Cross-species translation of the Morris maze for Alzheimer's disease

Katherine L. Possin,[1,2] Pascal E. Sanchez,[3] Clifford Anderson-Bergman,[3] Roland Fernandez,[4] Geoffrey A. Kerchner,[5] Erica T. Johnson,[1] Allyson Davis,[3] Iris Lo,[3] Nicholas T. Bott,[1] Thomas Kiely,[1] Michelle C. Fenesy,[5] Bruce L. Miller,[1] Joel H. Kramer,[1] and Steven Finkbeiner[2,6,7,8]

[1]Memory and Aging Center, Department of Neurology, UCSF, San Francisco, California, USA. [2]Hellman Family Foundation Alzheimer's Disease Research Program, San Francisco, California, USA. [3]Gladstone Institute of Neurological Disease, San Francisco, California, USA. [4]Microsoft Research, Redmond, Washington, USA. [5]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, California, USA. [6]Departments of Neurology and Physiology, UCSF, San Francisco, California, USA. [7]Keck Program in Brain Cell Engineering, Gladstone Institutes, San Francisco, California, USA. [8]Taube/Koret Center for Neurodegenerative Disease Research and Roddenberry Stem Cell Program, San Francisco, California, USA.

Analogous behavioral assays are needed across animal models and human patients to improve translational research. Here, we examined the extent to which performance in the Morris water maze — the most frequently used behavioral assay of spatial learning and memory in rodents — translates to humans. We designed a virtual version of the assay for human subjects that includes the visible-target training, hidden-target learning, and probe trials that are typically administered in the mouse version. We compared transgenic mice that express human amyloid precursor protein (hAPP) and patients with mild cognitive impairment due to Alzheimer's disease (MCI-AD) to evaluate the sensitivity of performance measures in detecting deficits. Patients performed normally during visible-target training, while hAPP mice showed procedural learning deficits. In hidden-target learning and probe trials, hAPP mice and MCI-AD patients showed similar deficits in learning and remembering the target location. In addition, we have provided recommendations for selecting performance measures and sample sizes to make these assays sensitive to learning and memory deficits in humans with MCI-AD and in mouse models. Together, our results demonstrate that with careful study design and analysis, the Morris maze is a sensitive assay for detecting AD-relevant impairments across species.

Introduction

Alzheimer's disease (AD) is becoming more prevalent in aging populations worldwide, and there are no effective treatments (1). Mouse models that recapitulate some aspects of AD have been useful for dissecting pathogenic mechanisms of AD and devising therapeutic strategies. However, clinical trials based on promising results from these models have not identified disease-modifying therapies (2, 3). There are many reasons why experimental discoveries in mice have not been translated successfully to humans (4). One reason is that the clinical tests used in patients minimally resemble the laboratory tests used to probe cognitive aspects of AD in animal models (5). Sensitive and analogous behavioral assays in animal models and humans are needed to improve the translational predictability of therapeutic strategies.

The Morris water maze (6) is the most frequently used behavioral assay of learning and memory in AD mouse models (5). Human amyloid precursor protein (hAPP) transgenic mice from line J20, which carry mutations that cause early-onset AD (7), show major deficits in this assay (8). Like humans with AD, hAPP mice have elevated levels of amyloid β (Aβ) peptides in the brain, network and synaptic dysfunction, and amyloid plaques (9). In some ways, hAPP mice better model the earlier mild cognitive impairment stage of AD (MCI-AD) than the dementia stage (10). Patients with AD or MCI-AD are impaired in real-space, 2D, and virtual human adaptations of the Morris maze (11, 12). Across species, performance in the Morris maze assay relies on hippocampal networks that are critically affected in AD (13, 14).

Although the Morris maze assay detects AD-related impairment and hippocampal dysfunction in mice and humans, its utility for translational research has been limited by major implementation differences across species. First, the typical mouse version of the Morris water maze includes visible-target training, hidden-target learning, and probe trials, but the human versions vary substantially in protocol design (15). Second, performance measures are inconsistent within and across species, and it is unclear which measures are most sensitive to deficits in MCI-AD and relevant mouse models. Third, the canonical statistical method used to analyze learning performances (repeated-measures ANOVA) violates critical statistical assumptions, leading to high rates of type I errors and unpredictability (16).

To address these limitations, we designed a virtual version of the Morris maze for humans that is analogous to the typical mouse version and includes visible-target training, hidden-target learning, and probe trials (Supplemental Video 1; supple-

▶ Related Commentary: p. 477

Authorship note: K.L. Possin and P.E. Sanchez are co–first authors.
Conflict of interest: The authors have declared that no conflict of interest exists.
Submitted: October 19, 2015; **Accepted:** December 3, 2015.
Reference information: *J Clin Invest*. 2016;126(2):779–783. doi:10.1172/JCI78464.

Table 1. Group differences in performance metrics for hAPP mice and MCI-AD patients

	Mice			Humans		
	Cohen's *d*	95% CI	*P* value	Cohen's *d*	95% CI	*P* value
Visible-target training						
Distance rank-summary score	1.00	(0.58, 1.41)	6.9×10^{-8A}	−0.33	(−0.76, 0.11)	1.3×10^{-1}
Hidden-target learning						
Distance rank-summary score	0.88	(0.48, 1.29)	2.8×10^{-6A}	0.75	(0.31, 1.20)	6.9×10^{-4A}
Latency rank-summary score	0.97	(0.56, 1.38)	8.6×10^{-8A}	0.52	(0.08, 0.96)	1.8×10^{-2}
CSE rank-summary score	0.96	(0.55, 1.38)	4.9×10^{-8A}	0.60	(0.16, 1.04)	6.1×10^{-3A}
Probe trial						
% time in target quadrant	1.0	(0.60, 1.43)	2.4×10^{-6A}	0.64	(−0.20, 1.08)	3.8×10^{-3A}
Mean proximity	1.2	(0.76, 1.61)	2.7×10^{-8A}	0.66	(0.21, 1.10)	3.1×10^{-3A}

Cohen's *d* effect size is presented with the 95% CI and *P* values were based on a 2-sample *t* test for each performance measure. Positive Cohen's *d* indicates larger values for the cases. [A]Significant after Bonferroni's correction for 6 comparisons per species ($\alpha = 0.05/6$).

mental material available online with this article; doi:10.1172/JCI78464DS1). We compared the sensitivity of performance measures and statistical methods for detecting impairments in hAPP mice and humans with MCI-AD. In addition, we present a novel sensitive summary measure to assess learning performance. Power analyses for group differences and treatment effects are presented for different sample sizes to guide preclinical and clinical study design.

Results and Discussion

Deficits in the Morris mazes were compared in hAPP mice and MCI-AD patients using consistent procedures and performance measures. In visible-target training, MCI-AD patients rapidly learned the task and navigated as directly to the target as the controls did (Table 1 and Figure 1A). Procedural learning is typically spared in mild AD (17), and the procedural demands of the assay appear to be minimal in humans. In contrast, hAPP mice had significant deficits (Table 1 and Figure 1E), as reported in this (18) and other transgenic lines (19, 20). The procedural demands of the mouse assay are substantial: mice must overcome their tendency to swim along the wall and instead learn to climb onto a platform to escape the water. A deficit in procedural learning can alter performance in the hidden-target task and confound the interpretation of spatial memory. We therefore recommend minimizing the procedural aspects of the task by training the mice to first locate a visible target.

In hidden-target learning, subjects had to learn the location of the target relative to extra-maze cues. Analysis of the distance, latency, and cumulative search error (CSE) rank-summary scores revealed that MCI-AD patients and hAPP mice were significantly impaired in this task (Table 1 and Figure 1, B–D and F–I), consistent with a deficit in hippocampus-dependent spatial learning. hAPP mice, but not MCI-AD patients, navigated more slowly than did their respective controls (Supplemental Figure 1), which may have impacted the latency rank-summary score. The distance rank-summary score was sensitive across species, with an estimated power of 80% to detect an impairment with a sample size of 14 mice and 19 humans per group (Figure 2, A and B). For this score to detect a theoretical treatment effect with 80% power, a sample size of 82 mice per group is required to detect a 0.5 effect, and a sample size of 447 humans per group is required to detect a 0.25 effect (Figure 2, C and D). This result implies that only large treatment effects can be detected by efficacy studies in hAPP mice, since the number of mice per group will likely not be greater than 50, for practical and ethical reasons. In humans, sample sizes this large are not unusual for phase III clinical trials (21). Thus, in principle, the virtual Morris maze could be used to detect the efficacy of therapeutic strategies in MCI-AD patients.

In the probe trial, the mean proximity and percentage of time in the target quadrant were both sensitive across species, but mean proximity was more sensitive than percentage of time in the target quadrant for mice, as previously reported (22) and as shown in Figure 1J and Table 1. For 80% power to detect impairment in mean proximity, a study would need 21 mice and 49 humans per group (Figure 2, A and B).

Results from the Morris water maze are usually analyzed using repeated-measures ANOVA, even though statistical assumptions required by this method are violated (Supplemental Materials). In contrast, the rank-summary method takes advantage of the identical training procedures for each subject in a given trial, greatly simplifies the methods, and does not require the assumptions of repeated-measures ANOVA (23). The rank-summary method also makes it easier to combine data from experimental cohorts that may differ slightly in subjects or procedures (Supplemental Figure 2), facilitating meta-analyses. We examined whether this new approach is as powerful as more complex methods proposed to address the limitations of repeated-measures ANOVA. Since the learning effects were linear, the Cox proportional-hazards model was the most appropriate alternative (Supplemental Materials). However, it did not increase statistical power (Supplemental Figure 3).

In both hidden-target learning and the probe trial, impairment levels were reproducible in both species but were greater in hAPP mice. This difference could reflect differences in the time between learning trials in the two assays. In mice, but not humans,

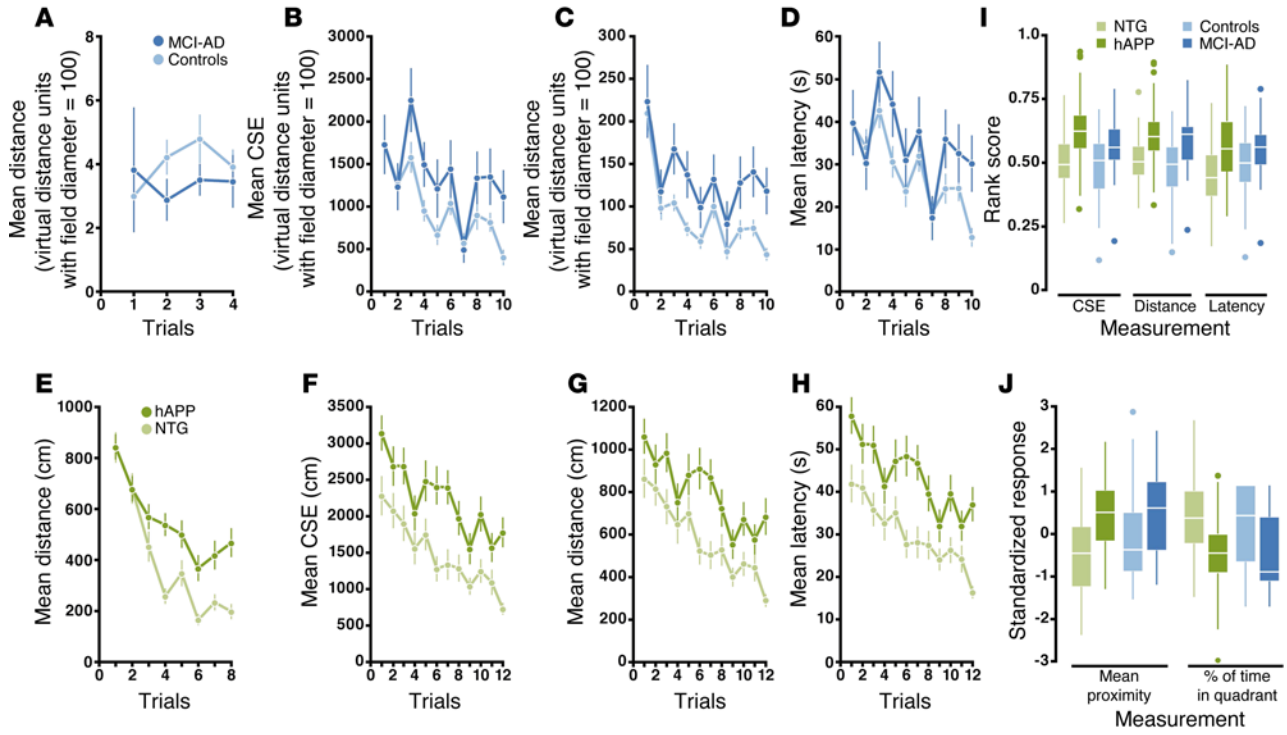

Figure 1. Morris maze deficits in MCI-AD patients and hAPP mice. (**A**–**H**) Learning curves for humans and mice in visible-target training (**A** and **E**) and hidden-target learning (**B**–**D** and **F**–**H**). Values represent the mean ± SEM. (**I**) Hidden-target learning rank-summary scores for CSE, distance, and latency across species. (**J**) Delayed probe trial standardized mean proximity and percentage of time spent in the target quadrant across species. (**I** and **J**) Boxes represent the median and first and third quartiles, whiskers represent non-outliers within the 1.5 × interquartile range from the edge of the boxes, and dots represent outliers. For all panels, *n* = 53 mice per genotype, 28 patients with MCI-AD, and 89 human controls. NTG, nontransgenic.

the trials were spaced to avoid the effects of hypothermia on performance. Spaced training enhances learning in healthy mice and humans (24, 25) but likely increases the difficulty of the task for hAPP mice and MCI-AD patients due to forgetting between trials (26, 27). Another possibility is species differences in incentive to complete the task. Since only the mouse assay was aversive, human controls may not have been as motivated as control mice to reach the target quickly.

Several recommendations and important considerations emerged from our translational study. First, the Morris maze can detect similar deficits in spatial learning and memory across species. Second, in mice, visible-target training should be conducted first to reduce the influence of procedural learning on subsequent hidden-target trials. Third, future research should examine whether spaced hidden-target training and stronger incentives yield greater impairments in MCI-AD patients. Fourth, the rank-summary score avoids the assumptions required of repeated-measures ANOVA and is as sensitive as more complex methods that are difficult to apply appropriately. Fifth, the rank-summary score for distance provides a sensitive cross-species measurement of learning and may be superior to latency. Sixth, adequate power can be obtained with these methods to detect clinically relevant treatment effects in human trials.

Methods

Mice. hAPPJ20 mice (7) were maintained on a C57BL/6J background by crossing heterozygous transgenic mice with nontransgenic C57BL/6J breeders (The Jackson Laboratory). Mice had access to food (PicoLab Rodent Diet 20; LabDiet) and water ad libitum. The same protocol was administered to 3 independent cohorts of sex-balanced 4- to 7-month-old nontransgenic (*n* = 53) and hAPP (*n* = 53) mice. All transgenic mice were heterozygous with respect to the transgene. Nontransgenic littermates served as controls.

Humans. Subjects were recruited from the UCSF Memory and Aging Center's longitudinal observational studies and from a Stanford University study on normal aging and mild cognitive impairment. Participants were diagnosed as neurologically healthy (*n* = 89) or as having MCI-AD (*n* = 28) (28).

Mouse and human Morris mazes. Mice were administered the Morris water maze test in an opaque pool with a submerged platform, and humans were administered a virtual version in a circular field with a buried treasure using a 30-in. monitor and a simple driving simulator. During visible-target training, no extra-maze landmarks were available; a visual cue indicated the target location. During hidden-target learning and probe trials, consistent extra-maze landmarks were presented. Learning trials ended when the target was found or a time limit was reached. In the probe trial, the target was removed, and subjects were allowed to navigate for 90 seconds. For details, see Supplemental Materials.

Performance measures. Rank-summary scores were computed for distance (visible-target training) and for distance, time, and CSE (hidden-target learning). Raw scores were replaced with quantile scores for each trial, which were averaged for each subject. CSE is the sum of the 1-second average proximities to the target (29). For

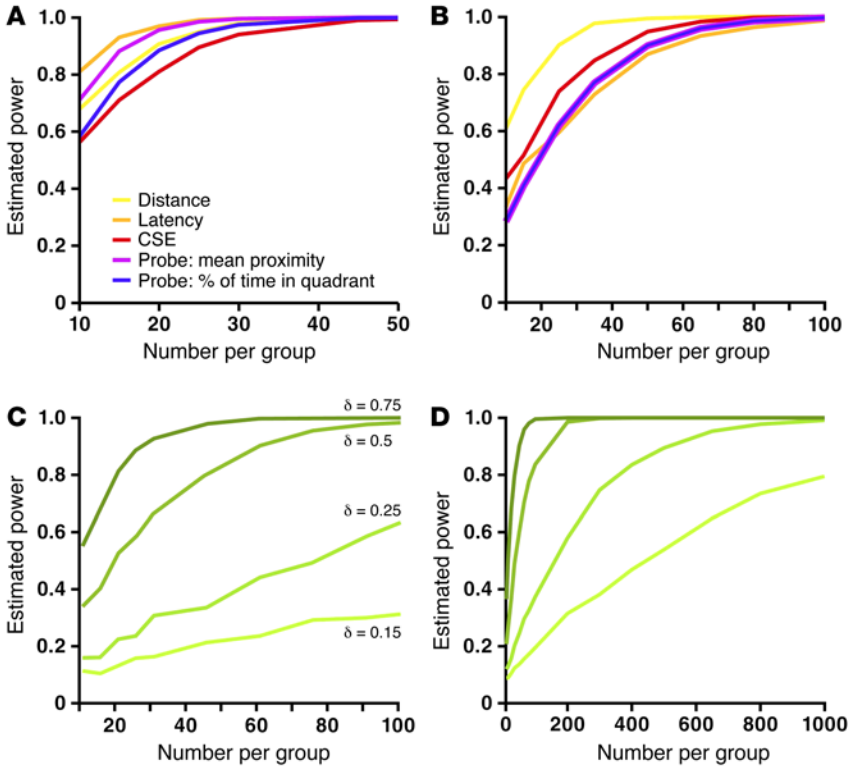

Figure 2. Sample size required to detect a deficit and a treatment effect in hAPP mice or MCI-AD patients across performance measures. Probability of detecting a difference between hAPP mice (**A**) or MCI-AD patients (**B**) and their respective controls and probability of detecting treatment effects of different magnitudes in hAPP mice (**C**) or MCI-AD patients (**D**). In all analyses, n = 53 mice per genotype, 28 patients with MCI-AD, and 89 human controls. The type I error rate was set at 0.05. Cases and controls were re-sampled according to a given sample size using bootstrap methods (30).

probe trials, we examined the percentage of time spent in the target quadrant and mean proximity to the target (29). Although the same protocol was used for the mouse cohorts, a linear regression model with indicators for genotype and cohort was used to adjust for possible training differences.

Statistics. All measures in cases versus controls were compared by 2-tailed t tests with Bonferroni's correction for 6 comparisons per species. A P value of less than 0.0083 was considered significant (Table 1). For power analyses of the treatment effect, untreated subjects were resampled from the cases; for a given treatment effect of δ, treated subjects were resampled from controls with a probability of δ and re-sampled from the cases with a probability of 1–δ (i.e., δ represents the proportion of subjects cured). For each configuration, 5,000 bootstrap samples were taken, leading to Monte Carlo standard errors below 0.007.

Study approval. The IACUC of the UCSF approved all mouse experiments. The UCSF and Stanford committees on human research approved the human study. Written informed consent was obtained for each subject.

Author contributions

KLP, PES, and SF conceived and supervised the study. KLP, PES, CAB, and ETJ conducted data analysis. RF programmed the virtual assay. KLP, GAK, ETJ, NB, MCF, and TK conducted human experiments. AD and IL conducted animal experiments. KLP, PES, CAB, BLM, JHK, and SF wrote the manuscript. KLP, GAK, and SF provided funding for the study.

Acknowledgments

We thank Lennart Mucke for providing behavioral data on hAPP-J20 mice; Mariel Finucane, Michael Gill, and Erik Johnson for their input on the manuscript; Charlie Toohey for generous programming support; Kelley Nelson for administrative support; Gary Howard and Stephen Ordway for editing; and Giovanni Maki and Teresa Roberts for graphical support. Steven Finkbeiner and Katherine Possin were supported by the Hellman Family Foundation. Steven Finkbeiner was also supported by the Taube/Koret Center for Neurodegenerative Diseases and Katherine Possin by a National Institute on Aging (NIA) grant (K23AG037566). Geoffrey Kerchner received funding from the NIA (K23AG042858); the American Federation for Aging Research; and the McKnight Endowment Fund for Neuroscience. Human studies were supported by the NIA (P50AG023501) and the Larry L. Hillblom Foundation. Mouse studies were supported by the NIH (P30NS065780) and an NIH Extramural Research Facilities Improvement Program Project grant (C06 RR018928).).

Address correspondence to: Katherine Possin, 675 Nelson Rising Lane, Ste 190, San Francisco, California 94158, USA. Phone: 415.476.1889; E-mail: kpossin@memory.ucsf.edu. Or to: Pascal Sanchez or Steven Finkbeiner, 1650 Owens Street, San Francisco, California 94158, USA. Phone: 415.734.2518; E-mail: pascal.sanchez@gladstone.edu (P. Sanchez). Phone: 415.734.2508; E-mail: sfinkbeiner@gladstone.edu (S. Finkbeiner).

1. World Health Organization Alzheimer's Disease International. *Dementia: A Public Health Priority*. Geneva, Switzerland: World Health Organization; 2010.
2. Finkbeiner S. Bridging the Valley of Death of therapeutics for neurodegeneration. *Nat Med*. 2010;16(11):1227–1232.
3. Zahs KR, Ashe KH. 'Too much good news' — are Alzheimer mouse models trying to tell us how to prevent, not cure, Alzheimer's disease? *Trends Neurosci*. 2010;33(8):381–389.
4. Pankevich, et al. Improving and accelerating drug development for nervous system disorders. *Neuron*. 2014;84(3):546–553.
5. Webster SJ, Bachstetter AD, Nelson PT, Schmitt FA, Van Eldik LJ. Using mice to model Alzheimer's dementia: an overview of the clinical disease and the preclinical behavioral changes in 10 mouse models. *Front Genet*. 2014;5:88.
6. Morris RG. Spatial localization does not require the presence of local cues. *Learn Motiv*. 1981;12(2):239–260.
7. Mucke L, et al. High-level neuronal expression of abeta 1-42 in wild-type human amyloid protein precursor transgenic mice: synaptotoxicity without plaque formation. *J Neurosci*. 2000;20(11):4050–4058.
8. Sanchez PE, et al. Levetiracetam suppresses neuronal network dysfunction and reverses synaptic and cognitive deficits in an Alzheimer's disease model. *Proc Natl Acad Sci*. 2012;109(42):E2895–E2903.
9. Huang Y, Mucke L. Alzheimer mechanisms and therapeutic strategies. *Cell*. 2012;148(6):1204–1222.
10. Ashe KH, Zahs KR. Probing the biology of Alzheimer's disease in mice. *Neuron*. 2010;66(5):631–645.
11. Burgess N, Trinkler I, King J, Kennedy A, Cipolotti L. Impaired allocentric spatial memory underlying topographical disorientation. *Rev Neurosci*. 2006;17(1–2):239–252.
12. deIpolyi A, Rankin K, Mucke L, Miller B, Gorno-Tempini M. Spatial cognition and the human navigation network in AD and MCI. *Neurology*. 2007;69(10):986–997.
13. O'Keefe J, Nadel L. *The Hippocampus As A Cognitive Map*. Oxford, United Kingdom: Clarendon Press; 1978.
14. Vlček K, Laczó J. Neural correlates of spatial navigation changes in mild cognitive impairment and Alzheimer's disease. *Front Behav Neurosci*. 2014;8:89.
15. Boccia M, Nemmi F, Guariglia C. Neuropsychology of environmental navigation in humans: review and meta-analysis of fMRI studies in healthy participants. *Neuropsychol Rev*. 2014;24(2):236–251.
16. Young ME, Clark MH, Goffus A, Hoane MR. Mixed effects modeling of Morris water maze data: advantages and cautionary notes. *Learn Motiv*. 2009;40:160–177.
17. Libon DJ, et al. Declarative and procedural learning, quantitative measures of the hippocampus, and subcortical white alterations in Alzheimer's disease and ischaemic vascular dementia. *J Clin Exp Neuropsychol*. 1998;20(1):30–41.
18. Roberson ED, et al. Reducing endogenous tau ameliorates amyloid β-induced deficits in an Alzheimer's disease mouse model. *Science*. 2007;316(5825):750–754.
19. King DL, Arendash GW. Behavioral characterization of the Tg2576 transgenic model of Alzheimer's disease through 19 months. *Physiol Behav*. 2002;75(5):627–642.
20. Janus C, Flores AY, Xu G, Borchelt DR. Behavioral abnormalities in APPSwe/PS1dE9 mouse model of AD-like pathology: comparative analysis across multiple behavioral domains. *Neurobiol Aging*. 2015;36(9):2519–2532.
21. Doody, et al. Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. *N Engl J Med*. 2014;370(4):311–321.
22. Maei HR, Zaslavsky K, Teixeira CM, Frankland PW. What is the most sensitive measure of water maze probe test performance? *Front Integr Neurosci*. 2009;3:4.
23. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. Boston, Massachusetts. USA: John Wiley & Sons; 2004.
24. Commins S, Cunningham L, Harvey D, Walsh D. Massed but not spaced training impairs spatial memory. *Behav Brain Res*. 2003;139(1):215–223.
25. Jackson CE, Maruff PT, Snyder PJ. Massed versus spaced visuospatial memory in cognitively healthy young and older adults. *Alzheimers Dement*. 2013;9(1 suppl):S32–S38.
26. Walsh CM, Wilkins S, Bettcher BM, Butler CR, Miller BL, Kramer JH. Memory consolidation in aging and MCI after 1 week. *Neuropsychology*. 2014;28(2):273.
27. Daumas S, et al. Faster forgetting contributes to impaired spatial memory in the PDAPP mouse: deficit in memory retrieval associated with increased sensitivity to interference? *Learn Mem*. 2008;15(9):625–632.
28. Albert MS, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. 2011;7(3):270–279.
29. Gallagher M, Burwell R, Burchinal MR. Severity of spatial learning impairment in aging: development of a learning index for performance in the Morris water maze. *Behav Neurosci*. 1993;107(4):618.
30. Beran R. Simulated power functions. *Ann Statist*. 1986;14(1):151–173.