Memorability of photographs in subjective cognitive decline and mild cognitive impairment: implications for cognitive assessment

Wilma A. Bainbridge[a*], David Berron[b,c], Hartmut Schütze[b,c], Arturo Cardenas-Blanco[b,c], Coraline Metzger[b,c,d], Laura Dobisch[c] , Daniel Bittner[c,e], Wenzel Glanz[c], Annika Spottke[f,g], Janna Rudolph[f], Frederic Brosseron[f,h], Katharina Buerger[i,j], Daniel Janowitz[j], Klaus Fliessbach[e], Michael Heneka[f,h], Christoph Laske[k,l], Martina Buchmann[k,l] , Oliver Peters[m,n], Dominik Diesing[n], Siyao Li[n], Josef Priller[m,o], Eike Jakob Spruth[o], Slawek Altenstein[m], Anja Schneider[f,h], Barbara Kofler[h], Stefan Teipel[p,q], Ingo Kilimann[p,q], Jens Wiltfang[r,s], Claudia Bartels[r,s], Steffen Wolfsgruber[f], Michael Wagner[f,h], Frank Jessen[f,t], Chris Baker[a], **Emrah Düzel**[b,c,u*]

[a] Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, United States

[b] Institute of Cognitive Neurology and Dementia Research, Otto-von-Guericke University Magdeburg, Leipziger Str. 44, 39120 Magdeburg, Germany

[c] German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Leipziger Str. 44, 39120 Magdeburg, Germany

[d] Department of Psychiatry and Psychotherapy, University Hospital Magdeburg, Medical Faculty, Leipziger Str. 44, 39120 Magdeburg, Germany

[e] Clinic for Neurology, University Hospital Magdeburg, Medical Faculty, Leipziger Str. 44, 39120 Magdeburg, Germany

[f] German Center for Neurodegenerative Diseases (DZNE), Bonn, Sigmund-Freud-Str. 27, 53127 Bonn, Germany

[g] Department of Neurology, University of Bonn, Sigmund-Freud-Str. 25, 53127 Bonn, Germany

[h] Department of Neurodegeneration and Geriatric Psychiatry, University Hospital Bonn, Sigmund-Freud-Str. 25, 53127 Bonn, Germany

[i] German Center for Neurodegenerative Diseases (DZNE), Munich, Feodor-Lynen-Str. 17, 81377 Munich, Germany

[j] Institute for Stroke and Dementia Research, University Hospital, LMU Munich, Feodor-Lynen-Str. 17, 81377 Munich, Germany

[k] German Center for Neurodegenerative Diseases (DZNE), Tübingen, Otfried-Müller-Str. 23, 72076 Tübingen, Germany

[l] Section for Dementia Research, Hertie Institute for Clinical Brain Research and Department of Psychiatry and Psychotherapy, University of Tübingen, Calwerstr. 14, 72076 Tübingen, Germany

[m] German Center for Neurodegenerative Diseases (DZNE), Berlin, Charitéplatz 1, 10117 Berlin, Germany

[n] Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Psychiatry and Psychotherapy, Hindenburgdamm 30, 12203 Berlin, Germany

[o] Department of Psychiatry and Psychotherapy, Charité,, Charitéplatz 1, 10117 Berlin, Germany

[p] German Center for Neurodegenerative Diseases (DZNE), Rostock, Gehlsheimer Str. 20, 18147 Rostock, Germany

[q] Department of Psychosomatic Medicine, Rostock University Medical Center, , Gehlsheimer Str. 20, 18147 Rostock, Germany

[r] German Center for Neurodegenerative Diseases (DZNE), Goettingen, Von-Siebold-Str. 3a, 37075 Goettingen, Germany

[s] Department of Psychiatry and Psychotherapy, University Medical Center Goettingen, University of Goettingen, Von-Siebold-Str. 5, 37075 Goettingen, Germany

[t] Department of Psychiatry, University of Cologne, Medical Faculty, Kerpener Str. 62, 50924 Cologne, Germany

[u] Institute of Cognitive Neuroscience, Univ. College London, London, UK

***Correspondence:**

Dr. Wilma A. Bainbridge, National Institute of Mental Health

NIH Building 10, Rm 4C-108, 10 Center Drive, Bethesda MD 20814, USA;
wilma.bainbridge@nih.gov

and

Prof. Dr. Emrah Düzel, German Center for Neurodegenerative Diseases

Leipziger Str. 44, 39120 Magdeburg, Germany; emrah.duezel@dzne.de

Abstract

INTRODUCTION: Impaired long-term memory is a defining feature of Mild Cognitive Impairment (MCI). We tested whether this impairment is item-specific, limited to some memoranda whereas some remain consistently memorable.

METHODS: We conducted item-based analyses of long-term visual recognition memory. 394 participants (healthy controls (HC), Subjective Cognitive Decline (SCD), and MCI) in the multicentric DZNE-Longitudinal Cognitive Impairment and Dementia Study (DELCODE) were tested with images from a pool of 835 photographs.

RESULTS: We observed consistent memorability for images in HCs, SCDs, and MCI, predictable by a neural network trained on another healthy sample. Looking at memorability differences between groups, we identified images that could successfully categorize group membership with higher success and a substantial image reduction than the original image set.

DISCUSSION: Individuals with SCD and MCI show consistent memorability for specific items, while other items show significant diagnosticity. Certain stimulus features could optimize diagnostic assessment, while others could support memory.

Keywords: Alzheimer’s disease (AD), subjective cognitive decline (SCD), mild cognitive impairment (MCI), memorability, diagnostic assessment, image analysis

1. Background

Recent work in healthy individuals has found that certain images are intrinsically memorable or forgettable across observers [1,2]; there are images of faces or scenes that most people remember or forget, regardless of their different individual experiences. This *memorability* of an image can be quantified and predicts 50% of the variance in people's performance on a memory test [2]. It is intrinsic to the image itself, stable across different image contexts [3], tasks [4,5], and timing [6,7]. Viewing memorable images automatically elicits specific neural signatures [8,9], and the memorability score of an image can be predicted by computational models [10,11]. However, image attributes such as aesthetics, emotionality, typicality, or what people believe will be memorable do not fully predict memorability [2,12], and memorability is an automatically processed image property that is resilient to the effects of attention [4]. This means that researchers can predict in advance what images a person is likely to remember or forget, and use such information to create memorable educational materials, or design well-balanced memory tests.

While memorability has so far been characterized based on healthy participants' memory behavior, it is unclear if memorability is also consistent in populations with memory impairments at increased risk for Alzheimer's Disease (AD), such as Mild Cognitive Impairment (MCI) or Subjective Cognitive Decline (SCD) [13]. Consistent memorability in SCD and MCI would enable better prediction of what images are likely to be remembered or forgotten. Furthermore, changes in memorability patterns across disease stages could improve cognitive staging and design of cognitive progression markers. By avoiding highly memorable images, cognitive tests could be made more time efficient and more sensitive. Understanding which stimulus features improve or impair memorability could provide insights into the cognitive processes that are impaired. Furthermore, knowledge about memorability could aid in the design of memorable environments or allow clinicians to focus on aiding memory for forgettable items.

In the current study, we analyzed the performance of 394 individuals, including those with SCD, MCI, and healthy controls (HC), on a visual recognition memory test in which each participant had to memorize a randomly selected subset of 88 photographs from a pool of 835. This randomization afforded us the possibility to assess memorability unconfounded by systematic effects of stimulus-selection or stimulus-order effects. First, we find significant similarities across groups in the images they remember and forget, and similarities to a convolutional neural network (CNN) trained on memorability, allowing the precise prediction of memory performance for each group. Second, we find a separate set of images that can reliably differentiate groups, with meaningful implications for diagnosis. Finally, using a large-scale online experiment to score the images, we analyze what image features might lead to the memorability and diagnosticity of different images.

2. Methods

2.1 Study design

Visual memory tests were analyzed from the DZNE-Longitudinal Cognitive Impairment and Dementia Study (DELCODE), an observational, longitudinal memory clinic-based study across 10 sites in Germany. Specific details about this study, the visual memory task, and data handling and quality control are reported in Jessen et al. [14] and Düzel et al. [15]. The data analyzed in this study were from the second data release from the DELCODE study comprising of 700 individuals of which 394 participants with complete datasets were analyzed, including 136 participants with SCD, 65 with MCI, and 193 HC. Individuals with SCD and MCI were recruited through referrals and self-referrals, while HC were recruited through public advertisements. Group membership was determined using the CERAD neuropsychological battery [16]. MCI individuals were defined as those with test performance under 1.5 standard deviations below the age-, sex-, and education-adjusted mean performance. SCD and HC individuals were defined as those with performance above this cutoff, but SCD individuals subjectively reported decline in cognitive functioning with concerns.

The study protocol was approved by all involved centers' institutional review boards and ethical committees, and all participants gave written informed consent. DELCODE is retrospectively registered at the German Clinical Trials Register (DRKS00007966), (04/05/2015).

2.2 Visual memory test

Participants performed an fMRI scene image encoding and retrieval task [17]. First, while in the fMRI scanner, participants studied 88 novel scene target images (44 indoor and 44 outdoor scenes) and 44 repetitions of two pre-familiarized images (one indoor and one outdoor, 22 times each). All images were 8-bit gray scale, presented on an MR-compatible LCD screen (Medres Optostim), scaled to 1250 x 750 pixel resolution and matched for luminance, with a viewing horizontal half-angle of 10.05° across scanners. Each image was presented for

2500ms (with an optimized jitter for statistical efficiency), and participants categorized them as "indoor" or "outdoor" with a button press. Outside of the scanner after a 70-minute delay, participants completed a recognition memory task with these 88 images and 44 novel foil images (22 indoor and 22 outdoor). Participants indicated their recognition memory with a 5-point scale: 1) *I am sure that this picture is new*, 2) *I think that this picture is new*, 3) *I cannot decide if this picture is new or old*, 4) *I think I saw this picture before*, or 5) *I am sure that I did see this picture before*. Results from the fMRI study are reported in [17].

While each participant was tested on 88 target images and 44 foil images, these images were randomly sampled from a larger set of 835 scene images, allowing us to conduct image-based analyses on a large set of images (see Figure 1 for example images). This randomization allowed us to avoid confounding effects of image selection and image order on memory performance. On average, each image served as a target image for 20.3 HC, 14.3 SCD, and 6.8 MCI individuals.

2.3 Analyzing similarity of MCI, SCD, and healthy individuals: Predicting performance

We first asked whether there are consistencies in memory performance for MCI and SCD just as there are for healthy individuals [1]; i.e., whether there are certain images that they tend to remember or forget, and, if such consistencies exist, to what degree they align with the images that tend to be remembered and forgotten by HCs.

To address this question, Spearman's rank correlations of hit rate (HR) performance on images in the visual memory task were calculated between the different groups. To assess memorability consistency, we conducted a *consistency analysis* as described in Isola et al. [1], where participants are split into random halves (across 1000 iterations) and their HRs are calculated for all images, and Spearman's rank correlated between the two halves. We also examined whether a convolutional neural network (CNN) that is significantly able to predict memory performance in healthy individuals [11] could predict memorability for SCD and MCI groups. MemNet is a CNN with the architecture and pretraining set of Hybrid-CNN [18], a CNN able to classify object and scene images, then trained to predict the memorability score of an image (i.e., the likelihood for that image to be remembered by any given person). The training of MemNet was originally conducted with a separate set of images in a separate set of healthy adults recruited online [11], and here we tested it with new images and data across participant groups from the current study. Specifically, we obtained MemNet scores for each of the 835 stimulus images and used Spearman's rank correlations to test the degree to which memorability CNN-predicted memory scores were correlated with participant group memory scores.

2.4 Analyzing dissimilarity of MCI, SCD and healthy individuals: Differentiating groups

An equally important question is whether there is a set of images in which consistencies in memory performance reliably differ between impaired populations and healthy individuals. If such images exist, then they could form an optimized test to distinguish memory impaired individuals from healthy controls with high efficiency.

To explore this question, we conducted an analysis we call the *Iterative Image Subset* (IIS) *Analysis* to compare the groups. Here, we describe the analysis comparing MCI to HC, however the same analysis was also conducted with SCD versus HC. First, the HC participant pool was randomly downsampled so that the same number of HC were used in the analysis as MCI individuals. The entire pool of participants was then split into two random halves (Group A and Group B). HR on the memory task was calculated for each image for the HC ($HR_{GroupA,Healthy}$) and for the MCI individuals ($HR_{GroupA,MCI}$) in Group A. Using this performance metric, we formed three subsets of images. The number of images used in each subset was selected iteratively for all possible subset sizes, ranging from 0% to 100% of images (835 images) in 1% increments, to determine the optimal image subset size. Only images with at least 4 individuals' data were included in the analysis. The three resulting subsets were:

1) **"H>M"**, the top set of images where HC outperformed MCI (i.e., maximizing $HR_{GroupA,Healthy} - HR_{GroupA,MCI}$; note that it is "H>S" for a comparison with SCD)
2) **"H<M"**, the top set of images where MCI outperformed HC (i.e., maximizing $HR_{GroupA,MCI} - HR_{GroupA,Healthy}$)
3) **"H=M"**, the top set of images where HC performed most similarly to MCI (i.e., minimizing $| HR_{GroupA,Healthy} - HR_{GroupA,MCI} |$)

We then assessed the performance of classifying subjects in Group B using each of the three subsets of images. Specifically, using just the images in a single subset (e.g., H>M), we determined the HR for each of the individuals in Group B (HR_{GroupB}). We then performed a Receiver Operating Characteristic (ROC) analysis to determine the diagnostic ability of this subset of images, applying a range of HR cutoffs from 0 to 1 to classify an individual from Group

B as either HC or MCI, using HR_{GroupB}. We calculated the accuracy of this test based on group membership, and contrasted successful MCI diagnosis (sensitivity, or true positive rate) with misclassification of HC (specificity, or 1 - false positive rate). We assessed classification performance by Area Under the Curve (AUC), where a score of 1 indicates perfect performance, while 0.5 indicates chance performance. This complete analysis was conducted across 100 random participant splits into Group A and B.

2.5 Finding image attributes that distinguish these image sets

To see what aspects of the images may determine their membership into different image sets, we conducted an experiment using the online crowd-sourcing platform Amazon Mechanical Turk (AMT). For each of the 835 images, 12 online participants rated the scene in the image on five relevant properties identified in previous scene perception and memorability research [12,19] using a 5-point Likert scale: size (the perceived size of the portrayed scene, not the image pixel size), clutter, aesthetics, interest, and whether they think they would remember the image (subjective memorability). They also indicated whether the image showed a natural or manmade scene and if there was a person present. 450 people anonymously participated in the study and provided consent, and this study was approved by the National Institutes of Health (NIH) Office of Human Subjects Research Protections. Two main comparisons were tested for each attribute, using paired samples t-tests: 1) forgettable versus memorable images with similar performance between HC and MCI/SCD individuals, 2) diagnostic versus non-diagnostic images, where HC and MCI/SCD individuals differed in their performance.

Forgettable and memorable images were identified as the top set of images where both HC and impaired individuals had average performance below or above (respectively) median performance, and the difference between groups was minimized (i.e., H=M). Diagnostic and non-diagnostic images were selected from the sets resulting from the IIS analysis (Section 2.4), e.g., H>M and H<M image sets, respectively. The number of images in each set was taken as the optimal number of images identified from the IIS analysis.

We also examined how memorability and diagnosticity relate to more meta-cognitive attributes: similarity to other images and confidence ratings of the participants. First, it is possible that the memorability or diagnosticity of an image is related to how similar that image is to other images in a set (e.g., memorable images are more visually unique). To assess image similarity, we used an object classification CNN called AlexNet CNN [20]. This classification CNN is often used as a model for the human visual system, showing similarities to the brain for visual processing of objects [21] and scenes [22]. This CNN can thus approximate the neural representations of an image at different levels of extraction (i.e., low-, mid-, and high-level visual features). For each classification CNN layer, we obtained the outputs for all 835 images and calculated their average Pearson correlation to all other images. Second, we also analyzed proportion of high confidence ratings given to each image by participants in the main experiment, to see if memory confidence is related to image diagnosticity.

3. Results

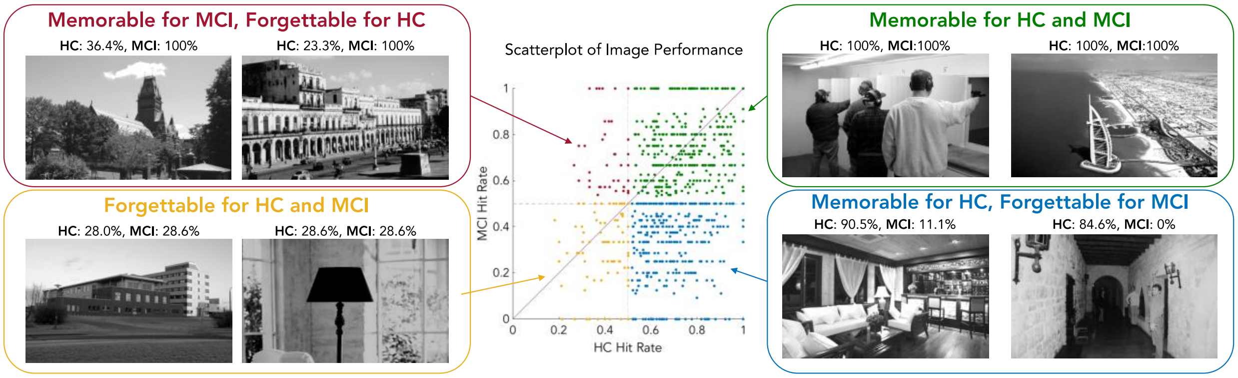

Figure 1: Example images and group performance. The scatterplot shows the distribution of memory performance (hit rate) for all 835 images for healthy controls (HC) versus individuals with Mild Cognitive Impairment (MCI). The diagonal line indicates the points at which performance is equal between both groups. Based on performance, images can be conceptually sorted into four quadrants: 1) images that are memorable to both HC and MCI individuals (green), 2) images that are memorable to HC but forgettable to MCI (blue), 3) images that are forgettable to both groups (yellow), and images that are memorable to MCI but forgettable to HC (red). Example images and performances at the extreme ends for each quadrant are arranged around the scatterplot. In the work that follows, we analyze these four groups of images and determine if they can be used meaningfully to predict memory performance.

3.1 Consistencies in the memories of participant groups

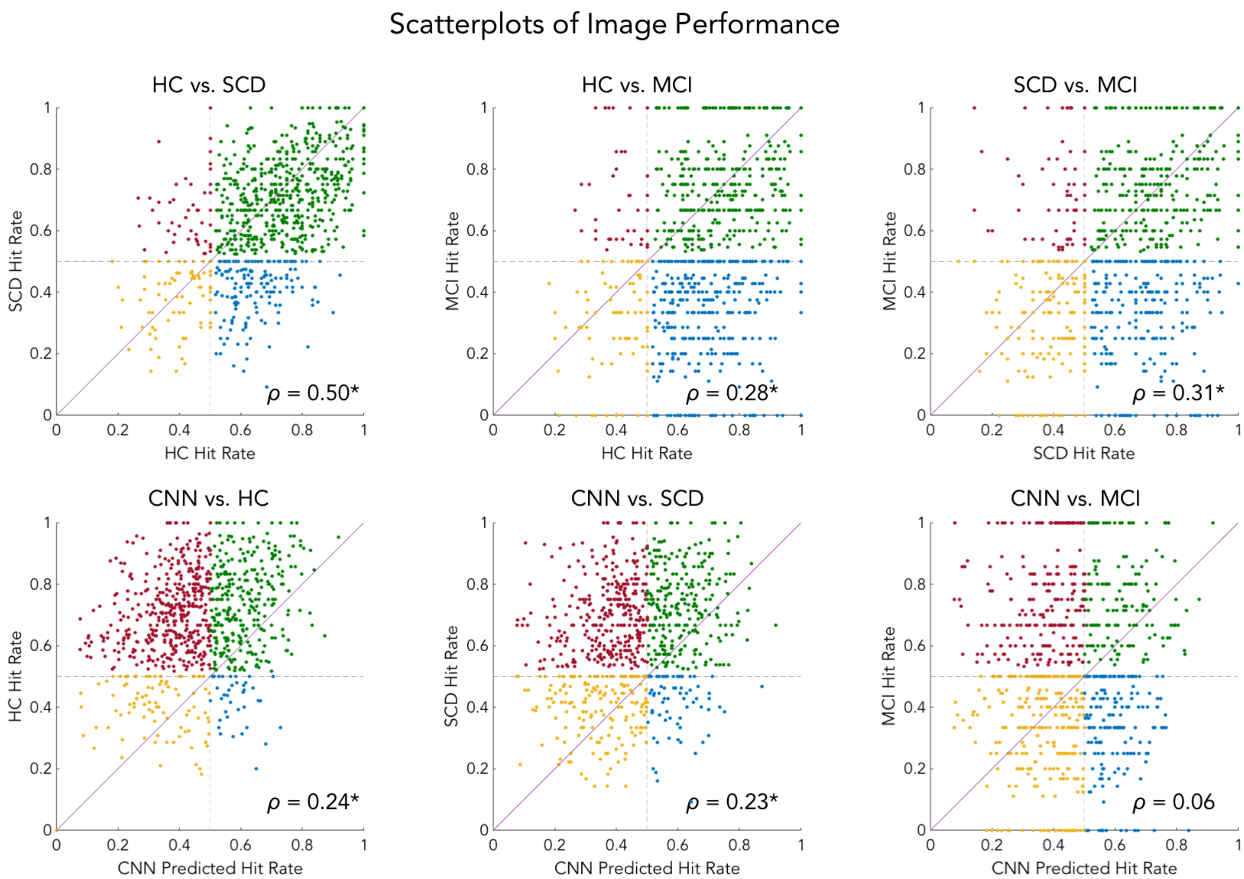

Figure 2: Consistencies across groups and the memorability neural network. The scatterplots show a comparison of hit rates for each of the 835 images between all pairings of the experimental groups (Healthy Controls, HC; Subjective Cognitive Decline, SCD; Mild Cognitive Impairment, MCI), as well as predicted hit rate from the memorability prediction convolutional neural network (CNN). Spearman's rank correlation (ρ) is shown for each plot, and asterisks (*) indicate significant correlations. Scatterplot points are colored by quadrant (as in Figure 1), and the diagonal line indicates points where both groups show equal performance.

As expected, participant groups with increasing memory impairment showed decreases in average memory performance (HC: M=0.68, SD=0.17; SCD: M=0.62, SD=0.18; MCI: M=0.53, SD=0.26). However, there were also impressive correlations across groups in the images they remembered best or worst (Figure 2). HC and SCD had a significant Spearman's rank correlation of ρ=0.50 (p=1.03 × 10^{-54}), while HC and MCI had a significant correlation of ρ=0.28 (p=1.34 × 10^{-16}), and SCD and MCI had a significant correlation of ρ=0.31 (p=2.12 × 10^{-19}). HC performance was significantly more similar to SCD performance than MCI performance (Z=6.13, p~0), and SCD performance was significantly more similar to HC performance than MCI performance (Z=5.42, p~0). These results indicate that all participant groups tended to remember the same images as each other. All groups were also internally consistent (HC: ρ=0.42; SCD: ρ=0.32; MCI: ρ=0.22; all $p < 0.0001$), meaning a memory impaired individual will still tend to remember similar images to someone else with the same diagnosis.

The MemNet CNN trained to predict image memorability showed significant correlations with HC (ρ=0.24, p=3.29 × 10^{-12}) and SCD behavior (ρ=0.23, p=1.84 × 10^{-11}), while MCI behavior correlations did not pass significance thresholds (ρ=0.06, p=0.080).

3.2 Differentiating memory impaired groups from healthy controls

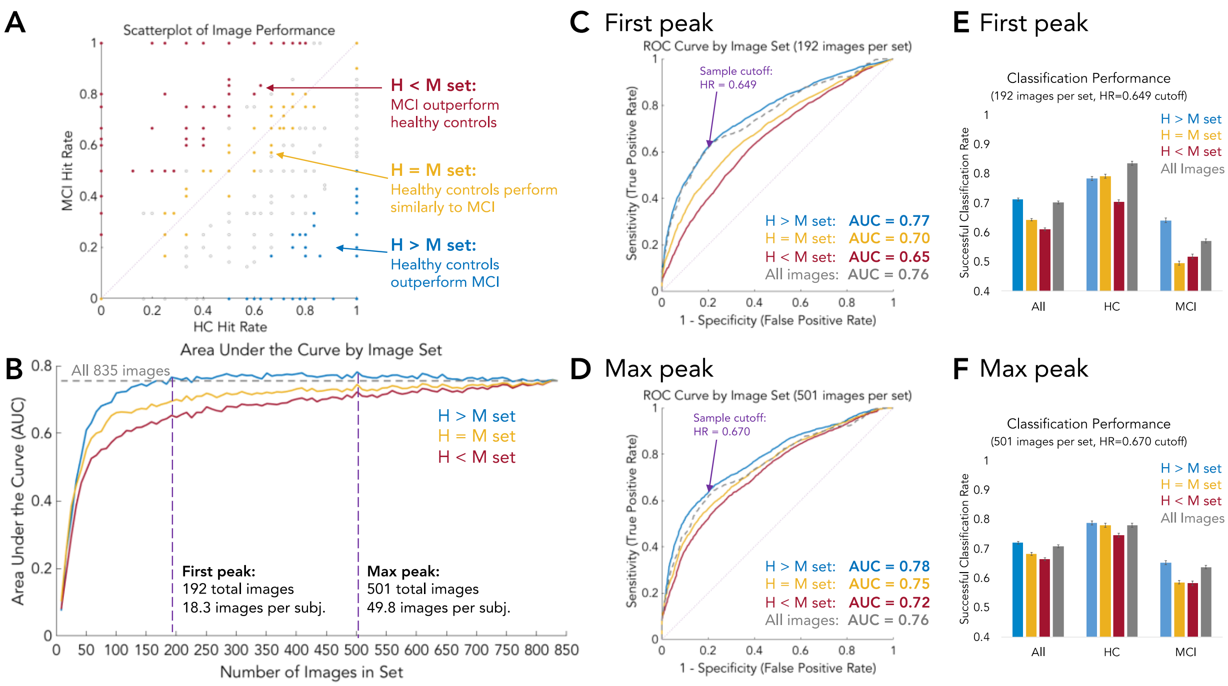

Figure 3: Finding the optimal number of images to diagnose MCI. A) This scatterplot of image performance shows an example of the three possible subsets the images can be divided into: H<M (red), H=M (yellow), and H>M (blue). B) Area Under the Curve (AUC) by image set and number of images in the set. Testing each of these subset types at different set sizes, we find that the H>M set (blue line) consistently outperforms the other image subsets at all set sizes. Importantly, the H>M set also outperforms the all-image set (gray dotted line) at a surprisingly small number of images, first overtaking the all-image set at only 192 images versus the 835 images used in the all-image set. From this set of 192 images, each participant saw on average only 18.3 images. C & D) Receiver Operating Characteristic (ROC) curves for two peaks – the first peak where H>M overtakes the all-image set, and the max peak where H>M has the largest difference from the all-image set. E & F) Participant classification performance, averaged across 100 iterations of participant split-halves, at a sample cutoff (determined as the point where the *sensitivity + specificity* is at its maximum), broken down by participant type for the different image sets. Error bars indicate standard error of the mean across the 100 iterations. Note that the optimized H>M image subset particularly shows a boost in MCI diagnosis sensitivity over all other image sets.

As a first test, we examined the ability to differentiate HC and MCI individuals. The IIS analysis shows that the H>M image subset consistently outperforms the H=M and H<M image subsets at all subset sizes, in diagnosing individuals as MCI versus HC (Figure 3). This means that images that are highly memorable to healthy controls but highly forgettable to MCI individuals are best able to distinguish these two groups. Surprisingly, H>M image subsets as small as 23% of the original image set were able to surpass the original image set in diagnostic ability. With only 192 total images (or 18.3 images seen per participant), the diagnosis AUC was 0.77, while using the full set of 835 images resulted in an AUC of 0.76. At this 192-image subset size, the difference between subsets is also clear: the H=M set only reaches an AUC of 0.70, while the H<M set performs worse with an AUC of 0.65.

Differentiating HC from SCD individuals shows similar results, even though the two groups have more similar memory performance. The AUC of the H>S set is higher than those of H=S and H<S at all image subset sizes, and the H>S subset first overtakes performance of the full image set at only 92 images in the subset. The AUC for the full image set is 0.59, while with the 92-image subset, the AUC of H>S is also 0.59. In regard to the other image subsets, the AUC for H=S is 0.57, and for H<S it is 0.55. H>S reaches a maximum of performance at a subset size of 367 images, with an AUC of 0.61.

We also determined if the image subsets generalized across groups. We performed the IIS analysis by training on MCI data to determine the image subsets, but then testing those images with SCD data. We find these subsets generalize to each other: the H>M image subset shows higher performance than the other image subsets (H=M, H<M), and first overtakes performance of all images (AUC=0.60) at a subset size of only 100 images (H>M: AUC=0.60; H=M: AUC=0.50; H<M: AUC=0.55). The H>M image subset reaches its peak in performance at 417 images, at an AUC of 0.63.

These results show that using a small, honed subset of images results in higher diagnostic performance than a large, exhaustive set of images, for both SCD and MCI populations. Additionally, using a poor set of images (e.g., H<M) could result in a high diagnosis failure rate. We also find that diagnostic images can successfully transfer across groups; using images that identify MCI can also successfully identify SCD. Since all of the above tests use separate halves of the participants to determine the diagnostic images and to predict group membership, this image diagnosticity is likely to translate to other participant samples as well as other experimental contexts.

3.3 Image attributes that distinguish these image sets

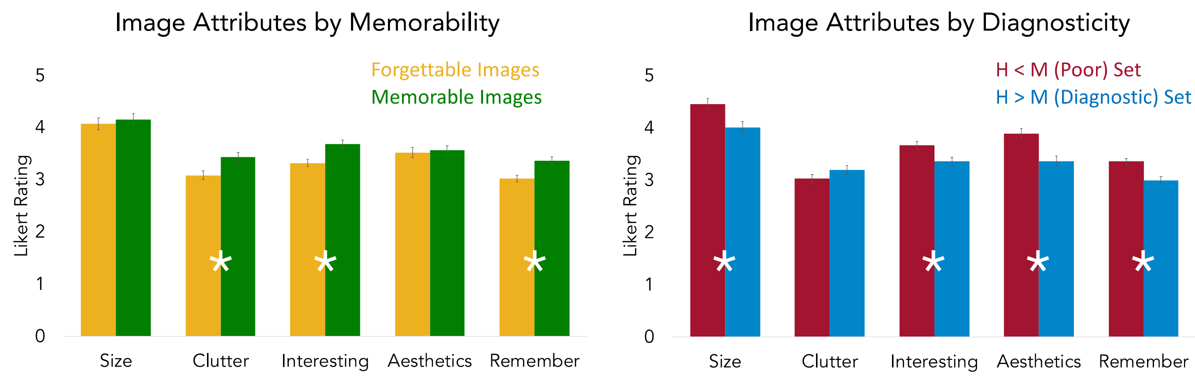

Figure 4: Average attribute ratings based on image set. (Left) Comparison of average attribute ratings between images that are forgettable versus memorable to both HC and individuals with MCI or SCD. (Right) Comparison of average attribute ratings between images from the poorly diagnostic image set (H<M) versus highly diagnostic set (H>M). (Both) All attributes are rated on a Likert scale of 1 (low) to 5 (high). “Remember” is a rating of how likely participants believed they’d be able to remember the image. Asterisks indicate significant differences in a paired samples t-test ($p < 0.05$). Error bars indicate standard error of the mean.

Finally, we investigated image attributes related to why an image is memorable to both groups, or why it is diagnostic (Figure 4). Focusing on images that have highly correlated performance between memory impaired individuals and healthy controls, memorable scene images tended to contain more clutter ($t(191)=2.84$, $p=0.005$), appeared more interesting ($t(191)=3.30$, $p=0.001$), and were subjectively more memorable to healthy controls ($t(191)=3.59$, $p=4.17 \times 10^{-4}$). However, they were not different in scene spatial size ($p=0.567$) or aesthetics ($p=0.752$). In terms of content, memorable versus forgettable images tended to be manmade rather than natural (forgettable: 76.6% manmade, memorable: 87.0%; $Z(191)=2.64$, $p=0.008$), but were equally likely to be indoors (forgettable: 52.1% indoors; memorable: 50.5%;

p=0.76) and contain people (forgettable: 7.8% contained people; memorable: 13.0%; *p*=0.09). Finally, memorable images showed no significant differences in across-image similarity based on responses across layers of a CNN trained on image classification, suggesting that memorable images are not more visually distinctive than forgettable images (Supplementary Table 1).

Focusing on images that show large differences between healthy controls and memory-impaired individuals, successfully diagnostic images versus non-diagnostic images tended to be of smaller spaces ($t(191)=3.05$, $p=0.003$), were less interesting ($t(191)=2.81$, $p=0.005$), less aesthetic ($t(191)=4.04$, $p=7.70 \times 10^{-5}$), and were judged to seem more forgettable by healthy controls ($t(191)=3.79$, $p=2.05 \times 10^{-4}$), but showed no difference in clutter ($p=0.153$). In terms of content, diagnostic images tended to be manmade (non-diagnostic: 72.4%; diagnostic: 83.9%; $Z(191)=2.72$, $p=0.007$), indoors (non-diagnostic: 37.5%; diagnostic: 55.7%; $Z(191)=3.58$, $p=3.40 \times 10^{-4}$), and contained people (non-diagnostic: 5.2%; diagnostic: 17.7%; $Z(191)=3.85$, $p=1.20 \times 10^{-4}$). Memorable images were significantly more interesting ($t(191)=2.80$, $p=0.006$) and seemed subjectively more memorable ($t(191)=3.55$, $p=4.86 \times 10^{-4}$) than diagnostic images. This shows that diagnostic images that SCD and MCI individuals forget but healthy controls remember tend to be those that are generally less aesthetic or interesting, yet are manmade, indoor scenes containing people. There were no significant differences in across-image similarity between diagnostic and non-diagnostic images as determined by the image classification CNN (Supplementary Table 1), suggesting that diagnostic images are not more visually distinctive. Additionally, a 2-way ANOVA (participant group × image diagnosticity) comparing proportion of high-confidence ratings found a main effect of participant group ($F=11.53$, $p=1.12 \times 10^{-5}$), but no significant effect of image diagnosticity (*p*=0.626), nor a significant interaction (*p*=0.350), suggesting no link between confidence and diagnosticity.

4. Discussion

While individuals with SCD and MCI have decreased memory performance in comparison to HC, there is a considerable overlap in the images that they remember and forget. Thus, there are images that are highly memorable and forgettable to everyone regardless of diagnosis. These consistencies in memorability exist not only between impaired memory groups and healthy controls, where consistencies in memorability are already well-established for controls [1,2], but also within the SCD and MCI groups themselves. Our questionnaire-based assessment of image attributes revealed that this common memorability is not related to aesthetics or spaciousness, but to being manmade scenes that contain more objects, and are subjectively more memorable and interesting. While previous work has reported that ratings of interestingness, subjective memorability, and aesthetics are ultimately not predictive of scene memorability at a fine-grained scale for healthy populations [7], such attributes may be important for guiding the selection of images that are broadly memorable across population types. We also find that memorable images are not necessarily the most visually distinctive, as determined by a CNN trained on image classification.

Additionally, we show that a publicly available convolutional neural network (MemNet [6]) trained to predict image memorability aligns with performance of HC as well as those with SCD and marginally with MCI. This raises the possibility that computational methods may guide the selection of images for diagnostic or therapeutic tools on the basis of memorability. Such tools may assist in creating or adapting environments to ease memory burdens on patients by avoiding low memorability items, or focusing strategies on rehearsing particularly forgettable information.

While memorability is generally consistent across HC, SCD, and MCI groups, we have also identified a specific set of images that significantly differ between groups. Namely, we find that there are images that are highly memorable to HC, yet highly forgettable to MCI and SCD individuals, and a certain subset of these images can be used to best determine if an individual is likely to be healthy or have MCI or SCD. The images generalize across impairments; images that differentiate MCI also successfully differentiate SCD, indicating that SCD may show similar cognitive impairments to those developed in MCI. This image set results in as much as a 10% improvement in diagnostic performance in comparison to a poorly chosen set of images (e.g., images memorable to MCI but forgettable to healthy controls). Further, this optimized image set reaches peak diagnostic performance with as few as 18.3 images seen per participant, classifying as well as the original set with 88 images per participant. This means that individuals with MCI or SCD can be identified with higher certainty, and in a quicker, easier test. In terms of content, these diagnostic images tended to be manmade, indoor scenes that contained people. However, in contrast to memorable images, they tended to be less aesthetic, less interesting, and seem subjectively less memorable. Scenes containing people tend to be the most memorable [12], however it is perhaps the combination of memorable image content (e.g., people, manmade objects) yet lack of memorable qualities (e.g., interestingness, aesthetics) that causes these images to be remembered by healthy controls but forgotten by SCD and MCI individuals.

Functional neuroimaging work with healthy individuals has found that viewing memorable images results in automatic, stereotyped activity patterns in the visual cortex and medial temporal lobe [8,9]. In future work, investigating the neural fate of memorable and forgettable images in older individuals and those with SCD or MCI may aid in understanding how patients may differentially process images at different processing stages of perception and memory encoding. In the DELCODE study, we have indeed obtained fMRI data alongside the behavioral data reported here [15] and will be able to address this question in the future. A related question is how Alzheimer's pathology is related to memorability. For instance, we have previously shown that increasing levels of CSF total-tau are related to decreasing novelty responses in the amygdala and the hippocampus [15]. These functional consequences of tau-pathology could influence memorability patterns in MCI or SCD. Indeed, activity in medial temporal lobe regions shows early and automatic sensitivity to the memorability of an image in healthy individuals [8]. Further, older adults at risk for MCI first show volume decrease in the entorhinal cortex, resulting in impairments in object location memory [23,24] and object discrimination [25]. The diagnostic images, with their higher scene complexity and several manmade objects, may be most affected by early object processing deficits. Image diagnosticity as calculated in this study could also be related to the biomarker status of individuals, a possibility that we will be able to address in the future with larger sample sizes. It will also be paramount to better understand the visual, semantic, and statistical features of an image that drive it to be forgettable, memorable, or diagnostic. Several studies are working to examine memorability with more varied image sets, in a variety of experimental image contexts, and using new computational methods ([26] for a review). Additionally, understanding the content that makes an image most sensitive to differences between groups will allow for better identification of early impairments. Using fine-grained confidence rating scales or an information-dense metric of recollection (such as drawing [27]) may provide a more nuanced understanding of the memory for these images. While the current work uses a memorability CNN trained on healthy participant memory data to predict participant memory, as larger-scale data from individuals with SCD, MCI, and Alzheimer's Disease is collected, a CNN could learn to identify images that would be particularly effective in diagnosis. Finally, while the current study does not find consistent diagnostic ability in images remembered by impaired individuals and forgotten by healthy controls, this set of images may be particularly interesting to investigate in future work.

In sum, we show the importance of images themselves in predicting what memory impaired individuals are likely to remember and differentiating them from healthy individuals. Such insights will have a meaningful impact in how we design cognitive assessment tools and tests for early diagnosis of memory impairments, and in understanding how and why we process and remember certain images over others in our complex, visual world.

Acknowledgements

The study was funded by the German Center for Neurodegenerative Diseases (Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE)), reference number BN012.

W.A. Bainbridge is supported by the Intramural Research Program of the National Institutes of Health (ZIA-MH-002909), under National Institute of Mental Health Clinical Study Protocol 93-M-1070 (NCT00001360).

Conflicts of Interest

E. Düzel and D. Berron are co-founders of neotiv GmbH.

5. References

[1] P. Isola, J.X. Xiao, A. Torralba, A. Oliva, What makes an image memorable? 24th IEEE Comput Soc Conf Comput Vis Pattern Recognit 2011;145-152.

[2] W.A. Bainbridge, P. Isola, A. Oliva, The intrinsic memorability of face photographs. J Exp Psychol Gen 2013;142:1323-1334.

[3] Z. Bylinskii, P. Isola, C.M. Bainbridge, A. Torralba, A. Oliva, Image memorability with fine-grained context. Vision Res 2015;116:165-178

[4] W.A. Bainbridge, The resiliency of memorability: A predictor of memory separate from attention and priming. arXiv 2018.

[5] Y. Mohsenzadeh, C. Mullin, A. Oliva, D. Pantazis, The perceptual neural trace of memorable unseen scenes. Sci Rep 2019;9:6033.

[6] N. Broers, M.C. Potter, M.R. Nieuwenstein, Enhanced recognition of memorable pictures in ultra-fast RSVP. Psychon Bull Rev 2018;25:1080-1086.

[7] L. Goetschalckx, P. Moors, J. Wagemans, Image memorability across longer time intervals. Memory 2017;26:581-588.

[8] W.A. Bainbridge, D.D. Dilks, A. Oliva, Memorability: A stimulus-driven perceptual neural signature distinctive from memory. Neuroimage 2017;149:141-152.

[9] W.A. Bainbridge, J. Rissman, Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. Sci Rep 2018;8:8679.

[10] A. Khosla, W.A. Bainbridge, A. Torralba, A. Oliva, Modifying the memorability of face photographs. Proc IEEE Int Conf Comput Vis 2013.

[11] A. Khosla, A.S. Raji, A. Torralba, A. Oliva, Understanding and predicting image memorability at a large scale. Proc IEEE Int Conf Comput Vis 2015:2390-2398.

[12] P. Isola, J. Xiao, D. Parikh, A. Torralba, A. Oliva, What makes a photograph memorable? IEEE Trans Pattern Anal Mach Intell 2014;36:1469-1482.

[13] C.R. Jack, D.A. Bennett, K. Blennow, M.C. Carrillo, B. Dunn, S.B. Haeberlein, D.M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, E. Liu, NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimers Dement 2018;14:535-562.

[14] F. Jessen, A. Spottke, H. Boecker, F. Brosseron, K. Buerger, C. Catak, K. Fliessbach, C. Franke, M. Fuentes, M.T. Heneka, D. Janowitz, I. Kilimann, C. Laske, F. Menne, P. Nestor, O. Peters, J. Priller, V. Pross, A. Ramirez, A. Schneider, O. Speck, E.J. Spruth, S. Teipel, R. Vukovich, C. Westerteicher, J. Wiltfang, S. Wolfsgruber, M. Wagner, E. Düzel, Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer’s disease (DELCODE). Alzheimers Res Ther 2018;10:15.

[15] E. Düzel, D. Berron, H. Schütze, A. Cardenas-Blanco, C. Metzger, M. Betts, G. Ziegler, Y. Chen, L. Dobisch, D. Bittner, W. Glanz, M. Reuter, A. Spottke, J. Rudolph, F. Brosseron, K. Buerger, D. Janowitz, K. Fliessbach, M. Heneka, C. Laske, M. Buchmann, P. Nestor, O. Peters, D. Diesing, S. Li, J. Priller, E.J. Spruth, S. Altenstein, A. Ramirez, A. Schneider, B. Kofler, O. Speck, S. Teipel, I. Kilimann, M. Dyrba, J. Wiltfang, C. Bartels, S. Wolfsgruber, M. Wawgner, F. Jessen, CSF

total tau levels are associated with hippocampal novelty irrespective of hippocampal volume. Alzheimers Dement (Amst) 2018;10:782-790.

[16] J.C. Morris, A. Heyman, R.C. Mohs, J.P. Hughes, G. van Belle, G. Fillenbaum, E.D. Mellits, C. Clark, and the CERAD investigators, The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. Neurology 1989;39:1159-1165.

[17] E. Düzel, H. Schütze, A.P. Yonelinas, H.-J. Heinze, Functional Phenotyping of Successful Aging in Long-Term Memory: Preserved Performance in the Absence of Neural Compensation. Hippocampus 2011;21:803-814.

[18] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database. Adv Neural Inf Process Syst 2014;487-495.

[19] S. Park, T. Konkle, A. Oliva, Parametric coding of the size and clutter of natural scenes in the human brain. Cereb Cortex 2015;25:1792-1805.

[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012;1097-1105.

[21] R.M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep 2016;6:27755.

[22] I.I.A. Groen, M.R. Greene, C. Baldassano, L. Fei-Fei, D.M. Beck, C.I. Baker, Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife 2018;7:e32962

[23] M.E. Wimmer, P.J. Hernandez, J. Blackwell, T. Abel, Aging impairs hippocampus-dependent long-term memory for object location in mice. Neurobiol Aging 2012;33:2220-2224.

[24] L.K. Yeung, R.K. Olsen, B. Hong, V. Mihajlovic, M.C. D'Angelo, A. Kacollja, J.D. Ryan, M.D. Barense, Object-in-place memory predicted by anterolateral entorhinal cortex and parahippocampal cortex volume in older adults. J Cogn Neurosci 2019;31:711-729.

[25] Z.M. Reagh, H.D. Ho, S.L. Leal, J.A. Noche, A. Chun, E.A. Murray, M.A. Yassa, Greater loss of object than spatial mnemonic discrimination in aged adults. Hippocampus 2015;26:417-422.

[26] W.A. Bainbridge, Memorability: How what we see influences what we remember. Psychol Learn Motiv 2019;70:1-27.

[27] W.A. Bainbridge, E.H. Hall, Chris I. Baker, Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. Nat Commun 2019;10:5.