over the years, of which many have addressed intriguing issues that I would never have thought of investigating. However, some have been reported studies which may be questionable with respect to procedure.

## Background

My attention has now been drawn to a series of 4 papers from a laboratory at Temple University in Pennsylvania whose data are, at best, somewhat odd. I will outline my comments:

### Paper 1 - Gain of function study

Phillip F. Giannopoulos, Jin Chu, Margaret Sperow, Jian-Guo Li, W. Haung Yu, Lynn G. Kirby, Mary Abood, and Domenico Praticò (2015) Pharmacologic Inhibition of 5-Lipoxygenase Improves Memory, Rescues Synaptic Dysfunction, and Ameliorates Tau Pathology in a Transgenic Model of Tauopathy. *Biological Psychiatry.*

There are numerous components to the paper reflecting the interdisciplinary nature of the projects reported. My exclusive focus has been on the watermaze task for which data is shown in Figure 2. The protocol consisted of visual platform training followed by 4 sessions of training and, 24 hrs later, a probe test (d5). The supplementary material indicates that animals (n=10; trained at about 10 months of age, p 694) were trained to a criterion of escaping in <20 sec, but only 4 days are plotted. I am confused by this as training to criterion and training for a set number of days/sessions are different, but perhaps day 4 simply went on for as many trials as necessary to reach this criterion. In any event, the latency data of Figure 2C is unusual in showing an almost linear decline in escape latency across groups and astonishingly small standard errors of the mean (SEMs). These are usually much larger on day 1 and then decline. On day 4, or whatever data is included in day 4, the escape latency of all groups was < 20 sec and, importantly, equivalent across groups. However, the data of Figure 2D shows that on day 5, the *htau* group has a latency to platform that jumps higher to approximately 25-27 sec, while the other groups jump lower to 10-12 sec. There is no intervening training. Both wild-type groups perform identically. Both facets of the data are odd, but might be explained by there being a dramatic within-session learning curve for the *htau* group on days 1-4 such that their first trial is slow and subsequent trials are much faster each day, but there is no mention of this. The Figure legend for panels D and E are transposed (merely a typo), but more importantly, a p value is given for the data of panel 2D (in text for 2E) indicating a significance level of p < 0.001. It is not clear what this refers to. Is it an ANOVA or something else? The two stars plotted above the blue bargraph of Panel 2D would appear to imply that it is the cause of the significant p value, but this is not justified in the text. Because no F value or degrees of freedom are given, we are obliged to accept this p value on trust and I cannot work out from the df what is compared with what. If it was an ANOVA with n=10 per group, the df of the Error mean square would be [a x (n-1)], namely, 36. But we are not told.

**Paper 2 - Gain of function study**

Phillip F. Giannopoulos, Jian Chiu, Domenico Pratico (2018) Antileukotriene therapy by reducing tau phosphorylation improves synaptic integrity and cognition of P301S transgenic mice. *Ageing Cell.*

A similar watermaze study is conducted in 2018 to that described in the 2015 paper with, as far as I can tell, the same protocol. The mice at testing were again about 10 months of age. Once again, the latency to escape declines in a linear fashion (Figure 1c), but here the SEMs do appear to decline appropriately across days - albeit to a very unusual level of consistency. The WT control group performs similarly to the 2015 group, but with (perhaps) a mean escape latency of 1-3 sec slower; this is fine. In Figure panel 1f, the latency to platform is measured during the session 5 probe test (when I assume the escape platform is absent), but here a similar but different pattern (to 2015) prevails. This is that the P301s transgenic group stays much the same in terms of latency from day 4 to day 5, but the other three groups dramatically improve. Indeed, the two wild-type groups - one drug-treated, one not - appear to have identical means and SEMs. Both improve a lot overnight. This is unusual by sessions 4-5 of training.

**Paper 3 - Loss of function study**

Phillip F. Giannopoulos & Domenico Praticò (2018) Overexpression of 5-Lipoxygenase Worsens the Phenotype of a Mouse Model of Tauopathy. *Molecular Neurobiology.*

Whereas a 5LO inhibitor appears (Paper 1) to make things better, this study tested whether overexpression of 5LO would make things worse. It is reported that it does. With respect to the watermaze, a similar study is again conducted with, as far as I can tell, the same protocol excepting that the mice are bit older (11-12 months). Once again, the latency to escape declines in an almost exactly linear fashion in all groups over days 1-3 (Figure 1c), but with the groups levelling up on day 4 (presumably because of the training to criterion). The SEMs are very small and within the size of the data points by the end of training, but the WT data otherwise identical to those of the other 2018 paper. One possibility is that the WT group of the two 2018 papers is exactly same group of animals, and that the two studies were run together as a 7 group study but then published separately. That would be fine, but this is not stated, nor is there any cross-referencing of the protocol between the studies (that I could find). We are therefore left with the notion that two separate studies were likely done and both reported in 2018 and both run blind, in which the 2 WT only groups (ns = 10 each) secured absolutely identical data. This is unusual. In this study, the learning data for the WT group is essentially identical to that of the other 2018 study (Paper 2 above). Once again, the WT groups show a striking overnight improvement between days 4 and 5, whereas the *htau* group with the transgenic overexpression of 5LO gets worse. This concords with the hypothesis under test.

**Paper 4 - cdk5 kinase study - gain of function after pathology present**
Phillip F. Giannopoulos, Jian Chiu & Domenico Pratico. (2019) Learning Impairments, Memory Deficits, and Neuropathology in Aged Tau Transgenic Mice Are Dependent on Leukotrienes Biosynthesis: Role of the cdk5 Kinase Pathway. *Molecular Neurobiology.*

This is not my scientific field and so I may not have a fully clear idea of the purpose of this study. My understanding is, however, that whereas Paper 1 was about a preventative treatment in animals that started the behavioural parts of the study at 10 months, this one is about a restorative treatment after pathology has developed at 16 months. Once again zileuton is beneficial. Once again, the watermaze data show a striking linear pattern, similar to if not identical to the earlier three papers. It thereby seems odd that there are no changes in escape latency in the WT groups over an age gap of up to 8 months - this is very unusual. I have conducted an aging study in mice and did not see this same consistency - to the contrary, I saw the usual age-related decline in the means (Chen et al, Nature, 2000). It is also a surprise to me to see such consistency of data within a study as well as between studies conducted over a 3-4 day period. Note that in this 2019 study, the latency data over days 1-4 have SEMs plotted that are within the size of the data points on days 1 and 2. I know of no previous study in the watermaze in which this has been observed - there is always variability on day 1 of spatial training even when, as in this case, preceded by visible platform training. Once again, the WT and *htau* groups reached the <20 sec criterion on day 4 and then showed an abrupt change in performance on day 5 despite no further training. Zileuton was strikingly beneficial, even in these 16 month old animals. As in paper 1, also in papers 2 and 3, all statistics is presented as p values to be taken on trust. It is impossible to check the statistics.

## Summary and implications

There are several facets of the data and their implications that are noteworthy:
1) I am not alone in being very worried about journals that permit statistics to be reported as p values without supporting ANOVA F or T-test t values, together with degrees of freedom. In their absence, the reader is obliged to accept the statistics on trust. With the df given, a reader can check - and that is a simple courtesy of one scientist to another. I served on the Board of Reviewing Editors (BORE) of Science from 2007 until 2018, and one campaign I had within the journal was to smarten up its act on statistics. I was not alone on the BORE in seeking this, and my sense is that the journal has responded. It still reports studies with just p values, much to my annoyance, but things are getting better. I attach part of the Figure Legend of one of my recent papers- Takeuchi et al, Nature (2016). Journals do allow proper reporting of statistics.

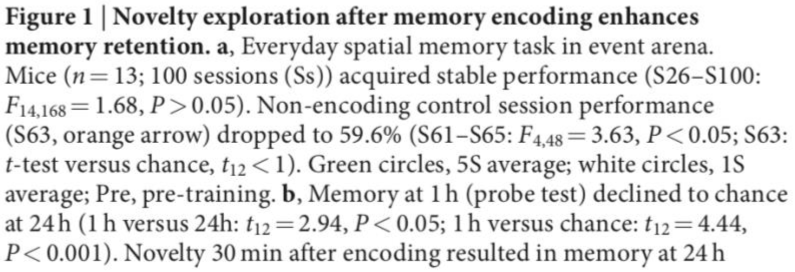

**Figure 1 | Novelty exploration after memory encoding enhances memory retention. a**, Everyday spatial memory task in event arena. Mice ($n = 13$; 100 sessions (Ss)) acquired stable performance (S26–S100: $F_{14,168} = 1.68$, $P > 0.05$). Non-encoding control session performance (S63, orange arrow) dropped to 59.6% (S61–S65: $F_{4,48} = 3.63$, $P < 0.05$; S63: *t*-test versus chance, $t_{12} < 1$). Green circles, 5S average; white circles, 1S average; Pre, pre-training. **b**, Memory at 1 h (probe test) declined to chance at 24 h (1 h versus 24h: $t_{12} = 2.94$, $P < 0.05$; 1 h versus chance: $t_{12} = 4.44$, $P < 0.001$). Novelty 30 min after encoding resulted in memory at 24 h

2) Many journals now require the individual data points to be plotted. I have mixed feelings about this as it makes figures very cluttered. But it is one way to go. One compromise might be for Figures to have a supplementary copy with the individual data points plotted, but continue to permit means and SEMs on the main figures.

3) The strikingly straight learning curves are very unusual, as is the stability across sessions and even the lack of change across studies between those conducted with young adult animals and those with relatively aged animals.

4) I do not really understand why the latencies to the platform in the session 5 probe test (when presumably the escape platform is not there) go down or up from those shown on day 4 exactly in accordance with the hypothesis being tested in each study. It is also astonishing that in 4 studies over 4-5 years, not a single hypothesis-testing comparison in the watermaze has failed to be upheld. I wish my experiments went like that - but of course they don't. That is part of the fun of science - telling it as it is.

Richard Morris

Richard G M Morris, CBE, FRS
The University of Edinburgh