Automating clinical assessments of memory deficits: Deep Learning based scoring of the Rey-Osterrieth Complex Figure

AUTHORS:
Nicolas Langer[1,2,3*], Maurice Weber[4], Bruno Hebling Vieira[1,2,3], Dawid Strzelczyk[1,2,3], Lukas Wolf[1], Andreas Pedroni[1,2,3], Jonathan Heitz[4], Stephan Müller[4], Christoph Schultheiss[4], Marius Tröndle[1,2,3], Juan Carlos Arango Lasprilla[5], Diego Rivera[6,7], Federica Scarpina[8,9], Qianhua Zhao[10], Rico Leuthold[11], Flavia Wehrle[12], Oskar G. Jenni[12], Peter Brugger[13], Tino Zaehle[14], Romy Lorenz[15,16,17], Ce Zhang[4]

AFFILIATIONS:
1: Methods of Plasticity Research, Department of Psychology, University of Zurich, Zurich, Switzerland
2: University Research Priority Program (URPP) Dynamics of Healthy Aging, Zurich, Switzerland
3: Neuroscience Center Zurich (ZNZ), University of Zurich and ETH Zurich, Zurich, Switzerland
4. Department of Computer Science, ETH, Zurich, Switzerland
5. Virginia Commonwealth University. Richmond, Virginia.
6. Department of Health Science, Public University of Navarre, Pamplona, Spain
7. Instituto de Investigación Sanitaria de Navarra (IdiSNA), Pamplona, Spain
8. "Rita Levi Montalcini" Department of Neurosciences, University of Turin, Italy
9. I.R.C.C.S. Istituto Auxologico Italiano, U.O. di Neurologia e Neuroriabilitazione, Ospedale San Giuseppe, Piancavallo (VCO), Italy
10. Huashan Hospital, Shanghai, China
11. Smartcode, Zurich, Switzerland
12. University Children's Hospital Zurich, Child Development Center, Zurich, Switzerland
13. Rehabilitation Center, Valens, Switzerland
14. University Hospital Magdeburg University Department of Neurology, Magdeburg, Germany
15. MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, United Kingdom
16. Department of Neurophysics, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
17. Stanford University, Stanford, CA

CORRESPONDING AUTHOR(*):
Nicolas Langer
Methods of Plasticity Research
Department of Psychology
University of Zurich
Andreasstrasse 15
CH-8050, Zürich, Switzerland
E-mail: n.langer@psychologie.uzh.ch

Word Count: 4362

Abstract

Background

Memory deficits are a hallmark of many different neurological and psychiatric conditions. The Rey-Osterrieth complex figure (ROCF) is the state–of-the-art assessment tool for neuropsychologists across the globe to assess the degree of non-verbal visual memory deterioration. To obtain a score, a trained clinician inspects a patient's ROCF drawing and quantifies deviations from the original figure. This manual procedure is time-consuming, slow and scores vary depending on the clinician's experience, motivation and tiredness.

Methods

Here, we leverage novel deep learning architectures to automatize the rating of memory deficits. For this, we collected more than 20k hand-drawn ROCF drawings from patients with various neurological and psychiatric disorders as well as healthy participants. Unbiased ground truth ROCF scores were obtained from crowdsourced human intelligence. This dataset was used to train and evaluate a multi-head convolutional neural network.

Results

The model performs highly unbiased as it yielded predictions very close to the ground truth and the error was similarly distributed around zero. The neural network outperforms both online raters and clinicians. The scoring system can reliably identify and accurately score individual figure elements in previously unseen ROCF drawings, which facilitates explainability of the AI-scoring system. To ensure generalizability and clinical utility, the model performance was successfully replicated in a large independent prospective validation study that was pre-registered prior to data collection.

Conclusions

Our AI-powered scoring system provides healthcare institutions worldwide with a digital tool to assess objectively, reliably and time-efficiently the performance in the ROCF test from hand-drawn images.

Introduction

Neurological and psychiatric disorders are among the most common and debilitating illnesses across the lifespan. In addition, the aging of our population, with the increasing prevalence of physical and cognitive disorders, poses a major burden on our society with an estimated economic cost of 2.5 trillion US$ per year (Trautmann, Rehm, & Wittchen, 2016). Currently, neuropsychologists typically use paper-pencil tests to assess individual neuropsychological functions and brain dysfunctions, including memory, attention, reasoning and problem-solving. Most neuropsychologists around the world use the Rey-Osterrieth complex figure (ROCF) in their daily clinical practice(Rabin, Barr, & Burton, 2005; Rabin, Paolillo, & Barr, 2016), which provides insights into a person's nonverbal visuo-spatial memory capacity in healthy and clinical populations of all ages, from childhood to old age (Shin, Park, Park, Seol, & Kwon, 2006).

Our estimation revealed that a single neuropsychological division (e.g. at the University Hospital Zurich) scores up to 6000 ROCF drawing per year. The ROCF test has several advantages as it does not depend on auditory processing and differences in language skills that are omnipresent in cosmopolitan societies (Osterrieth, 1944; Somerville, Tremont, & Stern, 2000). The test exhibits adequate psychometric properties (e.g. sufficient test–retest reliability (Levine, Miller, Becker, Selnes, & Cohen, 2004; Meyers & Volbrecht, 1999) and internal consistency (Berry, Allen, & Schmitt, 1991; Fastenau, Bennett, & Denburg, 1996)). Furthermore, the ROCF test has demonstrated to be sensitive to discriminate between various clinical populations (Alladi, Arnold, Mitchell, Nestor, & Hodges, 2006) and the progression of Alzheimer's disease (Trojano & Gainotti, 2016).

The ROCF test consists of three test conditions: First, in the *Copy condition* subjects are presented with the ROCF and are asked to draw a copy of the same figure. Subsequently, the ROCF figure and the drawn copy are removed and the subject is immediately instructed to reproduce the figure from memory (*Immediate Recall condition*). After a delay of 30 minutes, the subject is required to draw the same figure once again (*Delayed Recall* condition). For further description of the clinical usefulness of the ROCF please refer to Shin and colleagues (Shin et al., 2006).

The current quantitative scoring system (Meyers, Bayless, & Meyers, 1996) splits the ROCF into 18 identifiable elements (see Figure 1A), each of which is considered separately and marked on the accuracy in both distortion and placement according to a set of rules and scored between 0 and 2 points (see supplementary Figure S1). Thus, the scoring is currently undertaken manually by a trained clinician, who inspects the reproduced ROCF drawing and tracks deviations from

the original figure, which can take up to 15 minutes per figure (individually *Copy*, *Immediate Recall* and *Delayed Recall* conditions).

One major limitation of this quantitative scoring system is that the criteria of what position and distortion is considered “accurate” or “inaccurate” may vary from clinician to clinician (Canham, Smith, & Tyrrell, 2000; Groth-Marnat, 2000; Shin et al., 2006). In addition, the scoring might vary as a function of motivation and tiredness or because the clinicians may be unwittingly influenced by interaction with the patient. Therefore, an automated system that offers reliable, objective, robust and standardized scoring, while saving clinicians’ time, would be desirable from an economic perspective and more importantly leads to more accurate scoring and subsequently diagnosing.

In recent years, computerized image analysis and machine learning methods have entered the clinical neuropsychological diagnosis field providing the potential for establishing quantitative scoring methods. Using machine-learning methods, such as convolutional neural networks and support vector machines, studies have successfully recognized visual structures of interest produced by subjects in the Bender Gestalt Test (BGT) (Nazar et al., 2017), and the Clock Draw Task (CDT) (Harbi, Hicks, & Setchi, 2016; Kim, Cho, & Do, 2010) - both are less frequently applied screening tests for visuospatial and visuo-constructive (dis-)abilities (Webb et al., 2021). Given the wide application of the ROCF, it is not surprising that we are not the first to take steps towards a machine-based scoring system. [Citation error] have developed an algorithm to automatically identify a selection of parts of the ROCF (6 of 18 elements) with great precision. This approach provides first evidence for the feasibility of automated segmentation and feature recognition of the ROCF. More recently, Vogt et al., 2019 (Vogt et al., 2019) developed an automated scoring using a deep neural network. The authors reported a r=0.88 Pearson correlation with human ratings, but equivalence testing demonstrated that the automated scoring did not produce strictly equivalent total scores compared to the human ratings. Moreover, it is unclear if the reported correlation was observed in an independent test data set, or in the training set. Finally, Petilli et al. (2021) have proposed a novel tablet-based digital system for the assessment of the ROCF task, which provides the opportunity to extract a variety of parameters such as spatial, procedural, and kinematic scores. However, none of the studies described have been able to produce machine-based scores according to the original scoring system currently used in clinics (Meyers et al., 1996) that are equivalent or superior to human raters. A major challenge in developing automated scoring systems is to gather a rich enough set of data with instances of all possible mistakes that humans can make. In this study, we have trained convolutional neural networks with over 20’000 digitized ROCFs from different

populations regarding age and diagnostic status (healthy individuals or individuals with neurological and psychiatric disorder (e.g. Alzheimer, Parkinson)).

Methods

Data

For our experiments, we used a data set of 20225 hand-drawn ROCF images collected from 90 different countries (see supplementary Figure S2) as well as various research and clinical settings. This large dataset spans the complete range of ROCF scores (Figure 1F) which allows the development of a more robust and generalizable automated scoring system. Our convergence analysis (see supplementary material), suggests that the error converges when approaching 10’000 digitized ROCFs. The data is collected from various populations regarding age and diagnostic status (healthy or with neurological and /or psychiatric disorder), shown respectively in Figure 1E and Figure 1D. The demographics of the participants and some example images are shown respectively in Figure 1B and Figure 1C. For a subset of the figures (4030), the clinician’s scores were available. For each figure only one clinician rating is available. The clinicians ratings derive from six different international clinics (University Hospital Zurich, Switzerland; University Children’s Hospital Zurich, Switzerland; BioCruces Health Research Institute, Spain; I.R.C.C.S. Istituto Auxologico Italiano, Ospedale San Giuseppe, Italy; Huashan Hospital, China; University Hospital Magdeburg, Germany).

The study was approved by the Institutional Ethics Review Board of the “Kantonale Ethikkommission” (BASEC-Nr. 2020-00206). All collaborators have written informed consent and/or data usage agreements for the recorded drawings from the participants. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. To ensure generalizability and clinical utility, the model performance was replicated in a large independent prospective validation study that was pre-registered prior to data collection (https://osf.io/82796). For the independent prospective validation study, an additional data set was collected and contained 2498 ROCF images from 961 healthy adults from and 288 patients with various neurological and psychiatric disorders. Further information about the participants demographics and performance in the copy and recall condition can be found in the supplementary Figure S3.

Crowdsourced human intelligence for ground truth score

To reach high accuracy in predicting individual sample scores of the ROCFs, it is imperative that the scores of the training set are based on a systematic scheme with as little human bias as possible influencing the score. However, our analysis (see results section) and previous work(Canham et al., 2000) suggested that the scoring conducted by clinicians may not be consistent, because the clinicians may be unwittingly influenced by the interaction with the patient/participant or by the clinicians factor (e.g. motivation and fatigue). For this reason, we have harnessed a large pool (~5000) of human workers (crowdsourced human intelligence) who scored ROCFs, guided by our self-developed interactive web application (see supplementary Figure S4). In this web application, the worker was first trained by guided instructions and examples. Subsequently, the worker rated each area of real ROCF drawings separately to guarantee a systematic approach. An average of 13.38 crowdsource-based scores were collected for each figure element. For each drawing, the ground truth score was obtained by computing the median scores for every item in the figure, which were then summed to yield the final total score of the image. Taking the median has the advantage of being robust to outliers. For details on the crowdsourcing please see the supplementary material.

Human MSE

As described in the previous section there are multiple independent scores (from different crowdsourcing workers) available for each image. It frequently happens that two people scoring the same image produce different scores. In order to quantify the disagreement among the crowdsourcing participants, we calculated the empirical standard deviation of the scores for each image. With *s1,...,sk* referring to the *k* crowdsourcing scores for a given image and $\bar{s}$ to their mean, the empirical standard deviation is calculated as

$$SD_{empirical} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(s_i - \bar{s})^2}$$

The mean empirical standard deviation is 3.25. Using as ground truth the sum of item score medians, we can also calculate a *human MSE*. Assuming that the sum of median scores of all items of an image is the ground truth, the average rating performance of human raters can be estimated by computing the mean squared error (MSE) and the mean absolute error (MAE) of the human ratings by first computing the mean error for each human rater, and then computing the mean of all individual MSEs. These metrics describe how close one assessment is to this ground truth on average and lets us make a statement on the difficulty of the task. Reusing above notation, we denote the k crowdsourcing scores for a given image by *s1,…,sk*. Let $\tilde{s}$ be their median. For an image or item, we define the *human MSE* as

$$MSE_{human} = \frac{1}{k}\sum_{i=1}^{k}(s_i - \tilde{s}_i)^2.$$

Similarly, we define the human MAE as

$$MAE_{human} = \frac{1}{k}\sum_{i=1}^{k}|s_i - \tilde{s}_i|.$$

Based on clinician ratings, we define both $MSE_{clinician}$ and $MAE_{clinician}$ in similar fashion. These metrics are used to compare individual crowdsourced scores, clinician scores and AI-based scores.

Convolutional Neural Network (CNN)

For the automated scoring of ROCF drawings, two variations of deep learning system were implemented: a regression-based network and a multilabel classification network. The multilabel classification problem is based on the assumption that one drawing can contain multiple classes. That is, the scores for a specific element correspond to four mutually exclusive classes (i.e. 0, 0.5, 1, 2), which corresponds to the original Osterrieth scoring system (Meyers et al., 1996) (see also supplementary Figure S1). Each class can appear simultaneously with score classes corresponding to other elements in the figure. While the scores for different elements share common features such as e.g., texture or drawing style, which are relevant for the element score, one can view the scoring of different elements as independent processes, given these shared features. These considerations are the starting point for the architectural design of our scoring system, shown in Figure 2A. The model consists of a shared feature extractor, followed by 18 element classifiers, each taking as input the shared features extracted by the primary feature extractor. The shared feature extractor is composed of three blocks, introducing

several model hyperparameters: Each block consists of a 3x3 convolutional layer, batch normalization, 3x3 convolutional layer, batch normalization and max pooling with 2x2 stride. The ReLU activation function is applied after each batch normalization and dropout is applied after max pooling. The output of each block is a representation with 64, 128 and 256 channels, respectively. Each of the eighteen per-item networks is applied on this representation. These consist of a similar feature extractor block, but with global max pooling instead, which outputs a 512-dimensional representation that is subsequently fed into a fully connected layer, that outputs a 1024-dimensional vector, followed by batch normalization, ReLU, dropout and an output fully connected layer. For a more detailed description of the CNN, we refer the reader to the supplementary material.

Training & Validation

To evaluate our model, we set aside 4045 of the 20225 ROCF drawings as a testing set, corresponding to 20% of the dataset. The remaining 16180 images were used for training and validation. Specifically, we used 12944 (80%) drawings for training and 3236 (20%) as a validation set for early stopping.

The networks and training routines were implemented in PyTorch 1.8.1 (Paszke et al., 2019). The training procedure introduces additional hyperparameters: Our main model was trained on drawings of size 232x300 (see Figure 2D and supplementary on details on the image resolution analysis), using batches of 16 images. Model parameters were optimized by the Adam optimizer (Kingma & Ba, 2014) with the default parameters $\beta 1 = 0.9$, $\beta 2 = 0.999$ and $\varepsilon = 1e\text{-}8$. The initial learning rate was set to 0.01 and decayed every epoch exponentially with decay rate $\gamma = 0.95$. To prevent overfitting dropout rates were set to 0.3 and 0.5 in the convolutional and fully-connected layers, respectively. Additionally, we used the validation split (20%) of the training data for early stopping. Since our dataset is imbalanced and contains a disproportionately high number of high score drawings, we sampled the drawings in a way that results in evenly distributed scores in a single batch. We trained the model for 75 epochs and picked the weights corresponding to the epoch with the smallest validation loss. The multilabel classification model was trained with cross entropy loss applied to each item classifier independently so that the total loss is obtained by summing the individual loss terms. The regression model was trained in an analogous manner except that we used the MSE for each item score, instead of the cross entropy loss.

We further investigated the effect of data augmentation and test time augmentation on the generalization and robustness of the neural networks (see supplementary for details on robustness analysis, the data and test-time augmentation procedures).

We evaluated the overall performance of the neural networks on the test set based on MSE, MAE and R-squared (see supplementary material for the equations of the performance metrics and final model selection). In addition, we evaluated the performance in a pre-registered (https://osf.io/82796) independent prospective validation study. Importantly, both data sets (original test set (i.e. 4045 images) and the independent prospective data set (i.e. 2498 images) were never seen by the neural networks.

Results

A human MSE and clinicians' scoring

We have harnessed a large pool (~5000) of human workers (crowdsourced human intelligence) to obtain ground truth scores and compute the *human MSE*. An average of 13.38 (sd = 2.23) crowdsource-based scores per figure were gathered. The *average human MSE* over all images is 16.3, and the *average human MAE* is 2.41.

For a subset (4030) of the 20225 ROCF images, we had access to scores conducted by different professional clinicians. This enabled us to analyze and compare the scoring of professionals to the crowdsourcing results. The clinician MSE over all images is 9.25 and the clinician MAE is 2.15, indicating a better performance of the clinicians compared to the average human rater.

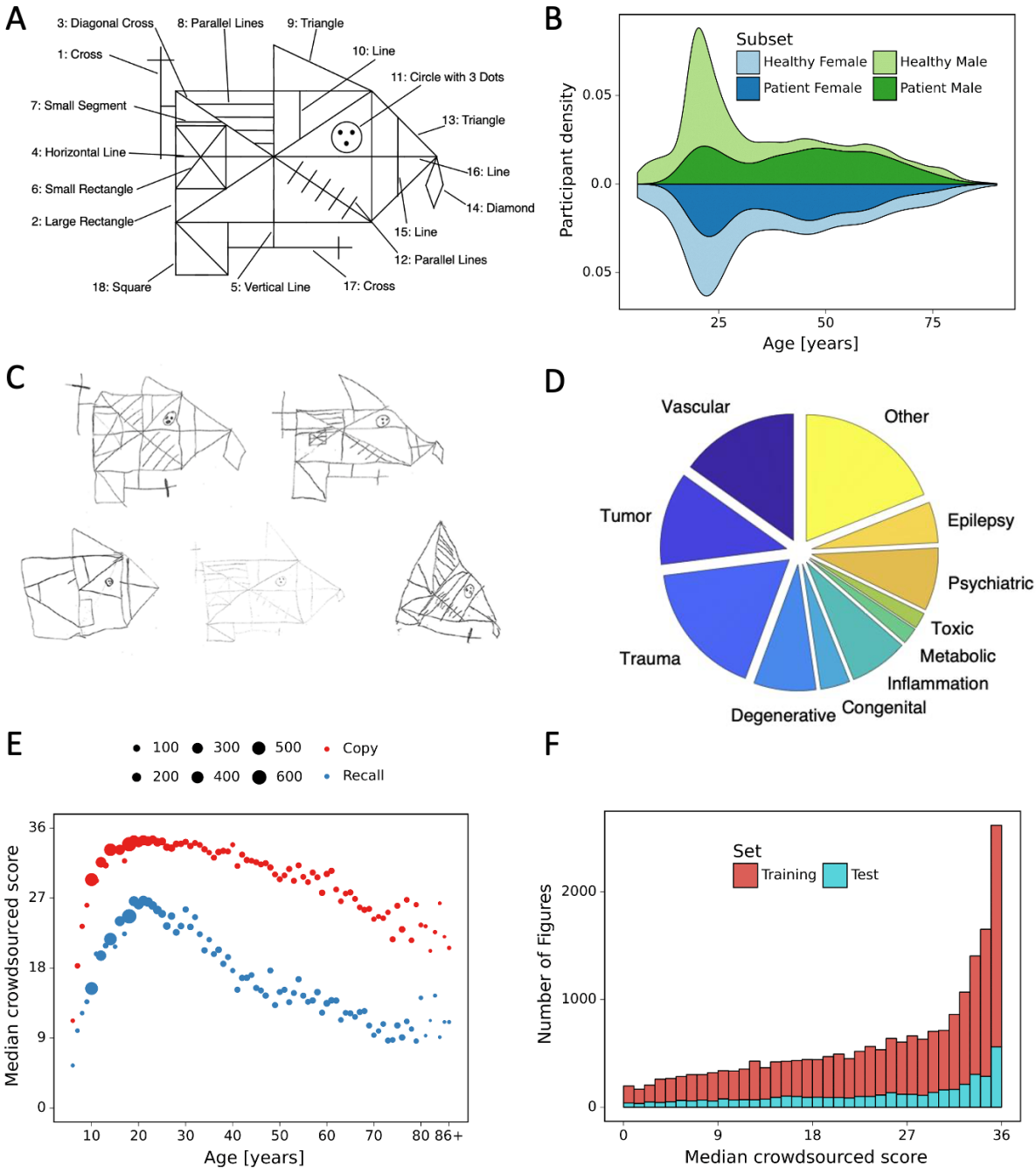

Figure 1. A: ROCF figure with 18 elements. B: demographics of the participants and clinical population. C: examples of hand-drawn ROCF images. D: Pie chart illustrates the numerical proportion of the different clinical conditions. E: performance in the copy and (immediate) recall condition across the lifespan in the present data set. F: distribution of number of images for each total score (online raters).

Machine-based scoring system

The multilabel classification network achieved a total score MSE of 3.56 and MAE of 1.22, which is already considerably lower than the corresponding human performance. Implementing additional data augmentation has led to a further improvement in accuracy with an MSE of 3.37 and MAE of 1.16. Furthermore, when combining data augmentation with test time augmentation leads to a further improvement of performance with an MSE of 3.32 and MAE of 1.15.

The regression based network variant led to a further slight improvement in performance, reducing the total score MSE to 3.07 and the MAE to 1.14. Finally, our best model results from combining the regression model with the multilabel classification model in the following way: For each item of the figure we determine whether to use the regressor or classifier based on its performance on the validation set (Figure 2B). Aggregating the two models in this way, leads to a MSE of 3.00 and a MAE of 1.11. Figure 2C presents the error for all combinations of data and test time augmentation with the multilabel classification and regression model separately and in combination. During the experiments it became apparent that for the multilabel classification network applying both data augmentation as well as test-time-augmentation jointly improves the model's performance (see Figure 2C). Somewhat surprisingly, applying these elements to the multihead regression model did not improve the performance compared to the non-augmented version of the model. The exact performance metrics (MSE, MAE and R-squared) of all model variants are reported in the supplementary table S1 and for each figure element in table S2. In addition, the model performance was replicated in the independent prospective validation study (i.e. MSE = 3.32; MAE = 1.13). The performance metrics of each figure element for the independent prospective validation study are reported in the supplementary table S4 and S5.

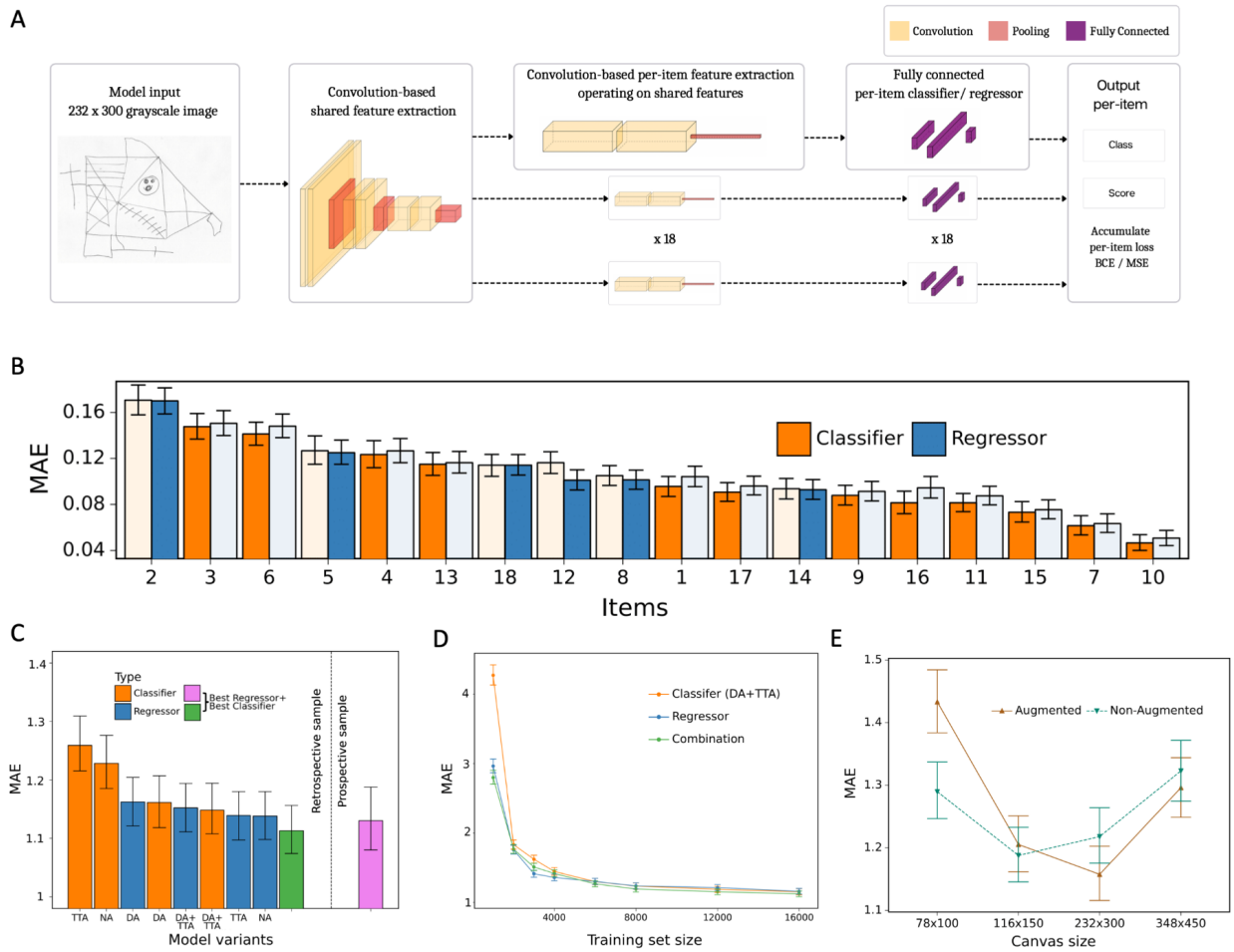

Figure 2. A: network architecture, constituted of a shared feature extractor and 18 item-specific feature extractors and output blocks. The shared feature extractor consists of three convolutional blocks, whereas item-specific feature extractors have one convolutional block with global max-pooling. Convolutional blocks consist of two convolution and batch-normalization pairs, followed by max-pooling. Output blocks consist of two fully connected layers. ReLU activation is applied after batch normalization. After pooling, dropout is applied. B. item-specific MAE for the regression-based network (blue) and multilabel classification network (orange). In the final model, we determine whether to use the regressor or classifier network based on its performance in the validation data set, indicated by an opaque color in the bar chart. In case of identical performance, the model resulting in least variance was selected. C: Model variants were compared and the performance of the best model in the original, retrospectively collected (green) and the independent, prospectively collected (purple) test set is displayed; Clf: multilabel classification network; Reg: regression-based network; NA: no augmentation; DA: data augmentation; TTA: test time augmentation. D. Convergence analysis revealed that after ~8000 images, no substantial improvements could be achieved by including more data (see also supplementary material). E. The effect of image size on the model performance measured in terms of MAE.

We further conducted a more fine-grained analysis of our best model. Figure 3A shows the score for each figure in the data set contrasted with the ground truth score (i.e. median of online raters). In addition, we computed the difference between ground truth score and predicted score revealing that errors made by our model are concentrated closely around 0 (Figure 3B), while the distribution of clinician's errors is much more spread out (Figure 3D and 3E). It is interesting to note that, like the clinicians, our model exhibits a slight bias towards higher scores, although much less pronounced. Importantly, the model does not demonstrate any considerable bias towards specific figure elements. In contrast to the clinicians, the MAE is very balanced across each individual item of the figure (Figure 3C and 3F, and supplementary table S2). Finally, a detailed breakdown of the expected performance across the entire range of total scores is displayed for the model (Figure 3G and supplementary table S3), the clinicians (3H) and average online raters (3I). As can be seen, the model exhibits a comparable MAE for the entire range of total scores, although there is a trend that high scores exhibit lower MAEs. These results were confirmed in the independent prospective validation study (see Figure 3G-3I, supplementary Figure S5 and supplementary table S5).

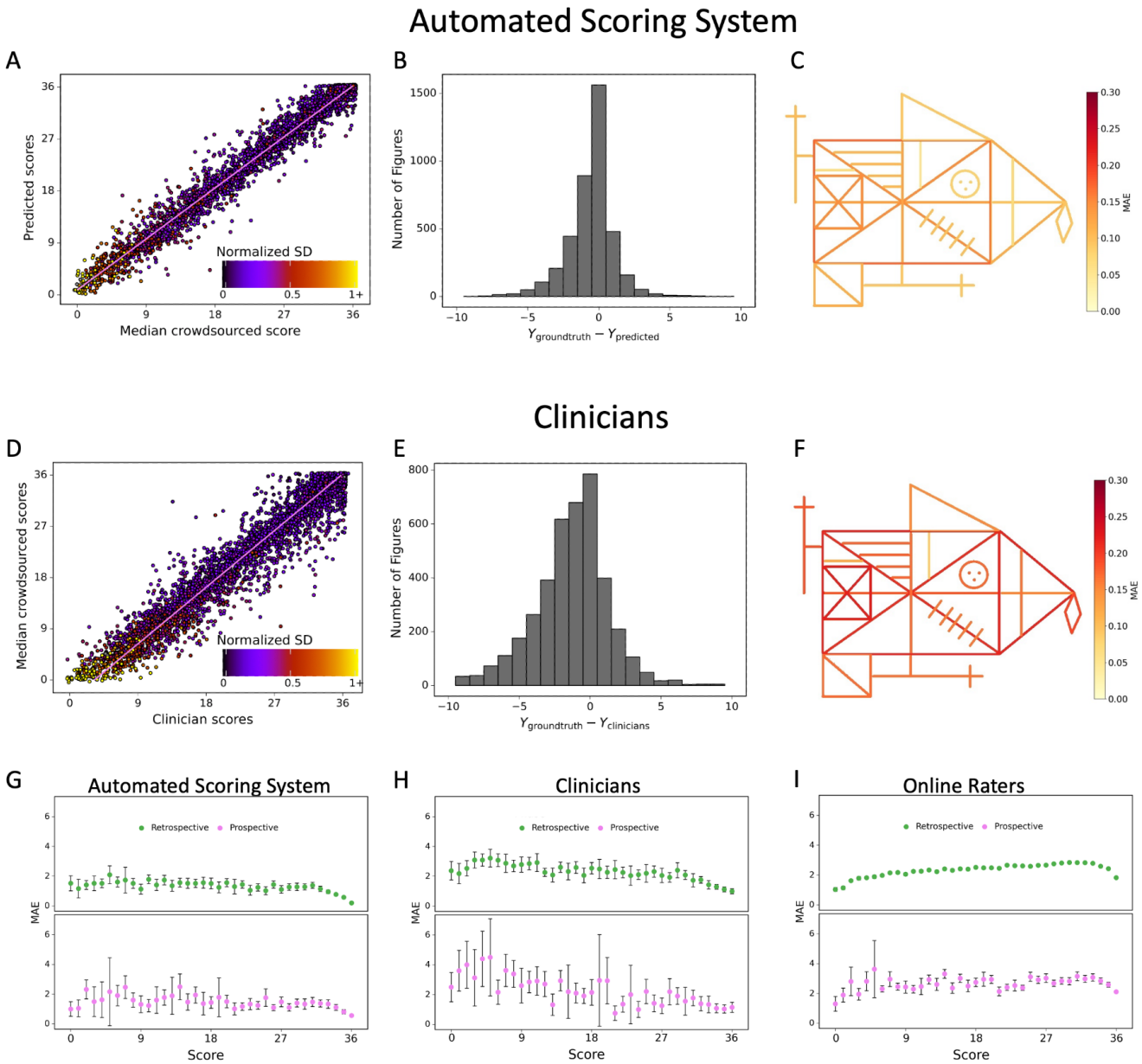

Figure 3. Contrasting the ratings of our model (A) and clinicians (D) against the ground truth revealed a larger deviation from the regression line for the clinicians. A jitter is applied to better highlight the dot density. The distribution of errors for our model (B) and the clinicians ratings (E) are displayed. The MAE of our model (C) and the clinicians (F) is displayed for each individual item of the figure (see also supplementary table S2). The corresponding plots for the performance on the prospectively collected data is displayed in the supplementary Figure S5. The model performance for the retrospective (green) and prospective (purple) sample across the entire range of total scores for model (G), clinicians (H) and online raters (I) are presented.

Robustness Analysis

Using data augmentation did not critically improve the accuracy, which is likely due to the fact that our dataset is already large and diverse enough. However, data augmentation does significantly improve robustness against semantic transformations, as can be seen from supplementary Figure S6. In particular, using our data augmentation pipeline, the model becomes much more robust against rotations and changes in perspective. On the other hand, for changes to brightness and contrast we could not find a clear trend. This is however not surprising as the data augmentation pipeline does not explicitly encourage the model to be more robust against these transformations. Overall, we demonstrate that our scoring system is highly robust for realistic changes in rotations, perspective, brightness, and contrast of the images (Figure 4). Performance is degraded only under unrealistic and extreme transformations.

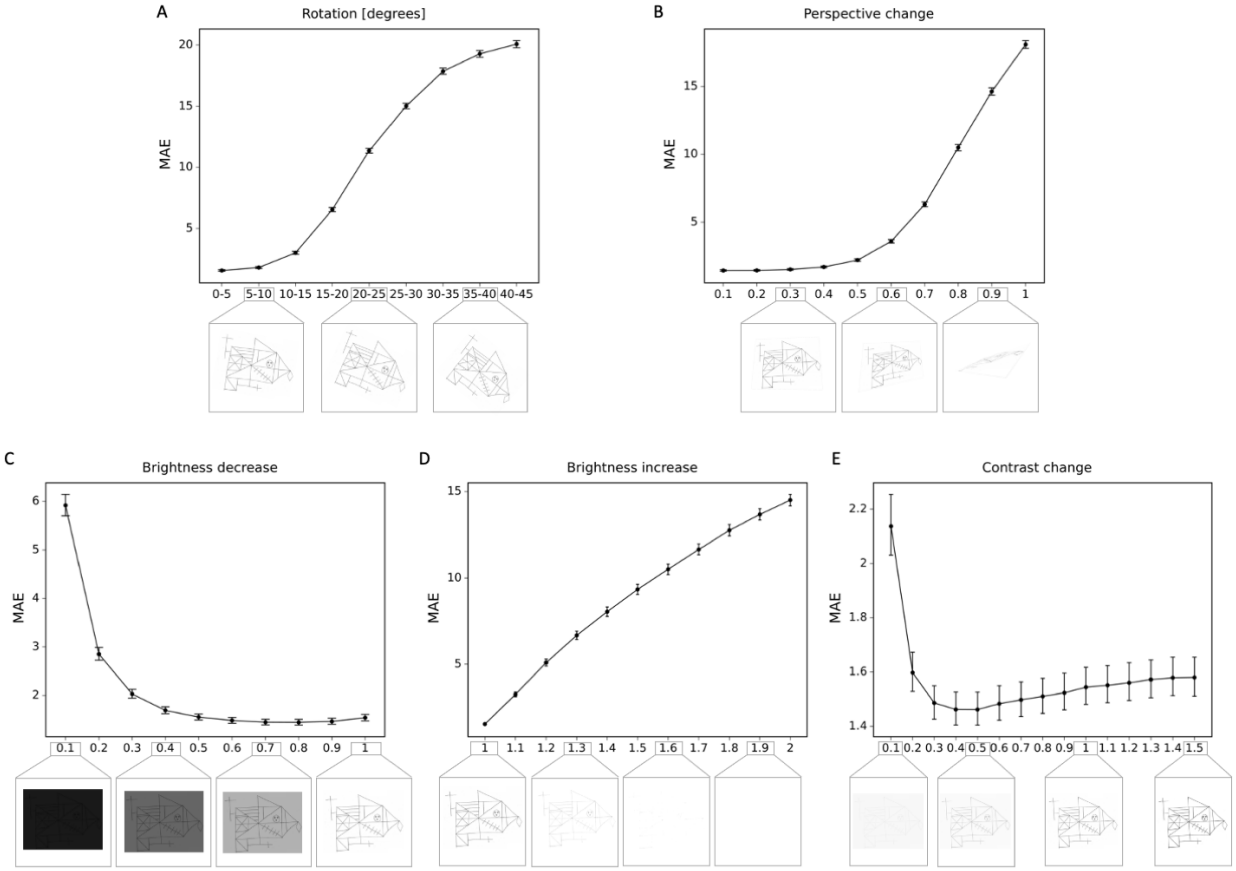

Figure 4. Robustness to geometric, brightness and contrast variations. The MAE is depicted for different degrees of transformations. In addition examples of the transformed ROCF draw are provided.

Discussion

In this study, we developed an AI-based scoring system for a nonverbal visuo-spatial memory test that is being utilized in clinics around the world on a daily basis. For this, we trained two variations of deep learning systems and used a rich data set of more than 20k hand-drawn ROCF images covering the entire life span, different research and clinical environments as well as representing global diversity. By leveraging human crowdsourcing we obtained unbiased high-precision training labels. Our best model results from combining a multihead convolutional neural network for regression with a network for multilabel classification. Data and test time augmentation were used to improve the accuracy and made our model more robust against geometric transformations like rotations and changes in perspective.

Overall, our results provide evidence that our AI-scoring tool outperforms both amateur raters and clinicians. Our model performed highly unbiased as it yielded predictions very close to the ground truth and the error was similarly distributed around zero. Importantly, these results have been replicated in an independent prospectively collected data set. The scoring system reliably identified and accurately scored individual figure elements in previously unseen ROCF drawings, which facilitates explainability of the AI-scoring system. Comparable performance for each figure element indicates no considerable bias towards specific figure compositions. Furthermore, the error of our scoring system is rather balanced across the entire range of total scores. An exception are the highest scored images displaying the smallest error, which has two explanations: First, compared to low score images, for which the same score can be achieved by numerous combinations of figure elements, the high score images by nature do not have as many possible combinations. Secondly, the data set contains a proportional larger amount of available training data for the high score images.

While the ROCF is one of the most commonly employed neuropsychological test to evaluate nonverbal visuo-spatial memory in clinical setting (Shin et al., 2006), the current manual scoring in clinics is time-consuming (5 to 15 minutes for each drawing), requires training, and ultimately depends on subjective judgements, which inevitably introduces human scoring biases (Franzen, 2000; Watkins, 2017) and low inter-rater reliability (Franzen, 2000; Huygelier, Moore, Demeyere, & Gillebert, 2020) (see also supplementary Figure S7). Thus, an objective and reliable automated scoring system is in great demand as it can overcome the problem of intra- and inter-rater variability. More importantly, such an automated system can remove a time-consuming, tedious and repetitive task from clinicians and thereby helps to reduce

workload of clinicians and/or allowing more time for more patient-centered care. Overall, automation can support clinicians to make healthcare decisions more accurate and timely.

Over the past years, deep learning has made significant headway in achieving above human-level performance on well-specified and isolated tasks in computer vision. However, unleashing the power of deep learning depends on the availability of large training sets with accurate training labels. We obtained over 20k hand-drawn ROCF images and invested immense time and labor to manually scan the ROCF drawings. This effort was only possible by orchestrating across multiple clinical sites and having funds available to hire the work force to accomplish the digitization. Our approach highlights the need to intensify national and international effort of the digitalization of health record data. Although electronic health records (EHR) are increasingly available (e.g. quadrupled in the US from 2007 to 2012 (Hsiao, Hing, & Ashman, 2014)), challenges remain in terms of handling, processing and moving such big data. Improved data accessibility and better consensus in organization and sharing could simplify and foster similar approaches in neuropsychology and medicine.

Another important difficulty is that data typically comes from various sources and thus exhibits large heterogeneity in terms of quality, size, format of the images, and crucially the quality of the labeled dataset. High-quality labeled and annotated datasets are the foundation of a successful computer vision system. A novelty of our approach is that we have trained a large pool of human internet workers (crowdsourced human intelligence) to score ROCFs drawings guided by our self-developed interactive web application. Each element of the figure was scored by several human workers (13 workers on average per figure). To derive ground truth scores, we took the median score, as it has the advantage of being robust to outliers. To further ensure high-quality data annotation, we identified and excluded crowdsourcing participants that have a high level of disagreement (>20% disagreement) with this rating from trained clinicians, who carefully scored manually a subset of the data in the same interactive web application. Importantly our two-fold innovative approach that combines the digitization of neuropsychological tests and the high-quality scoring using crowdsourced intelligence can provide a roadmap for how AI can be leveraged in neuropsychological testing more generally as it can be easily adapted and applied to various other neuropsychological tests (e.g. Clock Drawing Test (Freedman et al., 1994), Taylor Complex Figure Test (Awad et al., 2004), Hamasch 5-point test (Regard, Strauss, & Knapp, 1982)).

Another important prerequisite of using AI for the automated scoring of neuropsychological tests is the availability of training data that is sufficiently diverse and obtains sufficient examples from all possible scores. Recent examples in computer vision have demonstrated that insufficient diversity in training data can lead to biases and adverse consequences (Buolamwini & Gebru, 23--24 Feb 2018; Ensign, Friedler, Neville, Scheidegger, & Venkatasubramanian, 2017). Given that our training data set included data from children and a representative sample of patients (see Figure 1B), exhibiting a range of neurological disorders, drawings were frequently extremely dissimilar to the target figure. Importantly, our scoring system delivers accurate scores even in cases where drawings are distorted (e.g. caused by fine motor impairments) or omissions (e.g. due to visuospatial deficits), which are typical impairment patterns in neurological patients. In addition, our scoring system is robust against different semantic transformations (e.g. rotations, perspective change, brightness) which are likely to occur in real life scenarios, when the examiner takes a photo of the ROCF drawing, which naturally will lead to varying viewpoints and illumination conditions. Thus, this robustness is a pivotal prerequisite for any potential clinical utility.

To improve usability of our system and guarantee clinical utility, we have integrated our model into a tablet- (and smartphone-) based application, in which users can take a photo (or upload an image) of an ROCF drawing and the application instantly provides a total score. Importantly, the automated scoring system also maintains explainability by providing visualization of detailed score breakdowns (i.e. item-specific scores), which is a highly valued property of AI-assisted medical decision making and key to help interpret the score and communicate them to the patient (see supplementary Figure S8). The scoring is completely reproducible and thus facilitates valid large-scale comparisons of ROCF data, which enable population-based cognitive screening and (assisted) self-assessments.

In summary, we have created a highly robust and automated AI-based scoring tool, which provides unbiased, reproducible and immediate scores for the ROCF test in research and clinical environments. Our current machine-based automated quantitative scoring system outperforms both amateur raters and clinicians.

The present findings demonstrate that automated scoring systems can be developed to advance the quality of neuropsychological assessments by diminishing reliance on subjective ratings and efficiently improving scoring in terms of time and costs. Our innovative approach can

be translated to many different neuropsychological tests, thus providing a roadmap for paving the way to AI-guided neuropsychology.

References

Alladi, S., Arnold, R., Mitchell, J., Nestor, P. J., & Hodges, J. R. (2006). Mild cognitive impairment: applicability of research criteria in a memory clinic and characterization of cognitive profile. *Psychological Medicine*, Vol. 36, pp. 507–515. https://doi.org/10.1017/s0033291705006744

Awad, N., Tsiakas, M., Gagnon, M., Mertens, V. B., Hill, E., & Messier, C. (2004). Explicit and objective scoring criteria for the taylor complex figure test. *Journal of Clinical and Experimental Neuropsychology*, *26*(3), 405–415.

Berry, D. T. R., Allen, R. S., & Schmitt, F. A. (1991). Rey-Osterrieth complex figure: Psychometric characteristics in a geriatric sample. *Clinical Neuropsychologist*, Vol. 5, pp. 143–153. https://doi.org/10.1080/13854049108403298

Buolamwini, J., & Gebru, T. (23--24 Feb 2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91). PMLR.

Canham, R. O., Smith, S. L., & Tyrrell, A. M. (2000). Automated scoring of a neuropsychological test: the Rey Osterrieth complex figure. *Proceedings of the 26th Euromicro Conference. EUROMICRO 2000. Informatics: Inventing the Future*. https://doi.org/10.1109/eurmic.2000.874519

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). Runaway Feedback Loops in Predictive Policing. arXiv. Retrieved from http://arxiv.org/abs/1706.09847

Fastenau, P. S., Bennett, J. M., & Denburg, N. L. (1996). Application of psychometric standards to scoring system evaluation: is “new” necessarily “improved”? *Journal of Clinical and Experimental Neuropsychology*, *18*(3), 462–472.

Franzen, M. D. (2000). Validity as Applied to Neuropsychological Assessment. *Reliability and Validity in Neuropsychological Assessment*, pp. 33–54. https://doi.org/10.1007/978-1-4757-3224-5_5

Freedman, M., Leach, L., Kaplan, E., Winocur, G., Shulman, K., Delis, D. C., & Others. (1994). *Clock drawing: A neuropsychological analysis*. Oxford University Press, USA.

Groth-Marnat, G. (2000). *Neuropsychological Assessment in Clinical Practice: A Guide to Test Interpretation and Integration*. Wiley.

Harbi, Z., Hicks, Y., & Setchi, R. (2016). Clock Drawing Test Digit Recognition Using Static and Dynamic Features. *Procedia Computer Science*, Vol. 96, pp. 1221–1230. https://doi.org/10.1016/j.procs.2016.08.166

Hsiao, C.-J., Hing, E., & Ashman, J. (2014). Trends in electronic health record system use among office-based physicians: United States, 2007-2012. *National Health Statistics Reports*, (75), 1–18.

Huygelier, H., Moore, M. J., Demeyere, N., & Gillebert, C. R. (2020). Non-Spatial Impairments Affect False-Positive Neglect Diagnosis Based on Cancellation Tasks. *Journal of the International Neuropsychological Society: JINS*, *26*(7), 668–678.

Kim, H., Cho, Y. S., & Do, E. Y.-L. (2010). Computational clock drawing analysis for cognitive impairment screening. *Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction*. https://doi.org/10.1145/1935701.1935768

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. https://doi.org/10.48550/arXiv.1412.6980

Levine, A. J., Miller, E. N., Becker, J. T., Selnes, O. A., & Cohen, B. A. (2004). Normative Data for Determining Significance of Test–Retest Differences on Eight Common Neuropsychological Instruments. *The Clinical Neuropsychologist*, Vol. 18, pp. 373–384. https://doi.org/10.1080/1385404049052420

Meyers, J. E., Bayless, J. D., & Meyers, K. R. (1996). Rey complex figure: memory error

patterns and functional abilities. *Applied Neuropsychology*, *3*(2), 89–92.

Meyers, J. E., & Volbrecht, M. (1999). Detection of Malingerers Using the Rey Complex Figure and Recognition Trial. *Applied Neuropsychology*, Vol. 6, pp. 201–207. https://doi.org/10.1207/s15324826an0604_2

Nazar, H. B., Moetesum, M., Ehsan, S., Siddiqi, I., Khurshid, K., Vincent, N., & McDonald-Maier, K. D. (2017). Classification of Graphomotor Impressions Using Convolutional Neural Networks: An Application to Automated Neuro-Psychological Screening Tests. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. https://doi.org/10.1109/icdar.2017.78

Osterrieth, P. A. (1944). *Le test de copie d'une figure complexe: Contribution à l'étude de la perception et de la mémoire*. Delachaux et Niestlé.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1912.01703

Petilli, M. A., Daini, R., Saibene, F. L., & Rabuffetti, M. (2021). Automated scoring for a Tablet-based Rey Figure copy task differentiates constructional, organisational, and motor abilities. *Scientific Reports*, *11*(1), 14895.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: a survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *20*(1), 33–65.

Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in Test-Usage Practices of Clinical Neuropsychologists in the United States and Canada Over a 10-Year Period: A Follow-Up Survey of INS and NAN Members. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *31*(3), 206–230.

Regard, M., Strauss, E., & Knapp, P. (1982). Children's Production on Verbal and Non-Verbal Fluency Tasks. *Perceptual and Motor Skills*, Vol. 55, pp. 839–844. https://doi.org/10.2466/pms.1982.55.3.839

Shin, M.-S., Park, S.-Y., Park, S.-R., Seol, S.-H., & Kwon, J. S. (2006). Clinical and empirical applications of the Rey–Osterrieth Complex Figure Test. *Nature Protocols*, Vol. 1, pp. 892–899. https://doi.org/10.1038/nprot.2006.115

Somerville, J., Tremont, G., & Stern, R. A. (2000). The Boston Qualitative Scoring System as a measure of executive functioning in Rey-Osterrieth Complex Figure performance. *Journal of Clinical and Experimental Neuropsychology*, *22*(5), 613–621.

Trautmann, S., Rehm, J., & Wittchen, H. (2016). The economic costs of mental disorders. *EMBO Reports*, Vol. 17, pp. 1245–1249. https://doi.org/10.15252/embr.201642951

Trojano, L., & Gainotti, G. (2016). Drawing Disorders in Alzheimer's Disease and Other Forms of Dementia. *Journal of Alzheimer's Disease*, Vol. 53, pp. 31–52. https://doi.org/10.3233/jad-160009

Vogt, J., Kloosterman, H., Vermeent, S., Van Elswijk, G., Dotsch, R., & Schmand, B. (2019). Automated scoring of the Rey-Osterrieth Complex Figure Test using a deep-learning algorithm. *Archives of Clinical Neuropsychology*, Vol. 34, pp. 836–836. https://doi.org/10.1093/arclin/acz035.04

Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: from alpha to omega. *The Clinical Neuropsychologist*, *31*(6-7), 1113–1126.

Webb, S. S., Moore, M. J., Yamshchikova, A., Kozik, V., Duta, M. D., Voiculescu, I., & Demeyere, N. (2021). Validation of an automated scoring program for a digital complex figure copy task within healthy aging and stroke. *Neuropsychology*, *35*(8), 847–862.

Financial Support

This work was supported by the URPP “Dynamics of Healthy Aging” and BRIDGE [40B2-0_187132], which is a joint programme of the Swiss National Science Foundation SNSF and Innosuisse. Furthermore, BHV is funded by the Swiss National Science Foundation [10001C_197480]. Finally, CL is supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number MB22.00036. None of the authors have been paid to write this article by a pharmaceutical company or other agency. Authors were not precluded from accessing data in the study, and they accept responsibility to submit for publication.

Declaration of Interests

The authors declare no conflict of interest.

Data sharing

The clinical dataset can not be shared publicly because the consent for data sharing was not obtained from study participants. The Prolific dataset is available on request after signing a data use agreement. All preprocessing and analysis scripts used for the analyses are uploaded on github (https://github.com/DS3Lab/rey-figure).

Authorship Contribution

Nicolas Langer: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Resources, Funding acquisition.
Maurice Weber: Conceptualization, Methodology, Software, Formal analysis, Validation, Data curation, Writing - original draft, Writing - review & editing, Visualization.
Bruno Hebling Vieira: Conceptualization, Methodology, Software, Formal analysis, Validation, Data curation, Writing - original draft, Writing - review & editing, Visualization.
Dawid Strzelcy: Conceptualization, Methodology, Software, Formal analysis, Data curation, Validation, Resources, Writing - original draft, Writing - review & editing, Visualization.
Lukas Wolf: Conceptualization, Methodology, Software, Formal analysis, Data curation, Validation, Writing - original draft, Writing - review & editing, Visualization.
Andreas Pedroni: Conceptualization, Methodology, Writing - review & editing
Jonathan Heitz: Conceptualization, Methodology, Writing - review & editing
Stephan Müller: Conceptualization, Methodology, Writing - review & editing
Christoph Schultheiss: Conceptualization, Methodology, Writing - review & editing
Marius Tröndle: Conceptualization, Methodology, Data curation, Resources, Writing - review & editing
Juan Carlos Arango Lasprilla: Data curation, Resources, Writing - review & editing
Diego Rivera: Data curation, Resources, Writing - review & editing
Federica Scarpina: Data curation, Resources, Writing - review & editing
Qianhua Zhao: Data curation, Resources, Writing - review & editing

Rico Leuthold: Conceptualization, Methodology, Software
Flavia Wehrle: Data curation, Resources, Writing - review & editing
Oskar G. Jenni: Data curation, Resources, Writing - review & editing
Peter Brugger: Data curation, Resources, Writing - review & editing
Tino Zaehle: Data curation, Resources, Writing - review & editing
Romy Lorenz: Conceptualization, Methodology, Writing - original draft, Writing - review & editing
Ce Zhang: Conceptualization, Methodology, Formal analysis, Writing - review & editing, Supervision, Resources, Funding acquisition.